Infection, Genetics and Evolution 72 (2019) 4–9

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Pathways and strategies followed in the genomic epidemiology of *Mycobacterium tuberculosis*

Darío García de Viedma[a,b,c,*,1]

[a] Servicio de Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Madrid, Spain
[b] Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain
[c] CIBER Enfermedades respiratorias CIBERES, Spain

A B S T R A C T

The progressive reduction of costs in next-generation sequencing is responsible for the speed with which new genomic epidemiological approaches are being used. However, this speed has meant a lack of consensus on the way the genomic pathway is being addressed. Alternative pathways, strategies, and shortcuts have been proposed during this initial period of the genomic epidemiology era in tuberculosis. The aim of this review is not to make a systematic analysis of these different approaches but to show how various lines of progression are being followed, each looking for different ways of integrating the language of genomics. This review covers several aspects, from paths that provide high-quality data from cultured isolates to strategies that attempt to shorten response times through challenging analyses directly on specimens or primary cultures. The review presents strategies proposed by several groups, ranging from those that focus on universal population-based systematic application to others proposing shortcuts by targeting selected relevant strains. Finally, the decision to analyze complete genomic content *vs* abbreviated analysis of preselected sets of genes is discussed. The reader is shown the exciting variety of efforts being made to find the best fit between genomics and the demands and challenges of the epidemiology of tuberculosis.

During recent decades, we have witnessed how molecular epidemiology has transformed our knowledge of the dynamics of transmission of *Mycobacterium tuberculosis* (MTB). Identification of transmission clusters has been based on various genotyping methods, which have been sequentially replaced in the search for higher discrimination, faster availability of data, or other, practical issues such as efficiency in sharing results for comparison or enabling automatic assignation of the genotyping pattern. Thus, the reference method IS6110 RFLP (van Embden et al., 1993) was displaced by a PCR-based approach, namely MIRU-VNTR (Supply et al., 2001; Supply et al., 2006), which underwent serial transformations from a limited discriminative 12-locus format to the improved high-throughput analysis based on the 15- and 24-locus versions. These latest genotyping tools even allowed us to obtain results directly from respiratory specimens (Alonso et al., 2012) or from the remnants of other commercial molecular tests (Mambuque et al., 2018; Alame-Emane et al., 2017), thus paving the way towards real-time molecular epidemiology and acquisition of data on transmission dynamics in settings where culture is not available.

Professionals working in molecular epidemiology were reasonably satisfied with the fingerprinting tools available for the study of MTB. However, in a relatively short period, molecular epidemiology has been replaced by a new approach to the understanding of transmission of TB, namely, genomic epidemiology, which is based on the unbeatable discriminatory power obtained from whole genome sequencing (WGS) data.

1. Genomic epidemiology based on WGS

The first steps in genomic epidemiology in TB resemble those of the molecular epidemiology era. Standard genotyping initially followed an "a la carte" scheme, which focused on the analysis of previously suspected outbreaks in order to confirm them or rule them out. Genotyping gradually moved on to systematic universal population-based analysis.

Similarly, the first application of WGS in TB was to analyze a major outbreak in Vancouver, Canada (Gardy et al., 2011). This made it possible to split an extensive MIRU-VNTR cluster into 2 transmission networks, thus demonstrating the high discriminatory power and epidemiological usefulness of genomic epidemiology. In 2013, Walker *et al* (Walker et al., 2013a) published their reference paper on the application of WGS to define outbreaks in communities and households, enriching the analysis with longitudinal within-patient isolates. This robust study was the first to propose thresholds of diversity for inferring recent transmission from WGS data. The authors identified epidemiological links for those cases differing in 5 or fewer SNPs, whereas no links were found for cases differing in > 12 SNPs. They also reported a rate of change of 0.5 SNPs per genome per year.

* Corresponding author.
E-mail address: dgviedma2@gmail.com.
[1] Phone: (+34) 914265104. Postal address C/Dr. Esquerdo 46, 28007 Madrid, Spain.

https://doi.org/10.1016/j.meegid.2019.01.027
Received 29 October 2018; Received in revised form 15 January 2019; Accepted 18 January 2019
Available online 22 January 2019
1567-1348/

The study yielded other relevant findings, such as the informative value of the topology of the networks of relationships derived from the distribution of SNPs in a cluster. For example, of extraordinary epidemiological relevance was the star-like distribution, which informed us about the presence of a superspreader at the origin of the transmission. Additionally, the identification of these superspreaders has been described to be aided by the identification of cases with mixed-base calls (Walker et al., 2013a), likely due to the emergence of subpopulations within those cases with longer than average infectiousness periods. Finally, as backward mutations are rare in MTB, the order in which SNPs accumulate in a cluster gives us valuable information about the sequence of transmission, which could never be inferred from standard molecular epidemiology data (Walker et al., 2013b).

Extensive outbreaks were also analyzed using WGS. The most extensive probably corresponds to an INH-monoresistant outbreak in London, UK (Casali et al., 2016). WGS of 344 isolates spanning 14 years revealed that none of the isolates differed in > 9 SNPs, thus confirming the high clonality of the cluster.

