Chapter 4. Genomic epidemiology of tuberculosis

Iñaki Comas

Institute of Biomedicine of Valencia (IBV-CSIC) and CIBER in Epidemiology and Public Health
Jaime Roig 11, 46010, Valencia, Spain
icomas@ibv.csic.es
Phone number: 0034 963393773

Abstract

The application of next generation sequencing technologies has opened the door to a whole new molecular epidemiology of tuberculosis in which we can now look at transmission at a resolution not possible before. At the same time, new technical and analytical challenges have appeared, and we are still exploring the wider potential of this new technology. Whole genome sequencing in tuberculosis still requires bacterial cultures. Thus, although whole genome sequencing has revolutionized the interpretation of transmission patterns, it is not yet ready to make a difference at the point-of-care. In this chapter, we will see which are the promises and challenges ahead of genomic epidemiology as well as the new open questions that the application of this new technology has already been generating. In addition, we will examine the role of molecular epidemiology within the general picture of global tuberculosis control and how genomic epidemiology can change this role in a world on its way towards eradication.

4.1 Introduction

The tuberculosis (TB) epidemic is not anymore like in 1993 when it was declared a "global health emergency" by the World Health Organization. At that time, TB control was insufficient almost in any country of the world, and large outbreaks occurred even in high-income countries (Comas and Gagneux 2009). Since 2015, eradication of TB is back in the public health agenda but at different speeds across the globe (Dye et al. 2013). Some countries like the United States are close to eradication among US-born patients, while TB incidence remains significantly higher among recent immigrants. On the contrary in six of the 22 high burden countries such as Mozambique or Pakistan TB is on the rise and the growing incidence of new TB cases already multidrug-resistant (MDR) is hampering TB control programs. Human immunodeficiency virus (HIV) is not anymore the only co-morbidity focusing attention of the public health community. Diabetes, alcohol abuse or smoking are also drivers of the disease in developing countries. Globally, TB incidence has only started to decline but at a pace insufficient to eradicate the disease in the next century. The DOTS program set up by WHO in 1996 barely can cope with this changing epidemic, and now that eradication is back on the agenda, there are concerns that we are reaching the limits of DOTS (Dye et al. 2013). Efforts like DOTS-plus has been deployed to deal with MDR-TB forms of the disease in resource limited countries but it dependence on culture facilities and in prompt diagnosis of the cases limits the impact on MDR-TB control (Sterling et al. 2003). Even though DOTS has saved six million lives, we will need to adapt global TB control to the specific epidemics in each country or geographic region. The "one-size-fits-all" approach cannot be applied anymore to a situation, in which many countries

are close to reaching eradication while in others the epidemic has still ongoing (WHO 2015).

The good news is that in the last five years, we have witnessed the development of new tools against TB that are slowly being integrated into the control programs. Rapid molecular tests, in particular Xpert MTB/RIF, have dramatically improved and accelerated the diagnosis of TB and is being used for drug resistance surveillance (Pai and Schito 2015). However, despite of these major achievements, many problems around the global rollout of these tests remain. For example, these tests are mainly used at referral centers and are yet to be further developed into point-of-care diagnostics. Moreover, increased detection of TB cases does not always lead to an increase of cases starting or completing treatment. Two new drugs, bedaquiline and delamanid, have recently been licensed for the treatment of TB for the first time in 40 years, but so far, they have primarily been used on a compassionate basis in multidrug resistance cases (Zumla et al. 2014). While still evaluating their future role on global TB control, we are still far from being able to substitute the current four drug regime lasting six months by a new and shorter regimen. Moreover, we have already witnessed the emergence of resistance to these new drugs (Bloemberg et al. 2015) and even cross-resistance to some of them (Andries et al. 2014). In 2013, the first novel vaccine candidate for prevention of TB infection has entered clinical trials (Tameris et al. 2013). The vaccine failed to enhanced protection compared to the current, almost century old BCG vaccine. However, the work showed that it is possible to carry out controlled TB vaccine clinical trials at a large scale. Many additional new vaccine candidates are in the pipeline, but none of them are expected to be licensed for clinical use anytime soon.

An additional dimension of global tuberculosis control is how to tackle on-going transmission. Part of the problem is that DOTS failed to deliver one of its primary aims. By treating patients early and providing full follow-up, DOTS was expected to limit both the emergence of drug resistance and transmission. But even under the umbrella of the most advanced public health systems in low-endemic countries, direct transmission of TB still accounts for up to 20% of cases, while in high-burden countries, it is the main contributor to disease incidence (up to 75% of cases) (Yates et al. 2016). As a result, even the success of DOTS limiting the number of new drug resistance cases is in danger as lack of control of transmission will leave room to MDR-TB to spread. Therefore, new epidemiological tools to inform new intervention strategies including advances towards real-time molecular epidemiology are needed to limit TB transmission. However real-time molecular epidemiology in tuberculosis is a challenge. This is because the nature of an outbreak in TB is very special. One could even argue whether outbreaks in TB exists at all. First, most individuals infected do not develop TB disease. For the remaining, it can take months, years or even decades for active TB to develop. Contrary to most acute infections, outbreaks of TB may span years before being noticed. For those cases, in which progression to active disease occurs within months there is room for infection control, but because the TB bacteria take between two to four weeks to grow, by the time molecular epidemiological data is generated, it is already too late for an intervention. This is the reason why molecular epidemiology has been usually confined to retrospective studies rather than represent a tool for infection control in real-time. Yet, we badly need such a tool to cut transmission.

Whole genome sequencing (WGS) shows great potential for becoming the ideal tool to tackle TB transmission. WGS has been shown to resolve with greater resolution transmission clusters and overcome the limitations of conventional typing techniques by avoiding false assignments to transmission clusters (Walker et al. 2013). It is still not an ideal marker due to the slow pace of mutation accumulation, but when combined with classical epidemiological analysis, it becomes a powerful tool. Furthermore, due to the

comprehensive nature of genomic data, it has already given important insights into different aspects of the biology of TB bacteria, including the emergence of drug resistance (Comas et al. 2012; Cohen et al. 2015), within-host bacterial variation (Eldholm et al. 2014), and responses to immune pressure (Comas et al. 2010). However, until now, the use of WGS has largely been confined to research environments, and in most cases, it has been applied retrospectively. In this chapter, I will review different aspects on the application of WGS to better understand the epidemiology of TB. In particular, I will review the use of WGS to resolve large TB outbreaks, define transmission clusters, monitor new interventions, and differentiate relapse from re-infection. Moreover, I will review recent insights gained from WGS into the biological factors behind the diversity within patients and between transmission cases, and how this diversity may impact epidemiological inference. I shall end by discussing the future role of WGS in the global control of TB.

