ELSEVIER

Enfermedades Infecciosas y Microbiología Clínica

www.elsevier.es/eimc

Continuing medical education: Mycobacterial infections

Molecular epidemiology of tuberculosis[☆],[☆☆]

Pere Coll[a],[b],[c],[*], Darío García de Viedma[d],[e],[f],[*]

[a] Servicio Microbiología, Hospital de Sant Pau, Barcelona, Spain
[b] Departament de Genètica i Microbiologia, UniversitatAutònoma de Barcelona, Bellaterra, Spain
[c] Institut de Recerca, Hospital de Sant Pau, Barcelona, Spain
[d] Servicio Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Madrid, Spain
[e] Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain
[f] CIBER Enfermedades Respiratorias, CIBERES, Madrid, Spain

ARTICLE INFO

Article history:
Received 27 December 2017
Accepted 13 January 2018
Available online 17 March 2018

Keywords:
Tuberculosis
Molecular epidemiology
Genotyping

ABSTRACT

The application of genotyping tools allowed us to discriminate between the *Mycobacterium tuberculosis* isolates obtained in the laboratory. The differentiation between single strains opened the door to molecular epidemiology studies, which had helped us to progress in our knowledge of how this pathogen is transmitted in the progressively more complex socio-epidemiological scenario. The genetic stability of this microorganism led to develop specific methodologies, which are thoroughly revised in this chapter. In addition to their application in epidemiology, we review, how they can offer a response to different diagnostic and clinical challenges. Finally, we focus on describing the novel genomic revolution we are experiencing in the analysis of tuberculosis, the methodology in which it is based and the novel possibilities it offers, including new routes of integrating both the molecular and genomic languages in innovative post-genomic proposals, better suited to our real-life context.

Epidemiología molecular de la tuberculosis

Palabras clave:
Tuberculosis
Epidemiología molecular
Genotipificación

RESUMEN

La aplicación de técnicas de genotipificación ha permitido discriminar entre los aislados de *Mycobacterium tuberculosis* obtenidos en el laboratorio. Esta singularización a nivel de cepa abrió las puertas a estudios de epidemiología molecular que han permitido progresar en nuestro conocimiento de la transmisión de este patógeno en entornos socio-epidemiológicos cada vez más complejos. La estabilidad genética de este microorganismo ha llevado al desarrollo de metodologías específicas, que son revisadas en detalle en este capítulo. Además de las aplicaciones epidemiológicas, repasamos cómo dan respuesta asimismo a diversos interrogantes diagnósticos y clínicos. Por último, nos ocupamos de describir la nueva revolución

DOI of original article: http://dx.doi.org/10.1016/j.eimc.2018.01.001

[☆] Please cite this article as: Coll P, García de Viedma D. Epidemiología molecular de la tuberculosis. Enferm Infecc Microbiol Clin. 2018;36:233–240.

[☆☆] Section accredited by the *Consell Català de Formació Continuada de les Professions Sanitàries* (Catalan Board of Continuing Education for Health Professionals). For questions on each article, visit: http://www.eslevier.es/eimc/formacion.

[*] Corresponding authors.
E-mail addresses: pcoll@santpau.cat (P. Coll), dgviedma2@gmail.com (D. García de Viedma).

genómica en el estudio de la tuberculosis, los métodos en los que descansa y las posibilidades inéditas que ofrece, incluyendo nuevas vías de integración de los lenguajes moleculares y genómicos en propuestas innovadoras de trabajo posgenómico, más adaptadas a la realidad de nuestro entorno.

Introduction

Molecular epidemiology includes a wide variety of techniques that have the objective of comparing nucleic acid sequences from two or more isolates. The term clone or clonal group refers to a group of isolates that are related because they descend from a more or less remote common ancestor. The related strains therefore originate from the clonal expansion of a single precursor and possess a degree of similarity between their genotypes and phenotypes that is significantly higher than that found between arbitrarily selected unrelated isolates of the same species. The definition of clone is probabilistic, and the degree of similarity required for its definition must take into account the taxon studied, the marker used and the duration of the epidemiological investigation. Therefore, before a molecular marker can be selected, we need to clearly and precisely formulate the question we are trying to answer in epidemiological terms, define the degree of genetic relatedness we need to establish in order to answer the previous question, choose those markers that are capable of discriminating this degree of relatedness and verify the efficacy of the chosen methods.

Comparative sequencing studies of *Mycobacterium tuberculosis (M. tuberculosis)* have found little genetic variation between strains, minimal horizontal gene transfer and a clonal population structure. In spite of this, there is some variation in the *M. tuberculosis* population, which may be due to single nucleotide polymorphisms (SNPs), long sequence polymorphisms (LSPs) or repeated sequence polymorphisms. This variation can be detected using markers. Some of these markers visualise variations that have arisen at more or less distant times and serve to classify the strains into “lineages” of phylogenetic interest. Other markers detect much more recent variations and allow the strains to be genotyped in order to recognise chains of transmission.