The robustness of proposed SNP thresholds has been called into question, owing to the finding of higher than expected within-patient diversity in certain cases. Casali et al. (Casali et al., 2016) performed a subanalysis on within-patient diversity by analyzing multiple single colonies and found that it reached a maximum of 10 SNPs. In their reference article, Walker et al. (Walker et al., 2013a) found intrapatient diversity in up to 9 SNPs. Herranz et al. (Herranz et al., 2017) also performed a more systematic analysis of within-patient diversity and found that it could be equivalent to that accumulated in patient-to-patient transmission. These findings seem to indicate that the magnitude of intrapatient diversity can modify the SNP reference cutoffs that had been defined to infer transmission. However, the robustness of the original thresholds proved to be valid in subsequent studies and even in challenging situations, such as extended prolonged clusters, clusters with the coincidence of host-to-host transmission and reactivations, and long latency periods for some cases in the cluster (Herranz et al., 2017). All of these situations could lead us to expect higher diversity *a priori*. Even under such challenging circumstances, the initially proposed diversity thresholds proved to be valid.

Studies that restricted the application of WGS to specific circumstances or relevant outbreaks were followed by more systematic population-based studies. Examples of the systematic application of universal WGS have been reported in Oxfordshire, UK, in 2007–12 (Walker et al., 2014) or in a long-term survey of a whole district in Malawi (Guerra-Assuncao et al., 2015). In both cases, the data obtained from genomics proved to be extremely useful and superior to standard molecular epidemiology data. Prospective efforts to implement WGS strategies have also been undertaken, as in the study by Pankhurst et al. (Pankhurst et al., 2016), where a centralized pipeline was applied to analyze the data produced over 8 months from 8 laboratories in Europe and North America, thus ensuring fast response times and proving to be financially feasible. Finally, the results of a prospective, nationwide genomic epidemiology program were recently published in The Netherlands (Jajou et al., 2018a). A study performed in parallel to standard MIRU-VNTR analysis and involving 535 isolates obtained in a complete year showed that WGS reduced the clusters defined by MIRU-VNTR by 50% without reducing the percentage of epidemiological links detected.

2. Core genome–based analysis

WGS is clearly superior to standard genotyping for ensuring a highly discriminative definition of clusters and for providing us with valuable additional data on the structure of the transmission links and the chronology of this transmission. However, the analytical procedures depend on in-house pipelines for mapping sequences and SNP calling, which are not easy to standardize.

Kohl et al. (2014, 2018) proposed core genome (cg) MLST as an alternative to the standard path of WGS data. This approach restricts the analysis to the core genome of MTB, which they defined as comprising 3257 genes (77% of the whole genome used as the reference). As cgMLST focuses on core genes, where the likelihood of false SNP calls is lower, however, it still requires a proper standardization of the pipeline used to call the SNPs. The restriction of the analysis to the core genes, makes cgMLST less computationally demanding, and the transference of the SNPs identified in the core genome to an allele numbering system facilitates a more standardized reporting.

cgMLST was evaluated in parallel with standard WGS to analyze an outbreak in Hamburg spanning the period 2001–10, which included 26 cases sharing identical genotypic patterns. Given the more limited genomic material analyzed by cgMLST, the approach offered slightly lower power of resolution than WGS, which led to a lower overall number of differences between isolates. However, both the topology of the genome-based trees and the grouping of the cases were equivalent for both strategies, thus validating the cgMLST proposal for epidemiological purposes. More recently (Meehan et al., 2018) it has been described that WGS-SNP and cgMLST based analyses have similar clustering/timing characteristics even for data obtained from a high incidence setting.

3. Targeted surveillance based on selected data extracted from WGS

This review has examined several approaches in which considerable technical transformations have been implemented, although these have all been based on strategies identical to those followed with our pregenomic molecular tools, that is, detection of transmission events through the identification of clustered strains according to their similarity after a non-biased population-based analysis. While this is undoubtedly the most appropriate way of proceeding, we must sacrifice the speed we had acquired with our primitive molecular tools, which provided fingerprinting data when intervention was still possible, even directly from clinical specimens (Alonso et al., 2012; Bidovec-Stojkovic et al., 2014). If we pursue high-quality WGS data, we must wait until subcultures are available, thus considerably delaying the availability of data. In addition, we were able to extend the applicability of molecular epidemiology out of high-resource countries thanks to non-technologically demanding and inexpensive approaches. Now, however, we seem to have lost that gain, owing to a new dependence on complex technology, which, despite being less expensive than before, remains out of the reach of many countries, especially those with the highest TB burden.

In order to address this issue, we propose an alternative line of progression, namely, one which tries to reconcile the discriminative power of WGS with the speed, low cost, and simplicity of PCR-based approaches. The tradeoff of this shortcut is that we must sacrifice complete knowledge of all the transmission clusters in a population, because it is necessary to monitor strains that, based on our previous knowledge taken from standard molecular epidemiology, deserve special attention because they are more actively transmitted or are high-risk MDR or XDR strains.

The scheme is based on identifying specific SNPs from the strain that is to be surveyed. These are identified by WGS of a selection of representative isolates. The SNPs are considered to be strain-specific once they have passed through a filtering process using a database of SNPs from more than 4000 strains representing all the lineages circulating worldwide (Álvaro Chiner-Oms et al., 2018). They are then used to design allele-specific (ASO)-PCRs that are tailored to target them. The ASO-PCRs, known as tailored regional allele-specific PCR (TRAP), are designed in a multiplex format to investigate the presence of several of these strain-specific SNPs and thus ensure high specificity. In the design, some of the primers are homologous with the allele present in the strain, while the remaining primers are designed to be homologous to the complementary allele. This ensures that amplification patterns are always obtained, regardless of whether or not the isolate

interrogated corresponds to the surveyed strain, thus ruling out misassignment due to inhibitors.

