Published in final edited form as:
Future Microbiol. 2011 February ; 6(2): 203–216. doi:10.2217/fmb.10.165.

Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies

Midori Kato-Maeda[1,†], **John Z. Metcalfe**[1], and **Laura Flores**[1]

[1]University of Colifornia, Son Francisco, Francis J Curry Notional Tuberculosis Center, Division of Pulmonary & Critical Care Medicine, Son Francisco General Hospital, 1001 Potrero Avenue, Building 100, Room 109, Mail box 0841, San Francisco, CA 94110-0111, USA

Abstract

Genotyping is used to track specific isolates of *Mycobacterium tuberculosis* in a community. It has been successfully used in epidemiologic research (termed ‘molecular epidemiology’) to study the transmission dynamics of TB. In this article, we review the genetic markers used in molecular epidemiologic studies including the use of whole-genome sequencing technology. We also review the public health application of molecular epidemiologic tools.

Keywords

IS*6110* RFLP; MIRU-VNTR; molecular epidemiology; public health; spoligotyping; TB; whole-genome sequencing

It is estimated that *Mycobacterium tuberculosis* causes latent infection in one-third of humanity. However, only 5–10% of infected individuals, 90% of whom live in low- or middle-income countries, will ultimately go on to develop active and infectious disease. Disease caused by *M. tuberculosis* still claims approximately 1.3 million lives per year [201] and represents a long-standing public health catastrophe. The transmission of *M. tuberculosis* from the index patient to his/her contacts is influenced by multiple factors. These include characteristics of the index patient (i.e., bacillary load [1]) and conditions of the environmental air shared by the patient and potential contacts [2]. The absolute risk of developing active TB among otherwise healthy persons is highest during the first 2 years following infection, when 3–10% of newly infected persons will develop active, infectious TB [3]. Differentiation of patients who rapidly evolve to active disease following recent infection from those who reactivate following remote infection is an important measure of regional TB control efforts.

[†]Author for correspondence: Tel +1 415 206 8121, Fax +1 415 695 1551, midori.kato-maeda@ucsf.edu.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Genotyping of *M. tuberculosis* isolates is primarily used to differentiate between recently transmitted and reactivation disease. In population-based studies, isolates that share the same genotype are considered clustered and are assumed to be epidemiologically linked, although the link may be indirect. By contrast, cases with isolates of a unique genotype not shared by other isolates within the population are considered to have resulted from reactivation of latent infection, presumably acquired either outside of the population or time period of interest. The assumption is that the rate of change of the molecular marker used to determine the genotype is rapid enough to show variation in a local community but slow enough that it is unlikely to change within a person in a shorter time period. If a patient has recurrent TB (more than one episode of active TB) and the isolates from the initial and the recurrent case are available, it is also possible to differentiate between relapse (TB caused by the same strain that caused the previous episode) and re-infection TB (TB caused by a different strain), Since the early 1990s, genotyping of *M tuberculosis* has been successfully used in epidemiologic research in a scientific field known today as molecular epidemiology [4]. This field has enabled TB control programs, mainly in high-income settings, to track specific isolates of *M. tuberculosis* in a community [5]. The obtained knowledge has high public health importance as it allows these programs to determine population-level risk factors for transmission, establish tailored public health strategies and gauge the success of control measures [6].

In contrast to genetic markers that change rapidly enough to trace recent transmission in a community, genotyping has also been used to characterize the bacterial evolution and phylogeny of *M. tuberculosis* [7]. This information has been used to gain greater understanding of pathogen-specific risk factors for transmission and pathogenesis. Until a decade ago, host and environmental factors were thought to account for variability in *M. tuberculosis* transmission and pathogenesis. For example, patients infected with HIV are 20–37-times more likely than people not infected with HIV to develop active TB [202]. However, recent studies suggest that different isolates of *M. tuberculosis* may contribute to different clinical outcomes [8,9], an observation made possible by recently discovered genetic markers representing a robust phylogenetic classification of *M. tuberculosis.*

The aim of this article is to review the genetic markers used in molecular epidemiologic studies of *M. tuberculosis* and their applications to public health. We also review the genetic polymorphisms used in phylogenetic analyses in the context of molecular epidemiology.

Genetic polymorphism in *M. tuberculosis*

Mycobacterium tuberculosis is part of the *M. tuberculosis* complex, which also includes *M. africanum* subtype 1, *M. pinnipedii, M. microti, M. bovis* subsp. *caprae* and *M. bovis* [10], all of which are mote than 99.9% similar at the DNA level. Genetic analysis suggests that the ancestor of the modern human-adapted *M. tuberculosis* complex is *M. canetti*. This is a rare bacillus that forms smooth colonies and has been isolated only in patients from Djibouti. East Africa. The genetic analysis of fragments of the complete 16S rRNA gene and six housekeeping genes (*katG, gyrB, gyrA, rpoB, hsp65* and *sodA*) in several clinical isolates of *M. canetti* showed none or one single nucleotide polymorphism (SNP) in the 1537 base pairs of the 16S rRNA sequence and greater diversity in other parts of the genome when compared with isolates of the *M. tuberculosis* complex [11]. These data suggest that *M. canetti* underwent an extreme bottleneck and that the current *M. tuberculosis* complex emerged in East Africa from a single successful lineage that expanded clonally to the rest of the world. Recently published papers have detailed the origin and evolution of *M. tuberculosis* [10,11].

Mycobacterium tuberculosis is considered genetically monomorphic because it has low levels of genetic diversity and homoplasies (independent mutational events that result in the same genotype among isolates with a different ancestry) and very rare homologous recombination events. However, relative to other monomorphic bacteria (i.e., *Salmonella enterica* or *Yersinia pestis)*, *M. tuberculosis* has substantial genetic variation [12], including large sequence polymorphisms (LSPs) [13] and SNPs, which are phylogenetically informative and useful for population genetic analyses. Further variation can be attributed to a variable number and location of the insertion element (IS) *6110* [14] and the polymorphic GC-rich repetitive sequences (PGRSs), as well as polymorphisms in the clustered regularly inters paced short palindromic repeats (CRISPRs) [15] and variable number tandem repeats (VNTR) [16], all of which have been commonly employed in molecular epidemiology.

Genetic markers in *M. tuberculosis*

Genetic markers have been used for molecular epidemiology and for phylogeny studies of *M. tuberculosis* (Table 1). For molecular epidemiology, it is important that the markers are highly discriminatory. For phylogeny and population genetic analyses, the markers should be phylogenetically robust, that is, exhibit low homoplasy and minimal rates of convergent evolution. It should also have a solid phylogenetic framework with a strong understanding of the genetic population structure in which specific strains can be positioned and could delineate biologically meaningful groups.

To date, no single genetic marker is equally suitable for both of these purposes. Therefore, it is important to select an adequate genetic marker and corresponding analysis only after clearly identifying the goals of genotyping.

Genetic markers for molecular epidemiologic studies

The ideal DNA market for the study of TB transmission dynamics is one that is polymorphic enough to distinguish among unrelated isolates, yet stable enough to make the connection between isolates that are indeed related. The ideal methodology to determine the genetic polymorphism should be simple, affordable, have a rapid turnaround time and the results should be in a format that can be easily shared between different laboratories. In this section, we will review the different DNA markers used in the study of TB transmission dynamics.

In the early 1990s, the restriction fragment length polymorphism (RFLP) method using IS*6110* was successfully used to identify and track individual isolates of *M. tuberculosis* in the community [17]. Later, PCR-based methods such as spoligotyping, or spacer oligonucleotide typing (based on the polymorphism, of the CRISPRs) [18], and mycobacterial interspersed repetitive units (MIRU)-VNTR typing [19] were added to the armamentarium of molecular epidemiological tools. Most recently, the availability of high-throughput technology and the dramatic decrease in costs have allowed for whole-genome sequencing as a viable method to study community transmission and microevolution of *M tuberculosis* [20,21].