Markers for phylogenetic studies of *M. tuberculosis*

Single nucleotide polymorphism

SNPs can be synonymous or non-synonymous. In synonymous SNPs, the nucleotide change does not involve an amino acid change and is therefore not subject to selective pressure. Its evolutionary clock runs slowly and it gives phylogenetic information. Non-synonymous polymorphisms, which make up approximately two thirds of the genetic variability observed in *M. tuberculosis*, potentially have functional consequences, undergo selective pressure and there is evidence that they can influence important characteristics such as transmissibility, virulence, drug resistance, immune response or clinical symptoms.[1]

In 1997, Sreevatsan et al.[2] compared the sequences of 26 structural genes and detected two synonymous SNPs, one in the gene that encodes for catalase peroxidase (katG463) and the other in DNA gyrase subunit A (gyrA95). These SNPs allow *M. tuberculosis* to be classified into three genetic groups. Group 1 includes the strains *Mycobacterium bovis (M. bovis)*, *Mycobacterium microti (M. microti)*, *Mycobacterium africanum (M. africanum)* and the evolutionarily older strains of *M. tuberculosis*, while groups 2 and 3 are made up of *M. tuberculosis* strains. The authors studied some 6000 strains of *M. tuberculosis* isolated in New York and Houston, and observed that the majority of the grouped cases of tuberculosis (transmission chains) corresponded to strains from group 1 or 2, while group 3 (the most recent in evolutionary terms) gave rise to sporadic cases, which led them to conclude that *M. tuberculosis* is evolving towards lower transmissibility or virulence. These three groups can be divided into six phylogenetic groups (with five subgroups) by studying 212 SNPs.[3] There is a strong association between these groups and the geographic origin of the strains or the birthplace of the patients.

Long sequence polymorphisms

The study of the deletion of 20 long sequences distributed along the length of the chromosome, regions of deletion (RDs),[4] allows us to establish an evolutionary framework for the *M. tuberculosis* complex, as well as to identify species within the complex. The *M. tuberculosis* complex is genetically homogeneous, but very diverse in relation to its habitat. *M. tuberculosis*, *M. africanum* and *Mycobacterium canettii* have an exclusively human habitat, *M. microti* is a pathogen in rodents, *Mycobacterium pinnipedii* in seals and *Mycobacterium caprae* in goats, while *M. bovis* has a wider range of hosts. The study of RDs has allowed us to establish that the species *M. tuberculosis* is the oldest ancestor from which the others evolved. Within the species *M. tuberculosis*, strains can be classified as ancient or modern based on whether or not they possess the TbD1 region. Deletion of the TbD1 region probably occurred prior to the eighteenth century, and the appearance of modern strains of *M. tuberculosis* may have contributed to the enormous spread of tuberculosis (TB) from the eighteenth century onwards. Gagneux et al.[5] analysed these RDs in 875 strains from 80 countries. The strains were subdivided into four lineages (Indo-Oceanic, East Asian, East African-Indian, Euro-American), plus two lineages traditionally identified as *M. africanum* (West African 1 and West African 2).

A study in San Francisco, where communities of diverse geographic origins live side by side, as do strains from the different lineages described, observed a close association between lineages and populations with a specific geographic origin.[5] In other words, there would appear to be adaptation or compatibility between the pathogen and a certain host population.

The studies based on both SNPs and LSPs are highly consistent, which is not unexpected given the clonal nature of the *M. tuberculosis* population. Both types of study suggest an association between lineage and geographic origin, speculating on the influence that these lineages might have on the performance of diagnostic tests, resistance to anti-tuberculosis drugs and vaccine response.[6]

Markers for molecular epidemiology studies

Spoligotyping: spacer oligonucleotide typing

The *M. tuberculosis* chromosome contains multiple copies of the direct repeat (DR) sequence located at a single chromosomal locus (DR region), forming an IS*6110* integration hot spot. DR sequences of 36 bp are separated by different, i.e. non-repeating, spacers of 34–41 bp. When comparing strains of *M. tuberculosis*, this region is polymorphic due to both homologous recombination of DRs and changes caused by the location of IS*6110* in this region. The spoligotyping technique[7] analyses this polymorphism. The first phase in this process is amplification, using two complementary oligonucleotides from the ends of the DR sequence as initiators, and, consequently, amplifying the various spacer sequences present in the strain. The initiators are marked with biotin. The second phase is the detection of the various amplified spacer sequences using PCR-product hybridisation on a filter on which complementary oligonucleotides from the various spacer sequences described have been immobilised. The binding of the PCR products to their specific sequences is detected using peroxidase marked with streptavidin, which binds to the biotin present in the PCR products. There are 94 spacer sequences, although most analysis protocols only use 43. Alternatives to membrane hybridisation have been developed, such as a technique based on microspheres and laser technology[8] or mass spectrometry (MALDI-TOF),[9] which make the technique quicker and easier. As it is the first phase in an amplification reaction technique, only a small quantity of DNA is required. In fact, spoligotyping can be used directly on the clinical sample.[10] The results are expressed as a binary number or as an eight-digit number following a simple conversion. This means that results are easy to export and there are global databases (SpolDB4; www.pasteur-guadeloupe.fr/tb/bd_myeo.html)

The main limitation of spoligotyping is lower discriminatory power than IS*6110*-RFLP or mycobacterial interspersed repetitive unit-variable number tandem repeats (MIRU-VNTR). Spoligotyping occupies a midpoint between techniques that give phylogenetic information and more discriminative techniques that can detect recent transmission chains. For this reason, it is considered to be more useful in phylogenetic studies (assignation of lineages, sublineages or families) than for true transmission chain identification for molecular epidemiology purposes.