Pérez-Lago et al. (Perez-Lago et al., 2015) showed this shortcut TRAP strategy to be efficient for active surveillance of transmitted strains in a population with a high proportion of immigrants in Almería, southeast Spain, by revealing secondary cases infected by the surveyed strains, even through direct analysis of the bacilli present in the respiratory specimens. The authors also demonstrated that the TRAP-based strategy could offer a quick response to a public health alert, as happened with the identification of two XDR Beijing strains imported from Russia to Spain (Perez-Lago et al., 2016a). Twenty-four days after receiving the primary cultures from these 2 cases, 2 strain-specific PCRs were implemented in the local laboratory investigating the transmission and provided results from stain-positive specimens and on cultures from those that were stain-negative.

The sensitivity, simplicity, and low cost of the TRAP approach also made it possible to perform fast, high-throughput updates of retrospective collections that are to be interrogated about the presence of a specific strain. In 1 week, Pérez-Lago *et al* (Perez-Lago et al., 2016b) analyzed 964 isolates from 2 collections from 2 major hospitals in Madrid. The authors aimed to identify secondary cases that could have been caused by a patient infected by the Beijing strain which caused a severe outbreak in Gran Canaria island and who had had active TB disease for the previous 8 years owing to lack of adherence to treatment. The same strain-specific PCR targeting specific SNPs was also applied to update the current situation of this strain in the context where it caused the outbreak. The approach enabled us to confirm the high prevalence of the strain more than 2 decades after the outbreak, as well as its widespread presence on other islands in the archipelago that had not previously been monitored (Pérez-Lago et al., 2019).

An equivalent shortcut strategy for fast update of a relevant event was followed to analyze a major outbreak in Bern in the 1990s (22 cases) (Stucki et al., 2015). WGS analysis of all isolates made it possible to identify strain-specific SNPs, which were targeted by a TaqMan probe in an RT-PCR format. The technique was applied to complete the description of the cluster over 2 decades by analyzing 1642 isolates, which revealed a further 46 members of the cluster. Targeted WGS analysis of all 68 cases shed light on the true structure of the cluster with 4 subclusters that had not been identified by standard genotyping. The screening of cases based on RT-PCR was much faster and 15 times less expensive than standard MIRU-VNTR analysis.

Finally, the TRAP approach proved useful for obtaining relevant information in settings where culture facilities are not available. A specific PCR applied directly on the Xpert remnants revealed a prevalent MDR strain in Equatorial Guinea that was responsible for a high percentage of the resistant cases in the country (Perez-Lago et al., 2017).

Despite the potential of the targeted surveillance based on selected data extracted from WGS presented in this section, we must accept that its application to survey transmission in high-burden countries still needs the support of well-equipped laboratories. Despite not requiring a systematic universal genomic analysis, which reduces severely the costs, the approach still needs a WGS-based study of a preselected sample of strains, which is not feasible to many high-burden countries.

4. Real-time genomic analysis

The absence of a fast response enabling transmission to be detected sufficiently early, which is shared by the developments presented in the previous section, has led various authors to attempt to shorten the delay in making genomic data available by analyzing primary cultures without waiting until subcultures are available, or even by directly examining bacteria in sputa.

The challenge for these approaches is the presence of contaminating human and bacterial DNA (other than that of MTB), which interferes with sequencing steps and makes it difficult to obtain a suitable coverage depth to call SNPs with confidence. One test (Deeplex^{R}-MycTB, GenoScreen) was developed with aim of applying it directly on specimens. However, it is not designed to identify all the potential SNPs that are necessary for epidemiological purposes. The test aims to obtain reliable calls for phylogenetic marker SNPs and those in 18 main targets associated with resistance to first- and second-line drugs. Instead of performing complete WGS, Deeplex^{R}-MycTB is based on targeted 24-plex amplicon deep sequencing (Tagliani et al., 2017), which, while ensuring greater depth of analysis, foregoes the information from regions other than those targeted. This test was recently applied in the description of an outbreak of MDR-TB in SouthAfrica caused by strains harbouring resistance mutations not detected by the Xpert test (Makhado et al., 2018).

In the case of real-time standard complete WGS applied for genomic epidemiology purposes, 2 alternative paths have been followed to minimize the interference of accompanying non-MTB DNA. These involve either deleting it from the specimens at the purification stage or, alternatively, enriching the presence of MTB-DNA just before sequencing.

The optimization of purification methods to minimize the presence of accompanying non-MTB DNA was proposed by Votintseva *et al* (Votintseva et al., 2015), who introduced a pretreatment saline wash to remove human DNA and a subsequent clean-up of the DNA with solid-phase reversible immobilization beads. This approach revealed whole genome sequences from 1 ml of early positive liquid cultures (MGIT), which were successfully mapped to the reference MTB genome with > 90% coverage. A modification of this method was applied directly on 40 stain-positive sputa, of which 62% yielded sufficient data to make predictions on susceptibility for first- and second-line drugs (Votintseva et al., 2017). Pretreatment procedures have also been undertaken to eliminate human DNA in other studies, in this case using differential lysis steps to first lyse human cells and degrade their DNA before applying shotgun metagenomics (Doughty et al., 2014).

The application of WGS directly on early-positive MGIT primary liquid cultures in a countrywide WGS-based model was evaluated prospectively over 20 months in Italy. The samples were prepared by an automatic procedure, and results were available within 72 h after delivery Thus, diagnosis, surveillance, and contact tracing were improved, although contact tracing was based on cgMLST (Cabibbe et al., 2018). In Australia, successful country-wide systematic prospective WGS was applied to identify resistance mutations, assign lineages, and identify transmission clusters and laboratory cross-contamination (Martínez et al., 2018).