IS*6110*-RFLP—IS*6110*-RFLP has been considered, until recently, to be the gold standard of molecular epidemiologic studies owing to its discriminatory power (ability to differentiate between two unrelated strains) and because it has been widely used since the early 1990s [5]. IS*6110* belongs to the IS*3* family of mobile elements. It has 1361 base pairs and is found only in organisms of the *M. tuberculosis* complex. *M. tuberculosis* has been shown to contain between 0 to 25 copies of IS*6110.* Although IS*6110* does not have a known target for insertion, it is believed that the insertion sites are not random. Several preferential integration loci or hotspots have been reported, some of them in recent years [22]. The molecular clock (rate of change for which two *M. tuberculosis* strains will diverge) of the IS*6110* element is estimated to be between 3.2 and 8.7 years [23–25]. The genotyping method is based on the variability of the number of copies of IS*6110* and the molecular weights of DNA fragments in which the insertions are found [17].

The main advantages of the IS*6110*-RFLP method are its high discriminatory power and the availability of studies for comparison. A main limitation of the IS*6110*-RFLP method is the low discriminatory power in isolates presenting five or fewer IS*6110* bands. In fact, studies have demonstrated that the frequency of clinically confirmed epidemiologic links between cases is decreased in clusters formed by isolates with five or fewer IS*6110* bands [26,27]. For these isolates, secondary markers such as PGRS and CRISPRs are used to increase the discriminatory power and to determine the epidemiological links among patients. Furthermore, IS*6110*-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.

PGRS-RFLP—Polymorphic GC-rich repetitive sequence RFLP has been used as a secondary typing method in isolates of *M. tuberculosis* with five or fewer copies of IS*6110* [28]. PGRS is a highly polymorphic GC-rich sequence present in multiple sets of copies on the *M. tuberculosis* chromosome. The evolution of PGRS containing regions is due to duplication, recombination and strand slippage [29], however, these mutations rarely induce changes that can be observed in the PGRS-RFLP image [30]. It has been estimated that the rate of change of PGRS is slower than that of IS*6110*, however, exact estimates are not available [31]. The PGRS is visualized using an RFLP-based methodology [27] and a probe based on the consensus sequence of the PGRS: ATCGGCAACGGCGGCAACGGCGGCAACGGCGG. The advantage of this technique is that it is highly discriminatory [28]. The disadvantages are similar to the technical limitations of IS*6110*-RFLP and PGRS-RFLP produces an image with many more bands of different intensities, complicating its reading and interpretation.

Spoligotyping—Spoligotyping has been used for secondary typing of isolates with five or fewer IS*6110* bands [26,32,33]. It has also been used as a primary genotyping method in combination with MIRU-VNTR [34,35]. Spoligotyping is based on the polymorphism in the direct repeat (DR) locus, which is a member of the CRISPRs [36]. This region consists of 36 base pair DR copies interspersed by non-repetitive 35–41 base pair sequences called spacers. The DR and the spacers together are called direct variable repeats (DVRs). There are 94 spacer sequences; however, 43 spacers are used in the most common typing methodology [18]. The DR region evolves through IS*6110*-mediated mutation, homologous recombination between repeat sequences that lead to the deletion of a DVR, strand slippage that lead to duplication of DVRs, and point mutations [37,38]. It has been estimated that the rate of change of the DR region is slower than that of IS*6110;* however, exact estimates are not available [39]. Spoligotyping is a PCR-based method that amplifies the spacer; the presence and absence of spacers will result in different polymorphisms. There are two methods currently used to obtain the spoligotype. The first one (the original method.) is based on reverse hybridization where the sequences of each of the spacers are attached to specific areas on a membrane and the PCR products are hybridized to the membrane [40]. The patterns (presence and absence of spacers) are represented as a binary number or, after a simple conversion, as an octal numeral. The second method is based on Luminex technology (Luminex Technology, TX, USA), where each of the spacer oligonucleotides is covalently attached to a microsphere, which serves as a capture probe. Each microsphere also contains two covalently attached fluorochromes. One laser will excite the fluorochromes to identify the spacer. The second laser will excite a reporter bound to the PCR product containing the spacer [41].

The advantages of 43 spacer-spoligotyping are that it is highly reproducible and, since it is based on PCR amplification, very little DNA (20–50 ng) is needed. It has been used with DNA extracted directly from sputum from smear-positive patients [42]. The simple reporting method (binary or octal numeral) has allowed the creation of a global database of spoligotyping patterns [43,203]. The main limitation is the inferior discriminatory power when compared with IS*6110*-RFLP and MIRU-VNTR typing. For example, isolates that have the spoligotype lacking the spacers 1–34 (also known as the Beijing family) may have different IS*6110*-RFLP and MIRU-VNTR polymorphisms [44]. To overcome this limitation, an extended panel of 68 spacer sequences was recently tested using the membrane format and Luminex technology [45]. The 68 panel included the original 43 spacers and 25 additional spacers. The extended panel of 68 spacer sequences was compared with the 43 original spacers in a convenience sample of 351 isolates. The 68 spacer-typing improved discrimination from 33 to 52 spoligotyping patterns.

Recently, we compared the performance of PGRS-RFLP with the 43 spacer-spoligotyping for secondary typing of *M. tuberculosis* isolates with five or fewer IS*6110* bands to examine the community epidemiology of TB in a population-based study [32]. Our data indicate that PGRS-RFLP and the 43 spacer-spoligotyping had similar discriminatory power for isolates with five or fewer IS*6110* bands and the cluster status of isolates were concordant in 84% of the cases. However, patients were included in different clusters depending on which genotyping method was employed. Given these data and lack of a clear gold standard, these methods should be considered non-interchangeable, and the same method should be used for longitudinal studies.

Over the years, signature 43 spacer-spoligotyping patterns have been identified and used to define strain families [43]. The families largely correspond to the geographic regions from which the isolates were collected, demonstrating that different spoligotypes predominate in different pans of the world [46]. Examples of these strain families include the Beijing family, which is highly prevalent in Eastern Asia [47], the Manila family in the Philippines [48], the Cameroon family [49], and the Central Asia/Delhi family [50]. Spoligotyping also discriminates between *M. tuberculosis* and *M. bovis*, which lacks spacers 39–43 [18]. Although the spacers used in spoligotyping exhibit a high rate of homoplasy (independent mutational events that result in the loss of the same spacer), we recently demonstrated that strains from a particular spoligotype family belonged to the same LSP/SNP lineage and that spoligotype families should be considered ‘sub-lineages’ within the main LSP/SNP lineages [51].

MIRU-VNTR typing—Mycobacterial interspersed repetitive unit VNTR is considered by some authors to be the new gold standard for *M. tuberculosis* genotyping, as it is highly discriminatory and reproducible [52]. This genotyping method is based on VNTRs of the genetic elements called MIRUs. The repetitive units are 40–100 base pairs in length and are located in 41 loci scattered throughout the genome of *M. tuberculosis* H37Rv [16]. The rate of mutation of each locus is variable [53–56]. The polymorphism is based on the variability in the number of copies of the repeat unit. The original methodology included 12 MIRU loci and was used together with spoligotyping for primary genotyping. However, the discriminatory power was less than that of IS*6110*-RFLP, especially in isolates with the Beijing spoligotype [53,57]. The current recommendations are to use a set of 15 MIRU loci for molecular epidemiologic studies and 24 MIRU loci for phylogenetic studies (the original set of 15 MIRU and nine additional loci) [56]. Special consideration is given to MIRU typing of Beijing isolates, where the set of 15 MIRUs are not sufficient to discriminate among unrelated Beijing isolates [53,58]. Therefore, hypervariable VNTR targets [53] or sets of 24 or more MIRU loci [59–61] are recommended to analyze this family of strains (Table 2).

Mycobacterial interspersed repetitive unit VNTR typing is based on PCR amplification using primers specific for the flanking regions of the different MIRUs. The original protocol includes the amplification of each locus and the visualization of the product in a gel. Because the length of the repeat unit is known, the size of the PCR product will reflect the number of copies of the repeat unit. The result is a numerical value that reflects the number of repeats in tandem at each locus [19]. At present, there is a high-throughput method based on multiplex PCR in which one primer of each primer set is tagged with a different fluorescent dye. The fluorescently labeled amplicons are subjected to electrophoresis using an automatic sequencer in order to estimate the PCR size [62]. The advantage of automated MIRU typing is that it is highly reproducible and fast because the results are obtained through a computerized analysis of the signals (in contrast with the visualization of each product in agarose gel, which is less reproducible and more time consuming). However, it requires a sequencer and specialized software packages [56]. A global epidemiological database is available [52,204]. The analysis of this data has led to insights into the distribution and evolution of *M. tuberculosis*, including the identification of clonally related *M. tuberculosis* families in specific geographic distributions [63–65].

