Genomic Analysis of Bacterial Outbreaks

Leonor Sánchez-Busó (1,3), Iñaki Comas (1,2, 4), Beatriz Beamud (1), Neris García-González (1), Marta Pla-Díaz (1) and Fernando González-Candelas (1,2)

(1) Unidad Mixta "Infección y Salud Pública" FISABIO/CSISP – Universidad de Valencia/Instituto Cavanilles de Biodiversidad y Biología Evolutiva. Valencia, Spain.
(2) CIBER en Epidemiología y Salud Pública. Valencia, Spain.
(3) Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridgeshire, United Kingdom.
(4) Instituto de Biomedicina, CSCIC, Valencia, Spain.

Author for Correspondence:

Prof. Fernando González Candelas

Evolución y Salud – Instituto Cavanilles Biodiversidad y Biología Evolutiva

Universidad de Valencia

c/Catedrático José Beltrán, 2. 46980 Paterna (Valencia). Spain.

Email: fernando.gonzalez@uv.es

Phone: + 34 961925961 , +34 963543653

FAX: +34 963543670

Introduction

Outbreaks of infectious diseases often produce social alarms. These can be very local or reach every corner of every village and city on Earth. But all they share a need for a quick control and remediation that ensures the safety of the population. The identification and control of the source of an outbreak becomes a health priority and many efforts are devoted to these activities in the first days and weeks after the detection and/or declaration of an outbreak (Mortimer, 2003).

Outbreaks come in many shapes and flavors. For epidemiologists, an outbreak is simply an unusual increase in the prevalence of a disease in time and space. Hence, some outbreaks may be declared and last for years while others are reduced to a few days or weeks; similarly, there might be an outbreak in a school or nursing home, but we talked a few years ago about an epidemic outbreak of "swine influenza" (Fraser et al, 2009;General Directorate of Epidemiology et al, 2009) and the WHO and other health organizations are currently worried about the spread of Zika virus. In some cases, the spread of the infectious pathogen occurs in a series of successive infections from one host to another thus producing transmission chains or networks, depending on the topology of the resulting connections among infected persons.

One of the first tasks when an outbreak is suspected is to establish the basic parameters for controlling it. This can depend on the detection of a source, and the application of actions that prevent it from spreading the pathogen, or the characterization of the vector, so it can be controlled with chemical or biological agents, or the identification of the hereditary factors that allow the pathogen eluding previous, successful treatments and originate nosocomial outbreaks of multi-resistant strains. The advent of faster and cheaper gene sequencing techniques lead to the first systematic and general proposal of using a universal typing scheme that was reproducible, cheap, objective and easily exchangeable among laboratories, known as Multi-Locus Sequence Typing or MLST (Maiden et al, 1998). In this method, the nucleotide sequence of 6-7 loci is determined and used to derive an array of allele profiles in these loci. A new combination of allele profiles corresponds to a new sequence type which is uploaded to a web-server for easy access. Typing schemes, with detailed laboratory protocols, proficiency tests, and full information on identified sequences types are available for tens of bacterial species in general and specific web-servers (see, for instance, http://www.PubMLST.org).

For many pathogens, the availability of a MLST scheme represented a more than significant change in the analysis of outbreaks. This method quickly became the new "gold-standard" for typing pathogens and replaced previous methods. However, for a few but important pathogens no MLST scheme revealing enough genetic variation for effectively distinguishing among non-epidemiologically linked isolates could be designed. These pathogens include the causative agents of plague (*Yersinia pestis*), anthrax (*Bacillus anthracis*), tuberculosis (*Mycobacterium tuberculosis*) and leprosy (*Mycobacterium leprae*), among others, and are collectively known as "genetically monomorphic bacteria" (Achtman, 2012). Specific typing methods such as insertion sequence RFLP and MIRU-VNTR were applied to *M. tuberculosis*, the pathogenic bacteria with the highest incidence and causing more deaths every year in the history of humankind. In these and other cases, the solutions adopted relied on very fast evolving markers, which are usually prone to homoplastic changes, thus resulting in some false positive identifications of phenotypic identities as indicative of very recent ancestry. Although this is not a problem in most settings, it became evident that the same logic applied in using MLST could be extended to the complete genome sequences to attain "perfect" accuracy by using all the genetic information in the isolates and not only a small sample from it.

This approach was first used in an outbreak setting in the investigation of the letters covered with anthrax spores in the aftermath of the 9/11 attacks in the USA. Complete genome sequences were obtained from a *B. anthracis* isolate derived from one of the victims and one reference strain, providing 60 SNPs that could be used subsequently to probe the common origin of the strain used in the bioterrorist attacks (Read et al, 2002). This work clearly showed that using the complete genome sequence was a more effective method for comparing isolates even in almost completely monomorphic species. However, Sanger sequencing is rather slow and painstaking as a result of the need to cut or amplify the genome in small pieces that are subsequently sequenced and assembled into a complete genome sequence. This situation changed dramatically with the introduction of new sequencing methods, then known as "next-generation sequencing" technologies. They offered several advantages over the traditional Sanger method (Medini et al, 2008). At the same time, other problems arose, such as the difficulties in handling and analyzing very large volumes of data, a myriad of programs and methods to analyze them, and new conceptual challenges in the interpretation of the results.

In this chapter we provide a brief overview of the different next-generation sequencing platforms and methods currently available for deriving complete genome sequences from bacteria, the main results in terms of the epidemiological and evolutionary advances that have resulted from their application to bacterial outbreaks and transmission networks, and provide a more detailed analysis of two cases, the analysis of *Legionella pneumophila* outbreaks and of *M. tuberculosis* transmission networks.

High throughput sequencing technologies in outbreak investigations

Several high throughput sequencing platforms have been applied to the genomic study of both bacterial and virus pathogens. Encouraged by the increasing need of sequencing human genomes, three technologies were almost simultaneously released from different companies: 454 (Roche, introduced in 2005 and discontinued in 2016), Solexa (Illumina, introduced in 2006), and SOLiD (Life Technologies, introduced in 2006). These platforms share a general workflow, based on the idea of performing billions of sequencing reactions simultaneously. These are produced through molecular amplification of DNA fragments that are previously attached to a solid surface. These have been enhanced in their subsequent updates to increase both sequencing quality and throughput (Figure 1).