The alternative path of enriching the MTB-DNA to minimize interference from contaminating DNA during sequencing was followed by Brown et al. (Brown et al., 2015), who used biotinylated RNA baits to specifically capture the MTB-DNA on 24 stain-positive sputa. The authors obtained sufficient coverage to call mutations associated with resistance to first- and second-line drugs in all but 4 specimens. The same group applied their approach in real time to individually adjust treatment in a case infected with a resistant strain, thanks to the extended information on resistances obtained from WGS directly on sputa, compared with that offered by the standard molecular commercial tests (Nimmo et al., 2017). More recently, Doyle et al. (Doyle et al., 2018) evaluated the feasibility of performing WGS directly on sputa, with the introduction of an enrichment step.

However, the above mentioned studies restrict the analysis to calling of SNPs with diagnostic value for identification of MTB or prediction of resistance. No data have been reported on the efficiency of performing an extensive SNP call with epidemiological purposes. However, data have been reported from a comparison of the diversity between the sputa and corresponding cultures (Votintseva et al., 2017) and from studies placing the isolates in a phylogenetic tree (Votintseva et al., 2017; Brown et al., 2015). Preliminary data to obtain the complete set of SNPs to perform real-time genomic epidemiology were obtained by applying a novel pan-genome capture platform (Lozano

et al., 2018), although the approach requires improvements to be made.

5. New models to address new challenges

In addition to the challenge of ensuring accurate identification in a highly refined analysis of transmission of TB, which has been addressed by novel genomic epidemiology approaches, new specific challenges require the application of new strategies. Global migratory movements have blurred the limits of the target population for epidemiological surveillance. We must now activate multinational cross-border surveillance to fully understand the complexity of transmission in the new global scenario. Consequently, we must be prepared to differentiate between overlapping phenomena, such as discriminating between cases involved in an international cross-border cluster. Such cases can be the result of i) recent transmission after arrival in the host country, ii) transmission in the country of origin, iii) independent importations of a strain prevalent in the country of origin, or iv) transmission along the migration route. Recent studies have demonstrated that standard molecular approaches cannot address this challenge and show the same fingerprinting pattern for all or most of the cases in the previous categories. Below, I review recent studies offering precise snapshots of some of these new complex situations, which are increasingly common owing to the new global nature of TB transmission.

5.1. Refined analysis of transnational outbreaks

Integrated multinational efforts are helping us to clarify potential cross-border events, which mostly continue to be identified based on standard molecular epidemiology data, but can now be examined in greater depth using WGS. Such was the case of a joint cross-border investigation of an MDR cluster involving 3 European countries (Fiebig et al., 2017). Five cases sharing the same MDR strain were identified in Austria in 2014. The finding that 3 of these were from Romanian migrants triggered an integrated analysis of a further 3 cases in Germany and 5 in Romania, also based on identical MIRU-VNTR or potential epidemiological links. The initial suspicion of an internationally spread common strain led to a new reinterpretation based on WGS data. In fact, of the 3 independent events identified, 2 involved different countries and the third involved transmission within Austria. Imported cases coexisted in one of the 2 multinational events, with recent transmission at the destination, whereas in the other, all cases were more likely independent importations from exposures in the city of origin in Romania.

Acosta et al. have been involved in a study on the analysis of another cross-border—in this case intercontinental—outbreak of an MDR strain involving Latin America and Europe (Acosta et al., 2019). The comparison of TB-Sprint (Molina-Moya et al., 2017) and MIRU-VNTR patterns from isolates in Lima, Madrid, Florence, and Milan revealed the existence of cases infected by an MDR strain that was circulating in Lima and was exported to Europe. WGS enabled a more precise analysis of the event. At least 2 variants were exported, likely owing to the diversity acquired by the strain in Lima after prolonged periods. One defined a branch in the network of relationships depicted from the SNPs identified by WGS, including a limited number of cases in Italy and Spain, while the other was responsible for a large outbreak in Florence that is currently active. This analysis led to the development of a specific PCR for this MDR strain, which was shared by all the countries involved to enable simplified simultaneous multinational prospective surveillance.

5.2. Transmission “en route”

The previous section described clusters involving migrant cases in various countries to show how a strain that is prevalent in the country of origin is transmitted in different settings once distributed. The application of WGS in an MDR-TB outbreak involving migrants from the Horn of Africa and Sudan in 7 European countries recently revealed a different transmission scenario (Walker et al., 2018). Twenty-nine cases were closely related (< 2SNPs). The study involved a thorough epidemiological investigation based on detailed questionnaires and interviews and revealed that most of the cases had been living in overcrowded conditions for prolonged periods in a Libyan town during the migratory route. These data indicate a novel likely explanation for a multinational cluster, ie, transmission due to exposure “en route”, which should be considered in the epidemiological investigation of migrant clusters from now on.

5.3. Discrimination between recent transmission and independent importations

In settings with a high percentage of migrants, transmission clusters rich in cases from a single nationality are common. Clusters could correspond to recent transmission in the host country after arrival or to independent importations of a strain that is prevalent in the country of origin. Discrimination between these 2 alternatives is highly relevant from a public health perspective.

Molecular epidemiology fails to differentiate between these alternatives and offers identical patterns for all the cases involved. A recent study in the UK investigated a cluster that was rich in Filipino migrants (Davidson et al., 2018), even in the absence of high discriminatory WGS. 24-Locus MIRU-VNTR grouped 53 cases over 6 years; 43 were born in the Philippines, and of these, 21 were health care workers. The additional analysis based on 8 additional VNTR loci split the cluster, that is, the initial interpretation of recent transmission in the UK was replaced by the alternative explanation of independent importations of a prevalent strain from the country of origin, with variants not detected by standard genotyping.

