Review Article

Indian J Med Res 141, June 2015, pp 761-774
DOI:10.4103/0971-5916.160695

Genetic markers, genotyping methods & next generation sequencing in *Mycobacterium tuberculosis*

Srinidhi Desikan & Sujatha Narayanan

Department of Immunology, National Institute of Research in Tuberculosis (ICMR), Chennai, India

Received April 17, 2014

Molecular epidemiology (ME) is one of the main areas in tuberculosis research which is widely used to study the transmission epidemics and outbreaks of tubercle bacilli. It exploits the presence of various polymorphisms in the genome of the bacteria that can be widely used as genetic markers. Many DNA typing methods apply these genetic markers to differentiate various strains and to study the evolutionary relationships between them. The three widely used genotyping tools to differentiate *Mycobacterium tuberculosis* strains are IS6110 restriction fragment length polymorphism (RFLP), spacer oligotyping (Spoligotyping), and mycobacterial interspersed repeat units - variable number of tandem repeats (MIRU-VNTR). A new prospect towards ME was introduced with the development of whole genome sequencing (WGS) and the next generation sequencing (NGS) methods, where the entire genome is sequenced that not only helps in pointing out minute differences between the various sequences but also saves time and the cost. NGS is also found to be useful in identifying single nucleotide polymorphisms (SNPs), comparative genomics and also various aspects about transmission dynamics. These techniques enable the identification of mycobacterial strains and also facilitate the study of their phylogenetic and evolutionary traits.

Key words DNA typing - genetic markers - molecular epidemiology - next generation sequencing

Introduction

Classified under the phylum Actinobacteria, the mycobacteria consist of over 150 different species that include certain pathogens known to cause severe diseases in mammals. The *Mycobacterium tuberculosis* complex (MTBC) involving *M. tuberculosis, M. bovis, M. bovis BCG, M. africanum, M. microti, M. cannetii, M. caprae* and *M. pinnipedii* are the cause for human and animal tuberculosis. The causative organism for tuberculosis (TB) is the *M. tuberculosis* (*Mtb*) which is one of the most widely spread disease causing mycobacteria in humans. The MTBC species are closely related taxonomic group of bacteria contributing to about 100 per cent chromosomal homology between each other[1]. The *Mtb* genome is about 4.4 Mbp and contains 0.01-0.3 per cent synonymic nucleotide polymorphisms[2]. Apart from the chromosomal homology, the similarity in the clinical presentation and treatment of these infections has made the study of these organisms more difficult[3]. These findings suggest that the MTBC may have evolved from a common progenitor[4].

TB is a growing public health problem among the HIV infected individuals. Around 7 to 8 million new cases of TB are recorded annually with 1.5 to 2 million deaths occurring in about one third of the infected population[5]. It is said that only 10 per cent of the individuals who are diagnosed with TB present a disease pattern that is heterogeneous, suggesting that host factors play a major role in disease vulnerability and natural history[6].

An understanding about the transmission of TB demands the analysis of distinct epidemiological populations based on universal molecular epidemiological techniques and long term surveillance programmes. The molecular epidemiology (ME) is a combination of both molecular biology and epidemiology, which involves the study of distribution of the diseases in human populations, identified at the molecular level[7]. It is a powerful technique for monitoring infectious diseases such as TB, where patients infected with a given strain may undergo relapse due to reactivation of the same strain or a different strain after cure[8]. Not only does ME help in improving our understanding of the pathogenesis of the disease, but also provides unique insights into the international dissemination of tuberculosis by the geographic comparison and evolutionary analysis of highly widespread pathogen populations[9]. In addition, cross-contamination in the laboratories and the risk factors associated with the TB transmission can also be conveniently traced[7].

There are several available applications of ME in control of TB[10]. The first important role of genotyping studies is on intensive use of clustering rate (%) to trace outbreaks[10-14]. Modern methods for molecular epidemiological typing of mycobacteria are usually based on the revelation and comparative characteristic of repetitive mycobacterial genome sequences evolving quickly[15]. Due to the high resolving capacity and the sensitivity of these methods, the differences and similarities between various strains can be easily revealed, even if the epidemiological information is completely absent. This review article throws light on the ME of *Mtb* with insights into the next generation sequencing (NGS) platforms to differentiate the strains more efficiently.

Polymorphisms within the mycobacterial genome

Earlier, the *Mtb* genome was considered to be stable that lacked polymorphisms[15]. Studies carried out in the mid-1990s[16] showed the discovery of monomorphic and polymorphic sites in the genome of the bacteria that can be possibly divided into three groups namely single nucleotide polymorphisms (SNPs), large sequence polymorphisms (LSPs), and polymorphisms in repetitive sequences. When the latter was subjected to subsequent characterization, two classes of DNA were revealed: *(i)* transposable DNA elements or scattered repeats such as insertion sequences (IS), and *(ii)* short tandem repeats[1,17]. Short tandem repeats are further classified into major polymorphic tandem repeats (MPTR); which are basically multiple repeats of 10bp in size that are separated by interspersed units of 5bp and are found in different positions in the *Mtb* genome and short tandem repeats (STR) containing 49 repeats of 36 bp each separated by short DNA sequences of 41bp in size called as the spacer DNA[18]. When compared to SNPs and LSPs, polymorphisms in repetitive sequences tend to evolve more quickly thereby resulting in analogous patterns in *Mtb* species that are completely unrelated[15]. Therefore, research on the polymorphisms in repetitive sequences has led to its usage in epidemiological investigations; the other two polymorphisms have also proven to be a useful marker in ME[19].

Genetic markers and genotyping methods of *M. tuberculosis*

IS6110: IS6110, first described by Thierry *et al*[17] is a mycobacterial insertion element that has been consistently used as a genetic marker for typing of *Mtb* species. Originally discovered in *Escherichia coli* and *Shigella* species, this 1.4kb size element belonging to the IS3 insertion sequence family contains three independent sequences that are virtually identical to each other and differ only in a few base pairs[17,18]. The number of copies of this widespread insertion element varies from strain to strain indicating that it is highly unstable. This instability induces genetic rearrangements at a very high frequency. It can also be attributed to the presence of innumerable amounts of IS6110 DNA sequences in the *Mtb* genome[19]. These properties result in fingerprint alterations after a few generations thereby validating its use in epidemiological studies[20]. Experimental evidences prove that though IS6110 is an unstable element, its transposition events are very rare[21]. Hence, these elements are useful in fingerprinting various strains with exceptions to those strains that lack this element or with those strains that contain only a fewer copies of such elements[22].

IS6110 restriction fragment length polymorphism (RFLP) typing: IS6110- RFLP has been widely used

since the early 1990s owing to its ability to differentiate between two unrelated strains[23]. The genotyping method is based on the variability of the number of copies of IS6110 and the molecular weights of DNA fragments in which the insertions are found[24].