4.2 Next-generation DNA sequencing applied to the tubercle bacilli

In 1905, a strain of *Mycobacterium tuberculosis* known as H37 was isolated from an outbreak in the United States. That strain is today known as H37Rv and is the reference strain used worldwide for experimental work on *M.tuberculosis*. In 1998, H37Rv was among the first bacterial genome sequenced, opening the era of TB genomics research (Cole et al. 1998). The genome of H37Rv and its experimental manipulation has represented a quantum leap on the development of new antibiotics, vaccines and diagnostics. It has allowed to understand the genetic determinants of clinical resistance to antibiotics, define regions appropriate for molecular typing or identifying virulence factors that now are the origin of many of the vaccine candidates under development.

In 1998, sequencing a bacterial genome required two years of work and cost around 4-5 million US dollars. Today, we could sequence *M. tuberculosis* H37Rv hundreds of times in less than a week and at price hundreds of thousands of times cheaper than in 1998. It is therefore not surprising that the whole genome sequence of bacterial pathogens is becoming the new gold standard used as an epidemiological marker (Loman and Pallen 2015). The first next-generation sequencing technology (also known as deep sequencing, massive sequencing and next generation sequencing) was introduced in 2005. Contrary to the previous techniques used for whole genome sequencing, known as shotgun sequencing, next generation sequencing technologies do not require a cloning step in *Escherichia coli*. The other main feature of these new technologies was the high throughput in terms of sequencing yield in a single run. For example, Illumina platforms are the most widely used for bacterial genome sequencing today. The mean genome size of *M. tuberculosis* is around 4.4 million base pairs (bp). If one were to run a single strain of *M. tuberculosis* on an Illumina MiSeq platform with v3 kits, one would read each nucleotide of the genome 3,400 times, i.e. the coverage of the strain would be 3,400x. With the NextSeq "rapid run" platform, the coverage would be around 27,200x, and on an Illumina HiSeq platform at full capacity 136,300x. On an Ion Torrent platform, one can read each base up to 455 times. Reading the same position of the genome hundreds of times is typically not necessary as for many studies we are interested in a coverage around 80x. Thus, all of these platforms generate much more data than what we really need to analyze a single genome. To take advantage of the high throughput of these new sequencing machines, several strains are usually multiplexed in a single run. The amount of multiplexing depends on the targeted coverage, but in theory one can sequence at 80x up to 1,700 strain in a single run of HiSeq 1500/2500. To allow multiplexing for each strain, the genomic DNA is sheared into fragments of 500-1000 bp and specific nucleotide tags are added that can be retrieved during the bioinformatics analyses and used to assign each sequenced read to its corresponding strain. By multiplexing, one not only maximizes the number samples analyzed in one run, but one also reduces the time and the cost per

sample. Multiplexing allows sequencing one strain for 80-100 euros if it is done in-house with recommended reagents. If customized reagents are used, the price can drop to below 50 euros, which is not more expensive than other methods for the molecular characterization of pathogens used in clinical microbiology units of hospitals in high-income countries.

The other main feature of next generation sequencing technologies that has to be taken into account is the length of the sequencing reads. Depending on the organism, one usually needs to choose a platform based on a trade-off between read length and overall throughput. In tuberculosis epidemiology, short read length technologies are mostly used that generate reads between 100-300 bp. There are several reasons for that. The throughput is much higher, and as we have seen above, this is essential for high multiplexing and lower price. In addition, for *M. tuberculosis,* the bioinformatic analysis does usually not reconstruct the strain genome (approach known as "de novo assembly") but maps the sequencing reads to a reference genome. This is because it is known that in terms of gene content and genome structure, all strains are very similar. Indeed, the average nucleotide identity is above 99% and thus many stretches of the genome are almost identical across strains. This reference mapping is known as re-sequencing and it is very useful to identify single nucleotide polymorphism (SNPs) among strains. The drawback is that short reads are difficult to map to repetitive regions and those regions must therefore be excluded from the analysis.

In more genetically diverse bacteria, the strategy of mapping to a reference can be misleading unless there is *a priori* knowledge that they have a very recent common ancestor (for example if they belong to the same outbreak than the reference genome). In the case of *E.coli,* up to 60% of the gene content may not be shared across strains, and a mapping to reference will therefore likely miss important information on the origin and other genomic characteristics of the strain of interest (Gordienko et al. 2013). In these cases, it is better to build a new genome from the sequencing reads based on those reads that are overlapping. In general, technologies that produce longer reads are more likely to generate a better assembly of the genome. These so-called third generation technologies like PacBio (Pacific Biosciences) not only generate reads between 10-15 kb, but also work with the DNA molecule directly, thus avoiding the PCR steps necessary when using the Illumina or Ion Torrent platforms. This is why PacBio and other technologies like Oxford Nanopore are considered single molecule technologies (Loman and Pallen 2015). In addition, single molecule sequencing can identify simultaneously the methylation pattern of the DNA molecule. The drawback is that because the throughput of third-generation technologies based on long reads is comparably low, one cannot multiplex many strains and thus the price per strain becomes too high for routine public health and diagnostics purposes. One common strategy is to combine both technologies so short-read sequenced strains can be mapped to accurate assemblies of representative strains generated with long-read technologies. In summary, short-read, highly multiplexed technologies are currently preferred for molecular epidemiological studies or diagnostics of bacterial diseases including tuberculosis. In the future, technologies that offer simultaneously high multiplexing, long reads and cheap sequencing will likely replace the current combination of different technologies.