The suspicion that MIRU-VNTR–based clusters including migrants in low-incidence countries might not always be indicators for recent transmission in the host countries was also supported by a study in Switzerland (Stucki et al., 2016). All 35 MIRU-VNTR–defined clusters during 2000-08 were revisited by applying WGS. Most of the clusters including Swiss-born patients were confirmed, although very few of those involving foreign-born patients were confirmed. The percentage of clustering among foreign-born patients decreased from 17% to 7% when MIRU-VNTR clusters were reanalyzed using WGS. Similar findings were obtained in The Netherlands and Denmark (Jajou et al., 2018b), where a large cluster of 40 migrants from the Horn of Africa sharing an identical MIRU-VNTR pattern was split by WGS into several subclusters, thus ruling out transmission in the country of destination.

This problem was analyzed more systematically by Abascal *et al* (Abascal et al., 2019), who selected clusters including migrants in Almería, southeast Spain, from 3 geographic areas (Eastern Europe, North Africa, and Sub-Saharan Africa) that were representative of clusters rich in a single nationality. The authors demonstrated how WGS could differentiate the cases corresponding to independent importations from those due to recent transmission after arrival. The differentiation was based on the number of SNPs detected for each of these 2 categories of patients. A low number of SNPs was detected for cases resulting from recent transmission after arrival, whereas a much higher number of SNPs was detected for cases resulting from independent importations of a strain that was prevalent in the country of origin and had likely been circulating for a long period, thus enabling the acquisition of higher diversity (> SNPs).

The findings reported by Abascal *et al* (Abascal et al., 2019) proved that the complexity behind transmission clusters in the new global TB scenario renders standard molecular epidemiology useless. Only WGS offers sufficient discriminatory power to enable accurate interpretation of the new complexity of transmission. However, again, systematic worldwide application of genomic epidemiology is far from realistic. The authors followed a shortcut based on the application of strain-specific PCRs to analyze the most complex cluster in our study in

Almería, which involved 11 cases, most of which were Moroccan migrants. They then identified the SNPs that were only shared by the 6 cases involved in recent transmissions and not shared by the remaining cases, which accumulated high diversity, indicating that they were independent importations from a strain that was prevalent in Morocco. A strain-specific PCR was developed to target the SNPs that were specific for the recent transmission subcluster. This was applied prospectively and enabled us to identify 2 new cases belonging to the recent transmission group. All these cases were indistinguishable according to standard MIRU-VNTR data. A simplified version of this PCR was designed to identify representatives of this strain in the country of origin, Morocco. The PCR was applied directly on inactivated aliquots of a retrospective collection of isolates from North Morocco and revealed, as expected, that the strain was prevalent in the country of origin of these migrants.

6. Challenges

Considerable efforts have been made to establish SNP-based diversity cutoffs that are shared by most studies on genomic epidemiology in TB. Despite the robustness of the study determining these cut-offs, they have been validated mainly in low-incidence settings. It is time to extend the evaluation of their usefulness to high-burden countries, to validate them in more challenging circumstances, where the distinction between 'outbreak' and persistent ongoing transmission might be more blurred.

It is pointless to strictly define thresholds to be shared if no equivalent efforts are made for the stages that lead to the identification of SNPs.

The first stage where standardization is required is the reference to be used to map sequences, because this will affect all subsequent analysis of SNPs. Most studies use H37Rv as the reference. However, this seems arbitrary, and the selection of another reference, with chronological significance, could help us to better establish time references and thus accurately order the acquisition of SNPs in our strains. In this sense, the most recent ancestor strain has been proposed as a reference. This ancestral MTB genome is identical to H37Rv in terms of structure but includes the maximum likelihood–inferred ancestral nucleotide positions from a virtual ancestor (Comas et al., 2013). By using this reference, the chronology of acquisition of strains will be more meaningful owing to the application of a reference that is a true evolutionary reference for the strains under study. In contrast, when H37Rv is used as an arbitrarily selected strain, SNPs shared with H37Rv are not called, thus leading us to miss valuable evolutionary information.

Despite the lack of consensus on a reference, various pipelines of analysis have been developed by different groups and we must now homogenize the criteria used to filter data in these pipelines, namely, coverage depth, percentage of reads requested to distribute calls as homozygous or heterozygous, requirements for interpreting a SNP as true or resulting from an incorrect call, and the addition of visual inspectors to supervise at least the most relevant calls. The absence of consensus on these levels of analysis is likely causing differences in the data obtained from each group.

WGS-based analysis of the transmission of TB has the added value of revealing the chronology and direction of a transmission event. However, although it may seem obvious, we still need to identify SNPs between clustered isolates to be able to establish the direction of transmission. The analysis of the isolates in a cluster frequently leads some isolates to be considered identical (0 SNPs), thus limiting this approach. It would be desirable to access the diversity that remains hidden to our systems of analysis because it lies on indels or SNPs in repetitive regions in MTB DNA. Since these regions are deleted systematically from the analysis owing to the technical limitations in mapping these regions properly, SNP calling in these regions is not feasible. These regions could account for 10% of the total genome (Casali et al., 2016). The limitation could be resolved by improving the ability to obtain longer reads and thus access the likely diversity that remains hidden with current procedures. Such an approach would allow us to complete the precise chronology of all the members in the cluster, especially in those contexts where the proportion of cases showing 0 SNPs is high (Casali et al., 2016).

Despite progress in performing WGS on early primary cultures and specimens, no efforts have been made to go beyond calling SNPs with diagnostic value. Advances need to be made to try to expand the analysis to a complete call of SNPs with epidemiological value.