The main advantages of the IS6110-RFLP method are its high discriminatory power and the availability of studies for comparison. The main limitation of this method is the low discriminatory power in isolates presenting five or fewer IS6110 bands. Studies have demonstrated that the frequency of clinically confirmed epidemiologic links between cases is decreased in clusters formed by isolates with six or fewer IS6110 bands[7,20,25]. For these isolates, secondary markers such as polymorphic GC-rich repetitive sequence (PGRS) and clustered regularly interspersed short palindromic repeats (CRISPRs) are used to increase the discriminatory power and to determine the epidemiological links among patients[26]. Furthermore, IS6110-RFLP has significant technical limitations, including the need for 2–3 µg of high quality DNA and, therefore, the need of prior culture of the isolates and the determination of results based on visual inspection of images of band patterns that are difficult to share between laboratories[7].

PGRS-RFLP: PGRS - RFLP has been used as a secondary typing method in isolates of *Mtb* with five or fewer copies of IS6110[27]. It was first described by Ross *et al*[28]. Being similar to IS6110 RFLP analysis, it is also a hybridization technique that utilizes the PGRS specific probe (a 3.4 kb fragment of the PGRS sequence) cloned in plasmid pTBN12[27-29.] This has helped in distinguishing strains from unrelated cases of TB and demonstrated identical banding patterns for isolates from epidemiologically related cases[27,29]. The isolates clustered by IS6110-based RFLP analysis were further discriminated by PGRS typing[30]. This is particularly the case when IS6110 low copy number strains are further analyzed by PGRS genotyping[29]. This method, like IS6110 genotyping, is resource intensive[16], but the disadvantages are similar to the technical limitations of IS6110-RFLP, and PGRS-RFLP produces an image with many more bands of different intensities, complicating its reading and interpretation[26]. It is no longer in use these days.

Mycobacterial interspersed repetitive units-variable number tandem repeat (MIRU-VNTR): Variable number tandem repeat is used as an invaluable genetic marker that provides data in a simple and format based structure on the number of repetitive sequences in polymorphic micro- and mini-satellite regions[31]. Mycobacterial interspersed repetitive units (MIRUs) are a VNTR introduced for *Mtb* by Supply *et al*[32]. Of the 41 different MIRU loci, 12 loci were identified as hypervariable repetitive units[10,31,33]. The repeated units are 52 to 77 nucleotides in length and, therefore, power of this method may be comparable to that of IS6110-RFLP.

MIRUs help in studying the population structure and also in epidemiologic studies by means of the number of repeats of each loci present in various strains[31]. Most important advantage of this method is that it can be applied to *Mtb* cultures without DNA purification[34]. On the other hand, the difficulty of this method is associated with accurate sizing of multiple small fragments[35].

A study based on a worldwide collection of tubercle bacillus isolates defined an optimized set of 24 MIRU-VNTR loci, including a subset of 15 discriminatory loci proposed to be used as a first-line typing method. These 15 and 24 loci sets reliably improved the discrimination of *Mtb* isolates compared to the original 12-locus set[10,36].

MIRU-VNTR typing: MIRU-VNTR typing is based on PCR amplification using primers specific for the flanking regions of the different MIRUs[26]. This high resolution typing method based on the VNTR of MIRUs has been successfully employed in typing the mycobacterial isolates yielding a resolution power close to IS6110-RFLP[7]. Of the 41 loci identified[31-33], 12 were selected and used for typing of *Mtb* isolates. In this method, each locus is amplified and the product is visualized in a gel. The size of the PCR product will reflect the number of copies of the repeat unit. The result yielded 12-digit number corresponding to the number of repeats at each MIRU loci, forming the basis of a coding system that facilitates inter-laboratory comparisons[7]. There is a technical difficulty of sizing the multiple small PCR fragments, which is overcome by combining multiplex PCR with a fluorescence-based DNA analyzer[31,32]. A global epidemiological database is available[37, 38] which has led to insights into the distribution and evolution of *Mtb*, including the identification of clonally related families in specific geographic distributions[39].

The discriminatory power of MIRU-VNTR analysis is typically proportional to the number of loci evaluated[16]. In general, when only the 12 loci are used, it is more discriminating for isolates with low copy

number of IS6110 insertions but less discriminating than IS6110-RFLP genotyping for isolates with high copy number of IS6110[7,16,26] especially with the Beijing spoligotypes (ST)[36,40]. When more than 12 loci are used, or MIRU analysis is combined with spoligotyping, the discriminatory power approximates that of IS6110 RFLP analysis[16]. A set of 15 and 24 MIRU loci is currently recommended for molecular epidemiologic phylogenetic studies[41].

It has been shown that 24 MIRU loci sets are specially recommended to discriminate the Beijing isolates[42] as the set of 15 MIRUs was not sufficient to discriminate them[40,42]. A recent study done by Allix-Béguec *et al*[43] showed that the set of 24 loci MIRU-VNTR lacked the sufficient discriminatory power to differentiate specifically Beijing isolates as these seemed to be hypervariable and were associated with the spread of multi drug resistant (MDR) strains. An additional seven hypervariable MIRU-VNTR loci were identified by this group, which produced a better resolution and reduced the clustering rate for the Beijing strains[43]. The MIRU-VNTR method is also considered to be the gold standard for typing the *Mtb* isolates[44].

Direct repeats (DR) or clustered regularly interspersed short palindromic repeats (CRISPR) locus: The DR region in MTBC strains is composed of multiple direct variant repeat sequences (DVRS) each of which is composed of 49 repeats of 36 bp DR and a non-repetitive spacer sequence of 35 - 41 bp in size[1,7]. This unique cluster of DR is located in a hot spot for integration of IS elements[1]. There is extensive polymorphism in the DR region by the variable presence of DVRS which is probably driven by homologous recombination between adjacent or distant chromosomal DRs[7,16]. As a result of these events, some spacers may be deleted from the genome[16].

Spacer oligonucleotide typing (Spoligotyping): Spoligotyping is the most commonly used PCR-based technique for differentiating *Mtb* strains[45]. This method is based on revealing the presence and order of location of spacers that separate DRs in a specific locus of the *Mtb* genome. Forty three types of mycobacterial spacers have been revealed, of which 37 types are typical *Mtb* and another six types additionally characterize the *M. bovis* BCG strain[15]. In practice, membranes are spotted with 43 synthetic oligonucleotides which hybridize the PCR amplified DR locus of the tested strain that is labelled. This results in a pattern that can be detected by chemiluminescence. These patterns reveal absence or presence of the spacers[7] and are read in the form of a binary that can be easily interpreted and computerized[16]. An edition of the international spoligotyping database namely SpolDB4/SITVIT[46] was introduced containing 1,939 different spoligotypes (ST) identified. The disadvantage of the database is that the actual version of SITVIT/SpolDB4 does not offer functionality for more complex analysis such as the tree based analysis[46]. Also, some important profiles such as Latin American Mediterranean (LAM), Haarlem (H), *etc*. are still assigned to wrong families even by improved decision rules of SITVIT database[46].

Apart from being simple, rapid, robust and an economical means for typing MTBC[34,35,47,48], this method has other similar advantages as that of the IS6110 RFLP method in that it can be performed with small amount of DNA and a little time after inoculation of bacteria into liquid culture[34]. Spoligotyping is useful for discrimination between isolates of *Mtb* with a few copy number of IS6110[49]. Although spoligotyping can be a method to study the molecular epidemiology of *Mtb*, but the differentiating power of spoligotyping is less than IS6110 typing when high copy number strains are being analyzed[7,48,50].