4.3 The genome as an epidemiological marker

In tuberculosis, the first evidence that the whole genome sequencing can lead to a higher epidemiological resolution was the analysis of two strains from Uzbekistan (Niemann et al. 2009). Those strains had almost exactly the same MIRU profile and RFLP pattern, the two most commonly used epidemiological markers (reviewed in Chapter 1). The only known difference between these two strains was the drug sensitivity profile, as one was

pan-susceptible and the other multidrug-resistant. Whole genome sequence data revealed a more complex picture with dozens of SNP differences between the strains. For the first time, it was shown that the genome can provide a greater resolution than other molecular markers. Moreover, this study provided first evidence that the *M. tuberculosis* complex is more diverse than previously anticipated, even at an epidemiological scale. This analysis was performed using one of the first next-generation sequencing platforms, the Solexa Genome Analyzer (known as Illumina today). Shortly thereafter, another technology, Roche 454, was used to analyze genomes from the first and last case of a suspected transmission cluster (Schurch et al. 2010). Again the genome showed a higher resolution compared to conventional methods.

In 2011, the first large-scale genome analysis of a tuberculosis outbreak was published, starting the era of genomic epidemiology for tuberculosis (Gardy et al. 2011). In this landmark publication, Gardy *et al.* analysed 36 cases belonging to an outbreak previously identified by MIRU and/or RFLP typing, and spanning several years in British Columbia. WGS was able to trace a clear picture about how the outbreak started. In particular, this analysis pointed to the importance of a superspreader in the outbreak. Superspeaders are patients that generate a disproportionally high number of secondary cases. Superspreaders rather than chains of transmission seem to be a common transmission topology in tuberculosis. Following this initial publication, other large outbreaks have been described, all of them have in common the complexity of the transmission network. In Hamburg, a large outbreak identified by MIRU-VNTR included 86 cases between 1996 and 2011 (Roetzer et al. 2013). However, not all cases were linked by epidemiological investigations and before whole genome sequencing, the true source of the outbreak remained unknown. WGS analyses closed the gap between molecular and epidemiological data. Two different outbreaks with two closely related but distinct strains were involved. One was linked to many of the cases of the first years and the other was still on-going in 2010. This explained the lack of epidemiological links between the patients involved in this outbreak, and why the index case was so difficult to trace. Thus, with the high resolution obtained from whole genome sequences now we can apply different evolutionary tools, including phylodynamics and phylogeography, to dissect the tempo and mode of transmission and drug resistance acquisition of successful tuberculosis clones (Eldholm et al. 2015).

Genomic analysis applied to specific strains of interest has also allowed to identify the true extend of an outbreak. This information can be of vital importance for public health authorities. Recent examples of that include TB outbreaks in Bern, Switzerland (Stucki et al. 2015) and in Almeria, Spain (Perez-Lago et al. 2015). In the Bernese outbreak, Stucki *et al.* (Stucki et al. 2015) applied WGS to three strains of an outbreak that started in Bern among homeless people, a typical high risk population for TB in high-income countries. These strains were thought to be representative of the diversity within the outbreak as they were chosen from different periods during the outbreak. By comparing to control strains, Stucki *et al.* (Stucki et al. 2015) were able to identify single nucleotide polymorphism (SNPs) that could be used to assign patients to the outbreak. Real-time PCR analyses designed based on these SNPs was then developed and applied to the retrospective collection of 1,642 TB cases in the canton of Bern between 1991 and 2011. The analyses allowed assigning 68 new cases to the Bernese outbreak. The RFLP pattern between all these strains was almost identical. In contrast, WGS comparison revealed the true complexity of the outbreak. Transmission network reconstruction based on genetic data detected three central nodes in the topology of transmission that combined with epidemiological data allowed to detect two index cases that had infected many others, i.e. two superspreaders. The Bernese outbreak is an example of how WGS data can illuminate epidemiological investigations in TB. A similar approach was used by Pérez-Lago *et al.* (Perez-Lago et al. 2015) to identify retrospectively and prospectively new cases due to an

M. tuberculosis strain that had already led to many secondary cases in Almeria, Spain. The authors developed a SNP-typing approach similar to Stucki *et al.* (Stucki et al. 2015) but based on a low-cost, low-tech, decentralized protocol with the aim to use it at the point-of-care or the closest referral center. The typing assay is called TRAP and can be run on a gel to quickly scan a large number of strain both from cultures and sputum samples.

These different works are examples of how genomic information can be used to identify transmission patterns, and how a technology that is not universally accessible can help to design other rapid and low-cost molecular assays. In summary, the application of WGS to specific outbreaks have shown its superiority when compared with conventional typing tools and have led to important new epidemiological insights (Walker et al. 2013).

4.4 Population scale analysis of TB transmission using WGS

Public Health England has led the way to implement WGS as both an epidemiological marker and as a diagnostic for public health systems. In a series of publications from 2012 to 2015, they have demonstrated the potential impact of WGS on the control of TB. In a landmark paper, Walker and collaborators (Walker et al. 2012) published the first large-scale, population-based study of TB transmission based on WGS. In that study, the superiority WGS over previous strain typing methods was corroborated by identifying more accurately those cases belonging to an outbreak. At the same time, the authors sequenced serial isolates from the same patient and isolates from different body parts of the same patient. Based on these data, the authors proposed a threshold to identify a transmission link between two TB cases. Specifically, a genetic distance separating patient isolates of five or fewer SNPs was used to define high-confident transmission clusters, as most of the time, these clusters were also supported by epidemiological links. A genetic distance of between 5 and 12 SNPs was considered a cluster in which recent transmission was very likely but often not supported by epidemiological data. A genetic distance of more than 12 SNPs was defined to classify epidemiologically unrelated cases. Using those thresholds, a study in Switzerland has shown that standard genotyping (MIRU) overestimates the rate of transmission among immigrant (Stucki et al. 2016). This is likely due to the high genetic similarity of strains circulating in high-burden countries. In addition, it is important to remember that identification of transmission clusters based on WGS data does not necessarily imply that transmission has occurred at the place of study. For example, transmission between immigrants may have occurred at their country of origin, followed by progression to active disease at the country of residence. This is best exemplified by the detection of transcontinental spread of *M. tuberculosis* clones like in the published case of Thai refugees in California (Coscolla et al. 2015).