Restriction-hybridisation patterns (IS6110-RFLP, PGRS-RFLP)

When a DNA molecule is broken down by a restriction enzyme, the number of fragments obtained is equivalent to the number of times the restriction locus is repeated along the length of the molecule. The size of the various restriction fragments corresponds to the separation between two neighbouring restriction loci. Strains are compared to detect restriction fragment length polymorphism (RFLP) by separating the fragments obtained using agarose gel electrophoresis. To simplify the interpretation of RFLPs of the entire DNA, we can focus on analysing a repeated sequence in the genome rather than studying the whole chromosome. To do this, the restriction fragments of the entire DNA obtained are transferred to a membrane (Southern-blot) which is analysed using hybridisation with a marked probe. Only those fragments containing all or part of the complementary sequence for the chosen probe will show up in the hybridisation results, giving us profiles with few bands. The number of bands will depend on the number of copies of the sequence that are present on the genome, while the polymorphism will depend on its distribution along the length of the genome, as well as the variations in the restriction loci within this sequence and in the neighbouring regions. Both the Southern transfer and hybridisation notably complicate the technical aspects of the analysis.

One of the limitations of these techniques is the need for a large quantity of DNA, which prevents them being used directly on the clinical sample, making it necessary to work from cultures. Similarly, in strains in which the chosen sequence is repeated a small number of times (5 or less for IS6110-RFLP), the technique's discriminatory power is reduced. Moreover, the results are obtained as a pattern of bands that is susceptible to electrophoretic distortions, which makes comparison of results between laboratories difficult. In spite of this, the restriction-hybridisation patterns associated with IS6110 have for years been the reference technique for the study of the molecular epidemiology of TB, both due to their discriminatory power and their stability (estimated molecular clock of between 3.2 and 8.7 years).[11] The use of a standardised protocol[12] with molecular weight markers included in each electrophoresis channel has allowed global databases to be established, improving our knowledge of the global epidemiology of TB.

Mycobacterial interspersed repetitive units-variable number tandem repeats (MIRU-VNTR)

Along the length of the *M. tuberculosis* chromosome, some sequences repeated in tandem are polymorphic in the different strains thanks to a variable number of tandem repeats (VNTR). These structures are similar to the minisatellites observed in eukaryotic cells. MIRUs are one of these VNTRs. They are tandem repeats of 46–101 bp, with some 41 MIRUs dispersed throughout the chromosome. MIRU-VNTR genotyping initially used 12 loci (the most polymorphic), but due to the low discrimination observed in comparison to other markers, the analysis of 24 loci became the standard and is the method used today.[13] For this, a specific PCR is performed for each locus, marking the initiators used with a fluorochrome. The length of the amplified fragment is later determined using capillary electrophoresis. This length depends on the number of repetitions of the core sequence present at the locus. The number of repetitions becomes the digit for this locus in the VNTR code, which will therefore have 24 digits.

MIRU-VNTR genotyping is fast, relatively simple from a technical point of view, highly reproducible and discriminatory (comparable to IS6110-RFLP). Its results are expressed as a 24-digit code and can therefore be exported easily, and there are global databases that allow strains to be compared worldwide (http://www.miru-vntrplus.org). One of these databases includes MIRU-VNTR and spoligotyping data for around 62,000 strains of *M. tuberculosis* isolated from 153 countries (http://www.pasteur-guadeloupe.fr:8081/SITVIT_online). For all these reasons, MIRU-VNTR has become the reference technique for genotyping *M. tuberculosis*. From the information obtained, we can infer phylogenetic hypotheses, as well as discriminate at strain level, enabling us to detect transmission chains.

Nevertheless, the discriminatory power of the loci used can vary depending on the *M. tuberculosis* lineage studied. Therefore, in view of the homoplasmy that exists between members of the Beijing family, standard 24-locus MIRU-VNTR has its limitations.[14,15] The inclusion of loci not present in the standardised scheme (VNTRs 3232, 3280 and 4120) has been proposed for the analysis of this family. In other words, for certain studies, it may be necessary to adapt the MIRU-VNTR scheme to the lineages studied.

Whole genome sequencing (WGS)

We have seen how, to date, the molecular markers used analyse the variations that exist in "hypervariable" regions, which represent only a minuscule part of the whole genome. When whole genome sequencing (WGS) is used, information about the microevolution of the entire genome becomes available. It has become possible to use WGS in the molecular epidemiology of TB thanks to the technological development of platforms for massive sequencing, as well as the appearance of commercial preparations that facilitate the preparation of genomic libraries for sequencing. Obviously, the main difficulty of this technique is the analysis of the information generated. In recent years, numerous IT tools have been developed that allow us not only to calculate the number of SNPs that exist between genomes, but also to extrapolate, from the whole genome, both the molecular markers used for genotyping (IS*6110*, MIRU-VNTR, spoligotyping) and those that provide phylogenetic information (SNP, LPS), as well as the molecular antibiogram (analysing resistance-associated mutations).[16]