It has been shown that strains having identical ST patterns with distinct IS6110 fingerprint profiles are often encountered[16,51]. For example, the W-Beijing family of strains which are large phylogenetically related group of *Mtb* isolates comprising hundreds of similar yet distinct IS6110 variations, has an almost identical spoligopattern lacking spacers 1 through 34[52]. In such cases, spoligotyping may be useful in identifying W-Beijing strains in a population. Kremer *et al*[48] have shown that spoligotyping together with IS6110 genotyping can provide an accurate and discriminatory genotyping system.

Single nucleotide polymorphisms (SNPs): Research in genetic polymorphisms at the nucleotide level has revealed certain genetic markers that not only help in differentiating various clinical isolates but also in studying the phylogenetic relationship between these strains. SNPs can be categorized into two types namely: nonsynonymous SNP (nsSNP) and synonymous SNP (sSNP)[16,53].

In general, nsSNPs cause changes in the amino acids and also in the genetic loci that determine the drug resistance, thereby resulting in changes in the internal

or external pressure and phenotypic drug resistance[15,16]. These nsSNP genes conferring in drug resistance can aid in understanding the spread and the nature of the drug resistant isolates within the populations.

In contrast, sSNPs are considered functionally neutral as these do not alter the amino acid profile. These neutral alterations, when in structural or housekeeping genes, can provide the basis to study genetic drift and evolutionary relationships among mycobacterial strains[16]. Apart from helping in the study of phylogenetic relatedness of clinical isolates, sSNP also serves to determine the different epidemiologies in a given population.

SNP analyses are amenable to targeting multiple polymorphisms that are informative in phylogenetic grouping, drug resistance, virulence, and other epidemiologically instructive markers, but their low discriminatory power limits their use[16]. In a study done by Homolka *et al*[53], 26 different genes were sequenced to find 161 polymorphisms, of which, 59 were genotype specific and 13 polymorphisms defined deeper phylogenetic branches. Study of the most variable set of 11 genes in a set of population from Germany, validated the SNP analysis with high accuracy.

Large sequence polymorphisms (LSPs): Comparative genomic analysis of strains H37Rv and CDC1551 has revealed LSPs in addition to SNPs[54]. LSPs are thought to mainly occur as a result of genomic deletions and rearrangements rather than through recombination following horizontal transfer[16,55]. In the absence of horizontal gene transfer, deletions are irreversible and, therefore, these have been proposed for genotyping as well as for constructing phylogenies[16,56].

A deletion occurring in the progenitor strain can serve as a genetic marker to differentiate various strains and to aid in the epidemiological analysis. Some chromosomal deletions are associated with IS transpositions and the deletions occurring in these regions are called as the regions of difference (RD). There are two types of deletions namely: ancient and recent. The ancient deletions occur at different stages in the speciation process and are widespread whereas the recent deletions have a more restricted distribution[57]. For instance, IS6110 mediated deletion of the 7 kb locus RvD2 in *Mtb* H37Rv is still present in the closely related avirulent derivative H37Ra[58]. This region undergoes great variability in clinical isolates of *Mtb* and seems to represent a hot-spot for IS6110 transposition events[59]. Other RD loci such as RD9, RD105, RD 207 and RD239 have also been studied[41]. It has been revealed that RD105 is a genetic marker to identify Beijing strains[60]. But, this was proved otherwise in another study that RD105 was a marker for East Asian lineage[41] whereas RD207 was the true marker for Beijing strains. Another interesting study done by Fenner *et al*[61], showed certain pseudo Beijing characteristics in three different strains that were previously classified as Beijing strains. Three different genetic markers were used to analyse them and the study concluded that phylogenetically robust markers should be used to differentiate clinical or experimental phenotypes. From close inspection of the DNA sequences bordering these RD regions it is apparent that deletions occurr within coding regions also.

Another example of LSPs is the deletion of TbD1, a 2,153bp *Mtb* specific deletion 1 fragment which was identified in all *Mtb* strains but was absent in the ancestral strains[62]. It is found that these deletions tend to aggregate and are not always randomly distributed in the chromosome[63]. The correlations between IS elements and the deletion regions such as TbD1 have not yet been determined[16]. Using these deleted fragments as genetic markers, this analysis can be performed by simple PCR-based methods[63]. A recent study performed in Northeast Thaliland by Faksri *et al*[64] showed the classification of *Mtb* based on LSP using multiplex real-time PCR which provided a simple, rapid and high performance tool for characterizing *Mtb* based on LSPs.

Deletion analysis: Deletion analysis can be very efficiently used for epidemiological investigations especially when the presence of a specific deletion (LSPs) associated with a single strain has been predetermined. In these cases, a single PCR may suffice to track down the spread of a single strain[65,66]. However, in studies where no particular clone or strain has been predetermined, simultaneous analysis of multiple deletion regions is required[67].

A high-throughput method for detecting large polymorphic deletions was developed[68]. In this method, 43 genomic regions for large scale deligotyping analysis were selected and PCR was performed. The PCR products generated from these 43 deligosites were hybridized to a membrane containing the target sequences of the 43 loci. Amplification of the selected deletion region is possible by using the flanking regions and usage of the flanking regions has shown to increase

the discriminatory power of the technique[68]. This approach proved to be highly sensitive and efficient for the rapid screening of clinical isolates[16].

The most recent method to differentiate the strains based on the deletions within the genome is the deletion microarray. In this method, the microarray used to compare the genome of a strain is compared against that of a sequenced reference strain[7]. This helps to identify the deletion that has occurred in the sample genome. The number and distribution of these deletions help in phylogenetic/evolutionary studies, facilitation of genome structure-function studies, host-pathogen interactions based on specific genomic deletions, in molecular epidemiology both on phylogenetic relationships and information about the various phenotypes[7,16]. Table I evaluates the various genotyping methods used in epidemiologic studies.

Whole genome sequencing (WGS)

The genotyping helps to determine the patients involved in the part of chain of transmission but does not allow to distinguish the chain of transmission of events[26]. One of the important breakthroughs in the field of ME is the WGS technology which involves the shearing of the DNA to several distinct sizes that are sub-cloned in plasmids. The sub-clones are then over sampled to generate sequencing reads which provide the necessary information to perform whole genome assembly algorithms. This method is relatively faster and affordable[69].

One of the widely used WGS methods is the Sanger's method of sequencing. Sanger sequencing is method of sequencing DNA which was developed by Frederick Sanger and his co-workers[70]. It uses the ABI 3730xL platform for sequencing and has been employed in sequencing large scale projects[71]. The main disadvantage of this method is that it includes poor quality in the first 15-40 bases of the sequence due to primer binding and deteriorates the quality of sequencing traces after 700-900 bases.