The SNP thresholds described above have been corroborated by other studies, but most examples are from low-burden countries (Bryant et al. 2013b; Hatherell et al. 2016). The complexity of the global TB epidemic suggests that the same threshold may not apply to all epidemiological scenarios. For example, in high-burden countries where there is a high rate of on-going transmission, several genetically similar clones may act as index cases of different transmission clusters, and thus differentiating between those clusters maybe difficult (Yates et al. 2016). The only example from a high-burden country we have until now is the analysis of 1,687 TB patient isolates collected in rural Malawi between 1995 and 2010 (Guerra-Assunção et al. 2015a). Genetic distances and genetic network analyses showed consistency with the thresholds described by Walker *et al.* (Walker et al. 2012). However, this is just one case, and it is likely that the situation may change in large urbanized African settings. Similarly, areas with a high burden of multidrug (MDR) resistant TB might also show different patterns. Due to the selection of drug resistance mutations, we can expect a higher number of mutations between epidemiologically linked strains as a result of genetic hitchhiking effects (further discussed in the following section)

(Sun et al. 2012; Eldholm et al. 2014; Liu et al. 2015). For example, by looking at serial isolates of a single patient _Eldholm et al. (2015) identifies more SNPs separating the isolates than the number expected between two transmission cases. To date, only one large population-based study has been published from a high- MDR-TB burden setting in Russia (Casali et al. 2014). The study included more than one thousand patient isolates, 50% of which were MDR. Most of these MDR isolates carried the mutation S450L in the *rpoB* gene. This mutation is generally the most common mutation conferring resistance to rifampicin. In addition, large transmission clusters also carried additional fitness compensatory mutations, mainly in the *rpoC* gene (Comas 2012), explaining the high transmissibility of these strains in the region. Unfortunately, because the associated epidemiological data was not published, it is difficult to evaluate SNP thresholds levels to delineate transmission clusters in this Russian dataset.

The epidemiological significance of SNP thresholds remains a matter of ongoing research. On one hand, we lack data from different epidemiological settings. On the other hand, calling SNPs from next generation sequencing data is not straightforward (O'Rawe et al. 2013). There are multiple steps in the bioinformatics analyses that can introduce false positive or false negative SNPs. It is thus very important to implement and follow multiple quality control checkpoints. In particular, mobile elements and repetitive regions of the *M. tuberculosis* genome are difficult to interrogate with short read length technologies such as Illumina. In a typical analysis, most of these loci are excluded. In addition, other loci can also be problematic. For example, insertion elements scars that led to incorrect mapping of reads and unknown deletions and/or insertions can complicate the analysis. Importantly, the parameters used for mapping and SNP calling are critical. Finally, initial quality of the sequencing data is key. Given that epidemiological inferences are based in small number of SNPs, it is recommended to corroborate results independently with a different analysis pipeline and/or by laboratory confirmation of a subset of SNPs. In addition, Illumina technology is currently the main platform used in genomic epidemiology of TB, but we do not yet know the potential epidemiological impact of using long-read sequencing platforms (Quail et al. 2012). These platforms, although expensive, should allow identifying SNPs in regions of the genome that are not accessible by short-read sequencing technologies. While some of these regions, like the PE/PPE genes, are clearly the most variable of the *M. tuberculosis* genome (Copin et al. 2014), they are also likely involved in gene conversion events and/or recombination with external sources (Phelan et al. 2016). Thus, even if interrogated with the appropriate technology, we first will need to understand if and how these loci can be exploited for epidemiological purposes.

4.5 Role of within-host diversity in transmission inference

As in any other pathogen, the bacteria belonging to the *M. tuberculosis* complex are in constant evolution, and it is therefore not surprising that the bacterial population infecting a patient is not genetically homogeneous (Pérez-Lago et al. 2013). Understanding this within-host diversity is crucial, as it can have important consequences for drug resistance diagnostics, epidemiology and disease outcomes (Didelot et al. 2016). One important open question is what is the magnitude of within-host diversity during TB infection? This question has been mainly studied in the context of drug resistance because of its clinical importance. Historically, clinical microbiologists have recognized the phenomenon of heteroresistance, which manifests in discordant results of drug susceptible testing using repeat testing or different isolates from the same patient (Van Rie et al. 2005). Heteroresistance suggests that there are two co-existing populations in the patient, one drug-resistant and the other susceptible to a particular drug. WGS has the potential to identify the genotypes of these co-existing populations. Moreover, we can follow how the relative frequency of these genotypes changes over time, based on serial samples from the

same patient (Sun et al. 2012, Eldholm et al. 2015). The precision of this technique is such that we can detect drug resistance minority variants in an otherwise homogenous population at the 10% or even 5% level when the sequencing coverage is high enough. The drawback is that many of these minority variants associated with drug resistance are identified based on cultured specimens, and thus we cannot easily trace back their origin in the lung. However, PET-CT scans have been able to correlate different lesions in the lung with different *M. tuberculosis* genotypes obtained from cultured isolates, suggesting that micro-geography of the lungs and the lung lesions are important for the selection of genetically distinct sub-populations (Liu et al. 2015).

The co-existence of several bacterial clones within an individual patient has been demonstrated beyond the case of heteroresistance (Pérez-Lago et al. 2013). But what is the origin of such genetic variation? Recent efforts using whole genome data have explored this variation both *in vivo* and *in vitro* (see Chapter 13). The mean whole genome mutation rate as derived from experiments with cynomolgus macaques is around 0.39 (0.16-0.80 95% CI) SNPs per genome per year (Ford et al. 2011). This rate was found to be similar between macaques that developed active disease and those that remained latent although the number of SNPs analyzed were too low to allow for strong statistical conclusions. In fact, given that the generation time of the bacteria is thought to be longer during latency, the authors concluded that the mutation rate was higher during latency than during active disease (Ford et al. 2011). The same research group was later able to define the *in vitro* rate of mutation of different strains belonging to different lineages in the presence and absence of different antibiotics (Ford et al. 2013). This *in vitro* rate was fairly similar to the rate seen *in vivo* in macaques, and slightly higher for Lineage 2 compared to Lineage 4.