Mycobacterial interspersed repetitive unit VNTR typing has similar or more discriminatory power than IS*6110*-RFLP, depending on the MIRU-VNTR locus number and combinations [44,66,67]. However, the concordance of the two methods to determine linked cases may vary depending on the genotyping method used (i.e., patients may be clustered by both methods, but the cases with whom they are linked are different) [35,44,54]. In some studies up to 85% of isolates were clustered with the same patients by both methods [35,54], but in some instances this concordance was as low as 40%, specifically when using 27 MIRU-VNTR ser genotypes among isolates of *M. tuberculosis* of the Beijing family [44]. Nevertheless. MIRU-VNTR typing is considered the new gold standard for molecular epidemiological studies owing to its acceptable discriminatory power and the easily exchangeable format of the data.

Drug resistance-associated mutations—Recently, drug resistance-associated mutations have been used together with other genetic markers to define clustering of drug-resistant *M. tuberculosis* isolates [68,69]. Drug resistance-associated mutations have also been useful in evaluating the impact of specific mutations on the transmission and secondary case generation of drug-resistant strains. For example, it has been found that isolates resistant to isoniazid and that have a serine to threonine substitution in codon 315 of the *katG* gene are more likely to cause secondary cases than strains with other mutations in *katG* [68–71]. Drug-resistance mutations are determined by amplification and sequencing of a target amplicon, such as *katG* for isoniazid resistance [72], or by reverse hybridization of the target amplicon to probes representative of the possible mutations causing drug resistance [73]. At present, there are no data about the molecular clock or the overall discriminatory power of the drug-resistance mutations.

Whole-genome sequencing—The most recent breakthrough in the field of molecular epidemiology is the availability of relatively affordable, high-throughput whole-genome sequencing (WGS) technology. The data generated by IS*6110*-RFLP, spoligotyping and VNTR-MIRU typing can determine which patients form part of a chain of transmission, but do not always allow us to distinguish the exact transmission chain of events (i.e., order in which patients were infected, or the presence of multiple index cases). Recently, Schurch *et al*. performed WGS in three isolates (from 1992, 2004 and 2006) of *M. tuberculosis*, that had the same IS*6110*-RFLP genotyping pattern [20,21]. These isolates were part of a prevalent cluster in the community that was composed of 104 cases identified between 1992 and 2008. Owing to the number of patients involved, the index patient(s) and the transmission events could not be determined. The authors found eight SNPs among the three sequenced strains. These SNPs were subsequently investigated in all 104 isolates of the cluster, resulting in the identification of secondary and tertiary index patients.

In another study, Niemman, *et al*. performed WGS of two clinical isolates of *M. tuberculosis* that had matching IS*6110* RFLP and spoligotype (Beijing family) and a similar MIRU-VNTR profile (1 of 24 MIRU-VNTR with one copy difference) [74]. These strains originated within a large cluster of prevalent strains in Karakalpakstan, Uzbekistan. One strain was susceptible to isoniazid, rifampin, ethambutol, pyrazinamide and streptomycin, while the other was resistant to all five drugs. The authors reported that the two isolates differed in 130 SNPs (including putative drug-resistance mutations) and one large deletion, suggesting that the epidemiological link between these strains may have been remote.

Whole-genome sequencing will be an important molecular epidemiologic tool in that it will have the capability to determine sequence variation at a real epidemiological scale, to identify the source(s) of infection and the transmission events among individuals that share the same *M. tuberculosis* isolate, and to determine the evolutionary relationship between isolates. Based on the few studies available, WGS may also provide additional resolution to strains with matching IS*6110* RFLP, spoligotype or MIRU-VNTR profiles to more accurately distinguish patients who are part of a recent chain of transmission from those with reactivation disease related to distant infection. Such enhanced resolution may prove important for drug and vaccine development in high burden areas with a large proportion of endemic strains. It is possible that WGS may become the gold standard for strain typing for molecular epidemiological studies [75], though efforts to address several limitations are needed. These limitations include cost (which has recently decreased substantially), the need for specialized software for data analysis, and incomplete understanding of the molecular clocks of SNPs and LSPs. Further, the potential impact of WGS on patient care or design of public health strategies remains to be determined.

Genetic markers for phylogenetic analysis

In this section, we briefly describe genetic markers used for phylogenetic analysis of *M. tuberculosis*. These markers are useful in classifying the isolates of *M. tuberculosis* into families or lineages with a robust phylogenetic structure. There is evidence suggesting that different lineages may have different phenotypic characteristics, including lineage-specific effects on the outcome of TB transmission and disease in clinical settings [76–78].

Large sequence polymorphisms—Large sequence polymorphisms are unique event polymorphisms that define a phylogeny with a geographic structure from which the *M. tuberculosis* lineage names were derived: Euro–American, East-Asian, Indo-Oceanic, East-African–Indian and the West-African lineages 1 and 2 (also known as *M. africanum*) [79]. The robustness of this phylogeographic classification has been recently confirmed using multilocus analysis of 89 complete genes in 108 strains, [80] as well as with WGS [81].

Multilocus sequence analysis—Multilocus sequence analysis is a technique for the typing of multiple loci in *M. tuberculosis*. It is based on the DNA sequence of multiple genes. Hershberg *et al*. recently performed multilocus sequence analysis using the complete coding sequences of 89 genes, corresponding to approximately 70 kb per strain [80].

Implications at different *M. tuberculosis* lineages on the use of molecular markers

There is evidence that the performance of genetic markers used in molecular epidemiologic techniques differs according to specific LSP-based lineages. Recently, we used a novel expectation-maximization algorithm to estimate the birth and death of IS*6110* elements in *M. tuberculosis*. We analyzed the IS*6110*-RFLP patterns of 196 patients with two or more isolates obtained 10 or more days apart. We found a substantial difference between death rates (deletions) of the IS*6110* from the Euro–American (0.036 units of changes per copy of transposon per year, 95% CI: 0.015–0.056) and East-Asian lineages (0.004 units of changes per copy of transposon per year, 95% CI: 0–0.011). In other words, if a hypothetical isolate with ten copies belongs to the Euro–American lineage, there will be 0.3 changes per year or one change in approximately 3 years. By contrast, if the hypothetical isolate with ten copies is East Asian, it will take ten times as long to observe one change. The birth rate (new insertions) was similar among these lineages. We did not find any evidence that the number of IS*6110* bands confounded our analysis [Doss CR *ET AL*., SUBMITTED MANUSCRIPT]. These data suggest that the molecular clock of IS*6110* may differ among the various lineages, though the implications of this finding on molecular epidemiologic studies is unclear.

The discriminatory power of loci used in the MIRU-VNTR technique also may differ according to lineage [82]. This has been studied mainly in *M. tuberculosis* from the Beijing family (East-Asian lineage) in which the discriminatory power is low in several of the MIRU-VNTR loci from the MIRU 12 and MIRU 15 sets. As a result, the use of MIRU 24 or a combination of different loci has been suggested when typing isolates from this lineage (Table 2). The 43 spacer-based spoligotype of *M. tuberculosis* isolates from the Beijing family lacks spacers 1–34; therefore, the probability of detecting polymorphisms is limited to the presence or absence of nine DVRs [60,83]. We recently analyzed the performance of spoligotyping in combination with the MIRU 12 set. Compared with IS*6110*-RFLP, the performance of spoligotyping/MIRU-12 was lower among isolates of East-Asian lineage (as expected), though its performance among the isolates of Euro–American lineage was similar.

Public health application of molecular epidemiologic tools

The evaluation of individuals in close contact with an infectious TB patient in order to identify secondary cases of active TB and latent TB infection (i.e., contact investigation) has been an integral component of the public health approach to TB control in high-income, low TB burden countries for decades [84]. Molecular epidemiology of *M. tuberculosis* was introduced as a research tool and supplement to traditional contact investigation in the early 1990s [85]. Multiple ‘universal’ regional genotyping programs have since been initiated in high-income settings [86–91] with the aim to enhance TB control activities through identification of previously unrecognized chains of transmission, monitor disease trends including drug resistance, and provide for more precise allocation of public health resources [92]. Molecular genotyping has also been crucial in estimating relative trends in reactivation and transmission of TB within and between native and immigrant subpopulations [84,91,93,94], and such findings have been useful in coordinating TB control activities [6,93]. In high burden settings where the majority of transmission is thought to occur outside the household in community settings [95], contact investigation, along with routine culture of isolates, is rarely available. In these settings regional genotyping programs have been incorporated into research programs but not evaluated as an adjunct to routine TB control activities [96,97].