Although 454 was the first released platform, its use has mainly been relegated to metagenomic studies (Schlüter et al, 2008b;Schlüter et al, 2008a;Ghai et al, 2010) because of its long reads and relatively high error rates, which complicates the study of transmission chains or related cases during outbreak investigations. However, it has been used as the main technology in several studies (Lewis et al, 2010;Kennemann et al, 2011;Loman and Constantinidou, 2013) and also following mixed strategies involving the usage of 454 reads as scaffolds and posterior error correction using Illumina (McAdam et al, 2012;Hasan et al, 2012). SOLiD has been the least used for outbreak investigations due to shorter and lower quality reads. As an example, it has been punctually applied in the investigation of *L. pneumophila* outbreaks in an endemic locality in Spain (Sánchez-Busó et al, 2014), *Mycobacterium abscessus* subsp. *bolletii* in Brazil and UK outbreaks (Davidson et al, 2013) or *Coccidioides immitis* producing coccidioidomycosis in transplanted patients in Los Angeles (Engelthaler et al, 2011). By far, Illumina has been the most

widely used platform because of its high quality and sensible sized reads, which allow more accurate mapping and SNP calling. A thorough summary of the application of different sequencing technologies to analyze different mainly bacterial outbreaks is shown in Table 1.

In 2010, the Ion Torrent (Life Technologies) platform, a new benchtop device with a different sequencing strategy was commercialized. This technology is based on monitoring pH changes in multi-well plates. A single reaction occurs per well so that when a hydrogen atom is released after the incorporation of each nucleotide during amplification, the pH in the media changes in a nucleotide-specific manner, so that the system is able to translate chemical into digital information. Reads produced by the Ion Torrent were of relatively good quality and was punctually applied to the study of *Escherichia coli* outbreaks (Mellmann et al, 2011;Holmes et al, 2015) and *Pseudomonas aeruginosa* (Snyder et al, 2013;Witney et al, 2014).

In early 2011, the PacBio RS system was also released, being the first platform performing Single Molecule Real Time (SMRT) sequencing, which is being increasingly applied to complete microbial genomes because of the long read lengths (Mutreja et al, 2011). But the definite current revolution in sequencing technologies with an impact in public health has been the release of the Oxford Nanopore MinION platform, currently in test mode, and scalable in the form of the GridION platform. These contain a membrane with millions of embedded nanopores coupled with a polymerase. Changes in the electrical conductivity in the membrane as the different four bases pass through the nanopore are measured, allowing sequencing in real time. Specifically, the MinION platform is an USB-like device which can be connected directly to a computer and provide the sequences from extracted DNA in real time after a very simple library preparation. The portable MinION platform has been shown to be useful in real-time outbreak investigations, such as the 2015 Ebola virus disease epidemic in West Africa (Quick et al, 2016).

The different platforms differ in their sequencing strategy, which yields different throughputs and sequence qualities. Currently, the highest throughput can be achieved with the HiSeq X Ten Illumina platform, which can yield up to 3 billion of paired-end 150 bp sequences. This high level throughput is mainly directed to population-scale human genome sequencing projects. In the case of microorganism sequencing, because their genomes are much smaller, sequencing throughput must depend on the depth of coverage required for each specific study. However, large-scale microbial sequencing projects can benefit from these high throughput platforms by multiplexing different strains in the same run. Coverage depths of 50X-100X are usually sought for base call error correction, minimizing the rate of false positive SNPs. Currently, the technologies with the lowest error rates are Illumina platforms, and the highest error rate from raw data is provided by Oxford Nanopore and PacBio platforms. However, bioinformatics pipelines for error correction during the post-processing of reads improve these rates, especially in the second case, in which the current final error rate can get as low as 1E-05. Multiple reviews on the characteristics of the different sequencing technologies, applications, advantages and drawbacks have been published in the literature up to now (Metzker, 2010;Casey et al, 2013;Ekblom and Wolf, 2014).

Choosing the most appropriate sequencing technology depends on the scope of the study. High throughput technologies can be applied in different steps during an outbreak investigation (Köser et al, 2012); from the detection and identification of the pathogen in direct uncultured samples (i.e. blood, sputum, etc.), epidemiological typing and detection of mutations associated to drug susceptibility to the study of transmission chains and potential super-spreaders.

Achievements and limitations of NGS in outbreak investigations

Initial results. Although NGS techniques and devices became available around 2005 (Loman and Pallen, 2015), it took a few more years until the new technologies were firstly applied to analyze an outbreak. This corresponded to an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) (Harris et al, 2010). They analyzed a set of 63 isolates from two origins, a global collection of 43 samples collected between 1982 and 2003, and 20 isolates from a Thai hospital sampled in a very short time period (months), suspected to correspond to a transmission chain. Their results provided evidence for the international spread of the resistant clone of *S. aureus* and the single origin of the samples from the hospital. But they also showed that bacteria can and do evolve rapidly. They estimated that in the core genome, the set of shared positions among all the studied isolates, the rate of divergence was about 1 SNP every 6 weeks. This explained the lack of identity among most hospital isolates, which differed in a few SNPs from each other, but it also revealed differences from the patterns of evolution revealed by other markers, such as *spa* and PFGE. Of note, the analysis of complete genomes showed that over a quarter of the homoplasies found among the isolates were directly related to the evolution of resistance to antibiotics.

At about the same time, Lewis et al. (2010) used complete genome sequences to establish relationships among otherwise indistinguishable strains of *Acinetobacter baumannii* which had cause a small outbreak at a British hospital. The SNPs found by WGS allowed the investigators to discriminate among alternative epidemiological hypotheses. These pioneering studies have been followed many studies (Table 1) which have dealt with outbreaks and transmission networks of over 30 different bacteria species infecting humans. An even larger number of works have been published about viral infections (not included in this review) and a few have dealt with fungal infections. Two particular bacteria, *M. tuberculosis* and *L. pneumophila*, the main etiological agents of tuberculosis and legionellosis, respectively, are analyzed in more detail below, but some general patterns and conclusions have started to emerge from the analysis of more than 30 pathogenic bacteria, that we briefly review next.

From retrospective to real time analysis of outbreaks. We have previously commented that the molecular analysis of outbreaks and transmission networks is necessarily a complement to the epidemiological investigations leading to the identification and control of the source(s), vectors or routes so to put a fast stop to ongoing processes. Hence, it is very important that the information obtained from the molecular analyses can be shared with the epidemiology team for a better evaluation of the total evidence available thus far and more appropriate and accurate decisions can be adopted. The initial methodologies available for WGS were very labor intensive and the shortest time since a sample was obtained until its complete sequence could be determined was in the order of weeks. Too long for a pressing demand of action. However, the advent of new technologies, such as Ion Torrent PGM and, more recently, MinION have changed this situation. Both methods can deliver sequence information within a few hours of gaining access to the sample, thus allowing a very rapid communication of results to field workers.