However, the mutation rate of a bacteria measured in short-term experiments is not necessarily the same than the substitution rate we observe in a clinical setting. This is because in addition to the rate of mutation generation, we have to consider the action of evolutionary forces. Most of the mutations arising *de novo* are either neutral or deleterious. Neutral mutations can increase in frequency unnoticed, whereas deleterious mutations will be removed, more or less efficiently, by natural purifying selection. Thus, the action of evolutionary forces uncouples the intrinsic bacterial mutation rate from the substitution rate. Importantly, the substitution rate is time-dependent because natural selection needs time to act. Hence, the shorter natural selection can act, the closer the substitution rate will be to the intrinsic mutation rate (Rocha et al. 2006; Biek et al. 2015). This explains why the long-term substitution rate of many pathogens is much lower than the mutation rate or the rate measured over short periods of time. At the epidemiological level, the substitution rate can be derived by comparing the number of differences between two transmission cases and the difference in time of diagnosis between both cases. This approach has been used to estimate the substitution rate of *M. tuberculosis* in clinical settings. Findings from several studies converged in a substitution rate of 0.3-0.5 SNPs per genome per year, thus very close to the *in vivo* and *in vitro* mutation rate discussed above. However, the variance around that estimate is so high that it has to be interpreted cautiously (Bryant et al. 2013b). One of the main "known unknowns" likely to impact our inferences of mutation- and substitution rates is latency. To date, only one study has used time of infection rather than time of diagnosis to measure the substitution rate in *M. tuberculosis*. The study showed that the substitution rate during latency was much lower than during active TB (Colangeli et al. 2014). Much more data is needed to draw conclusions about the substitution rate throughout the life cycle of *M. tuberculosis*. In the meantime, we are limited to model the substitution rate based on estimates during infection and transmission.

How is the mutation rate modulated by different evolutionary forces within the host? Again, this is best exemplified in the context of antibiotic treatment. As antibiotics are a strong selective force, the mutations causing drug resistance are positively selected. Because *M. tuberculosis* is clonal, other mutations present in a particular genetic background experiencing such positive selection are also selected, even when they have nothing to do with antibiotic resistance. This phenomenon is called genetic hitchhiking and have been elegantly described in several publications (Sun et al. 2012; Eldholm et al. 2014). In this context, the drug resistance conferring mutation can also be referred to as "driver" mutations, and the hitchhiking mutations as "passenger" mutations. However, it is also important to understand the bacterial variation expected in a patient that does not develop drug resistance. In these cases, purifying selection is mainly acting in the form of antibiotic purification (Black et al. 2015). However, variation maybe present because of neutral processes (e.g. genetic drift) or because of selection due to an unidentified evolutionary force. Thus, as a consequence of the equilibrium between different evolutionary forces, the amount of bacterial variation seen in different patients may vary substantially. For example, a recent study has shown that bacterial variation can be observed in most consecutive sputum samples obtained from a given patient (Pérez-Lago et al. 2013). By contrast, other studies have reported no genetic differences between bacterial samples from the same patient, even when the cultures were separated by more than one year (Pérez-Lago et al. 2015). The causes of these discrepancies are not well understood, but they might be linked to differences in treatment efficacy, variation in the site of infection, or the differential control of the bacterial load by the immune system. However, what is clear is that within-host bacterial diversity exists both inside and outside the context of drug resistanceFrom a practical standpoint, the potential overlap between within-patient diversity and the diversity seen between transmission cases will complicate the epidemiological interpretation (Pérez-Lago et al. 2013). From a biological standpoint it is important to identify the small fraction of SNPs detected that are linked to natural selection pressures apart from antibiotics.

4.6 Special cases of within-host diversity: relapse, re-infection and co-infection

One major outcome measured by clinical trials of new drugs and drug regimens the number of relapses occurring after successful completion of the treatment (Johnston et al. 2015). In high-burden countries, it is difficult to differentiate between a true case of relapse, indicating failure of the drug/regimen, or reinfection after treatment completion. Thus, molecular tools that can easily identify these two situations are key to evaluate new interventions. Several recent examples have highlighted the potential of WGS data to distinguish between relapses and secondary infections (Bryant et al. 2013a; Guerra-Assunção et al. 2015b). However, these studies also highlight the complexity associated to the interpretation of the data

As discussed before, a difference of 10-12 SNPs is often used as the cutoff to define a transmission link. This same cutoff can be used to determine if two TB episodes in the same patient are due to the same strain or not. Applying this logic, Bryant *et al.* (Bryant et al. 2013a) identified the number of true relapses *versus* the number of re-infections in the context of a clinical trial of a new treatment regimen (ReMOX-TB) (Gillespie et al. 2014). Forty-seven patients had a second episode of TB after completing treatment. Using the SNP-threshold approach, the authors found that 33 of the 47 patients had less than 12 SNPs between the first and the second isolate, indicating cases of true relapse. Only three cases harboured very different strains, indicated by a pairwise difference ofmore than 1,000 SNPs; accordingly, these cases were classified as re-infections. In addition, the authors also reported a mix of two strains during the second episode of one patient. One of the isolate was almost identical to the isolate of the first episode while the other was clearly different. This single patient therefore likely represents case of relapse due to the

presence of the original strain, plus a re-infection as a new strain was detected in the second episode. Similar results were obtained after analysis of relapse cases in a long-term epidemiological study carried out in Malawi (Guerra-Assunção et al. 2015b). However, there are limitations detecting such mixed infections. Identifying a re-infection when the second strain belongs to a different lineage is straightforward, as the strains will be separated by hundreds or thousands of SNPs. However, in high-burden countries, where many of the strains circulating are closely related, classifying strains as "the same" or "different" is much more challenging (Guerra-Assunção et al. 2015b).