Niemann *et al*[72] performed a study using WGS on two *Mtb* Beijing family clinical isolates that had matching IS6110 RFLP pattern, ST pattern and a

Table I. Evaluation of various genotyping methods used in molecular epidemiology (ME) of tuberculosis

Marker	Principle	Advantages	Disadvantages
IS6110-RFLP typing[7]	Based on the number of copies of IS6110 elements present and the molecular weights of the DNA fragments in which the elements are found.	High discriminatory power for isolates having higher IS6110 copies, widely used till date.	Low discriminatory power in isolates containing five or fewer IS6110 copies, lengthy process, inter-laboratory comparisons are difficult.
PGRS-RFLP[16]	Based on the number and the location of the PGRS regions within the genome.	Discriminates isolates having ≤5 copies of IS6110 elements, resource intensive.	Analysis is difficult, lengthy process, limited data available using this technique.
Spoligotyping[34]	Based on the spacers present between the 36bp direct repeat sequences in the genome.	Simple, rapid and robust, highly reproducible PCR based method, requires low quantities of DNA, data available in exchangeable format.	Limited discriminatory power.
MIRU-VNTR[41]	Based on the polymorphisms of MIRU loci within the genome.	Highly discriminatory, data available in exchangeable format, 24 loci used for ME and phylogenetic studies.	Analysis of the DNA by electrophoresis is less reproducible than sequencer based method. Sequencer based method is expensive.
Deletion analysis[7,16,68]	Based on the detection of deletions in the selected 43 genomic regions.	High throughput with microarray analysis, data available in exchangeable format, irreversible genetic marker used.	The target deletions need to be pre-determined, technique has to be evaluated in different settings.

RFLP, restriction fragment length polymorphism; PGRS-RFLP, polymorphic GC-rich repetitive sequence-restriction fragment length polymorphism; MIRU-VNTR, mycobacterium interspersed repetitive units-variable number of tandem repeats.

similar MIRU-VNTR profile. With the help of WGS results, the authors were able to discriminate the two isolates that differed in 130 SNPs and one large deletion thereby suggesting that the epidemiological link between the two isolates may have been remote[72].

WGS is considered as an important tool to determine sequence variation at a real epidemiological scale, to determine the evolutionary relationship of the strains and also to determine the source of infection and the transmission of the disease between various patients[26]. WGS may become the gold standard for typing various strains for ME in the near future[26, 73] but, there are certain limitations such as the need for specialized software to analyze the various sequence reads produced and the incomplete understanding of the various polymorphisms such as the SNPs and LSPs[26]. Hence, the Next Generation Sequencing (NGS) technologies are introduced which produce millions of short reads of the entire genome that are then analyzed by specialized softwares. These do not require any cloning of the template DNA into the bacterial vectors and are optimally suited for re-sequencing[73].

Next generation sequencing (NGS) technologies

Several NGS technologies have emerged which generate a magnitude of 3 to 4 times more sequence and are considerably less expensive than the Sanger's method of sequencing[70]. Various platforms of NGS for the production of massively parallel DNA reads are being used widely [Roche/(454) FLX, Illumina/ Solexa Genome Analyzer, Applied Biosystems SOLiD™ System, Helicos Heliscope™, Ion Personal Genome Machine (Ion PGM), Pacific Biosciences Single Molecule Real Time (SMRT) and nanopore sequencing instruments]. These instruments allow simple and stepwise preparation of the sample prior to DNA sequencing thereby saving time[69]. The performance characteristics and the primary advantages and the disadvantages of the various NGS platforms are described in Tables II and III, respectively.

The strains can be easily distinguished from each other due to small variations in sequence of the DNA. Moreover, more complex determination of SNPs and LSPs can be effortlessly identified. In southern India, NGS has paved way in identifying new *Mtb* isolates having completely different DNA sequences from the ones that have already been characterized. For example, WGS of clinical isolates of *Mtb* from Kerala, Andhra Pradesh and other parts of south India were characterized[79-81].

Apart from these, NGS is also used in the field of comparative genomic studies and in transmission dynamics. Ever since the sequencing of the entire mycobacterial genome is done, investigators have constantly used low and high resolution comparative genomic techniques to identify small differences between various strains[82]. With the evolution of NGS, this application has made differentiation between various strains much easier. This has been proved in a recent study where NGS platform was used to reveal the genetic heterogeneity of *Mtb* in extra pulmonary TB patients[83]. Transmissison dynamics has become an important aspect to understand the pathogenesis of the disease. A retrospective observational study was done using WGS to establish the relapse and the reinfection of TB in patients. This study concluded that WGS has more resolving power than any other genotyping methods and also may define end points for clinical studies[84]. Complex evolutionary patterns of MDR *Mtb* Beijing strains in patients have been revealed using WGS[85].

Table II. Comparing performance characteristics of various next generation sequencing (NGS) platforms

Properties	Roche (454)	Illumina	ABI SOLiD	ION PGM	Heliscope	Pacific Biosciences
Sequencing chemistry	Pyrosequencing	Sequencing by synthesis	Ligation based sequencing	Semiconductor sequencing	Single molecule approach	Real time single molecule approach
Accuracy (%)	99	99.9	99.94	99	99.5	99.999
Millions of reads/run	1	3.4	>700	0.10	800	0.01
Run time[a]	10 h	26 h	~ 1-2 wk	~2 h	~1 day	~2 h

[a] Run time depends on the length of the genome to be sequenced. Here, it is considered to be the length of the human genome.
Source: Refs 74-78

Table III. Advantages and disadvantages of the various next generation sequencing (NGS) platforms

NGS platform	Primary advantages	Primary disadvantages
Roche (454)	Produces maximum read length.	Very expensive – high cost per Mb
Illumina	Versatile instrument and scalable in future.	Relatively few reads and higher cost / Mb
ABI SOLiD™	Low cost of the instrument and high accuracy.	Takes a long time for sequencing and cost per read is high.
ION PGM	Simple and a low cost instrument can be easily upgraded.	New platform with low accuracy.
Heliscope™	Single molecule approach produces large number of reads.	Accuracy and longevity of this approach remain questionable.
Pacific Biosciences	Real time sequencing through single molecule approach produces large number of reads; the run time is very low thereby enabling sequencing of large number of samples.	High capital cost.

Mb, mega base pair
Source: Refs 74,75,78

Roche/454 FLX pyrosequencer: The 454 GS20 pyrosequencing platform has been replaced by GS FLX platform in 2005. It is the first generation high throughput sequencing technology[73]. The Roche 454 sequencer (Roche, USA) is based on the principle of pyrosequencing which involves the release of the pyrophosphates due to the incorporation of nucleotides during the PCR reaction. The pyrophosphate so released sets out a series of downstream reactions that lead to the production of light by means of the enzyme luciferase[73].

In this method, single stranded DNA library is constructed by fragmentation of the genomic DNA and ligation of the fragments to the adaptor sequences. The ligated products are selected based on avidin-biotin purification and the library is created. This step takes approximately 4.5 h and does not involve the cloning of the genomic DNA into plasmids. The fragments are then mixed with agarose beads whose surfaces carry oligonucleotides complementary to the adaptor sequence present in the fragment library. PCR amplification is carried out by subjecting individual fragment : bead mixture in oil : water emulsion containing all the PCR reactants. The emulsion PCR step takes about eight hours to complete. The fragments are amplified as a whole on a picotitre plate where each bead is placed in a single well. With the addition of nucleotides during PCR amplification, the pyrophosphates are released and the light is produced[69]. The produced light is recorded as an image for analysis[73].