The first case in which these new technologies were applied during the investigation of the source of an outbreak was that an enteroaggregative *Escherichia coli* O104:H4 strain that affected several European countries in the spring of 2011 (Mellmann et al, 2011). Complete genome sequences were obtained from a representative isolate of the outbreak and a reference strain which produced similar clinical features in just 62 hours. The comparison revealed key differences in plasmid and gene contents between the strains, indicating that the outbreak was

due to a new and not a previously circulating strain of the bacterium. It also allowed the design of a test to be applied for quick diagnostic in any lab.

Loss of identity as hallmark of relatedness. One consequence of using complete genome sequences for the analysis of outbreaks and transmission chains is the necessary dismissal of complete identity as the proof of charge in considering two or more isolates as linked to the same transmission event or episode. This was usually the case for most previous markers which explored only a minor fraction of the nucleotides in the genome of the pathogenic bacteria. Except for a few rapidly evolving markers, usually associated to tandem repeats, the number of differences expected between two isolates depends on three factors: the mutation rate per site, the number of sites being compared and the time since they diverged from their last common ancestor. When the number of generations since divergence is relatively small, as in outbreaks and most transmission networks, and the number of sites being sampled is also small, the probabilities of finding a SNP (or a different allele in the case of MLST) are also very small. However, using complete genome sequences, and assuming that the previous assumptions remaining identical, will increase those probabilities in a three-fold factor or more, because the number of sites interrogated is now in the order of millions instead of tens or hundreds.

Within-host evolution. In addition, the exploration of complete genome sequences of long- or chronically-infecting bacteria has shown that evolution does occur within hosts at relevant rates for being reflected in some nucleotide changes (Didelot et al, 2016). Even for pathogens that produce acute infections, a low per site mutation rate is compensated by the large number of nucleotides present in a genome and the different random and directional processes that occur in an infected individual, thus leading to some new mutations arising in many newly replicated genomes (Kennemann et al, 2011;Mathers et al, 2015). If the infection last longer or becomes chronic, the chances that changes occur in the pathogen are very high and additional evolutionary processes such as compartmentalization may contribute to within patient differentiation of bacterial sub-populations.

These processes have important consequences at different levels. On the one hand, a variable population can adapt more rapidly to new environmental conditions which might include new treatments or an adaptive immune response by the host (Mwangi et al, 2007). On the other hand, a variable population will result in different initial compositions in successive transmission events, which will be reflected in differences among the populations established in the new hosts. The analysis of transmission networks becomes more complicated because using a single genome sequence per host cannot reveal the whole range of variation present within it (Worby et al, 2014). Under these circumstances, the use of evolutionary methods to reveal the common ancestry of isolates derived from patients presumably included in the same network becomes an absolute necessity.

Mutation patterns and processes. Apart from revealing larger amounts of variation than anticipated from previous studies with just a few gene sequences, whole genome sequences have also informed about the types and distribution of mutational changes occurring at different time-scales. A few years ago, the contribution of homologous recombination and horizontal gene transfer to genetic variation in bacterial genomes was found to be considerably more important than previously thought (Doolittle, 1998). But this was thought to be the result of millions of generations in which a generally rare process might have been acting. In shorter time-scales, months or years, the impact of processes generating variation other than point mutation was thought to be negligible except for loci including repeat units, such as in MIRU-VNTRs in *M. tuberculosis*, in which slippage-and-mispairing during replication often lead to new alleles.

Recent analyses at the complete genome level have shown that this view is incorrect, at least for some bacteria such as *Neisseria gonorrhoeae*, *Salmonella enterica* or *L. pneumophila* (Didelot and Maiden, 2010;Sánchez-Busó et al, 2014). In fact, a comparison of the relative effects of recombination and point mutation in almost 50 bacterial species revealed variation of three orders of magnitude (Vos and Didelot, 2009). Although there are not quantitative estimates yet, horizontal gene transfer, with or without final stabilization in the receiving genome, is also known to play a significant role in the short term evolution of many bacteria, as unfortunately shown by the ease of spread of many antibiotic resistance genes across species. The additional variation introduced by these processes has to be considered when analyzing large transmission networks or long-lasting outbreaks, because the incorporation of these new variants may confound inferences of recent ancestry based on overall similarity or on a few loci.

Rates of evolution. The increased availability of complete genome sequences from bacteria with a more or less direct epidemiological link has also provided an opportunity for a more detailed study of evolutionary processes at the population genomic level. Apart from the different types of variants introduced in these populations, the access to asynchronically sampled isolates allows the application of Bayesian methods to estimate evolutionary rates (Drummond et al, 2006). These methods can accommodate strict and relaxed clock models, different demographic regimes, as well as variation in rates among lineages, thus allowing the estimation of relevant evolutionary parameters from organisms with different natural and evolutionary histories. Most often they are applied to rapidly evolving organisms, collectively known as measurably evolving populations (Drummond et al, 2003;Biek et al, 2015), which mainly include viruses along with some bacteria. But the methods are also valid for more slowly evolving organisms with sampling dates different enough as to provide estimates of the evolutionary rate. Recently, this approach has been used with bacterial genomes obtained from ancient samples (Schuenemann et al, 2013;Bos et al, 2014;Mendum et al, 2014;Rasmussen et al, 2015;Bos et al, 2016;Maixner et al, 2016).

One apparent feature of the estimates of bacterial evolutionary rates is the negative correlation between the time to the most recent common ancestor of the sample studied and the inferred evolutionary rate (Figure 1). Higher evolutionary rates at short times can be explained by the relative inefficiency of natural selection and/or genetic drift in the removal of neutral or quasi-neutral polymorphisms which are continuously arising in bacterial populations. Hence, transitional polymorphisms contribute significantly to the apparent acceleration of evolutionary rates in short time-scales. At the same time, they also provide a wealth of variation what might have an adaptive value if the circumstances are appropriate. On the long run, many of these transient variants will have disappeared and evolutionary rates are reduced correspondingly. This negative correlation has to be taken into account when comparing rates across studies, even for the same species, and in the inference of other evolutionary parameters (Biek et al, 2015).