Acknowledgments

We thank Thomas O'Boyle for proofreading the manuscript. This project was funded by ERANET-LAC (ELAC2015/T08-0664, E035-ERAnet-LAC/J110-2016/FONDECYT, PER-2012-ELAC2015/T08-0664) and ISCIII (AC16/00057, FIS15/01554, FIS13/01207) and cofunded by ERDF Funds from the European Commission: "A way of making Europe".

References

Abascal, E., Perez-Lago, L., Martínez-Lirola, M., Chimer-Oms, A., Herranz, M., Chaoui, I., Comas, I., El Messaoudi, M.D., Garrido Cárdenas, J.A., Santantón, S., Bouza, E., García de Viedma, D., 2019. Whole-Genome Sequencing–Based Analysis of tuberculosis in Migrants: Rapid Tools for Cross-Border Surveillance and to Distinguish between Recent Transmission in the Host Country and New Importations. Euro Surveill. 24 (4) (Jan).

Acosta, F.A., Cabibbe, A.M., Cáceres, T., Sola, C., Pérez-Lago, L., Abascal, E., Herranz, M., Meza, E., Klotoe, B., Muñoz, P., Rossolini, G.M., Bartoloni, A., Tortoli, E., Cirillo, D.M., Gotuzzo, E., García de Viedma, D., 2019. Exportation of Multidrug-Resistant Tuberculosis to Europe from a Setting with Actively Transmitted Persistent Strains in Peru. Emerg. Infect. Dis. 25 (3) (March).

Alame-Emane, A.K., Pierre-Audigier, C., Aboumegone-Biyogo, O.C., Nzoghe-Mveang, A., Cadet-Daniel, V., Sola, C., Djoba-Siawaya, J.F., Gicquel, B., Takiff, H.E., 2017. Use of GeneXpert remnants for drug resistance profiling and molecular epidemiology of tuberculosis in Libreville, Gabon. J. Clin. Microbiol. 55, 2105–2115.

Alonso, M., Herranz, M., Martinez Lirola, M., I.-T. Group, Gonzalez-Rivera, M., Bouza, E., Garcia de Viedma, D., 2012. Real-time molecular epidemiology of tuberculosis by direct genotyping of smear-positive clinical specimens. J. Clin. Microbiol. 50, 1755–1757.

Bidovec-Stojkovic, U., Seme, K., Zolnir-Dovc, M., Supply, P., 2014. Prospective genotyping of *Mycobacterium tuberculosis* from fresh clinical samples. PLoS One 9, e109547.

Brown, A.C., Bryant, J.M., Einer-Jensen, K., Holdstock, J., Houniet, D.T., Chan, J.Z., Depledge, D.P., Nikolayevskyy, V., Broda, A., Stone, M.J., Christiansen, M.T., Williams, R., McAndrew, M.B., Tutill, H., Brown, J., Melzer, M., Rosmarin, C., McHugh, T.D., Shorten, R.J., Drobniewski, F., Speight, G., Breuer, J., 2015. Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. J. Clin. Microbiol. 53, 2230–2237.

Cabibbe, A.M., Trovato, A., De Filippo, M.R., Ghodousi, A., Rindi, L., Garzelli, C., Baretti, S., Allodi, G., Mannino, R., Rossolini, G.M., Bartoloni, A., Tortoli, E., Cirillo, D.M., 2018. Countrywide implementation of whole genome sequencing: an opportunity to improve tuberculosis management, surveillance and contact tracing in low incidence countries. Eur. Respir. J. 51.

Casali, N., Broda, A., Harris, S.R., Parkhill, J., Brown, T., Drobniewski, F., 2016. Whole genome sequence analysis of a large isoniazid-resistant Tuberculosis Outbreak in London: a retrospective observational study. PLoS Med. 13, e1002137.

Chiner-Oms, A., Sánchez-Busó, L., Corander, Jukka, Gagneux, Sebastien, Harris, Simon, Young, Douglas, González-Candelas, Fernando, Comas, Iñaki, 2018. Genomic Determinants of Sympatric Speciation of the *Mycobacterium tuberculosis* Complex across Evolutionary Timescales. bioRxiv 314559.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S.R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D., Gagneux, S., 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat. Genet. 45, 1176–1182.

Davidson, J.A., Fulton, N., Thomas, H.L., Lalor, M.K., Zenner, D., Brown, T., Murphy, S., Anderson, L.F., 2018. Investigating a tuberculosis cluster among Filipino health care workers in a low-incidence country. Int. J. Tuberculosis Lung Dis. 22, 252–257.

Doughty, E.L., Sergeant, M.J., Adetifa, I., Antonio, M., Pallen, M.J., 2014. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and M. Africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ 2, e585.

Doyle, R.M., Burgess, C., Williams, R., Gorton, R., Booth, H., Brown, J., Bryant, J.M., Chan, J., Creer, D., Holdstock, J., Kunst, H., Lozewicz, S., Platt, G., Romero, E.Y., Speight, G., Tiberi, S., Abubakar, I., Lipman, M., McHugh, T.D., Breuer, J., 2018. Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. J. Clin. Microbiol. 56.

Fiebig, L., Kohl, T.A., Popovici, O., Muhlenfeld, M., Indra, A., Homorodean, D., Chiotan, D., Richter, E., Rusch-Gerdes, S., Schmidgruber, B., Beckert, P., Hauer, B., Niemann, S., Allerberger, F., Haas, W., 2017. A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt. Euro Surveill. 22.