As is to be expected, whole genome analysis is more discriminatory than the markers traditionally used to detect transmission chains. The parameter used is the calculation of the SNPs that exist between the sequences compared. Various studies have analysed the SNPs that exist between strains isolated sequentially in the same patient or strains belonging to well-defined transmission chains, both for molecular epidemiology and for field studies. The objective of these studies is to estimate the short-term genetic derivation of *M. tuberculosis* in order to establish a cut-off value that will enable us to differentiate epidemiologically related and unrelated strains. Comparisons of the genomes of strains belonging to a single transmission chain almost never exceed 3–5 SNPs.[17,18] However, the average change in the DNA sequence of the *M. tuberculosis* genome has been calculated at 0.5 SNPs per genome per year.[18] The phylogenetic trees obtained through the study of SNPs in the whole genome correlate better with epidemiological field data than the trees obtained with traditional markers. In other words, the study of SNPs throughout the genome allows us to separate members of a transmission chain from related strains that do not belong to that chain. This is especially helpful in the analysis of complex outbreaks.[18,19]

Furthermore, WGS data allows us to determine the distribution over time and in space of TB cases as it is able to differentiate the index case from secondary cases in the transmission chain. In phylogenetic trees, the secondary cases (clonal variants with very few differentiating SNPs) will give rise to star-like structures with the index case as the nucleus. Likewise, in a long-lasting epidemic, the unidirectional accumulation of SNPs allows us to more clearly associate new cases with earlier cases.[18,19]

Applications of genotyping and genomic analysis strategies for *M. tuberculosis*

Strictly speaking, mycobacterium tuberculosis (MTB) genotyping techniques find their natural niche in molecular epidemiology studies that aim to identify clusters of patients infected by the same strain and which therefore form part of the same transmission chain. These grouped cases are differentiated from those not belonging to the cluster, or orphan cases, which, as they are infected with a strain with a unique genotype in that population, are considered to potentially be reactivations of more or less remote past exposures.

In this review, seeking to cover the broadest spectrum of applicability of MTB genotyping techniques, we have opted to go beyond the narrow concept of the application of these techniques for molecular epidemiology purposes and include other analysis settings where these same methods have been found to be useful. We will therefore revisit a series of questions to which the characterisation of the different genetic markers of MTB has been able to provide an answer.

Can I identify a false diagnosis of tuberculosis?

In the laboratory, the suspicion of possible cross-contamination (CC) arises when two or more positive cultures are identified in samples processed on the same day. The suspicion of CC is strengthened if the following factors are also present: only one of the cases has additional positive samples, only the sample from one of the cases under suspicion is a positive stain, and the growth in the sample in the case suspected of being a false positive takes longer than average to give a positive result. However, in order to irrefutably document that we are dealing with a laboratory CC event, we must demonstrate that the cultures in the cases under suspicion are of the same strain. For this, the recommended technique is MIRU-VNTR, due to the greater speed with which it yields results. In addition, since in this case it is not necessary to determine the precise genotype of the strains and a qualitative analysis of their similarities (demonstrates CC) or differences (rules out CC) is sufficient, rapid qualitative analysis methods have been developed that allow rapid documentation of the similarities/differences between genotypes by comparing their mobility patterns in conventional agarose gel electrophoresis, using triplex (MLP3 technique)[20] as well as duplex[21] PCR products.

Is this a microepidemic or outbreak?

Before MTB genotyping techniques were developed, suspicion of outbreaks or microepidemics was based purely on identifying epidemiological links between several TB cases. However, verification of an outbreak requires confirmation that the strains infecting the theoretically linked cases share an identical genotype. Documenting such outbreaks was handled initially using IS6110-RFLP and in more recent years with MIRU-VNTR. Thanks to the on-demand application of these techniques, it has been possible to reveal that some microepidemics involved different strains.[22,23] These were clearly identified by epidemiological studies of the cases, which was the first indication that epidemiological research should be extended to contact settings beyond the workplace, family or home environments.

Similarly, as with the documentation of laboratory contaminations, the molecular epidemiology analysis of suspected outbreaks does not in principle require the exact genotype of the strains to be obtained, since its aim is merely to establish similarities or differences between the isolates concerned. MLP3 or derived techniques[20,21] may therefore be sufficient for laboratories with limited resources. However, obtaining the genotype through a conventional full analysis, in the past with RFLP and now with MIRU-VNTR, allows us to identify real contacts with other cases in the population which helps to redefine the extent of the outbreak.

How can I identify mixed or complex infections?

One of the fields other than molecular epidemiology where genotyping techniques have led to notable advances is in the identification of cases infected with more than one strain, whether in a single episode (mixed infections, where these involved different strains, or polyclonal infections, in cases of clonal variants originating from the same parent strain) or in successive episodes (reinfections).