The analysis of (almost) complete genome data. One of the main advantages of MLST or SBT over alternative methods for the analysis of pathogenic bacteria in the context of outbreaks and transmission chains is the objectivity and simplicity in the specification of the variants found in any isolate. The nucleotide sequences obtained for each locus are compared to a predetermined database in which previous homologous sequences have been deposited. If there is a perfect match, the newly determined variant received the same identifier as the pre-existing one. If that is not the case, curators of the database will assign a new code to the variant. The combination of allele codes in the loci included in the typing scheme is summarized in a sequence type (ST)

with a different number of each combination of variants. This procedure is easily communicated because it requires the identification of nucleotide variants, usually through Sanger sequencing, in just 6 or 7 loci. However, the advent of NGS and the determination of complete genome sequences makes this procedure of denoting the variants impractical.

Several alternative have already been proposed for the identification of complete genome sequences for epidemiological analysis. One method consists in extending the MLST naming scheme to more loci, eventually all the loci in the genome of the corresponding species, thus leading to "whole genome MLST" (wgMLST) schemes (Cody et al, 2013). The first proposal of wgMSLT was done for *Campylobacter* isolates and the initial MLST scheme based on 7 loci was extended to 1667 loci, although this number was reduced to 1026 when only those present in all the isolates analyzed were considered. This represents the "core genome" of the species, which is complemented by the "auxiliary genome", the set of loci which are present in some but not all the isolates of a species. In light of the very large genome plasticity of many bacterial species, fixed compositions of the core and auxiliary genomes are almost impossible, which creates an additional problem for the stability of the scheme. Nevertheless, this approach has gained some popularity and cgMLST ("core genome MLST", a reduced version of wgMLST as described above) schemes are now available for several pathogens including *S. aureus*, *Listeria monocytogenes*, *Enterococcus faecium* (de Been et al, 2015), and *S. enterica* (Taylor et al, 2015), among others.

To prevent the proliferation of STs which inevitably accompanies wgMLST or cgMLST, a first level classification of STs into clusters or clonal groups is usually performed (Cody et al, 2013;Qin et al, 2016). These can be based on an extension of the BURST method (Feil et al, 2001;Feil et al, 2004), which considers as variants of the same clonal group to those that differ in one single locus of the original MLST scheme, or use more sophisticated approaches based on the population genetic analysis of the actual SNPs detected in the loci included in the wgMLST or cgMLST (Qin et al, 2016) with different molecular population methods such as BAPS (Corander and Tang, 2007) or STRUCTURE (Rosenberg et al, 2002). These methods share the advantage of portability thus allowing comparisons among different laboratories and needs. However, they also discard important information, eventually crucial, contained in the auxiliary genome. Hence, although standard typing schemes are useful, whole genome sequence information should not be reduced to a ST number or complex under a wgMLST and the complete data should still be available for future use by the scientific community.

Outbreak investigation in *Mycobacterium tuberculosis*: the genome as an epidemiological marker

Mycobacterium tuberculosis is the main causative agent of human tuberculosis in the world. Every year more than 1.5 million persons die of tuberculosis, more than of any other infectious disease (WHO, 2014). The epidemiology of the disease has to take into account the natural history of the bacteria. It is an obligate human pathogen with very effective airborne transmission and that typically infects the lungs. It is estimated that one third of the human population is infected by the bacilli and this explains why every year around 9 million new cases are declared. In most cases the initial infection derives in an asymptomatic state called latency in which the bacteria have not been eliminated but are controlled by the immune system. In 5-

10% of the latent cases the disease progresses to an active state in which the bacteria actively replicate and cause pulmonary disease. Only an active tuberculosis case can transmit the disease and thus in tuberculosis, disease and transmission are linked. The typical window of progress to active disease after infection is two years but the bacteria may remain latent for years or even decades.

Mycobacterium tuberculosis has been traditionally regarded as a monomorphic organism due to the low genetic diversity found among representative strains datasets (Achtman, 2008). Thus epidemiological tools were developed based on fast evolving genetic elements (Barnes and Cave, 2013). Typing of the insertion sequence IS6110 by RFLP and of minisatellites, called MIRU-VNTR, are the two gold standards in tuberculosis molecular epidemiology and, together with spoligotyping, based on the CRISPR region of the bacteria, have allowed to define successful *M. tuberculosis* clones. Among these clones, the identification of an hypervirulent clade, called Beijing family, has attracted much attention (Parwati et al., 2010). Strains from the Beijing family are more common in East Asia but can be identified across the globe. Experimental and epidemiological research have identified Beijing strains as hypervirulent in the mice model of infection and with frequent association to drug resistance in humans. In South Africa Beijing strains have been on the rise for the last 40 years (Cowley et al., 2008). Beijing strains belong to one of the seven lineages of human tuberculosis strains (Comas et al., 2013). The most common is lineage 4, which is highly frequent in Africa, Europe and America. There is a strong association between lineages and their geographic origin, being the most extreme cases the two lineages of *Mycobacterium africanum,* that can only be found in West Africa (De Jong et al, 2010), and Lineage 7 recently described in Ethiopia (Comas et al., 2013). Regardless the lineage, drug resistance to first and second line treatments have been identified (Farhat et al., 2013). The mutations responsible for drug resistance are always chromosomal mutations because there is no ongoing horizontal gene transfer in *M. tuberculosis*. Although ecological theory predicts that drug resistance mutations have a fitness cost, experimental evolution and molecular epidemiology have shown that different drug resistance mutations have different fitness costs (Comas et al., 2012). As a consequence, multidrug-resistance cases (MDR-TB) among people never treated before, and therefore due to transmission, are on the rise and in some particular areas represent more than 50% of the tuberculosis burden of the region. Although not part of this review whole genome sequencing is allowing to define the set of mutations associated to resistance to the different antibiotics but also the genotype of highly successful MDR-TB strains.

The first study that showed the potential of the genome as an epidemiological marker dates back to 2009 (Niemann et al, 2009). In this study, three strains which looked almost identical using traditional molecular epidemiology markers such as restriction fragment length polymorphisms (RFLP) and minisatellite (MIRU-VNTR) were shown to differ in more than 100 SNPs. Later on, Jennifer Gardy and collaborators (2011) used genome comparison techniques to solve an on-going outbreak in British Columbia suspected to have started in the early 1990s. By combining genomic, epidemiological and social contact data the authors showed that it can be gained get a better resolution of the transmission events within transmission clusters. Such events are very difficult to identify with traditional molecular epidemiology markers. This work already defined index cases associated to multiple secondary cases, also denoted as super-spreaders. Super-spreaders are becoming a common topic when analyzing large transmission clusters (Walker et al, 2013b) instead of the traditional view of a stepwise "chain" of transmission.