Applications in patient management

Genotyping has been shown to have added value in routine patient management and contact investigation in establishing epidemiologic linkages, especially within nontraditional groups or settings [98–100], and in outbreak situations [101–105]. In addition, molecular epidemiologic tools have been used to identify and quantify laboratory cross-contamination [106], as well as to distinguish recurrent disease from exogenous re-infection [107,108]. This latter distinction, along with the differentiation of strain-specific elicited immune response, has important implications for vaccine development. The impact of genotyping programs on patient-important outcomes (e.g., case mortality, treatment delay or case prevention) has not been studied in any setting.

Epidemiologic measure of transmission dynamics

Within a closed population, the proportion of clustered *M. tuberculosis* strains directly estimates the proportion of cases attributable to recent transmission and rapid progression to active disease [85]. Thus, a decrease in the number of clustered cases is considered an important measure of the success of regional TB control [6]. Population-level risk factors for recent transmission are typically assessed through comparison of ‘clustered’ (shared *M. tuberculosis* genotypes among two or more patients) and ‘unique’ cases (unmatched genotypes, assumed to have resulted from reactivation of distantly acquired infection). Demographic and clinical risk factors for transmission and pathogenesis of both drug-susceptible and drug-resistant *M. tuberculosis* have been well described in multiple populations [109,110], though recently, molecular techniques have also sought to clarify the contribution of pathogen-specific factors in the generation of secondary TB cases [68,69,111].

Influence of strain lineage on *M. tuberculosis* transmission & pathogenesis

Molecular epidemiologic tools have demonstrated that *M. tuberculosis* strains may vary in their ability to transmit and rapidly progress to active disease (clustered cases) or be associated with major outbreaks [112]. Differential clinical outcomes have traditionally been considered to be predominantly associated with host and environmental differences among cases. More recently, strain diversity has also been implicated as a possible causal factor influencing clinical outcomes [7].

The evidence suggesting that *M. tuberculosis* strains from different lineages have different phenotypic properties comes from a variety of sources. Several *in vitro* and animal studies demonstrate strain-specific differences in immunogenicity and virulence [113–115]. Recently, clinical studies have demonstrated lineage-specific effects on the outcome of TB infection and disease in various settings [76–78]. In the Gambia, transmission of *M. tuberculosis* to household contacts (measured using skin test conversions) was similar among *M. tuberculosis* strains from different lineages. However, the proportion of contacts developing active TB within the 2 year follow-up period varied; 1% for those exposed to strains of *M. africanum* (Lineage West Africanum 1 and 2), 5.6% for those exposed to strains from the East-Asian family, and 1.2–3.9% for strains from the different sublineages that compose the Euro–American lineage [77]. In San Francisco, USA, we found evidence that strains from different East-Asian sublineages of *M. tuberculosis* have varying frequencies of genotypic clustering [116]. In a population-based cohort study of all incident multidrug-resistant TB cases occurring in California over a 4-year period, the East-Asian lineage was associated with an elevated proportion of genotypic clustering while Indo-Oceanic strains produced no secondary cases [69]. Taken together, these studies suggest that strains from different lineages may have an impact on clinical outcomes. However, owing to effects of the complex interaction of host, pathogen and environment on the transmission and pathogenesis of TB, a systems epidemiology approach (i.e., multidisciplinary approach combining systems biology with epidemiology) may be needed to determine the contribution of each element to the TB epidemic [117].

In addition to variable transmissibility and pathogenicity, distinct clinical outcomes have been associated with certain lineages. In Vietnam, *M. tuberculosis* strains of the Euro–American lineage were positively associated with pulmonary TB and less likely to cause tuberculous meningitis [76]. Patients with tuberculous meningitis caused by strains of the-Asian lineage had a shorter duration of disease and fewer lymphocytes in their cerebrospinal fluid at presentation, suggesting a differential intracerebral inflammatory response [78].

Interpretation of molecular epidemiologic studies

The limitations of genotypic methods in the setting of endemic strains, low copy number and need for culture isolates has been discussed [118]. The analysis and interpretation of molecular epidemiologic studies require at least two special considerations. First, molecular epidemiologic studies often have a hierarchical data structure (e.g., individual clustered cases occurring within households). A fundamental assumption of traditional statistical methods (e.g., logistic regression) is that the analyzed events are independent. Because this assumption is often violated in molecular epidemiologic studies, additional statistical adjustments, such as generalized estimating equations [119], may be necessary to provide unbiased estimates of risk. Second, molecular epidemiology studies must take into consideration sampling bias. Misclassification may occur when clustered cases are falsely interpreted as unique (e.g., when other cases in a particular cluster are not included due to undersampling) or when unique cases are falsely interpreted as clustered (e.g., as would occur in epidemiologic settings with a high proportion of endemic strains). Simulations have demonstrated that the extent to which sampling of the population base is incomplete affects the proportion of cases truly identified as clustered decreases and the variance of the estimate increases, with the effect of such incomplete sampling increasing as median cluster size decreases [120,121]. Similarly, increasing the length of study duration would be expected to increase cluster proportion, though the probability of clustering truly reflecting recent transmission and rapid progression to disease decreases with time. ‘Cluster windows’ (usually 1–2 years to reflect the increased probability of recently infected individuals developing active TB during this time) define the maximum allowable time between two matching strains in molecular epidemiology studies, and are meant to improve specificity in established chains of transmission. The effect of misclassification on specificity is dependent on the population studied, pathogen factors and the molecular clock of the specific genotyping element. Specificity is reduced in the presence of endemic strains occurring either within the local population [122,123] or among immigrants from other populations, or in settings where the genetic diversity of *M. tuberculosis* is known to be low

[89,124]. In addition, there have been several reports of strains with small variations in one [37] or multiple genotyping methods [125], which were concluded to be clonal after careful consideration of the clinical and epidemiological data. Genetic differences in these cases were likely the product of microevolutionary events of *M. tuberculosis* occurring during transmission [37] or during dissemination within an individual patient [125]. Therefore, clinical and epidemiological data is important in some cases in order to corroborate molecular genotypes.

Future perspective

Molecular epidemiologic studies of TB are intended to identify risk factors for transmission and for active TB in order to design effective preventive interventions both at the individual and population level. The translation of molecular epidemiologic studies of *M. tuberculosis* into ‘real-time’ clinical management, however, remains a challenge. Information regarding cluster/unique status of the bacteria is often not available until 2–3 months following patient diagnosis, and, consequently, such data has been mainly employed in retrospective analyses. ‘Real-time’ molecular epidemiologic data may allow for early identification of *M. tuberculosis* clusters and tailored contact investigation strategies. Therefore, it will be important to continue to shorten the time between patient diagnosis and the availability of molecular epidemiologic data. Strain classification through robust phylogenetic polymorphisms has recently made association of *M. tuberculosis* genetic markers with important clinical consequences, such as high transmissibility, treatment failure or elevated propensity to acquire primary drug resistance possible. This information may ultimately also prove useful for the clinical management of patients. A remaining challenge is how to translate data obtained from molecular epidemiologic studies into a public health database that could be integrated through statistical modeling with host information, other pathogen characteristics, and environmental and sociological data to aid prediction of future epidemics involving specific strains or groups of strains.

Some of the challenges discussed here have the potential to be overcome with data obtained through WGS. The comparison of the whole-genome sequences of well-characterized isolates may result in identification of genetic markers associated with important clinical outcomes. Also, a more thorough understanding of the rate of acquisition of SNPs among the clinical isolates of *M. tuberculosis* will allow determination of temporal evolutionary relationships of isolates and dynamics of epidemics involving specific strains. The expectations for the information that WGS technology can provide are very high. Thus, well designed epidemiologic studies will be required to affirm that this technology can demonstrate patient- and population-level benefit.