The Roche/454 FLX instrument provides 100 flows of each nucleotide during an 8 h run, which incorporates an average of 2.5 bases per flow thereby producing an average read length of 400 nucleotides. These raw reads analyzed by the software and are screened to remove poor quality sequences. Finally, 100 MB of quality data on average are obtained[69]. The main disadvantage of this method is that it incorporates more than one nucleotide to the complementary strand in one cycle and lacks reversible nucleotide terminator to stop the addition of nucleotide. It is also said to have certain problems in resolving homopolymeric stretches of sequences[73].

Illumina/solexa Genome analyzer: Illumina genome analyzer (Illumina, USA) uses sequencing by synthesis approach that involves amplification of a single nucleotide. In this method, the single stranded genomic DNA library is created where the fragments are ligated with the adaptor sequences. The creation of the library takes place on the oligo derivatized solid support of a flow cell which is done automatically by a device called as a cluster station[69]. The flow cell is an 8-channel sealed glass device where the DNA polymerase enzyme allows the generation of *in situ* copies of DNA molecule on the oligo decorated surface (Bridge Amplification). Fluorescent labelled dNTPs are added where the 3'OH group is blocked such that each addition occurs as a unique event[69]. An image is taken after every addition of the base to the growing strand. This is followed by the chemical removal of 3'

blocking group which prepares the strand for the next base incorporation.

The sequencing is carried for approximately four days and the computed data filters to remove the poor quality reads. Around 40-50 million reads are produced for every run[71]. The amplification step used in this method primarily reduces the background noise and increases the levels of the signals produced. On the other hand, inaccuracies can be caused if the simultaneous stepwise sequencing of the molecules is disturbed due to non-synchronous behaviour of individual molecules[73].

Applied Biosystems SOLiD system: The SOLiD platform uses adaptor ligated fragment library and an emulsion PCR technique similar to other NGS platforms[69]. The DNA amplified on the beads undergoes two cycles. The first cycle involves the deposition of the beads on to a glass slide which is eventually subjected to hybridization with random fluorescent labelled oligonucleotides containing known 3' dinucleotide. The synthesized DNA is removed from the template after five cycles and the process is repeated. The second cycle starts at the upstream region on the template where the synthesis began in the previous cycle. This type of repeated cycle process tends to reduce the error rates and allows for multi-colour analyses of the bases in the DNA sequence[73]. Complicated DNA variations and deletions such as SNPs and LSPs can be easily detected. AB SOLiD system can detect read lengths between 25-50 nucleotides and the run yields a data of 2-4GB[69,73].

The advantage of SOLiD system over the other platforms is that it has slightly better performance in terms of accuracy as the bases are sequenced twice by dinucleotide detection[86]. Apart from having similar limitations as that of the Illumina platform[73], a complicated algorithm is required to interpret the raw data[86].

Helicos Heliscope: Helicos Heliscope (Applied Biosystems, USA) uses single molecule approaches. This kind of approach avoids amplification and errors and bias and other intensity and phasing related problems[73]. Similar to the above mentioned platforms, Heliscope sequencing also involves the creation of adaptor ligated DNA fragment library. These fragments remain unamplified and are attached to a solid substrate. Fluorescent labelled oligonucleotides are added to the unamplified DNA fragment by the DNA polymerase enzyme to create a second strand DNA. A virtual terminator helps in preventing the addition of extra nucleotides in a cycle. Image is taken with the amount of fluorescence sensed and the nucleotide from each DNA sequence can be determined. The fluorescent molecule is then cut away, and the process is repeated until the fragments have been completely sequenced[87]. The background noise is greatly reduced due to the use of fluorophore at the start of each cycle. Currently, reads have a final length of 35 nucleotides[73].

Ion Personal Genome Machine (Ion PGM): One of the recently introduced NGS systems includes the Ion PGM (Life Technologies, USA). This method is based on the principle of release of hydrogen ions from nucleotides which is sensed by an array of semiconductor chips that are capable of sensing minor changes in *p*H[86]. Generally, when a nucleotide is added to the growing DNA template by the polymerase enzyme, a proton is released. The PGM recognizes the addition of nucleotide by the release of the proton which induces a *p*H change. When there in an incorrect base added, there is no voltage difference observed whereas when two nucleotides are added at the same time, a double voltage is detected[88].

This platform is the first NGS technology that does not require fluorescence and imaging techniques resulting in higher speed, lower cost, and smaller instrument size. The preparation of the sample takes about six hours for eight samples in parallel and the read length produced is about 200bp[86,88].

Pacific Biosciences Single Molecule Real Time (SMRT): The Pacific Biosciences SMRT platform (USA) also uses single molecule approach and is considered to be the most revolutionary technique[73]. In this method, the DNA polymerase is immobilized in a zeptolitre wells (10^{-21} wells)[73]. The single stranded DNA is added to the wells containing the immobilized enzymes along with the phospho-linked nucleotides and other PCR reactants. The complementary strand is synthesized by the addition of nucleotides. Since the wells are only nanometers in size, they accumulate the nucleotides leading to a highly focused and continuous detection process.

This platform gives various advantages in that the newly introduced base can be read in a short period of time, has high accuracy and low background noise due to the use of phospho-linked nucleotides[73], and large inserts can be used[86]. This method detects the addition of nucleotides in real time thereby paving way for novel applications[86].

Computational resources for data analysis

The computational softwares necessary to analyse the data vary tremendously for every platform. High end softwares are required to analyse the short reads and most of the platforms produce millions of short reads. To analyse these short reads, generally high end computers are used. Some softwares can work only on LINUX based systems whereas a few others can work in Windows or in Mac systems. This variability poses a great disadvantage for the users to analyse the short reads. For the analysis of high amounts of data, high performance computational clusters produced by Illumina can be used which is less expensive and can be operated on any computer[74].

Nanopore sequencing

Nanopore sequencing is an upcoming sequencer and is classified under third generation sequencing. The biological role of a nanopore is to facilitate the ion exchange through the protein channels which are embedded on the lipid layer[89]. The basic principle of nanopore sequencing method is to insert a thread of single stranded DNA through the nanopore such as α-haemolysin (αHL) which is isolated from *Staphylococcus aureus*[90]. This 33kDa protein undergoes self-assembly to form heptameric ion channel[74] which can tolerate a large voltage of up to 100mV[90]. By the exploitation of this unique property, external current is applied constantly and the disruption of the current is detected by electrophysiological technique. The results are based on the size difference between deoxyribonucleoside monophosphates (dNMP). For every size of dNMP, there is a modulation in the current. The ionic current is resumed when the trapped nucleoside comes out of the pore[88].