The first molecular documents regarding reinfections[24] alerted us to an unexpectedly high proportion (75%) of cases of recurrent TB in which the strain isolated in the second episode was different to the first, although these were in settings with a high incidence where overexposure to infectious cases was more likely. In these settings, it was even demonstrated that reinfection could occur shortly after the end of treatment for the first episode.[25] It was later documented that reinfection was the cause of a considerable number of recurrences, including in moderate or low incidence settings.[26–28]

With regard to the simultaneous presence of more than one strain or clonal variant in the same patient, only the arrival of the MIRU-VNTR technique has led to noticeable advances in this field.[29,30] This is because the earlier strategies, spoligotyping or RFLP, were not capable of identifying the presence of more than one strain. In such cases, a fictitious genotype pattern was obtained as a result of the superimposed patterns of each of the strains, without this additive pattern alerting us to the presence of two strains or clonal variants. Nevertheless, the fact that the MIRU-VNTR technique gives a single amplification product for each locus analysed means that, where we are observing more than one product, this can only be due to the presence of more than one strain (if the number of loci with more than one allele is sufficiently high) or clonal variant (if there are 1–2 loci with more than one allele). Systemic application of the MIRU-VNTR technique has therefore enabled us to more accurately determine the proportion of mixed or polyclonal infections in a population,[31,32] and to be able to establish that the urgency of clonal variants is not restricted to extreme situations such as long delays in diagnosis but can occur even in conventional infections.[33] Effective identification of complex infections has indisputable epidemiological value; meanwhile, failure to take it into account can lead to difficulties in the treatment of patients in the event of mixed infections with strains with differing patterns of susceptibility.[34] Lastly, the presence of a complex infection can lead to complications in microbiological diagnosis, even in aspects as apparently unrelated as the analysis of laboratory contamination.[35]

How can I tackle tuberculosis transmission dynamics in my population?

In contrast to the documentation on outbreaks, which is within the scope of any laboratory with the capacity to implement genotyping techniques, the analysis of transmission within a population requires us to lay down more solid strategic foundations. The effective identification of transmission chains requires universal genotyping; in other words, we need to have a population-based sample which guarantees that we have access to all TB cases, and, therefore, that links in the transmission chain are not going to be left unanalysed. In addition, given the variable time between exposure in a case and development of the disease, for rigorous identification of transmission chains, a minimum time of more than 2–3 years is needed to systematically characterise all cases. The first population-based molecular epidemiology studies relied on the RFLP technique; however, the greater discrimination of the MIRU-VNTR technique, its shorter turn-around time and the ease of exchanging results between laboratories, due to formatting the genotyping results as a numeric code, have meant that it has completely replaced RFLP.

Universal molecular epidemiology studies allow us to find out the percentage of TB cases that are the result of recent transmission chains; this is an excellent indicator of the efficacy of control programmes, which should lead to a reduction in these percentages. They have also facilitated the uncovering of transmission dynamics in complex epidemiological settings, such as those derived from populations with a high immigration rate, in which it has been possible to identify settings with significant cross-transmission between indigenous and immigrant cases.[36,37] Maintaining systematic genotyping systems has allowed us to identify the prevalent strains that are responsible for a large part of the disease burden in a population.[38] Likewise, these long-term molecular monitoring programmes facilitate detailed dissection of the potential impact of the importation of a strain that is new to a population, but that years later may have come to be responsible for a significant portion of TB cases in that population.[39]

Is it possible to achieve very high-quality characterisation of tuberculosis transmission?

Despite the valuable information obtained from the systematic use of RFLP initially and MIRU-VNTR in recent years, we must be aware of the inevitable limitations that derive from determining relationships of similarities or differences between strains when we are only looking at the analysis of a very small part (<0.1%) of the MTB chromosome. This means that we could be classifying strains that differ in areas of the chromosome we are not observing as identical, and, therefore, that we could be overestimating some of the transmission clusters we propose.

Fortunately, the marked reduction in cost and simplification of bioinformatic procedures have led to the introduction of WGS in MTB characterisation. This means that we have access to an enormous quantity of information about the genome, which gives us a much higher discriminatory power between strains. This is ideal for guaranteeing maximum precision in molecular epidemiology studies, which are beginning to be replaced by genomic epidemiology.

The application of WGS to clusters defined by conventional genotyping means that some of them, as expected, have been subdivided into smaller groups after identifying differences between strains that, under earlier levels of resolution, appeared to be related as a cluster.[40] This means that the information derived from genomic epidemiology more closely reflects the geographic distribution of the cases and their epidemiological links.[19] The systematic application of WGS in situations which are difficult to control has revealed the enormous complexity of some of these scenarios, such as the coexistence of different overlapping outbreaks that previously would have been interpreted as one.[41]

In the field of population-based epidemiology, there are still few settings in which WGS has been able to be applied systematically. England stands out in this regard, having implemented an analysis system for all TB cases, beginning in the Midlands region[42] and recently extended to much of the country. These efforts to systematically apply WGS are paralleled by the development of a methodology to accelerate the extraction of results, applying WGS directly to primary cultures in an attempt to eliminate interference derived from human DNA present in the sample.[43] Likewise, efforts are beginning to be made to obtain genomic information directly from clinical samples.[44] The results are variable, and still of lower quality than those obtained from cultures, but this approach may offer a first line of analysis providing useful information.