Bibliography

Papers of special note have been highlighted as:

▪ of interest

▪▪ of considerable interest

1. Rauillon A, Perdrizet S, Parrot R. Transmission of tubercle bacilli: the effects of chemotherapy. Tubercle. 1976; 57(4):275–299. [PubMed: 827837]

2. Houk VN, Baker JH, Sorensen K, Kent DC. The epidemiology of tuberculosis infection in a closed environment. Arch Environ Health. 1968; 16(1):26–35. [PubMed: 5638222]

3. Centers for Disease Control and Prevention. Targeted tuberculin testing and treatment of latent tuberculosis infection. American Thoracic Society. MMWR Recomm Rep. 2000; 49(RR-6):1–51.

4. Foxman B, Riley L. Molecular epidemiology: focus on infection. Am J Epidemiol. 2001; 153(12): 1135–1141. [PubMed: 11415945]

5. 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(8):892–909. [PubMed: 19702595]

6. Cattamanchi A, Hopewell PC, Gonzalez LC, et al. A 13-year molecular epidemiological analysis of tuberculosis in San Francisco. Int J Tuberc Lung Dis. 2006; 10(3):297–304. [PubMed: 16562710]

7. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog. 2009; 5(10):E1000600. [PubMed: 19855821]

8. Malik AN, Godfrey-Faussett P. Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease. Lancet Infect Dis. 2005; 5(3):174–183. [PubMed: 15766652]

9▪. Nicol MP, Wilkinson RJ. The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. Trans R Soc Trop Med Hyg. 2008; 102(10):955–965. Review of the clinical consequences of strain diversity in *Mycobacterium tuberculosis*. [PubMed: 18513773]

10. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol. 2009; 7(7):537–544. [PubMed: 19483712]

11. Gutierrez C, Brisse S, Brosch R, et al. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog. 2005; 1:1–7.

12▪▪. Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol. 2008; 62:53–70. Review of genetically monomorphic bacteria. [PubMed: 18785837]

13. Tsolaki AG, Hirsh AE, DeRiemer K, et al. Functional and evolutionary genomics of *Mycobacterium tuberculosis:* insights from genomic deletions in 100 strains. Proc Natl Acad Sci USA. 2004; 101(14):4865–4870. [PubMed: 15024109]

14. McEvoy CR, Falmer AA, Gey van Pittius NC, Victor TC, van Helden PD, Warren RM. The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. Tuberculosis (Edinb). 2007; 87(5):393–404. [PubMed: 17627889]

15. van Embden JD, van Gorkom T, Kremer K, Jansen R, van Der Zeijst BA, Schouls LM. Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J Bacteriol. 2000; 182(9):2393–2401. [PubMed: 10762237]

16. 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(3):762–771. [PubMed: 10844663]

17. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol. 1993; 31(2):406–409. [PubMed: 8381814]

18. Kamerbeek J, Schouls L, Kolk A, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol. 1997; 35(4):907–914. [PubMed: 9157152]

19. Mazars E, Lesjean S, Banuls AL, 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(4):1901–1906. [PubMed: 11172048]

20▪▪. Schurch AC, Kremer K, Daviena O, et al. High resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol. 2010; 48(9):3403–3406. Use of whole-genome sequencing in three isolates with the same IS*6110*-restriction fragment length polymorphism genotype to determine the molecular evolution of *M. tuberculosis*. [PubMed: 20592143]

21. Schurch AC, Kremer K, Kiers A, et al. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. Infect Genet Evol. 2010; 10(1):108–114. [PubMed: 19835997]

22. Kim EY, Nahid P, Hopewell PC, Kato-Maeda M. Novel hot spot of IS*6110* insertion in *Mycobacterium tuberculosis*. J Clin Microbiol. 2010; 48(4):1422–1424. [PubMed: 20147648]

23. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. Analysis of rate of change of IS*6110*-RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J Infect Dis. 1999; 180(4):1238–1244. [PubMed: 10479153]

24. Warren RM, van der Spuy GD, Richardson M, et al. Evolution of the IS*6110*-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. J Clin Microbiol. 2002; 40(4):1277–1282. [PubMed: 11923345]

25. Warren RM, van der Spuy GD, Richardson M, et al. Calculation of the stability of the IS*6110* banding pattern in patients with persistent *Mycobacterium tuberculosis* disease. J Clin Microbiol. 2002; 40(5):1705–1708. [PubMed: 11980946]

26. Soini H, Pan X, Teeter L, Musser JM, Graviss EA. Transmission dynamics and molecular characterization of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110*. J Clin Microbiol. 2001; 39(1):217–221. [PubMed: 11136774]

27. Yang ZH, Ijaz K, Bates JH, Eisenach KD, Cave MD. Spoligotyping and polymorphic GC-rich repetitive sequence fingerprinting of *Mycobacterium tuberculosis* strains having few copies of IS*6110*. J Clin Microbiol. 2000; 38(10):3572–3576. [PubMed: 11015365]

28. Rhee JT, Tanaka MM, Behr MA, et al. Use of multiple markers in population-based molecular epidemiologic studies of tuberculosis. Int J Tuberc Lung Dis. 2000; 4(12):1111–1119. [PubMed: 11144452]

29. Karboul A, Gey van Pittius NC, Namouchi A, et al. Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE_PGRS duplicated gene pair. BMC Evol Biol. 2006; 6:107. [PubMed: 17163995]

30. Richardson M, van der Spuy GD, Sampson SL, Beyers N, van Helden PD, Warren RM. Stability of polymorphic GC-rich repeat sequence-containing regions of *Mycobacterium tuberculosis*. J Clin Microbiol. 2004; 42(3):1302–1304. [PubMed: 15004103]

31. Yeh RW, Ponce de Leon A, Agasino CB, et al. Stability of *Mycobacterium tuberculosis* DNA genotypes. J Infect Dis. 1998; 177(4):1107–1111. [PubMed: 9534994]

32. Flores L, Jarlsberg LG, Kim EY, et al. Comparison of restriction fragment length polymorphism with the polymorphic guanine-cytosine-rich sequence and spoligotyping for differentiation of *Mycobacterium tuberculosis* isolates with five or fewer copies of IS*6110*. J Clin Microbiol. 2010; 48(2):575–578. [PubMed: 20032250]

33. Yang ZH, Bates JH, Eisenach KD, Cave MD. Secondary typing of *Mycobacterium tuberculosis* isolates with matching IS*6110* fingerprints from different geographic regions of the United States. J Clin Microbiol. 2001; 39(5):1691–1695. [PubMed: 11325975]

34. CDC. Notice to readers: new CDC program for rapid genotyping of *Mycobacterium tuberculosis* isolates. MMWR Morb Mortal Wkly Rep. 2005; 54(2):47. [PubMed: 16177693]

35. Oelemann MC, Diel R, Vatin V, 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(3): 691–697. [PubMed: 17192416]

36. Zhang J, Abadia E, Refregier G, et al. *Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of ‘spoligotyping’ with new spacers and a microbead-based hybridization assay. J Med Microbiol. 2009; 59(Pt 3):285–294. [PubMed: 19959631]

37. Aga RS, Fair E, Abernethy NF, et al. Microevolution of the direct repeat locus of *Mycobacterium tuberculosis* in a strain prevalent in San Francisco. J Clin Microbiol. 2006; 44(4):1558–1560. [PubMed: 16597893]

38. Warren RM, Streicher EM, Sampson SL, et al. Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. J Clin Microbiol. 2002; 40(12):4457–4465. [PubMed: 12454136]

39. Niemann S, Richter E, Rusch-Gerdes S. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J Clin Microbiol. 1999; 37(2):409–412. [PubMed: 9889229]

40. Driscoll JR. Spoligotyping for molecular epidemiology of the *Mycobacterium tuberculosis* complex. Methods Mol Biol. 2009; 551:117–128. [PubMed: 19521871]

41. Cowan LS, Diem L, Brake MC, Crawford JT. Transfer of a *Mycobacterium tuberculosis* genotyping method, spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J Clin Microbiol. 2004; 42(1):474–477. [PubMed: 14715809]

42. Cafrune PI, Possuelo LG, Ribeiro AW, et al. Prospective study applying spoligotyping directly to DNA from sputum samples of patients suspected of having tuberculosis. Can J Microbiol. 2009; 55(7):895–900. [PubMed: 19767863]