Another biological nanopore being investigated for DNA sequencing is the *Mycobacterium smegmatis* porin A (MspA). This particular molecule has been identified as an alternative to αHL molecule due to its structure[91]. Generally it has a negative core that blocks the translocation of single stranded DNA but this has been modified by replacing three negatively charged aspartic acids with neutral asparagines thereby allowing the translocation process to take place[92]. The detection of the nucleotides and the identification of the bases by the application of the electric current across the membrane have shown to be ten folds higher than αHL[91]. Further studies are currently carried out in the United States to improve its specificity and also the base recognition of the MspA molecule[93].

Nanopore sequencing will enable sequencing of large DNA molecules in minutes without modifying or preparing samples. It is expected to offer solutions to limitations of other NGS technologies[86].

Current applications of molecular epidemiology

Apart from a few applications described earlier, several other applications for ME are available in the control of TB. Epidemiological data are necessary to understand the mechanism of outbreaks of any disease. But, based on the epidemiological data alone, one cannot have an insight into the transmission of TB. In these cases, ME allows the use of the genotyping studies which are extensively used to predict the clustering rates of species, identification of the risk groups and to monitor the outbreaks[12-14,23,47]. Evaluating the spread of drug resistant strains in patients can also be studied with the help of ME[94].

HIV infection exerts immense influence on the natural course of TB disease. Individuals with latent *Mtb* infection who contract HIV are at risk of developing active TB at a rate of 7 to 10 per cent per year, compared to approximately 8 per cent per lifetime for HIV-negative individuals[95]. With the help of ME, diseases involving co-infection with HIV can be improved[26].

Finally, with the advent of WGS technologies, various polymorphisms in the genome of strains are detected in no time, thereby making ME an easier method to detect the phylogenetic and evolutionary relationships between strains[26]. Even though *Mtb* strains lacks genetic diversity due to low mutation rate, it is surprising that in the face of rapid emergence of MDR and extensively drug resistant (XDR) strains, there is evidence of diversity of *Mtb in vivo*. The WGS analyses are providing insights into the ongoing evolution of *Mtb* during infection, treatment and acquisition of drug resistance[96]. When WGS was done on serial isolates of patient, it has revealed that resistance mutations to the drugs are independently acquired several times by different isolates[97]. By sequencing to a high depth of coverage we could identify mutations at specific loci that are present in a proportion of the bacteria sequenced[98].

Conclusion

ME has contributed extensively to our current knowledge of TB through various observational studies done since the early 1990s. These methods

have provided better accuracy and understanding in the global epidemiology of TB. Furthermore, various characteristics specific to phylogenetic lineages or strains comprising virulence properties, different replication rates and differential pathogenesis, have been widely found and suggested. With the advent and widespread availability of NGS systems, DNA sequencing has become a universal readout for a wide variety of front-end assays. There is a need for the development of novel designs and other mathematical models for analyzing the epidemiologic results obtained. An understanding of host pathogen interactions and identification of host susceptibility genes will be continuously integrated due to the current advances in designing ME studies. This will enhance the value of *Mtb* biology and epidemiology and perhaps unravel the mystery behind this elusive pathogen.

References

1. Van Embden JDA, van Soolingen D, Small PM, Hermans PMW. Genetic markers for the epidemiology of tuberculosis. *Res Microbiol* 1992; *143* : 385-91.
2. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth N, Graviss EA, *et al*. Single nucleotide polymorphism based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis* 2006; *193*: 121-8.
3. Habib NI., Warring FCJ. A fatal case of infection due to *Mycobacterium bovis*. *Am Rev Respir Dis* 1966; *93* : 804-10.
4. Gutierrez MC, Ahmed N, Willery E, Narayanan S, Hasnain SE, Chauhan DS, *et al*. Pre dominance of ancestral lineages of *Mycobacterium tuberculosis* in India. *Emerg Infect Dis* 2006; *12* : 1367-74.
5. WHO global tuberculosis report 2014. Available from: *http:// apps.who.int/iris/bitstream/10665/137094/1/978924156480 9_ eng.pdf?ua=1*, accessed on June 29, 2015.
6. Stein CM. Genetic epidemiology of tuberculosis susceptibility: Impact of study design. *PLoS Pathog* 2011*; 7* : e1001189.
7. Narayanan S. Molecular epidemiology of *Mycobacterium tuberculosis*. *Indian J Med Res* 2004; *20* : 233-47.
8. Van Rie A, Warren R, Richardson M, Victor TC, Gie RP, Enarson DA, *et al*. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. *N Engl J Med* 1999; *341* : 1174-9.
9. 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* : 3563-71.
10. Asgharzadeh M, Kafil HS. Current trends in molecular epidemiology studies of *Mycobacterium tuberculosis. Biotech Mol Biol Rev* 2007; *2* : 108-15.
11. Narayanan S, Sahadevan R, Narayanan PR, Krishnamurthy PV, Paramasivan CN, Prabhakar R. Restriction fragment length polymorphism of *Mycobacterium tuberculosis* strains from various regions of India using direct repeat probe. *Indian J Med Res* 1997; *106* : 447-54.
12. Bifani PJ, Plikaytis BB, Kapuru V, Sock bauerk W, Lutfey ML, Pan X, Lutfey ML, *et al*. Origin and interstate spread of a New York City multidrug resistant *M. tuberculosis* clone family identification of a variant outbreak of *Mycobacterium tuberculosis* TB via population based molecular epidemiology. *JAMA* 1996; *282* : 2321-7.
13. Frieden TR, Sherman LF, Maw KL, Fujiwara PI, Crawfold JT, Nivin B,Sharp V, *et al*. A multi-institution outbreak of highly drug resistant tuberculosis: epidemiology and clinical outcomes. *JAMA* 1996; *276* : 1229-32.
14. Moss AR, Alland D, Telzak E, Hewlett DJR, Sharp V, Chillade P. A city wide outbreak of a multiple drug resistant strains of *Mycobacterium tuberculosis* in New York. *Int J Tuberc Lung Dis* 1997; *1* : 115-21.
15. Kontsevaya IS, Nikolayevsky VV, Balabanova YM. Molecular epidemiology of tuberculosis: Objectives, methods and prospects. *Mol Genet Microbiol Virol* 2011; *26* : 3-10.
16. Mathema B, Kurepina NE, Kreiswirth BN. Molecular epidemiology of tuberculosis: Current insights. *Clin Microbiol Rev* 2006; *19* : 658-85.
17. Thierry D, Brisson-Noël A, Vincent-Lévy-Frébault V, Nguyen S, Guedson JL, Giequel B. Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis. *J Clin Microbiol* 1990; *28* : 2668-73.
18. Hermans PWM, van sollingen D, Dale JW, Schuitema AR, McAdam RA, Catty D, *et al*. Insertion element IS986 from *Mycobacterium tuberculosis*; a useful tool for diagnosis and epidemiology of tuberculosis. *J Clin Microbiol* 1990; *28* : 2051-8.
19. Van Soolingen D, Hermans PWM, de Haas PEW, Soll DR, van Embden JDA. The occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains; evaluation of IS-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol* 1991; *29* : 2578-86.
20. Warren RM, vander Spay GD, Richardson M, Beyers N, Borgdorff MW, Behr MA, *et al*. Calculation of the stability of the IS6110 banding pattern in patients with persistent *Mycobacterium tuberculosis* disease. *J Clin Microbiol* 2002*; 40* : 1705-8.
21. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* 2009; *4* : 7815.
22. Wiid IJ, Werely C, Beyers N, Donald P, van Helden PD. Oligonucleotide (GTG)5 as a marker for *Mycobacterium tuberculosis* strain identification. *J Clin Microbiol* 1994; *32* : 1318-21.
23. Valway SE, Greifinger RB, Papania M, Kiburn JO, Woodley C, Diferdinando GT, *et al*. Multidrug resistance tuberculosis in the New York State prison system, 1990-1991. *J Iinfect Dis* 1994; *170* : 151-6.
24. Houben RM, Glynn JR. A systematic review and meta analysis of molecular epidemiological studies of tuberculosis:

development of a new tool to aid interpretation. *Trop Med Int Health* 2009; *14* : 892-909.