Gardy, J.L., Johnston, J.C., Ho Sui, S.J., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M.K., Elwood, K., Jones, S.J., Brinkman, F.S., Brunham, R.C., Tang, P., 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med. 364, 730–739.

Guerra-Assuncao, J.A., Crampin, A.C., Houben, R.M., Mzembe, T., Mallard, K., Coll, F., Khan, P., Banda, L., Chiwaya, A., Pereira, R.P., McNerney, R., Fine, P.E., Parkhill, J., Clark, T.G., Glynn, J.R., 2015. Large-scale whole genome sequencing of M. Tuberculosis provides insights into transmission in a high prevalence area. elife 4.

Herranz, M., Pole, I., Ozere, I., Chiner-Oms, A., Martinez-Lirola, M., Perez-Garcia, F., Gijon, P., Serrano, M.J.R., Romero, L.C., Cuevas, O., Comas, I., Bouza, E., Perez-Lago, L., Garcia-de-Viedma, D., 2017. *Mycobacterium tuberculosis* acquires limited genetic diversity in prolonged infections, reactivations and transmissions involving multiple hosts. Front. Microbiol. 8, 2661.

Jajou, R., de Neeling, A., van Hunen, R., de Vries, G., Schimmel, H., Mulder, A., Anthony, R., van der Hoek, W., van Soolingen, D., 2018a. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: a population-based study. PLoS One 13, e0195413.

Jajou, R., de Neeling, A., Rasmussen, E.M., Norman, A., Mulder, A., van Hunen, R., de Vries, G., Haddad, W., Anthony, R., Lillebaek, T., van der Hoek, W., van Soolingen, D., 2018b. A predominant variable-number tandem-repeat cluster of *Mycobacterium tuberculosis* isolates among Asylum seekers in the Netherlands and Denmark, Deciphered by Whole-Genome Sequencing. J. Clin. Microbiol. 56.

Kohl, T.A., Diel, R., Harmsen, D., Rothganger, J., Walter, K.M., Merker, M., Weniger, T., Niemann, S., 2014. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. J. Clin. Microbiol. 52, 2479–2486.

Kohl, T.A., Harmsen, D., Rothganger, J., Walker, T., Diel, R., Niemann, S., 2018. Harmonized genome wide typing of Tubercle Bacilli using a web-based gene-By-gene nomenclature system. EBioMedicine 34, 131–138.

Lozano, N.H., Fernández, V., Alcaide, F., Tudó, G., Suárez, J., Chiner-Oms, A., Comas, I., Muñoz, P., García de Viedma, D., Pérez-Lago, L., 2018. Approaches to Whole Genome Sequencing of *Mycobacterium tuberculosis* from Clinical Samples: First Steps for a New Specific TB-DNA Capture Platform. ECCMID, Madrid.

Makhado, N.A., Matabane, E., Faccin, M., Pincon, C., Jouet, A., Boutachkourt, F., Goeminne, L., Gaudin, C., Maphalala, G., Beckert, P., Niemann, S., Delvenne, J.C., Delmee, M., Razwiedani, L., Nchabeleng, M., Supply, P., de Jong, B.C., Andre, E., 2018. Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. Lancet Infect. Dis. 18, 1350–1359.

Mambuque, E.T., Abascal, E., Venter, R., Bulo, H., Bouza, E., Theron, G., Garcia-Basteiro, A.L., Garcia-de-Viedma, D., 2018. Direct genotyping of *Mycobacterium tuberculosis* from Xpert((R)) MTB/RIF remnants. Tuberculosis 111, 202–206.

Martínez, E.M., Hennessy, D., Crighton, T., Wang, Q., Sintchenko, V., 2018. Clinical and Public Health Benefits of Prospective Genome Sequencing of *Mycobacterium tuberculosis*. ECCMID, Madrid.

Meehan, C.J., Moris, P., Kohl, T.A., Pecerska, J., Akter, S., Merker, M., Utpatel, C., Beckert, P., Gehre, F., Lempens, P., Stadler, T., Kaswa, M.K., Kuhnert, D., Niemann, S., de Jong, B.C., 2018. The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology. EBioMedicine 37, 410–416.

Molina-Moya, B., Gomgnimbou, M.K., Lafoz, C., Lacoma, A., Prat, C., Refregier, G., Samper, S., Dominguez, J., Sola, C., 2017. Molecular characterization of *Mycobacterium tuberculosis* strains with TB-SPRINT. Am. J. Trop. Med. Hygiene 97, 806–809.

Nimmo, C., Doyle, R., Burgess, C., Williams, R., Gorton, R., McHugh, T.D., Brown, M., Morris-Jones, S., Booth, H., Breuer, J., 2017. Rapid identification of a *Mycobacterium tuberculosis* full genetic drug resistance profile through whole genome sequencing directly from sputum. Int. J. Infect. Dis. 62, 44–46.

Pankhurst, L.J., Del Ojo Elias, C., Votintseva, A.A., Walker, T.M., Cole, K., Davies, J., Fermont, J.M., Gascoyne-Binzi, D.M., Kohl, T.A., Kong, C., Lemaitre, N., Niemann, S., Paul, J., Rogers, T.R., Roycroft, E., Smith, E.G., Supply, P., Tang, P., Wilcox, M.H., Wordsworth, S., Wyllie, D., Xu, L., Crook, D.W., C.-T.S. Group, 2016. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. The Lancet. Respiratory medicine 4, 49–58.

Perez-Lago, L., Martinez Lirola, M., Herranz, M., Comas, I., Bouza, E., Garcia-de-Viedma, D., 2015. 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. 21, 249 e1.