From 2010, NGS has been successfully applied to deeply resolve tuberculosis outbreaks. Considerably attention has been paid to understand those outbreaks that have been on-going over years. For example, a large outbreak in Hamburg, Germany, was identified by classical genotyping data in 1996 (Roetzer et al, 2013). However, clustering data not always correlated with epidemiological and geographical information leading to the suspicion that the outbreak was more complex than previously anticipated. By whole genome sequencing of 86 strains from the outbreak (1996-2011), Roetzer et al. (2013) were able to identify an independent transmission network, thus confirming the non-clonality of the outbreak. Two clusters were determined, one starting in 1997 and the other starting in 2010, much more in agreement with epidemiological investigations. Therefore, one important application of whole genome sequencing to investigate tuberculosis outbreaks is to ability to assign with higher confidence cases to the outbreak and exclude those that, albeit genetically close, correspond to a different chain of events.

Similarly, in Bern, Switzerland, a genotype detected by RFLP profiling caused a large number of tuberculosis cases during the 1990's (Stucki et al, 2015). The cases were associated to the typical risk factors in local populations found in European cities such as HIV infection or alcoholism. Stucki et al. (2015) sequenced the complete genome of strains belonging to the original outbreak along with local control strains. By comparing outbreak and control strains they designed a real-time SNP typing assay based on the detection of genome position with a polymorphism specific to the outbreak strains. Next, they typed a retrospective collection of isolates of the Canton of Bern from 1993 to 2011. They were able to identify 68 additional cases of the outbreak based on the presence of the mentioned SNP including cases from 2011. Therefore, the combination of whole genome sequencing and SNP typing allowed them to identify cases associated to the outbreak and find that the outbreak that started in early nineties was still on-going at the time of investigation. In addition, they obtained the whole genome sequence of all the isolates assigned to the outbreak. With this information, they were able to resolve the individual transmission patterns for 75% of the strains. Importantly, 66 out of the 68 strains had exactly the same RFLP pattern. Furthermore, the analysis of the transmission network together with the epidemiological information revealed two different sub-outbreaks initiated by two different "super-spreaders".

Therefore, next generation sequencing of the Hamburg (Roetzer et al, 2013), the Bern outbreak (Stucki et al, 2015) and others (Török and Peacock, 2012;Smit et al, 2015;Lee et al, 2015) have revealed the complexity of tuberculosis outbreaks. Given that tuberculosis is not an acute disease and that a tuberculosis case can be latent, asymptomatic for years, the true extent of tuberculosis outbreaks can only be revealed by a sustained genotyping efforts over years. Furthermore, as in the case of the Bern outbreak, whole genome sequence data can be used to design new diagnostics and/or surveillance tools. A similar approach has been used to prospectively identify new outbreak-associated cases in sputum samples (Pérez-Lago et al, 2015).

Apart from specific outbreaks, genomic epidemiology has been used in a population-based scale to evaluate its utility for surveillance and diagnostics. In a series of publications starting in 2012, Public Health England has applied next generation sequencing to incorporate whole genome sequencing as the default typing method of *Mycobacterium tuberculosis* in the United Kingdom (Walker and Beatson, 2012;Walker et al, 2014). They have shown that the genome data allow to delineate outbreaks better than MIRU-VNTR analyses. Furthermore, in an attempt to derive a rule of thumb to identify a transmission event between two cases they also sequence several

isolates from the same patient and known household contacts. They were able to identify a threshold of five SNPs when the cases had a confirmed epidemiological link and they proposed a threshold of up to 12 SNPs for casual transmission in the community (Walker and Beatson, 2012). Other studies have found a similar distribution of SNPs when analyzing transmission events in populations (Bryant et al, 2013a;Casali et al, 2014).

However, we are still blinded about how these thresholds apply to different clinical settings than the low-burden countries of Europe. In high-burden countries delineating transmission clusters should be more difficult if public health interventions cannot stop transmission events (Yates et al, 2016). Thus, the circulating strains may be participating at the same time in several clusters. The only population-based study published in a high-burden country shows that the threshold described in (Pérez-Lago et al, 2015) may be useful, although more work will be needed to generalize the results to, for example, large urban areas.

There are several factors that may distort the proposed threshold values. One of these factors is mixed infections. The true extent of co-infections in high-burden countries is not clear and there is hope that whole genome data can distinguish between relapses and re-infections (Bryant et al, 2013a;Guerra-Assunção et al, 2015). This issue is critical to delineate transmission in high burden countries but also for clinical trials investigations because relapse is one of the end points of those investigations. However, it is the diversity that can be found during infection from a single strain what is attracting more research and attention. From drug susceptibility clinical data, it has been clear for decades that several populations may co-exist in the same patient. These subpopulations were flagged due to inconsistent results in drug resistance susceptibility tests between isolates of the same patient (Rinder et al, 2001). Whole genome sequencing has shown that, in fact, this is the case and what is recovered from a sputum sample is often a mix of different sub-populations (Sun et al, 2012). These sub-populations can be revealed by looking at positions in which a mutant and a wild-type allele can be identified at the same time. In the context of drug resistance, it has been shown that several drug resistant sub-populations may co-exist and compete and that their frequencies may change over time (Liu et al, 2015). A similar phenomenon has been shown outside the context of drug resistance. The issue of within patient diversity not only has clinical and diagnostic implications. If several sub-populations co-exist and accumulate a different number of SNPs then chances are that the epidemiological investigation of outbreaks may be distorted by the isolate chosen for the analysis (Walker et al, 2013a;Walker et al, 2013b). An analysis of cases in which higher than expected diversity was expected confirmed that, although the thresholds proposed to delineate a transmission event are in general valid, there are epidemiologically cases in which a larger than expected number of SNPs can be found (Pérez-Lago et al, 2014). How frequent are those "outliers" is a matter of on-going investigation.

High throughput investigation of *Legionella pneumophila* outbreaks

High throughput sequencing can also be used to study organisms with higher level of polymorphism and strictly environmental, contrary to Mycobacterium tuberculosis. This is the case of *L. pneumophila*, causative agent of Legionellosis, and for which there is only one report of a possible person-to-person transmission (Correia et al, 2016) up to date. This opportunistic pathogen can produce pneumonia after inhalation of aerosols with enough bacterial load, with

the highest burden in warm water-related environments. The first reported outbreak dates from 1976 when more than hundred legionnaires were infected in a convention in Philadelphia (Fraser et al, 1977). A legionellosis outbreak is defined as a cluster of more than three cases occurring at the same place and time and the epidemiological investigation is crucial to find the environmental sources.