WGS data also gives us the opportunity to detect co-infection in a single sample, i.e. cases in which two genetically distinct *M. tuberculosis* strains (more than 12 SNPs) co-exist in the same patient. These cases where difficult to identify until now as one had to rely on e.g. identifying evidence of mixed number of alleles in MIRU loci. However, this approach tends to miss many instances of co-infection, as single locus data has low resolution. By contrast, because WGS achieves high coverage, the presence of two different isolates in the same culture becomes evident when looking at SNP positions where the reference allele and the variant allele coexist as low frequency variants (i.e. none of these alleles are fixed in the population). SNP positions showing this mix of reference and variant allele are known as "heterozygous calls", and are indicative of two different sub-populations present in the culture. Heterozygous calls maybe related to clonal diversification from a single infecting strain within the lung of a patient. In this case, the number of heterozygous calls is expected to be low given the low mutation rate of *M. tuberculosis*. On the other hand, if two different strains infected the same patient, the number of heterozygous calls will be higher. Thus, as in the case of relapse and re-infection, the number of heterozygous calls detected can be used to differentiate between these two scenarios. To be compatible with clonal diversification, the number of heterozygous calls must be below the 12 SNP threshold used to identify transmission clusters. If more than 12 SNPs are seen, one can safely assume that two different strains are co-infecting the same patient. However, although in theory differentiating between clonal diversification and co-infection is straight forward, interpreting heterozygous calls is very challenging (Hatherell et al. 2016). Due to errors in mapping and SNP calling, much noise can be introduced during the bioinformatics analyses. As a result, it is often difficult to differentiate between true and false heterozygous calls. This is particularly relevant when the sequencing is done directly from the diagnostic sample, in which DNA from unrelated organisms (i.e. contaminants and/or commensals) might be present. This contaminating DNA maybe sequenced together with the DNA of interest, which can contribute to a percentage of the heterozygous calls. In practice, bioinformatics analyses can easily detect a co-infection when the number of heterozygous calls is larger than 100, and clonal diversification when the number is below 12; however, anything in between, is difficult to interpret (Guerra-Assunção et al. 2015b). In summary, high-throughput WGS is an ideal tool to identify co-existing variants, but more work is needed to develop analytical tools than can reveal co-infection when the two infecting strains are evolutionary close (between 20-100 SNPs). This is especially relevant in high-burden countries where transmission is often due to highly similar strains (Kay et al. 2015).

4.7 Reconstructing transmission

The high variance in terms of genetic changes accumulated over time seen during both between host transmission and within a single patient leads to a very weak correlation between time and accumulated sequence diversity in *M. tuberculosis* (Bryant et al. 2013b). This weak correlation between time and accumulated number of SNPs is known as overdisperssion of the molecular clock. A practical consequence is that the low correlation limits our ability determine the exact time of infection using genetic data (Hatherell et al. 2016). Only when large transmission clusters spanning several years or even decades are

analyzed, researchers have been able to correlate the dating based on genomic substitution data with the epidemiological records (Roetzer et al. 2013). Thus, new analytical approaches are needed to model the genetic diversity seen during infection and transmission, and can use the information to build accurate genealogies on how and when transmission has occurred. Several such approaches have recently been developed, although not all of them have been tested in large-scale, population-based studies. Most approaches, like Outbreaker (Jombart et al. 2014), use the genetic data to infer the most likely network of transmission between isolates of a cluster, minimizing the number of genetic changes between isolates. However, as usually happens for this kind of approximations, clusters must be defined *a priori,* meaning that the researcher needs to select those strains suspected to be involved in a transmission group. As a consequence, transmission can only be defined among suspected cases, limiting the *de novo* discovery of cases linked to transmission in population-based samples. In addition, the "carriage" state represented by the latency period is usually not explicitly modeled. Hence, these approximations are better suited for acute diseases where infection, disease and transmission are directly linked (Jombart et al. 2014). Didelot *et al.* (Didelot et al. 2013) have developed a new way of interpreting transmission in bacterial infections including TB. Their approximation is based on using a dated phylogenetic topology, i.e. one that incorporates epidemiological data like date of diagnosis, and convert it into transmission events in a Bayesian framework. Furthermore, other epidemiological data like geography can be incorporated to delineate the most likely epidemiological scenario. The approach takes into account the amount of genetic diversity that may have accumulated within the patient during latency. It may be the case that within-host diversity can be neglected in most of the cases, particularly if the first isolate of each patient is analyzed. However overall, there remains a general lack of data to evaluate the true role of within-patient diversity. Certainly in the context of drug resistance, one expects a potentially misleading effect of this diversity on epidemiological inference (Eldholm et al. 2015).

In parallel, tools to understand the epidemic dynamics from WGS data are being applied to *M. tuberculosis* (see Chapter 15 for a review). Based on the principles of Bayesian phylodynamics, those approximations allow to infer important epidemiological parameters like incidence, prevalence, and changes in the basic reproductive number (R0) over time (Grenfell et al. 2004; Stadler et al. 2012). The integration of the phylodynamic and epidemiological frameworks is a research field in constant development, and thus different approaches are available, ranging from classical coalescence to birth-death models (Stadler et al. 2012). Compartmental epidemiological models like SIR (susceptible–infected–removed) (Rasmussen et al. 2011; Kühnert et al. 2014) can be also incorporated as well as geographical data (du Plessis and Stadler 2015). These different approaches are based on different assumptions, and it remains to be seen which one will perform best in the case of TB (Stadler et al. 2012). Again, the effect of latency and how best to model it remains a "black box" that needs to be addressed in the future. Nevertheless, these efforts will play an important role in the future, as we will be able to evaluate the impact of new interventions using genetic data as well as predict epidemic trends based on the evolution of R0 over time.

4.8 Challenges of genomic epidemiology

WGS is becoming the new gold standard for molecular epidemiology of TB. It provides higher resolution than previous molecular markers, which allows for both a better delineation of transmission clusters and the potential to establish the direction of transmission. However, we still have many knowledge gaps to resolve (van Soolingen 2014).

Understanding the biology. To properly model what happens within a patient, we need to understand how bacterial genetic diversity is generated within a patient and how much of it is transmitted. Linked to that, we need to determine the contribution of latency to the diversity seen, and how to account for it in transmission models. Interpretation of transmission network is also not straightforward, and while algorithms have been developed, they are not always easy to implement. Moreover, they still have to be evaluated at a larger scale. This is particularly true if we want to move towards routine WGS in the clinical environment.