The greater accessibility of WGS analysis has meant that it is not restricted merely to epidemiological studies but has been used to seek answers to some of the challenges discussed above, that had already been analysed using conventional genotyping. WGS has recently been used to tackle the documentation on the percentage of reinfections in a population.[45] Likewise, the capacity of next-generation sequencing to discriminate between different sequences present in a single sample has allowed cases of complex infection to be identified.[46] WGS could be considered to provide a level of resolution that is not necessary to solve some of these problems, especially when an initial search based on low discrimination techniques may be capable of identifying differences between strains without needing to search further. However, its lowered cost actually does make it a competitive alternative, which, in addition to answering these questions, for the same effort and cost, gives us an enormous amount of additional genotyping information that is of interest, going beyond epidemiological needs.

Can I find an alternative to the path opened up by genomic epidemiology?

Although the discriminatory power and the volume of useful epidemiological, therapeutic and phylogenetic information offered by WGS analysis are indisputable, its systematic application is still not feasible in many contexts. In particular, if we look at settings with limited resources, which also have the greatest disease burden, these genomic strategies seem hard to implement.

It is in this gap that emerges between the high resolution offered by WGS-based analysis and the need for monitoring in settings with transmission challenges (that cannot yet be tackled by genomic epidemiology) that an alternative line of progress has taken shape. It consists of developing simple, low-cost, easy-to-implement molecular methods for targeted monitoring of specific strains that are responsible for the most relevant transmission problems in each population.

Specific monitoring of individual strains using simple molecular techniques is not a new proposal. PCRs designed to trace the transmission of the W strain, which caused a large outbreak in New York, that targeted genetic features specific to this strain, were already being developed in the 1990s.[47] The disadvantage of these proposals is that it was only possible to design these PCRs following a comprehensive genetic study of the strains, and only if this uncovered a distinguishing feature. Now, with the accessibility of WGS analysis, we can identify distinctive features of any strain by using whole genome sequencing on a small number of representatives of that isolate, which leads us to identify SNPs specific to that strain.

Based on this premise, allele-specific PCRs have been developed for monitoring strains that are prevalent in a certain population.[48] Similarly, this analysis approach has been applied to enable fast, massive-scale searching of retrospective collections of isolates with the aim of identifying the presence of a strain with a high epidemiological risk in a population through its having been involved in a large outbreak.[49] Finally, the possibility of integrating this strategy to give an early response in the event of a real alert has been demonstrated. The diagnosis of two cases of XDR-TB imported from Russia to Spain activated a system that led to the rapid implementation of specific PCR pathways for SNP markers of these strains, uncovered by WGS analysis and comparison with global SNP databases. Their prospective local use on respiratory samples allowed secondary cases caused by these high-risk strains to be ruled out at an early stage.[50]

How can we tackle the distribution of tuberculosis strains in a global setting?

Up to this point, we have focused exclusively on the usefulness of the different molecular and genomic epidemiology approaches when dealing with challenges in defined populations or specific patients. However, we must not forget that TB is a global phenomenon and has become even more so as a result of international migratory movements. This means that there are also questions of a "macropopulational" nature that need answers. In this sense, as we discussed in the first part of this review, there is a level of MTB cataloguing that does not get into discriminating markers with the aim of discriminating at strain level. Rather, it uses a lower degree of discrimination to catalogue isolates more generically in each of the lineages described for the TB complex. Analysis of the distribution of these lineages provides highly relevant information not only from a global point of view, in that it allows us to determine their international distribution, but also from an evolutionary or phylogenetic point of view.

The first studies on the global distribution of lineages, as well as phylogenetic studies, relied on spoligotyping,[51] which has poor discriminatory power and a tendency towards homoplasmy that do make it ideal for use in field epidemiology, but it is still adequate for analysing the distribution of lineages, families or subfamilies of MTBC. However, the data obtained from WGS analysis also offer an alternative for studies of this type. A set of SNP markers has been proposed for the different lineages that have been found to be consistent and which have simplified the methodology used.[52] Various proposals have sought to simplify the identification of these SNPs, making use of the most widely available laboratory technologies and using different approaches such as real-time PCR,[52] allele-specific PCRs,[53] as well as other analysis formats which are more limited to those laboratories with greater resources, such as Luminex[52] or SNaPshot.[53]

Final comments

M. tuberculosis presents a genetic stability that has forced us to make developments to adapt our methodologies specifically to this disease. Efforts to optimise appropriate techniques for this pathogen have given us an array of methods that cover a wide range of discriminatory powers. Thanks to this, we now have a variety of customisable approaches that we can adapt to suit the needs of each study. We have low-discrimination analysis for studies on phylogenetics, evolution and the global distribution of strains, and high-discrimination methods to precisely discern active transmission chains in a certain population, identify laboratory contamination, characterise recurrences or document microepidemics. The fall in the cost of the analysis of whole chromosomes and the development of more affordable bioinformatic analysis systems have resulted in a natural transition of molecular epidemiology towards a new genomic epidemiology of TB. With the help of these strategies, we now have a much higher discriminatory power, both for the precise identification of transmission chains with solid epidemiological support and for determining the chronology of said chains. WGS analysis has also enabled us to identify SNP markers with a high degree of consistency, both for phylogenetic purposes and for use as markers of relevant strains. Specifically, these SNP strain markers, identified using WGS, may represent a new path for post-genomic advances in the development of specific low-cost, transferable PCRs for prioritised monitoring of high-risk strains in each population. This type of work may give rise to a new model for decentralised, multinodal TB transmission monitoring.