The investigation of legionellosis outbreaks has traditionally been conducted by using biochemical or molecular methods that allows comparing the clinical isolates with the strains obtained from the environment (Fields et al, 2002). Broad techniques such as serogrouping benefited from genetic methods that provided improved resolution in the so-called Sequence-Based Typing (SBT) (Gaia et al, 2003;Gaia et al, 2005), based on Multi-Locus Sequence Typing (MLST) approach (Urwin and Maiden, 2003) but incorporating virulence genes in the scheme to increase the discrimination power among strains.

However, although SBT provided researchers with a tool that allowed the classification of strains into groups (Sequence Types, STs), the introduction of high-throughput sequencing techniques for microbial analysis and outbreak investigations in other species derived in its application to legionellosis outbreaks because of its increased discrimination power. The first published work was indeed a pilot study to test the potential of whole-genome sequencing (WGS) on the discrimination between isolates from an outbreak produced in the UK in 2003 and non-outbreak related strains (Reuter et al, 2013b). From this point, a number of other outbreaks have been analyzed using WGS, as for example an outbreak of ST62 associated to a cooling tower in Quebec City in 2012 (Lévesque et al, 2014) or a massive outbreak that occurred in Edinburgh (UK, 2012) related to multiple STs and including mixed infections (McAdam et al, 2014). WGS has also been used to investigate the persistent infection history of ST23 in a hotel in Spain in 2012 (Sánchez-Busó et al, 2016) and the eradication of *L. pneumophila* associated to a hospital in Australia that have been responsible of nosocomial cases (Bartley et al, 2016).

The environmental source of legionellosis cases has been historically difficult to trace, and because of the high social and economic impact of this kind of outbreaks on the affected populations, public health interventions are obliged to be rapid and accurate. WGS has shown further variability within many STs (Underwood et al, 2013a;Sánchez-Busó et al, 2014), showing evidence that at least some of them are not clonal. This observation complicates the study of legionellosis outbreaks and was the leading aim in the study by Sánchez-Busó et al. (2014). In this work, 69 isolates including strains associated to 13 different outbreaks and sporadic cases occurred in a single locality (Alcoy, Spain) during more than 10 years (1999-2010) were analyzed by high throughput sequencing. Different STs were included, with special interest on ST578 cases, which had been recurrently reported as the causing ST of most of those outbreaks (Coscollá et al, 2010).

The analysis showed two main lineages within the endemic ST578, more than 1,000 SNPs apart from each other. Not all the strains from the same outbreak clustered together, revealing the non-clonality of the isolates, as these were phylogenetically grouped independently of their source (clinical or environmental), sampling date or outbreak. Because ST578 is known to be endemic in the area of Alcoy, these results suggest that it is indeed very complicated to find an infectious source using just molecular data in endemic areas. These should be used together with the epidemiological investigation to be able to draw the accurate conclusions that public health interventions require.

Other interesting fact that this work shows is that the genomic data can reflect public health actions along time. As an example, using Bayesian inference, an estimate of the ST578 population dynamics revealed a decreased population size between 2006 and 2008, which correlated with a moment in which public health measures were taken in the city by removing high-risk installation from the city center.

In the case of organisms where person-to-person transmission is very rare or even inexistent, whole genome sequencing can provide the most discriminant tool to link clinical cases with environmental sources, providing the accuracy that public health interventions require in these cases. But, moreover, it can help understand how outbreaks occur, which is the starting line to be able to predict and even prevent their occurrence.

Conclusion

Complete genome analysis of bacterial pathogens is still far from being the usual method for analyzing outbreaks and transmission networks, although it will not take long before it does so. The increasing speed, ease and reliability as well as the reduced costs associated to new high-throughput sequencing technologies point to that direction. But gaining information is only a part of the process. More data also mean an increased need for interpretative tools at all levels, from the mere analysis of reads to the inference of the evolutionary and genealogical relationships among the isolates. Progress is still pending at all levels, from the technology to obtain, fast and cheap, complete genome sequence data of a specific pathogen from an infected individual or a potential vector o source to analytical tools capable of extracting the relevant information from the deluge of data generated by high-throughput sequencers and for the integration of this information with the clinical, epidemiological and evolutionary information which are needed when they have to be interpreted in the appropriate context.

Acknowledgements

We thank Dr. Pierre Pontarotti for his kind invitation to write this chapter. This work has been funded by project BFU2014-58656-R from MINECO (Spanish Government) to FGC. IC is supported by Ramón y Cajal Spanish research grant RYC-2012-10627, MINECO research grant SAF2013-43521-R, and the European Research Council (ERC) (638553-TB-ACCELERATE). BB has been recipient of a Beca de Colaboración from the Spanish Ministerio de Educación y Cultura.

References

Literature Cited

Achtman M. (2008) Evolution, Population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 62:53–70.

Achtman M (2012). Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Phil Trans R Soc B 367:860-867.

Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, Broomall S, Chain PSG, Chertkov O, Chokoshvili O, Coyne S, Davenport K, Detter JC, Dorman W, Erkkila TH, Folster JP, Frey KG, George M, Gleasner C, Henry M, Hill KK, Hubbard K, Insalaco J, Johnson S, Kitzmiller A, Krepps M, Lo CC, Luu T, McNew LA, Minogue T, Munk CA, Osborne B, Patel M, Reitenga KG, Rosenzweig CN, Shea A, Shen X, Strockbine N, Tarr C, Teshima H, van Gieson E, Verratti K, Wolcott M, Xie G, Sozhamannan S, Gibbons HS, Threat Characterization Consortium (2012). Genomic Comparison of *Escherichia coli* O104:H4 Isolates from 2009 and 2011 Reveals Plasmid, and Prophage Heterogeneity, Including Shiga Toxin Encoding Phage stx2. PLoS ONE 7:e48228.

Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW (2012). High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13:1.

Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW (2013). On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. PLoS ONE 8:e55254.

Azarian T, Cook RL, Johnson JA, Guzman N, McCarter YS, Gomez N, Rathore MH, Morris JGJ, Salemi M (2015). Whole-Genome Sequencing for Outbreak Investigations of Methicillin-Resistant *Staphylococcus aureus* in the Neonatal Intensive Care Unit: Time for Routine Practice? Infection Control & Hospital Epidemiology 36:777-785.

Barnes PF, Cave MD. (2003). Molecular epidemiology of tuberculosis. N Engl J Med 349:1149–1156.

Bartley PB, Ben Zakour NL, Stanton-Cook M, Muguli R, Prado L, Garnys V, Taylor K, Barnett TC, Pinna G, Robson J, Paterson DL, Walker MJ, Schembri MA, Beatson SA (2016). Hospital-wide eradication of a nosocomial *Legionella pneumophila* serogroup 1 outbreak. Clin Infect Dis 62:273-279.

Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G, Fournier E, Doualla-Bell F, Levac E, Gaulin C, Ramsay D, Huot C, WAlker M, Sieffert C, Tremblay C (2016). Usefulness of High-Quality Core Genome Single-Nucleotide Variant Analysis for Subtyping the Highly Clonal and the Most Prevalent *Salmonella enterica* Serovar Heidelberg Clone in the Context of Outbreak Investigations. J Clin Microbiol 54:289-295.

Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, Maiden MCJ (2012). A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. Microbiology 158:1570-1580.

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

Blouin Y, Cazajous G, Dehan C, Soler C, Vong R, Hassan MO, Hauck Y, Boulais C, Andriamanantena D, Martinaud C, Martin É, Pourcel C, Vergnaud G (2014). Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, Djibouti. Emerging Infectious Diseases 20:21-28.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514:494-497.

Bos KI, Herbig A, Sahl J, Waglechner N, Fourment M, Forrest SA, Klunk J, Schuenemann VJ, Poinar D, Kuch M, Golding GB, Dutour O, Keim P, Wagner DM, Holmes EC, Krause J, Poinar HN (2016). Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. eLife Sciences12994.

Bryant J, Schurch A, van Deutekom H, Harris S, de Beer J, de Jager V, Kremer K, van Hijum S, Siezen R, Borgdorff M, Bentley S, Parkhill J, van Soolingen D (2013a). Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infectious Diseases 13:110.

Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA (2013b). Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. The Lancet 381:1551-1560.

Brzuszkiewicz E, Th++rmer A, Schuldes J+, Leimbach A, Liesegang H, Meyer FD, Boelter J+, Petersen H, Gottschalk G, Daniel R (2011). Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). Archives of microbiology 193:883-891.

Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M (2013). Phylogenetics and differentiation of *Salmonella* Newport lineages by Whole Genome Sequencing. PLoS ONE 8:e55687.

Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F (2014). Evolution and transmission of drug resistant tuberculosis in a Russian population. Nat Genet 46:279-286.

Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, Bentley SD, Parkhill J, Nejentsev S, Hoffner SE, Horstmann RD, Brown T, Drobniewski F (2012). Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Research 22:735-745.

Casey G, Conti D, Haile R, Duggan D (2013). Next generation sequencing and a new era of medicine. Gut 62:920-932.

Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, Salter SJ, Harris D, Nosten F, Goldblatt D, Corander J, Parkhill J, Turner P, Bentley SD (2014). Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 46:305-309.

Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK (2011). The origin of the Haitian cholera outbreak strain. New England Journal of Medicine 364:33-42.

Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler ICJW, Jolley KA, Maiden MCJ (2013). Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol 51:2526-2534.

Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S (2012). Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 44:106–10.

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, 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.

Corander J, Tang J (2007). Bayesian analysis of population structure based on linked molecular information. Mathematical Biosciences 205:19-31.

Correia AM, Ferreira JS, Borges V, Nunes A, Gomes B, Capucho R, Gonçalves J, Antunes DM, Almeida S, Mendes A, Guerreiro M, Sampaio DA, Vieira L, Machado J, Simões MJ, Gonçalves P, Gomes JP (2016). Probable person-to-person transmission of Legionnaires' disease. New England Journal of Medicine 374:497-498.

Coscollá M, Barry PM, Oeltmann JE, Koshinsky H, Shaw T, Cilnis M, Posey J, Rose J, Weber T, Fofanov VY, Gagneux S, Kato-Maeda M, Metcalfe JZ (2015). Genomic Epidemiology of Multidrug-Resistant *Mycobacterium tuberculosis* During Transcontinental Spread. Journal of Infectious Diseases.

Coscollá M, Fenollar J, Escribano I, González-Candelas F (2010). Legionellosis outbreak associated with asphalt paving machine, Spain, 2009. Emerging Infectious Diseases 16:1381-1387.

Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, Wilkinson RJ, Nicol MP (2008) Recent and Rapid Emergence of W-Beijing Strains of Mycobacterium tuberculosis in Cape Town, South Africa. Clinical Infectious Diseases 47:1252–9.

Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M (2013). Population genomics of post-vaccine changes in pneumococcal epidemiology. Nature Genetics 45:656-663.

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD (2011). Rapid pneumococcal evolution in response to clinical interventions. Science 331:430-434.

Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, Weinert LA, Wang Z, Guo Z, Xu L, Zhang Y, Zheng H, Qin N, Xiao X, Wu M, Wang X, Zhou D, Qi Z, Du Z, Wu H, Yang X, Cao H, Wang H, Wang J, Yao S, Rakin A, Li Y, Falush D, Balloux F, Achtman M, Song Y, Wang J, Yang R (2013). Historical

variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. Proceedings of the National Academy of Sciences USA 110:577-582.

Davidson RM, Hasan NA, de Moura VCN, Duarte RS, Jackson M, Strong M (2013). Phylogenomics of Brazilian epidemic isolates of *Mycobacterium abscessus* subsp. *bolletii* reveals relationships of global outbreak strains. Infection, Genetics and Evolution 20:292-297.

de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen D, Willems RJL (2015). Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. J Clin Microbiol 53:3788-3797.

De Jong BC, Antonio M, Gagneux S (2010). *Mycobacterium africanum*—Review of an important cause of human tuberculosis in West Africa. PLoS Negl Trop Dis;4:e744.

Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, Shi M, Fisman DN, Dhody AN, Forrest S, Bos KI, Earn DJD, Holmes EC, Poinar HN (2014). Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. New England Journal of Medicine 370:334-340.

Didelot X, Eyre D, Cule M, Ip C, Ansari A, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty E, Piazza P, Wilson D, Bowden R, Donnelly P, Dingle K, Wilcox M, Walker S, Crook D, Peto T, Harding R (2012a). Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. Genome Biology 13:R118.

Didelot X, Maiden MCJ (2010). Impact of recombination on bacterial evolution. Trends in Microbiolog 18:315-322.

Didelot X, Meric G, Falush D, Darling A (2012b). Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13:256.

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ (2016). Within-host evolution of bacterial pathogens. Nat Rev Micro 14:150-162.

Doolittle WF (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14:307-311.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006). Relaxed phylogenetics and dating with confidence. PLoS Biology 4:e88.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003). Measurably evolving populations. Trends in Ecology & Evolution 18:481-488.

Ekblom R, Wolf JBW (2014). A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026-1042.