Perez-Lago, L., Martinez-Lirola, M., Garcia, S., Herranz, M., Mokrousov, I., Comas, I., Martinez-Priego, L., Bouza, E., Garcia-de-Viedma, D., 2016a. 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. 54, 2969–2974.

Perez-Lago, L., Herranz, M., Comas, I., Ruiz-Serrano, M.J., Lopez Roa, P., Bouza, E., Garcia-de-Viedma, D., 2016b. Ultrafast assessment of the presence of a high-risk *Mycobacterium tuberculosis* strain in a population. J. Clin. Microbiol. 54, 779–781.

Perez-Lago, L., Izco, S., Herranz, M., Tudo, G., Carcelen, M., Comas, I., Sierra, O., Gonzalez-Martin, J., Ruiz-Serrano, M.J., Eyene, J., Bouza, E., Garcia de Viedma, D., 2017. A novel strategy based on genomics and specific PCR reveals how a multidrug-resistant *Mycobacterium tuberculosis* strain became prevalent in Equatorial Guinea 15 years after its emergence. Clin. Microbiol. Inf. 23, 92–97.

Pérez-Lago, L., Campos-Herrero, M.I., Cañas, F., Copado, R., Sante, L., Pino, B., Lecuona, M., Díez Gil, O., Martín, C., Muñoz, P., García de Viedma, D., Samper, S., 2019. A *Mycobacterium tuberculosis* Beijing strain persists at high rates and extends its geographic boundaries 20 years after importation. Sci. Rep (in press).

Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A.M., Droz, S., Butz, C., Borrell, S., Langle, C., Feldmann, J., Furrer, H., Mordasini, C., Helbling, P., Rieder, H.L., Egger, M., Gagneux, S., Fenner, L., 2015. Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J. Infect. Dis. 211, 1306–1316.

Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., Battegay, M., Droz, S., Bruderer, T., Coscolla, M., Borrell, S., Zurcher, K., Janssens, J.P., Calmy, A., Mazza Stalder, J., Jaton, K., Rieder, H.L., Pfyffer, G.E., Siegrist, H.H., Hoffmann, M., Fehr, J., Dolina, M., Frei, R., Schrenzel, J., Bottger, E.C., Gagneux, S., Fenner, L., 2016. Standard genotyping overestimates transmission of *Mycobacterium tuberculosis* among immigrants in a low-incidence country. J. Clin. Microbiol. 54, 1862–1870.

Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., Locht, C., 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J. Clin. Microbiol. 39, 3563–3571.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

Tagliani, E., Hassan, M.O., Waberi, Y., De Filippo, M.R., Falzon, D., Dean, A., Zignol, M., Supply, P., Abdoulkader, M.A., Hassangue, H., Cirillo, D.M., 2017. Culture and Next-generation sequencing-based drug susceptibility testing unveil high levels of drug-resistant-TB in Djibouti: results from the first national survey. Sci. Rep. 7, 17672.

van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., et al., 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31, 406–409.

Votintseva, A.A., Pankhurst, L.J., Anson, L.W., Morgan, M.R., Gascoyne-Binzi, D., Walker, T.M., Quan, T.P., Wyllie, D.H., Del Ojo Elias, C., Wilcox, M., Walker, A.S., Peto, T.E., Crook, D.W., 2015. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. J. Clin. Microbiol. 53, 1137–1143.

Votintseva, A.A., Bradley, P., Pankhurst, L., Del Ojo Elias, C., Loose, M., Nilgiriwala, K., Chatterjee, A., Smith, E.G., Sanderson, N., Walker, T.M., Morgan, M.R., Wyllie, D.H., Walker, A.S., Peto, T.E.A., Crook, D.W., Iqbal, Z., 2017. Same-day diagnostic and surveillance data for Tuberculosis via whole-genome sequencing of direct respiratory samples. J. Clin. Microbiol. 55, 1285–1298.

Walker, T.M., Ip, C.L., Harrell, R.H., Evans, J.T., Kapatai, G., Dedicoat, M.J., Eyre, D.W., Wilson, D.J., Hawkey, P.M., Crook, D.W., Parkhill, J., Harris, D., Walker, A.S., Bowden, R., Monk, P., Smith, E.G., Peto, T.E., 2013a. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect. Dis. 13, 137–146.

Walker, T.M., Monk, P., Smith, E.G., Peto, T.E., 2013b. Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. Clinical microbiology and infection : the official publication of the European. Soc. Clin. Microbiol. Infect. Dis. 19, 796–802.

Walker, T.M., Lalor, M.K., Broda, A., Ortega, L.S., Morgan, M., Parker, L., Churchill, S., Bennett, K., Golubchik, T., Giess, A.P., Del Ojo Elias, C., Jeffery, K.J., Bowler, I., Laurenson, I.F., Barrett, A., Drobniewski, F., McCarthy, N.D., Anderson, L.F., Abubakar, I., Thomas, H.L., Monk, P., Smith, E.G., Walker, A.S., Crook, D.W., Peto, T.E.A., Conlon, C.P., 2014. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir. Med. 2, 285–292.

Walker, T.M., Merker, M., Knoblauch, A.M., Helbling, P., Schoch, O.D., van der Werf, M.J., Kranzer, K., Fiebig, L., Kroger, S., Haas, W., Hoffmann, H., Indra, A., Egli, A., Cirillo, D.M., Robert, J., Rogers, T.R., Groenheit, R., Mengshoel, A.T., Mathys, V., Haanpera, M., Soolingen, D.V., Niemann, S., Bottger, E.C., Keller, P.M., Consortium, M.-T.C., 2018. A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. Lancet Infect. Dis. 18, 431–440.