43▪▪. Brudey K, Driscoll JR, Rigouts L, et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 2006; 6:23. Description of the fourth international spoligotyping database, SpolDB4. [PubMed: 16519816]

44. Hanekom M, van der Spuy GD, Gey van Pittius NC, et al. Discordance between mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing and IS*6110* restriction fragment length polymorphism genotyping for analysis of *Mycobacterium tuberculosis* Beijing strains in a setting of high incidence of tuberculosis. J Clin Microbiol. 2008; 46(10):3338–3345. [PubMed: 18716230]

45. Zhang J, Abadia E, Refregier G, et al. *Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of ‘spoligotyping’ with new spacers and a microbead-based hybridization assay. J Med Microbiol. 2010; 59(Pt 3):285–294. [PubMed: 19959631]

46. Filliol I, Driscoll JR, van Soolingen D, et al. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J Clin Microbiol. 2003; 41(5):1963–1970. [PubMed: 12734235]

47. van Soolingen D, Qian L, de Haas PE, et al. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. J Clin Microbiol. 1995; 33(12):3234–3238. [PubMed: 8586708]

48. Douglas JT, Qian L, Montoya JC, et al. Characterization of the Manila family of *Mycobacterium tuberculosis*. J Clin Microbiol. 2003; 41(6):2723–2726. [PubMed: 12791915]

49. Niobe-Eyangoh SN, Kuaban C, Sorlin P, et al. Genetic biodiversity of *Mycobacterium tuberculosis* complex strains from patients with pulmonary tuberculosis in Cameroon. J Clin Micrbiol. 2003; 41(6):2547–2553.

50. Kulkarni S, Sola C, Filliol I, Rastogi N, Kadival G. Spoligotyping of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mumbai, India. Res Microbiol. 2005; 156(4):588–596. [PubMed: 15862459]

51. Kato-Maeda M, Gagneux S, Flores LL, et al. Strain classification of *M. tuberculosis:* congruence between large sequence polymorphisms and spoligotypes. Int J Tuberc Lung Dis. 2011; 15(1): 131–133. [PubMed: 21276309]

52. 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(Web Server issue):W326–W331. [PubMed: 20457747]

53. Iwamoto T, Yoshida S, Suzuki K, 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(1):67–74. [PubMed: 17302938]

54▪▪. Allix-Beguec C, Fauville-Dufaux M, Supply P. Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol. 2008; 46(4):1398–1406. Evaluation of mycobacterial interspersed repetitive unit variable number tandem repeats typing based on 15 and 24 loci in a population-based study. [PubMed: 18234864]

55. Velji P, Nikolayevskyy V, Brown T, Drobniewski F. Discriminatory ability of hypervariable variable number tandem repeat loci in population-based analysis of *Mycobacterium tuberculosis* strains, London, UK. Emerg Infect Dis. 2009; 15(10):1609–1616. [PubMed: 19861054]

56. Supply P, Allix C, Lesjean S, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol. 2006; 44(12):4498–4510. [PubMed: 17005759]

57. Christianson S, Wolfe J, Orr P, et al. Evaluation of 24 locus MIRU-VNTR genotyping of *Mycobacterium tuberculosis* isolates in Canada. Tuberculosis (Edinb). 2010; 90(1):31–38. [PubMed: 20056488]

58. Alonso M, Alonso Rodriguez N, Garzelli C, et al. Characterization of *Mycobacterium tuberculosis* Beijing isolates from the Mediterranean area. BMC Microbiol. 2010; 10:151. [PubMed: 20500810]

59. Mokrousov I, Narvskaya O, Vyazovaya A, et al. *Mycobacterium tuberculosis* Beijing genotype in Russia: in search of informative variable-number tandem-repeat loci. J Clin Microbiol. 2008; 46(11):3576–3584. [PubMed: 18753356]

60. Jiao WW, Mokrousov I, Sun GZ, et al. Evaluation of new variable-number tandem-repeat systems for typing *Mycobacterium tuberculosis* with Beijing genotype isolates from Beijing, China. J Clin Microbiol. 2008; 46(3):1045–1049. [PubMed: 18199785]

61. Valcheva V, Mokrousov I, Narvskaya O, Rastogi N, Markova N. Utility of new 24-locus variable-number tandem-repeat typing for discriminating *Mycobacterium tuberculosis* clinical isolates collected in Bulgaria. J Clin Microbiol. 2008; 46(9):3005–3011. [PubMed: 18614651]

62. 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(10):3563–3571. [PubMed: 11574573]

63. Ferdinand S, Valetudie G, Sola C, Rastogi N. Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. Res Microbiol. 2004; 155(8):647–654. [PubMed: 15380552]

64▪▪. Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis. 2007; 7(5):328–337. Review of the global phylogeography of *M. tuberculosis*. [PubMed: 17448936]

65. Mulenga C, Shamputa IC, Mwakazanga D, Kapata N, Portaels F, Rigouts L. Diversity of *Mycobacterium tuberculosis* genotypes circulating in Ndola, Zambia. BMC Infect Dis. 2010; 10:177. [PubMed: 20565802]

66. Smittipat N, Billamas P, Palittapongarnpim M, et al. Polymorphism of variable-number tandem repeats at multiple loci in *Mycobacterium tuberculosis*. J Clin Microbiol. 2005; 43(10):5034–5043. [PubMed: 16207958]

67. Yokoyama E, Kishida K, Uchimura M, Ichinohe S. Improved differentiation of *Mycobacterium tuberculosis* strains, including many Beijing genotype strains, using a new combination of variable number of tandem repeats loci. Infect Genet Evol. 2007; 7(4):499–508. [PubMed: 17398165]

68. Gagneux S, Burgos MV, DeRiemer K, et al. Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*. PLoS Pathog. 2006; 2(6):E61. [PubMed: 16789833]

69. Metcalfe JZ, Kim EY, Lin SY, et al. Determinants of multidrug-resistant tuberculosis clusters, California, USA, 2004–2007. Emerg Infect Dis. 2010; 16(9):1403–1409. [PubMed: 20735924]

70. van Doorn HR, de Haas PE, Kremer K, Vandenbroucke-Grauls CM, Borgdorff MW, van Soolingen D. Public health impact of isoniazid-resistant *Mycobacterium tuberculosis* strains with a mutation at amino-acid position 315 of katG: a decade of experience in The Netherlands. Clin Microbiol Infect. 2006; 12(8):769–775. [PubMed: 16842572]

71. Hu Y, Hoffner S, Jiang W, Wang W, Xu B. Extensive transmission of isoniazid resistant *M. tuberculosis* and its association with increased multidrug-resistant TB in two rural counties of eastern China: a molecular epidemiological study. BMC Infect Dis. 2010; 10:43. [PubMed: 20187977]

72. Dalla Costa ER, Ribeiro MO, Silva MS, et al. Correlations of mutations in *katG*, *oxyR-ahpC* and *inhA* genes and *in vitro* susceptibility in *Mycobacterium tuberculosis* clinical strains segregated by spoligotype families from tuberculosis prevalent countries in South America. BMC Microbiol. 2009; 9:39. [PubMed: 19228426]

73. Morgan M, Kalantri S, Flores L, Pai M. A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis:* a systematic review and meta-analysis. BMC Infect Dis. 2005; 5:62. [PubMed: 16050959]

74. Niemann S, Koser CU, Gagneux S, et al. Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS ONE. 2009; 4(10):E7407. [PubMed: 19823582]

75. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol. 2009; 7(4):287–296. [PubMed: 19287448]

76. Caws M, Thwaites G, Dunstan S, et al. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog. 2008; 4(3):E1000034. [PubMed: 18369480]

77. de Jong BC, Hill PC, Aiken A, et al. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. J lnfect Dis. 2008; 198(7):1037–1043.