25. Van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, *et al*. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993; *31* : 406-9.
26. Kato-Maeda M, Metcalfe JZ, Flores L. Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies. *J Future Microbiol* 2011; *6* : 203-16.
27. Van Soolingen D, de Hass PEW, Hermans PWM, van Embden JDA. DNA fingerprinting of *Mycobacterium tuberculosis. Methods. Enzymol* 1994; *236* : 196-205.
28. Ross BC, Raios K, Jackson K, Dwyer B. Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. *J Clin Microbiol* 1992; *30* : 942-6.
29. Rhee JT, Tanaka MM, Behr MA, Agasino CB, Paz EA, Hopewell PC, *et al*. Use of multiple markers in population based molecular epidemiologic studies of tuberculosis. *Int J Tuberc Lung Dis* 2000; *4* : 1111-9.
30. Chaves F, Yang Z, Haji H, Alonso M, Burman WJ, Eisenach KD, *et al*. Usefulness of the secondary probe pTBN12 in DNA fingerprinting of *Mycobacterium tuberculosis. J Clin Microbiol* 1996; *34* : 1118-23.
31. Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent V, Gicquel B, *et al*. High resolution minisatellite based typing as a portable approach to global analysis *of Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci USA* 2001; *98* : 1901-6.
32. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol* 2000; *36* : 762-71.
33. Magdalena J, Vanchee A, Supply P, Comille L. Identification of a new DNA region specific for members of *Mycobacterium tuberculosis* complex. *J Clin Microbiol* 1998; *36* : 937-43.
34. Barnes PF, Cave MD. Molecular epidemiology of tuberculosis. *N Engl J Med* 2003; *349* : 1149-56.
35. Burgos MV, Pym AS. Molecular epidemiology of tuberculosis. *Eur Respir J* 2002; *20* : 545-655.
36. Oelemann MC, Diel R, Vatin V, Haas W, Rüsch-Gerdes S, Locht C, Niemann S , *et al*. Assessment of an optimized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. *J Clin Microbiol* 2007; *45* : 691-7.
37. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. *Nucleic Acids Res* 2010; *38* : 326-31.
38. MIRU-VNTRplus. Available from: *www.miru-vntrplus.org/*, accessed on August 29, 2014.
39. Mulenga C, Shamputa IC, Mwakazanga D, Kapata N, Portaels F, Rigouts L. Diversity of *Mycobacterium tuberculosis* genotypes circulating in Ndola, Zambia. *BMC Infect* 2010; *10* : 177.
40. Iwamoto T, Yoshida S, Suzuki K, Tomita M, Fujiyama R, Tanaka N, *et al*. Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on *Mycobacterium tuberculosis* strains predominated by the Beijing family. *FEMS Microbiol Lett* 2007; *270* : 67-74.
41. Flores L, Van T, Narayanan S, DeRiemer K, Kato-Maeda M, Gagneux S. Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. *J Clin Microbiol* 2007; *45* : 3393-5.
42. Mokrousov I, Narvskaya O, Vyazovaya A, Millet J, Otten T, Vishnevsky B, *et al*. *Mycobacterium tuberculosis* Beijing genotype in Russia: in search of informative variable number tandem repeat loci. *J Clin Microbiol* 2008; *46* : 3576-84.
43. Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, *et al*. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit–variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. *J Clin Microbiol* 2014; *52* : 164-72.
44. Singh UB, Arora J, Suresh N, Pant H, Rana T, Sola C, *et al*. Genetic biodiversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in India. *Infect Genet Evol* 2007; *7* : 441-8.
45. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol* 1993; *10* : 1057-65.
46. Weniger T, Krawczyk J, Supply P, Harmsen D, Niemann S. Online tools for polyphasic analysis of *Mycobacterium tuberculosis* complex genotyping data: now and next. *Infect Genet Evol* 2012; *12* : 748-54.
47. Kanduma E, McHugh TD, Gillespie SH. Molecular methods for *Mycobacterium tuberculosis* strain typing: a users guide. *J Appl Microbiol* 2003; *94* : 781-91.
48. Kremer K, van Soolingen D, Frothingham R, de Hass WH, Hermans PW, Martin C, *et al*. Comparision of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: inter-laboratory study of discriminatory power and reproducibility. *J Clin Microbiol* 1999; *37* : 2607-18.
49. Goguet dela Salmoniere YO, Li HM, Torrea G, Bunschoten A, Embden JDA, Gicquel B. Evaluation of spoligotyping in a study of the transmission of *Mycobacterium tuberculosis. J Clin Microbiol* 1997; *35* : 2210-4.
50. Doroudchi M, Kremer K, Basiri EA, Kadivar MR, van Soolingen D, Ghaderi AA. IS6110-RFLP and spoligotyping of *Mycobacterium tuberculosis* isolates in Iran. *Scand J Infect Dis* 2000; *32* : 663-8.
51. Mathema B, Bifani PJ, Driscoll J, Steinlein L, Kurepina N, Moghazeh SL, *et al*. Identification and evolution of an IS6110 low-copy-number *Mycobacterium tuberculosis* cluster. *J Infect Dis* 2002; *185* : 641-9.
52. Kremer K, Glynn JR, Lillebaek T, Niemann S, Kurepina NE, Kreiswirth BN, *et al*. Definition of the Beijing/W lineage of *Mycobacterium tuberculosis* on the basis of genetic markers. *J Clin Microbiol* 2004; *42* : 4040-9.

53. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nubel U, *et al*. High resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based on single nucleotide polymorphisms. *PLoS One* 2012; *7* : e39855.

54. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, *et al*. Whole genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 2002; *184* : 5479-90.

55. Brosch R, Pym AS, Gordon SV, Cole ST. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* 2001; *9* : 452-8.

56. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, *et al*. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA* 2002; *99* : 3684-9.

57. Cole ST. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. *J Microbiol* 2002; *148* : 2919-28.

58. Brosch R, Philipp W, Stavropolous E, Colston MJ, Cole ST, Gordon SV. Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra. *Infect Immun* 1999; *67* : 5768-74.

59. Ho TBL, Robertson BD, Taylor GM, Shaw RJ, Young DB. Comparison of *Mycobacterium tuberculosis* genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. *Yeast* 2001; *4* : 272-82.