Conflicts of interest

The authors declare that they have no conflicts of interest.

References

1. Niemann S, Supply P. Diversity and evolution of *Mycobacterium tuberculosis*: moving to whole-genome-based approaches. Cold Spring Harb Perspect Med. 2014;4:a021188.
2. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A. 1997;94:9869–74.
3. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, et al. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide

polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol. 2006;188:759–72.

4. 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 U S A. 2002;99:3684–9.
5. 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 U S A. 2006;103:2869–73.
6. Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis. 2007;7:328–37.
7. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol. 1997;35:907–14.
8. 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;42:474–7.
9. Honisch C, Mosko M, Arnold C, Gharbia SE, Diel R, Niemann S. Replacing reverse line blot hybridization spoligotyping of the *Mycobacterium tuberculosis* complex. J Clin Microbiol. 2010;48:1520–6.
10. Cafrune PI, Possuelo LG, Ribeiro AW, Ribeiro MO, Unis G, Jarczewski CA, et al. Prospective study applying spoligotyping directly to DNA from sputum samples of patients suspected of having tuberculosis. Can J Microbiol. 2009;55:895–900.
11. Warren RM, van der Spuy GD, Richardson M, Beyers N, Borgdorff MW, Beh 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.
12. 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.
13. Jagielski T, Minias A, van Ingen J, Rastogi N, Brzostek A, Zaczek A, et al. Methodological and clinical aspects of the molecular epidemiology of *Mycobacterium tuberculosis* and other mycobacteria. Clin Microbiol Rev. 2016;29:239–90.
14. Mokrousov I. *Mycobacterium tuberculosis* Beijing genotype and mycobacterial interspersed repetitive unit typing. J Clin Microbiol. 2006;44:1614 [author reply 1614–15].
15. Hanekom M, van der Spuy GD, Streicher E, Ndabambi SL, McEvoy CR, Kidd M, et al. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. J Clin Microbiol. 2007;45:1483–90.
16. Faksri K, Tan JH, Chaiprasert A, Teo YY, Ong RT. Bioinformatics tools and databases for whole genome sequence analysis of *Mycobacterium tuberculosis*. Infect Genet Evol: J Mol Epidemiol Evol Genet Infect Dis. 2016;45:359–68.
17. Schurch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, et al. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. Infect Genet Evol: J Mol Epidemiol Evol Genet Infect Dis. 2010;10:108–14.
18. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13:137–46.
19. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013;10:e1001387.
20. Sislema-Egas F, Ruiz-Serrano MJ, Bouza E, Garcia-de-Viedma D. Qualitative analysis to ascertain genotypic identity of or differences between *Mycobacterium tuberculosis* isolates in laboratories with limited resources. J Clin Microbiol. 2013;51:4230–3.
21. Yasmin M, le Moullec S, Siddiqui RT, de Beer J, Sola C, Refregier G. Quick and cheap MIRU-VNTR typing of *Mycobacterium tuberculosis* species complex using duplex PCR. Tuberculosis. 2016;101:160–3.
22. Martin A, Inigo J, Chaves F, Herranz M, Ruiz-Serrano MJ, Palenque E, et al. Re-analysis of epidemiologically linked tuberculosis cases not supported by IS6110-RFLP-based genotyping. Clin Microbiol Infect. 2009;15:763–9.
23. Verver S, Warren RM, Munch Z, Richardson M, van der Spuy GD, Borgdorff MW, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet. 2004;363:212–4.
24. 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.
25. Uys P, Brand H, Warren R, van der Spuy G, Hoal EG, van Helden PD. The risk of tuberculosis reinfection soon after cure of a first disease episode is extremely high in a hyperendemic community. PLOS ONE. 2015;10:e0144487.
26. Caminero JA, Pena MJ, Campos-Herrero MI, Rodriguez JC, Afonso O, Martin C, et al. Exogenous reinfection with tuberculosis on a European island with a moderate incidence of disease. Am J Respir Crit Care Med. 2001;163 Pt 1:717–20.
27. Bandera A, Gori A, Catozzi L, degli Esposti A, Marchetti G, Molteni C, et al. Molecular epidemiology study of exogenous reinfection in an area with a low incidence of tuberculosis. J Clin Microbiol. 2001;39:2213–8.
28. Garcia de Viedma D, Marin M, Hernangomez S, Diaz M, Ruiz Serrano MJ, Alcala L, et al. Tuberculosis recurrences: reinfection plays a role in a population whose clinical/epidemiological characteristics do not favor reinfection. Arch Intern Med. 2002;162:1873–9.
29. Garcia de Viedma D, Alonso Rodriguez N, Andres S, Ruiz Serrano MJ, Bouza E. Characterization of clonal complexity in tuberculosis by mycobacterial interspersed repetitive unit-variable-number tandem repeat typing. J Clin Microbiol. 2005;43:5660–4.