Beyond distance thresholds. Since the publication of Walker *et al.* 2012, using SNP threshold has become the gold standard to define transmission clusters. However, these thresholds are likely not universal, and have to be tested in a wider range of epidemiological settings. For example, due to the action of natural selection, it is very likely that in high MDR-TB burden settings, we will find epidemiologically linked cases with more than five or twelve SNPs. Furthermore, distance thresholds are useful but can be misleading. The absence of an intermediate transmission link, an un-sampled index case for example, may separate two strains that are in fact epidemiologically related. As we have seen, algorithms that capture the diversity and complexity of transmission trees are being developed. These methods should be tested at a larger scale to understand how to interpret transmission at the population level.

Towards routine genomic epidemiology. Interpretation of WGS data is not straightforward, and remains mostly confined to the research settings. Efforts have been done in certain places, most notably Public Health England/NHS and the US CDC, but we are still far from democratizing WGS among medical institutions. Implementation of WGS must be also integrated in local, regional, and national health systems, and the results shared across the globe to accelerate diagnostic and epidemiological research (Yozwiak et al. 2015). However, the reality in high-burden countries is very different. This is best illustrated by the low number of genomic epidemiological studies published to date from high-burden regions. In most of these countries, the only routine diagnostic remains sputum microscopy, and there is usually no access to bacterial cultures. Thus, the genomic revolution will likely take time to get to these high-burden, low-income countries given that the benefits remain limited.

Genomic sequencing from diagnostic samples. The real genomic revolution in TB clinical practice will be the direct sequencing from complex sputum samples. The advantages of such an approach are multiple. It will eliminate the need for culture, generate a positive diagnostic result in less than a week or perhaps even hours, and will be used simultaneously for infection control. However, the challenges of obtaining an *M. tuberculosis* genome from a sputum samples are enormous. Some advances have been achieved recently, for example by doing diagnostic metagenomics (Doughty et al. 2014). This approach consists in isolating total DNA from sputum and subject it to direct sequencing. However, for samples with low bacterial load,*M. tuberculosis* enrichment strategies must be applied (Brown et al. 2015). An alternative is targeted sequencing of certain regions of the genome, including known drug resistance loci, an approach also referred to as high-throughput amplicon sequencing (Colman et al. 2015). While still in early days, there is no doubt that new approaches will be developed, some probably linked to real-time, portable genomic technologies like those based on nanopores (Bradley et al. 2015) that have been successfully applied during the recent Ebola epidemic (Quick et al. 2016).

4.9 Practical implications of genomic epidemiology

WGS has the potential to, and to some extent already is, revolutionizing molecular epidemiology of TB. Once more powerful analytical tools are developed, we will be able to establish a high resolution picture of TB transmission in different epidemiological scenarios, which will help developed tailored control strategies. Such a high resolution picture will allow determining the role of different host, bacterial and environmental factors in the transmission of TB. In addition, together with geopositioning data, we will be able to develop new tools for spatial epidemiology of TB. Thus, genomic epidemiology will improve our view not only of the various drivers of TB transmission, but also of the specific foci of transmission. If factors driving TB transmission can be established, particularly those involving human and bacterial variation, new avenues of research will be open that will contribute to a better understanding of the complex interplay between the host and the pathogen. In addition, WGS offers the opportunity to identify the genetic changes and the evolutionary forces behind infection, disease and transmission. Indeed, the study of mutations associated with drug resistance is currently leading the way, but we need to identify the genetic loci of both the host and the pathogen that are under different selective pressures. If we are able to identify such loci, a whole new opportunity for research will be created for the development of new drugs, vaccines and host directed therapies. Finally, if we are able to position WGS analysis close to point-of-care and to the diagnostic sample, we will transition from a tool used in retrospective studies to a tool used for infection control and monitoring treatment efficacy in real-time. While there is still a lot of work to do in terms of standardization and interpretation, genomic epidemiology can have a central role on the new strategies for global TB control and help pave the way towards TB elimination.

Acknowledgements

I thank the members of my group for stimulating discussion. Work in my laboratory is supported by the Spanish National Foundation (MINECO SAF2013-43521-R) and the European Research Council (638553-TB-ACCELERATE).

References

Andries K, Villellas C, Coeck N, et al (2014) Acquired resistance of *Mycobacterium tuberculosis* to bedaquiline. PLoS One 9:e102135.

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X (2015) Measurably evolving pathogens in the genomic era. Trends Ecol Evol 30:306–313.

Black PA, de Vos M, Louw GE, et al (2015) Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. BMC Genomics 16:857.

Bloemberg G V, Keller PM, Stucki D, et al (2015) Acquired resistance to bedaquiline and delamanid int for tuberculosis. N. Engl. J. Med. 373:1986–1988.

Bradley P, Gordon NC, Walker TM, et al (2015) Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and Mycobacterium tuberculosis. Nat Commun 6:10063.

Brown AC, Bryant JM, Einer-Jensen K, et al (2015) Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. J Clin Microbiol 53:2230–7.

Bryant JM, Harris SR, Parkhill J, et al (2013a) Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. lancet Respir Med 1:786–92.

Bryant JM, Schürch AC, van Deutekom H, et al (2013b) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis 13:110.

Casali N, Nikolayevskyy V, Balabanova Y, et al (2014) Evolution and transmission of drug-