60. Tsolaki AG, Gagneux S, Pym AS, Goguet de la Salmoniere YO, Kreiswirth BN, van Soolingen D, *et al*. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis. J Clin Microbiol* 2005; *43* : 3185-91.

61. Fenner L, Malla B, Ninet B, Dubuis O, Stucki D, Borrell S, *et al*. Pseudo-Beijing: evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PLoS One* 2011; *6* : e24737.

62. Sun YJ, Bellamy R, Lee AS, Ng ST, Ravindran S, Wong SY, *et al*. Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to examine genetic diversity of *Mycobacterium tuberculosis* in Singapore. *J Clin Microbiol* 2004; *42* : 1986-93.

63. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, *et al*. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci USA* 2004; *101* : 4865-70.

64. Faksri K, Hanchaina R, Sangka A, Namwat W, Lulitanond V. Development and application of single-tube multiplex real-time PCR for lineage classification of *Mycobacterium tuberculosis* based on large sequence polymorphism in Northeast Thailand. *Tuberculosis* 2015; *95* : 404-10.

65. Freeman R, Kato-Maeda M, Hauge KA, Horan KL, Oren E, Narita M, *et al*. Use of rapid genomic deletion typing to monitor a tuberculosis outbreak within an urban homeless population. *J Clin Microbiol* 2005; *43* : 5550-4.

66. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, De Jong BC, Narayanan S, *et al*. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 2006; *103* : 2869-73.

67. Narayanan S, Gagneux S, Lalitha H, Tsolaki AG, Suganthi R, Narayanan PR, *et al*. Genomic interrogation of ancestral *Mycobacterium tuberculosis* from south India. *Infect Genet Evol* 2008; *8* : 474-83.

68. Goguet de la Salmoniere YO, Kim CC, Tsolaki AG, Pym AS, Siegrist MS, Small PM. High-throughput method for detecting genomic-deletion polymorphisms. *J Clin Microbiol* 2004; *42* : 2913-8.

69. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genom Hum Genet* 2009; *9* : 387-402.

70. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci USA* 1997; *74* : 5463-7.

71. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, *et al*. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009; *10* : R32.1-R32.13.

72. Niemann S, Köser CU, Gagneux S, Plinke C, Homolka S, Bignell H, *et al*. Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. *PLoS One* 2009; *4* : e7407.

73. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 2009; *7* : 287-96.

74. Glenn TC. Field guide to next-generation DNA sequencers. *Mol Eco Resour* 2011; *11* : 759-69.

75. *De novo* assembly using illumina beads. Available from: *http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf*, accessed on June 30, 2015.

76. Helicos announces improved performance of its heliScope™ Single molecule sequencer and tSMS ™ chemistry. Available from: *http://www.businesswire.com/news/home/20090204005323/en/Helicos-Announces-Improved-Performance-HeliScope%E2%84%A2-Single-Molecule#.VZJbFPmqpBc*, accessed on June 30, 2015.

77. Korlach J. Understanding accuracy in SMRT® sequencing. Available from: *http://www.pacificbiosciences.com/pdf/Perspective_UnderstandingAccuracySMRTSequencing.pdf*, accessed on June 30, 2015.

78. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009; *27* : 847-50.

79. Thomas SK, Iravatham CC, Moni BH, Kumar A, Archana BV, Majid M, *et al*. Modern and ancestral genotypes of *Mycobacterium tuberculosis* from Andhra Pradesh. *PLoS One* 2011; *6* : e27584.

80. Madhavilatha GK, Joseph BV, Paul LK, Kumar RA, Hariharan R, Mundayoor S. Whole-genome sequences of two clinical isolates of *Mycobacterium tuberculosis* from Kerala, South India. *J Bacteriol* 2012; *194* : 4430.

81. Narayanan S, Deshpande U. Whole-genome sequences of four clinical isolates of *Mycobacterium tuberculosis* from Tamil Nadu, south India. *Genome Announc* 2013; *1* : pii:e00186-13.

82. Dippenaar A and Warren RM. Fighting an old disease with next-generation sequencing. *eLife* 2015; *4* : 1-3.

83. Das S, Roychowdhury T, Kumar P, Kumar A, Kalra P, Singh J, Singh S, *et al*. Genetic heterogeneity revealed by sequence analysis of *Mycobacterium tuberculosis* isolates from extra-pulmonary tuberculosis patients. *BMC Genomics* 2013; *14* : 404.

84. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, *et al*. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med* 2013; *1* : 786-92.

85. Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, Rüsch-Gerdes S, *et al*. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. *PLoS One* 2013; *8* : e82551.

86. Myllykangas S, Buenrostro J, Ji HP. Overview of sequencing technology platforms. In: Rodriguez-Ezpeleta, N, Hackenberg M, Aransay AM, editors. *Bioinformatics for high throughput sequencing.* New York: Springer-Verlag; 2012. p. 255.

87. Thompson JF, Steinmann KE. Single molecule sequencing with a heliscope genetic analysis system. *Curr Protoc Mol Biol* 2010; Chapter 7 : 7.1.

88. Liu L, Li Y, Li S, Hu N, He Y, Pong R, *et al*. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012; *2012* : 25136-40.

89. Song L, Hobaugh MR, Shustak C, Cheley S, Bayley H, Gouaux JE. Structure of staphylococcal a-hemolysin, a heptameric transmembrane pore. *Science* 1996; *274* : 1859-66.

90. Deamer DW, Akeson M. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol* 2000; *18* : 147-51.

91. Manrao E, Derrington I, Pavlenok M, Niederweis M, Gundlach J. Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS One* 2011; *6* : e25723.

92. Butler TZ, Pavlenok M, Derrington I, Niederweis M, Gundlach J. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc Natl Acad Sci USA* 2008; *106* : 20647-52.

93. Current research on nanopore sequencing. Available from: *http://www.genome.gov/27545107*, accessed on June 29, 2015.

94. Anastasis D, Pillai G, Rambiritch V, Abdool Karim SS. A retrospective study of human immunodeficiency virus infection and drug resistant tuberculosis in Durban, south Africa. *Int J Tuberc lung Dis* 1997; *1* : 220-4.

95. Bishai WR, Graham NM, Harrington S, Pope DS, Hooper N, Astemborski J, *et al*. Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. *JAMA* 1998; *280* : 1679-84.

96. Iorger,T.R. Genome analysis of multi and extensively drug resistant tuberculosis from Kwazulu-natal South Africa. *PLoS One* 2009; *4* : e7778.

97. Casali N Micro-evolution of extensively drug resistant tuberculosis in Russia. *Genome Res* 2012; *22* : 735-45.

98. Fortune S.M. The surprising diversity of *Mycobacterium tuberculosis* - Change you can believe in. *J Iinfect Dis* 2012; *206* : 1642-4.

Reprint requests: Dr Sujatha Narayanan, National Institute of Research in Tuberculosis (ICMR),
1, Sathyamurthy Road, Chetput, Chennai 600 031, Tamil Nadu, India
e-mail: sujatha.sujatha36@gmail.com