78. Thwaites G, Caws M, Chau TT, et al. Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. J Clin Microbiol. 2008; 46(4):1363–1368. [PubMed: 18287322]

79. Gagneux S, DeRiemer K, Van T, et al. Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA. 2006; 103(8):2869–2873. [PubMed: 16477032]

80. Hershberg R, Lipatov M, Small PM, et al. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. 2008; 6(12):E311. [PubMed: 19090620]

81. Comas I, Chakravartti J, Small PM, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet. 2010; 42(6):498–503. [PubMed: 20495566]

82. 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(11):E7815. [PubMed: 19915672]

83. Wada T, Iwamoto T, Maeda S. Genetic diversity of the *Mycobacterium tuberculosis* Beijing family in East Asia revealed through refined population structure analysis. FEMS Microbiol Lett. 2009; 291(1):35–43. [PubMed: 19054072]

84. Bolotin S, Alexander D, Guthrie J, Drews S, Jamieson F. The Ontario universal typing of tuberculosis (OUT-TB) surveillance program – what it means to you. Can Respir J. 2010; 17(3): 51–54. [PubMed: 20422057]

85. Small PM, Hopewell PC, Singh SP, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med. 1994; 330(24): 1703–1709. [PubMed: 7910661]

86. Clark CM, Driver CR, Munsiff SS, et al. Universal genotyping in tuberculosis control program, New York City, 2001–2003. Emerg Infect Dis. 2006; 12(5):719–724. [PubMed: 16704826]

87. van Soolingen D, Borgdorff MW, de Haas PE, et al. Molecular epidemiology of tuberculosis in The Netherlands: a nationwide study from 1993 through 1997. J Infect Dis. 1999; 180(3):726–736. [PubMed: 10438361]

88. Kulaga S, Behr M, Musana K, et al. Molecular epidemiology of tuberculosis in Montreal. CMAJ. 2002; 167(4):353–354. [PubMed: 12197688]

89. Bauer J, Yang Z, Poulsen S, Andersen AB. Results from 5 years of nationwide DNA fingerprinting of *Mycobacterium tuberculosis* complex isolates in a country with a low incidence of *M. tuberculosis* infection. J Clin Microbiol. 1998; 36(1):305–308. [PubMed: 9431975]

90. Cowan LS, Diem L, Monson T, et al. Evaluation of a two-step approach for large-scale, prospective genotyping of *Mycobacterium tuberculosis* isolates in the United States. J Clin Microbiol. 2005; 43(2):688–695. [PubMed: 15695665]

91. Jasmer RM, Hahn JA, Small PM, et al. A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991–1997. Ann Intern Med. 1999; 130(12):971–978. [PubMed: 10383367]

92. Mathema B, Kurcpina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev. 2006; 19(4):658–685. [PubMed: 17041139]

93▪▪. Borgdorff MW, van den Hof S, Kremer K, et al. Progress towards tuberculosis elimination: secular trend, immigration and transmission. Eur Respir J. 2010; 36(2):339–347. Discusses use of molecular epidemiologic data to evaluate the progress towards TB elimination. [PubMed: 19996188]

94. Dahle UR, Eldholm V, Winje BA, Mannsaker T, Heldal E. Impact of immigration on the molecular epidemiology of *Mycobacterium tuberculosis* in a low-incidence country. Am J Respir Crit Care Med. 2007; 176(9):930–935. [PubMed: 17673698]

95. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet. 2004; 363(9404):212–214. [PubMed: 14738796]

96. Glynn JR, Crampin AC, Yates MD, et al. The importance of recent infection with *Mycobacterium tuberculosis* in an area with high HIV prevalence: a long-term molecular epidemiological study in Northern Malawi. J Infect Dis. 2005; 192(3):480–487. [PubMed: 15995962]

97. Verver S, Warren RM, Munch Z, et al. Transmission of tuberculosis in a high incidence urban community in South Africa. Int J Epidemiol. 2004; 33(2):351–357. [PubMed: 15082639]

98▪▪. McNabb SJ, Kammerer JS, Hickey AC, et al. Added epidemiologic value to tuberculosis prevention and control of the investigation of clustered genotypes of *Mycobacterium tuberculosis* isolates. Am J Epidemiol. 2004; 160(6):589–597. Outlines impact of additional epidemiologic investigations of cases with genotypically matched *M. tuberculosis* isolates (duster cases). [PubMed: 15353420]

99. Sebek M. DNA fingerprinting and contact investigation. Int J Tuberc Lung Dis. 2000; 4(2 Suppl 1):S45–S48. [PubMed: 10688148]

100. van Deutekom H, Hoijng SP, de Haas PE, et al. Clustered tuberculosis cases: do they represent recent transmission and can they be detected earlier? Am J Respir Crit Care Med. 2004; 169(7): 806–810. [PubMed: 14684559]

101. Tuberculosis outbreak in a low-incidence state – Indiana, 2001–2004. MMWR Morb Moral Wkly Rep. 2004; 53(48):1134–1135.

102. Centers for Disease Control and Prevention (CDC). Tuberculosis outbreak in a community hospital – District of Columbia, *2002*. MMWR Morb Mortal Wkly Rep. 2004; 53(10):214–216. [PubMed: 15029115]

103. Dewan PK, Banouvong H, Abernethy N, et al. A tuberculosis outbreak in a private-home family child care center in San Francisco, 2002 to 2004. Pediatrics. 2006; 117(3):863–869. [PubMed: 16510668]

104. Johnson R, Warren R, Strauss OJ, et al. An outbreak of drug-resistant tuberculosis caused by a Beijing strain in the western Cape, South Africa. Int J Tuberc Lung Dis. 2006; 10(12):1412–1414. [PubMed: 17167961]

105. McElroy PD, Southwick KL, Fortenberry ER, et al. Outbreak of tuberculosis among homeless persons coinfected with human immunodeficiency virus. Clin Infect Dis. 2003; 36(10):1305–1312. [PubMed: 12746777]

106. Lai CC, Tan CK, Lin SH, et al. Molecular evidence of false-positive cultures for *Mycobacterium tuberculosis* in a Taiwanese hospital with a high incidence of TB. Chest. 2010; 137(5):1065–1070. [PubMed: 19965955]

107. van Rie A, Warren R, Richardson M, et al. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N Engl J Med. 1999; 341(16):1174–1179. [PubMed: 10519895]

108. Dwyer B, Jackson K, Raios K, Sievers A, Wilshire E, Ross B. DNA restriction fragment analysis to define an extended cluster of tuberculosis in homeless men and their associates. J Infect Dis. 1993; 167(2):490–494. [PubMed: 8093624]

109. Nava-Aguilera E, Andersson N, Harris E, et al. Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis. Int J Tuberc Lung Dis. 2009; 13(1):17–26. [PubMed: 19105874]

110. Kliiman K, Altraja A. Predictors of extensively drug-resistant pulmonary tuberculosis. Ann Intern Med. 2009; 150(11):766–775. [PubMed: 19487711]

111. Dye C. Doomsday postponed? Preventing and reversing epidemics of drug-resistant tuberculosis. Nat Rev Microbiol. 2009; 7(1):81–87. [PubMed: 19079354]

112. Valway SE, Sanchez MP, Shinnick TF, et al. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. N Engl J Med. 1998; 338(10):633–639. [PubMed: 9486991]

113. Williams A, James BW, Bacon J, et al. An assay to compare the infectivity of *Mycobacterium tuberculosis* isolates based on aerosol infection of guinea pigs and assessment of bacteriology. Tuberculosis (Edinb). 2005; 85(3):177–184. [PubMed: 15850755]

114. Reed MB, Gagneux S, Deriemer K, Small PM, Barry CE 3rd. The W-Beijing lineage of *Mycobacterium tuberculosis* overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated. J Bacteriol. 2007; 189(7):2583–2589. [PubMed: 17237171]

115. Lopez B, Aguilar D, Orozco H, et al. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. Clin Exp Immunol. 2003; 133(1): 30–37. [PubMed: 12823275]

116. Kato-Maeda M, Kim EY, Flores L, Jarlsberg LG, Osmond D, Hopewell PC. Differences among sublineages of the East-Asian lineage of *Mycobacterium tuberculosis* in genotypic clustering. Int J Tuberc Lung Dis. 2010; 14(5):538–544. [PubMed: 20392345]

117. Coscolla M, Gagneux S. Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov Today Dis Mech. 2010; 7(1):E43–E59. [PubMed: 21076640]

118. Daley CL, Kawamura LM. The role of molecular epidemiology in contact investigations: a US perspective. Int J Tuberc Lung Dis. 2003; 7 (12 Suppl 3):S458–S462. [PubMed: 14677838]

119. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42(1):121–130. [PubMed: 3719049]

120. Glynn JR, Vynnycky E, Fine PE. Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. Am J Epidemiol. 1999; 149(4):366–371. [PubMed: 10025480]

121. Murray M. Sampling bias in the molecular epidemiology of tuberculosis. Emerg Infect Dis. 2002; 8(4):363–369. [PubMed: 11971768]

122. Braden CR, Templeton GL, Cave MD, et al. Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. J Infect Dis. 1997; 175(6):1446–1452. [PubMed: 9180185]

123. Chin DP, Crane CM, Diul MY, et al. Spread of *Mycobacterium tuberculosis* in a community implementing recommended elements of tuberculosis control. JAMA. 2000; 283(22):2968–2974. [PubMed: 10865275]

124. Shamputa IC, Lee J, Allix-Beguec C, et al. Genetic diversity of *Mycobacterium tuberculosis* isolates from a tertiary care tuberculosis hospital in South Korea. J Clin Microbiol. 2010; 48(2): 387–394. [PubMed: 20018816]

125. Al-Hajoj SA, Akkerman O, Parwati I, et al. Microevolution of *Mycobacterium tuberculosis* in a tuberculosis patient. J Clin Microbiol. 2010; 48(10):3813–3816. [PubMed: 20686077]

Websites

201. WHO. Tuberculosis. Fact Sheet No. 104. www.who.int/mediacentre/factsheets/fs104/en/

202. WHO. WHO Report 2009. WHO; Geneva: 2009. Global tuberculosis control – epidemiology, strategy, financing. www.who.int/tb/publications/global_report/2009/en/lindex.html

203. Institut Pasteur De La Guadeloupe. Tuberculose et Mycobactéries. www.pasteur-guadeloupe.fr/tb/bd_myeo.html

204. MIRU-VNTRplus. www.miru-vntrplus.org/

Executive summary

- Genotyping of *Mycobacterium tuberculosis* has been successfully used in molecular epidemiology. Genotyping is used to study the transmission dynamics of specific isolates of *M. tuberculosis* in a community by tracking their distribution and spread. The DNA marker for the study of TB transmission dynamics should be polymorphic enough to distinguish among unrelated isolates, yet stable enough to make the connection between isolates that are indeed related. By contrast, the DNA marker for phylogeny and population genetics should be phylogenetically robust (i.e., all isolates that share the marker inherited from the same common ancestor) and with a solid phylogenetic framework in which specific strains can be positioned and biologically meaningful groups delineated.
- The most recent breakthrough in the field of molecular epidemiology is the use of whole-genome sequencing technology. The advantages of this technology are its ability to determine the sequence variation at a real epidemiological scale, identify the exact source(s) of infection and the transmission events among individuals that share the same *M. tuberculosis* isolate, and to determine the evolutionary relationship between isolates.
- Molecular epidemiology has enhanced TB control activities through identification of previously unrecognized chains of transmission, by monitoring of disease trends, including drug resistance, and by providing information critical for more precise allocation of public health resources.
- The challenges facing this field are to obtain real-time molecular epidemiologic data and to translate genotyping data into a predictive public health database to help elucidate future epidemics of specific strains or groups of strains.

Table 1

Genetic markers in *Mycobacterium tuberculosis*.

Marker	Principle	Technical requirements	Advantages	Disadvantages	Ref.
IS*6110*-RFLP	Based on the variability in the number of copies of IS*6110* and the molecular weights of DNA fragments in which the insertions are found	PCR amplification and probe labeling 2–3 µg of pure DNA Southern blotting and x-ray film developer Visual comparison or software-based comparison	High discriminatory power Widely used	Limited discriminatory power in isolates with five or less IS*6110* bands Lengthy process Difficult to compare results between laboratories	[14]
PGRS-RFLP	Based on the variability in the number and location of the PGRS regions	PCR amplification and probe labeling 3 µg of pure DNA Southern blotting and x-ray film developer Visual comparison or software-based comparison	High discriminatory power	Limited data using this technique Lengthy process Difficult to analyze Not possible to compare results between laboratories	[116]
Spoligotyping	Based on polymorphism in the direct repeat locus which is a member of the CRISPRs	PCR amplification Spoligotyping® blotter and membrane or Luminex technology® and software	PCR-based It can be done in DNA extracted directly from sputum Data in an exchangeable format Highly reproducible	Limited discriminatory power	[37,38]
MIRU-VNTR	Based on polymorphisms of MIRU loci Set of 15 loci used for molecular epidemiology studies Set of 24 for phylogenetic studies	PCR amplification and gel electrophoresis or sequencer and probes to detect amplicons Manual comparison of patterns or software to analyze and compare amplicon lengths	High discriminatory power Data in an exchangeable format	Set of 12 loci less discriminatory than IS*6110* RFLP Electrophoresis method: determination of band size is less reproducible than the sequencer-based method Sequencer method: high cost	[16,56,58]
Drug-resistance mutations	Based on drug resistance-associated mutations	Real-time PCR Sequence of the region with the possible mutations Reverse hybridization of the target amplicon to probes representative of the possible mutations causing drug resistance	Combination with other genetic markers will increase the discrimination power Useful to determine the impact of specific mutations on the transmission of drug-resistant strains	Just for isolates with drug resistance	[65,66]
Large sequence polymorphism	Based on the presence or absence of specific segments of DNA	Real-time PCR using probes for detection PCR and gel electrophoresis for size amplicon comparison	Robust marker for phylogenetic classification	Not useful to track specific strains in the community	[76]
Multilocus sequence typing	Based on DNA sequences of multiple loci	PCR amplification and sequencing	Robust marker for phylogenetic classification	Expensive Depending on the loci used, probably limited discriminatory power to be used to track specific strains in the community	[77]
Whole-genome sequencing	Based on the analysis of the whole-genome sequence results	Next-generation sequencing	Gold standard for phylogenetic classification	Expensive, need for specialized technology and software	[17]

CRISPR: Clustered regularly interspaced short palindromic repeats; IS: Insertion element; MIRU: Mycobacterial interspersed repetitive unit; PGRS: Polymorphic GC-rich repetitive sequence; RFLP: Restriction fragment length polymorphism; VNTR: Variable number tandem repeat.

Table 2

MIRU loci included in the different MIRU sets.

Locus[†]	VNTR name	MIRU 12[‡]	MIRU 15[§¶]	MIRU 24[§#]	Beijing[††]
H37Rv_4348	MIRU 39	X		X	X
H37Rv_4156	QU8-4156		X	X	X
H37Rv_4052	QUB-26		X	X	X
H37Rv_3239	ETR-F				X
H37Rv_1982	QUB-18				X
H37Rv_0154	MIRU 2	X		X	
H37Rv_0580	MIRU 4 (ETR-D)	X	X	X	
H37Rv_0959	MIRU 10	X	X	X	
H37Rv_1644	MIRU 16	X	X	X	
H37Rv_2059	MIRU 20	X		X	
H37Rv_2531	MIRU 23	X		X	
H37Rv_2687	MIRU 24	X		X	
H37Rv_2996	MIRU 26	X	X	X	
H37Rv_3006	MIRU 27 (QUB-5)	X		X	
H37Rv_3192	MIRU 31 (ETR-R)	X	X	X	
H37Rv_0802	MIRU 40	X	X	X	
H37Rv_0424	Mtub04		X	X	
H37Rv_0577	ETR-C		X	X	
H37Rv_2156	ETR-A		X	X	
H37Rv_2401	Mtub30		X	X	
H37Rv_3690	Mtub39		X	X	
H37Rv_2163b	QUB-11b		X	X	X
H37Rv_1955	Mtub21		X	X	X
H37Rv_2347	Mtub29			X	
H37Rv_2461	ETR-B			X	
H37Rv_3171	Mtub34			X	

[†] Defined as position in H37Rv.

[‡] Data taken from [13].

[§] Data taken from [52].

[¶] Recommended for molecular epidemiologic studies.

[#] Recommended for phylogenetic studies.

[††] Data taken from [49,55,64].

ETR: Exact tandem repeats; MIRU: Mycobacterial interspersed repetitive unit; QUB: Queen’s University of Belfast; VNTR: Variable number tandem repeat.