resistant tuberculosis in a Russian population. Nat Genet 46:279–286.
Cohen K a., Abeel T, Manson McGuire A, et al (2015) Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. PLoS Med 12:1–22.
Colangeli R, Arcus VL, Cursons RT, et al (2014) Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. PLoS One 9:e91024.
Cole ST, Brosch R, Parkhill J, et al (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–44.
Colman RE, Schupp JM, Hicks ND, et al (2015) Detection of low-level mixed-population drug resistance in *Mycobacterium tuberculosis* using high fidelity amplicon sequencing. PLoS One 10:e0126626.
Comas I, Borrell S, Roetzer A, et al (2012) Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 44:106–110.
Comas I, Chakravartti J, Small PM, et al (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet 42:498–503.
Comas I., Gagneux S (2009) The past and future of tuberculosis research. PLoS Pathog 5:e1000600.
Copin R, Coscolla M, Seiffert SN, et al (2014) Sequence diversity in the pe_pgrs genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. MBio 5:e00960–13.
Coscolla M, Barry PM, Oeltmann JE, et al (2015) Genomic epidemiology of multidrug-resistant *Mycobacterium tuberculosis* during transcontinental spread. J Infect Dis 15:383–388
Didelot X, Gardy J, Colijn C (2013) Bayesian inference of infectious disease transmission from whole genome sequence data. Mol Biol Evol 31:1869–1879.
Didelot X, Walker AS, Peto TE, et al (2016) Within-host evolution of bacterial pathogens. Nat Rev Microbiol 14:150–162.
Doughty EL, Sergeant MJ, Adetifa I, et al (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.
du Plessis L, Stadler T (2015) Getting to the root of epidemic spread with phylodynamic analysis of genomic data. Trends Microbiol 23:383–386.
Dye C, Glaziou P, Floyd K, Raviglione M (2013) Prospects for tuberculosis elimination. Annu Rev Public Health 34:271–86.
Eldholm V, Monteserin J, Rieux A, et al (2015) Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. Nat Commun 6:7119.
Eldholm V, Norheim G, von der Lippe B, et al (2014) Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. Genome Biol 15:490.
Ford CB, Lin PL, Chase MR, et al (2011) Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. Nat Genet 43:482–486.
Ford CB, Shah RR, Maeda MK, et al (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet 45:784–790.
Gardy JL, Johnston JC, Sui SJH, et al (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364:730–739.
Gillespie SH, Crook AM, McHugh TD, et al (2014) Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis. N Engl J Med 371:1577–1587.
Gordienko EN, Kazanov MD, Gelfand MS (2013) Evolution of pan-genomes of Escherichia

coli, Shigella spp., and Salmonella enterica. J Bacteriol 195:2786–2792.
Grenfell BT, Pybus OG, Gog JR, et al (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science (80-) 303:327–332.
Guerra-Assunção J a, Crampin a C, Houben RMGJ, et al (2015a) Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife 4:1–17.
Guerra-Assunção JA, Houben RMGJ, Crampin AC, et al (2015b) 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 211 :1154–1163.
Hatherell H-A, Colijn C, Stagg HR, et al (2016) Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med 14:21.
Johnston JC, Khan FA, Dowdy DW (2015) Reducing relapse in tuberculosis treatment: is it time to reassess WHO treatment guidelines? Int. J. Tuberc. Lung Dis. 19:624.
Jombart T, Cori A, Didelot X, et al (2014) Bayesian eeconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol 10: e1003457.
Kay GL, Sergeant MJ, Zhou Z, et al (2015) Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat Commun 6:6717.
Kühnert D, Stadler T, Vaughan TG, Drummond AJ (2014) Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. J R Soc Interface 11:20131106.
Liu Q, Via LE, Luo T, et al (2015) Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. Sci Rep 5:17507.
Loman NJ, Pallen MJ (2015) Twenty years of bacterial genome sequencing. Nat Rev Microbiol 1–9.
Niemann S, Köser CU, Gagneux S, et al (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS One 4:e7407.
O'Rawe J, Jiang T, Sun G, et al (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5:28.
Pai M, Schito M (2015) Tuberculosis diagnostics in 2015: landscape, priorities, needs, and prospects. J Infect Dis 211 Suppl :S21–8.
Pérez-Lago L, Comas I, Navarro Y, et al (2013) Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. J Infect Dis 1–11.
Perez-Lago L, Martinez Lirola M, Herranz M, et al (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.e1–9.
Pérez-Lago L, Navarro Y, Montilla P, et al (2015) Persistent infection by a *Mycobacterium tuberculosis* strain that was theorized to have advantageous properties, as it was responsible for a massive outbreak. J Clin Microbiol 53:3423–9.
Phelan JE, Coll F, Bergval I, et al (2016) Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. BMC Genomics 17:151.
Quail M, Smith ME, Coupland P, et al (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13:1.
Quick J, Loman NJ, Duraffour S, et al (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232.
Rasmussen DA, Ratmann O, Koelle K (2011) Inference for nonlinear epidemiological models using genealogies and time series. PLoS Comput Biol 7: e1002136.
Rocha EPC, Smith JM, Hurst LD, et al (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239:226–235.
Roetzer A, Diel R, Kohl T a., et al (2013) Whole genome sequencing versus traditional

genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med 10:e1001387.
Schurch AC, Kremer K, Daviena O, et al (2010) High resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol 48: 3403-3406.
Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ (2012) Birth – death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci USA. 110: 228-233.
Sterling TR, Lehmann HP, Frieden TR (2003) Impact of DOTS compared with DOTS-plus on multidrug resistant tuberculosis and tuberculosis deaths: decision analysis. BMJ 326: 574.
Stucki D, Ballif M, Bodmer T, et al (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, et al (2015) Standard genotyping overestimates transmission of Mycobacterium tuberculosis among immigrants in a low-incidence country. J Clin Microbiol 7:1862–1870. Sun G, Luo T, Yang C, et al (2012) Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. J Infect Dis 206:1724–1733.
Tameris MD, Hatherill M, Landry BS, et al (2013) Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial. Lancet 6736:1–8.
Van Rie A, Victor TC, Richardson M, et al (2005) Reinfection and mixed infection cause changing *Mycobacterium tuberculosis* drug-resistance patterns. Am J Respir Crit Care Med 172:636–642.
Van Soolingen D (2014) Whole-genome sequencing of *Mycobacterium tuberculosis* as an epidemiological marker. Lancet Respir Med 4: 251-252.
Walker TM, Ip CL, Harrell RH, et al (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 13:137-146
Walker TM, Monk P, Smith EG, Peto TE a (2013) Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing. Clin Microbiol Infect 19:796–802.
World Health Organization. Global tuberculosis report 2015
Yates T a, Khan PY, Knight GM, et al (2016) The transmission of *Mycobacterium tuberculosis* in high burden settings. Lancet Infect Dis 16:227–238.
Yozwiak NL, Schaffner SF, Sabeti PC (2015) Data sharing: Make outbreak research open access. Nature 518:477–479.
Zumla A, Memish ZA, Maeurer M, et al (2014) Emerging novel and antimicrobial-resistant respiratory tract infections: New drug development and therapeutic options. Lancet Infect. Dis. 14:1136–1149.