30. Shamputa IC, Jugheli L, Sadradze N, Willery E, Portaels F, Supply P, et al. Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. Respir Res. 2006;7:99.
31. Navarro Y, Herranz M, Perez-Lago L, Martinez Lirola M, Indal TB, Ruiz-Serrano MJ, et al. Systematic survey of clonal complexity in tuberculosis at a populational level and detailed characterization of the isolates involved. J Clin Microbiol. 2011;49:4131–7.
32. Perez-Lago L, Herranz M, Lirola MM, Group I-T, Bouza E, Garcia de Viedma D. Characterization of microevolution events in *Mycobacterium tuberculosis* strains involved in recent transmission clusters. J Clin Microbiol. 2011;49:3771–6.
33. Perez-Lago L, Rodriguez Borlado AI, Comas I, Herranz M, Ruiz-Serrano MJ, Bouza E, et al. Subtle genotypic changes can be observed soon after diagnosis in *Mycobacterium tuberculosis* infection. J Med Microbiol. 2016;306: 401–5.
34. Perez-Lago L, Lirola MM, Navarro Y, Herranz M, Ruiz-Serrano MJ, Bouza E, et al. Co-infection with drug-susceptible and reactivated latent multidrug-resistant *Mycobacterium tuberculosis*. Emerg Infect Diseases. 2015;21:2098–100.
35. Perez-Lago L, Herranz M, Navarro Y, Ruiz Serrano MJ, Miralles P, Bouza E, et al. Clonal complexity in *Mycobacterium tuberculosis* can hamper diagnostic procedures. J Clin Microbiol. 2017;55:1388–95.
36. Alonso Rodriguez N, Andrés S, Bouza E, Herranz M, Ruiz Serrano MJ, García de Viedma D, et al. Transmission permeability of tuberculosis involving immigrants, revealed by a multicentre analysis of clusters. Clin Microbiol Infect. 2009;15:435–42.
37. Borrell S, Espanol M, Orcau A, Tudo G, March F, Cayla JA, et al. Tuberculosis transmission patterns among Spanish-born and foreign-born populations in the city of Barcelona. Clin Microbiol Infect. 2010;16:568–74.
38. Lillebaek T, Andersen AB, Rasmussen EM, Kamper-Jorgensen Z, Pedersen MK, Bjorn-Mortensen K, et al. *Mycobacterium tuberculosis* outbreak strain of Danish origin spreading at worrying rates among greenland-born persons in Denmark and Greenland. J Clin Microbiol. 2013;51:4040–4.
39. Pena MJ, Caminero JA, Campos-Herrero MI, Rodriguez-Gallego JC, Garcia-Laorden MI, Cabrera P, et al. Epidemiology of tuberculosis on Gran Canaria: a 4 year population study using traditional and molecular approaches. Thorax. 2003;58:618–22.
40. Gurjav U, Outhred AC, Jelfs P, McCallum N, Wang Q, Hill-Cawthorne GA, et al. Whole genome sequencing demonstrates limited transmission within identified *Mycobacterium tuberculosis* clusters in New South Wales, Australia. PLOS ONE. 2016;11:e0163612.
41. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011;364:730–9.
42. Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir Med. 2014;2:285–92.
43. Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D, Walker TM, et al. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. J Clin Microbiol. 2015;53: 1137–43.
44. Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J Clin Microbiol. 2017; 55:1285–98.
45. Guerra-Assuncao JA, Houben RM, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis. 2015;211:1154–63.
46. Ssengooba W, de Jong BC, Joloba ML, Cobelens FG, Meehan CJ. Whole genome sequencing reveals mycobacterial microevolution among concurrent isolates from sputum and blood in HIV infected TB patients. BMC Infect Dis. 2016;16:371.
47. Plikaytis BB, Marden JL, Crawford JT, Woodley CL, Butler WR, Shinnick TM. Multiplex PCR assay specific for the multidrug-resistant strain W of *Mycobacterium tuberculosis*. J Clin Microbiol. 1994;32:1542–6.
48. Perez-Lago L, Martinez Lirola M, Herranz M, Comas I, Bouza E, Garcia-de-Viedma D. Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: a pilot study. Clin Microbiol Infect. 2015;21, 249 e241-249.
49. Perez-Lago L, Herranz M, Comas I, Ruiz-Serrano MJ, Lopez Roa P, Bouza E, et al. Ultrafast assessment of the presence of a high-risk *Mycobacterium tuberculosis* strain in a population. J Clin Microbiol. 2016;54:779–81.
50. Perez-Lago L, Martinez-Lirola M, Garcia S, Herranz M, Mokrousov I, Comas I, et al. Urgent Implementation in a hospital setting of a strategy to rule out secondary cases caused by imported extensively drug-resistant *Mycobacterium tuberculosis* strains at diagnosis. J Clin Microbiol. 2016;54:2969–74.
51. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, 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.

52. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, et al. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. PLOS ONE. 2012;7: e41253.

53. Carcelén MA, Abascal E, Herranz M, Santantón S, Zenteno R, Ruiz Serrano MJ, et al. Optimizing and accelerating the assignation of lineages in *Mycobacterium tuberculosis* using novel alternative single-tube assays. PLOS ONE. 2017;12, e0186956.