Engelthaler DM, Chiller T, Schupp JA, Colvin J, Beckstrom-Sternbergg SM, Driebe EM, Moses T, Tembe W, Sinari S, Beckstrom-Sternbergg JS, Christoforides A, Pearson JV, Capten J, Keim P, Peterson A, Tersahita D, Arunmozhi B (2011). Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. Emerging Infectious Diseases 17:227-232.

Espedido BA, Steen JA, Ziochos H, Grimmond SM, Cooper MA, Gosbell IB, van Hal SJ, Jensen SO (2013). Whole Genome Sequence Analysis of the First Australian OXA-48-Producing Outbreak-

Associated *Klebsiella pneumoniae* Isolates: The Resistome and i*n Vivo* evolution. PLoS ONE 8:e59920.

Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TEA, Walker AS, Wilson DJ (2013a). Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. PLoS Comput Biol 9:e1003059.

Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TEA, Walker AS (2013b). Diverse sources of *C. difficile* infection identified on whole-genome sequencing. New England Journal of Medicine 369:1195-1205.

Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L (2012). A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open 2:e001124.

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nature Genetics 45:1183–1189.

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NPJ, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG (2001). Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. Proceedings of the National Academy of Sciences USA 98:182-187.

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from Multilocus Sequence Typing data. J Bacteriol 186:1518-1530.

Fields BS, Benson RF, Besser RE (2002). *Legionella* and Legionnaires' Disease: 25 Years of Investigation. Clinical Microbiology Reviews 15:506-526.

Fittipaldi N, Tyrrell GJ, Low DE, Martin I, Lin D, Hari KL, Musser JM (2013). Integrated whole-genome sequencing and temporospatial analysis of a continuing Group A *Streptococcus* epidemic. Emerg Microbes Infect 2:e13.

Fitzpatrick MA, Ozer EA, Hauser AR (2016). Utility of whole-genome sequencing in characterizing *Acinetobacter* epidemiology and analyzing hospital outbreaks. J Clin Microbiol 54:593-612.

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324:1557-1561.

Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE, Shepard CC, Brachman PS (1977). Legionnaires' disease: description of an epidemic of pneumonia. N Engl J Med 297:1189-1197.

Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, Peduzzi R, Harrison TJ (2005). Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. J Clin Microbiol 43:2047-2052.

Gaia V, Fry NK, Harrison TJ, Peduzzi R (2003). Sequence-based typing of *Legionella pneumophila* serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. J Clin Microbiol 41:2932-2939.

Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P (2011). Whole-genome squencing and social-network analysis of a tuberculosis outbreak. New England Journal of Medicine 364:730-739.

General Directorate of Epidemiology MoHM, Pan American Health Organization, World Health Organization, Public Health Agency of Canada, CDC (United States) (2009). Outbreak of swine-origin Influenza A (H1N1) virus infection–Mexico, March-April 2009. Morbidity and Mortality Weekly Report 58:467-470.

Ghai R, Martin-Cuadrado AB, Molto AG, Heredia IG, Cabrera R, Martin J, Verdu M, Deschamps P, Moreira D, Lopez-Garcia P, Mira A, Rodriguez-Valera F (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J 4:1154-1166.

Gilmour M, Graham M, Van Domselaar G, Tyler S, Kent H, Trout-Yakel K, Larios O, Allen V, Lee B, Nadon C (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. BMC Genomics 11:120.

Grad YH, Godfrey P, Cerquiera GC, Mariani-Kurkdjian P, Gouali M, Bingen E, Shea TP, Haas BJ, Griggs A, Young S, Zeng Q, Lipsitch M, Waldor MK, Weill FX, Wortman JR, Hanage WP (2013). Comparative genomics of recent shiga toxin-producing *Escherichia coli* O104:H4: short-term evolution of an emerging pathogen. mBio 4.

Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, Weinstock H, Parkhill J, Hanage WP, Bentley S, Lipsitch M (2014). Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. The Lancet Infectious Diseases 14:220-226.

Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, FitzGerald M, Godfrey P, Haas BJ, Murphy CI, Russ C (2012). Genomic epidemiology of the *Escherichia coli* O104: H4 outbreaks in Europe, 2011. Proceedings of the National Academy of Sciences USA 109:3065-3070.

Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RPA, McNerney R, Fine PEM, Parkhill J, Clark TG, Glynn JR (2015). Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. eLife Sciences 4:e05166.

Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. The Lancet Infectious Diseases 13:130-136.

Harris SR, Clarke IN, Seth-Smith HMB, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW, Lewis DA, Spratt BG, Unemo M, Persson K, Bjartling C, Brunham R, de Vries HJC, Morre SA, Speksnijder A, Bebear CM, Clerc M, de Barbeyrac B, Parkhill J, Thomson NR (2012). Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet 44:413-419.

Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD (2010). Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469-474.

Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, Haley BJ, Taviani E, Hine E, Su Q, Tallon LJ, Prosper JB, Furth K, Hoq MM, Li H, Fraser-Liggett CM, Cravioto A, Huq A, Ravel J, Cebula TA, Colwell RR (2012). Genomic diversity of 2010 Haitian cholera outbreak strains. PNAS 109:E2010-E2017.

He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D, Bamford KB, D'Arc S, Brazier J, Brown D, Coia JE, Douce G, Gerding D, Kim HJ, Koh TH, Kato H, Senoh M, Louie T, Michell S, Butt E, Peacock SJ, Brown NM, Riley T, Songer G, Wilcox M, Pirmohamed M, Kuijper E, Hawkey P, Wren BW, Dougan G, Parkhill J, Lawley TD (2013). Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet 45:109-113.

Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. mBio 2:e00157-11.

Holden MTG, Hauser H, Sanders M, Ngo TH, Cherevach I, Cronin A, Goodhead I, Mungall K, Quail MA, Price C, Rabbinowitsch E, Sharp S, Croucher NJ, Chieu TB, Thi Hoang Mai N, Diep TS, Chinh NT, Kehoe M, Leigh JA, Ward PN, Dowson CG, Whatmore AM, Chanter N, Iversen P, Gottschalk M, Slater JD, Smith HE, Spratt BG, Xu J, Ye C, Bentley S, Barrell BG, Schultsz C, Maskell DJ, Parkhill J (2009). Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. PLoS ONE 4:e6072.

Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlickova H, Coombs G, Kearns AM, Hill RLR, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramirez S, Feil EJ, Hudson LO, Enright MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M, Bentley SD, Nübel U (2013). A genomic portrait of the emergence, evolution and global spread of a methicillin resistant *Staphylococcus aureus* pandemic. Genome Research 23:653-664.

Holmes A, Allison L, Ward M, Dallman TJ, Clark R, Fawkes A, Murphy L, Hanson M (2015). Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. J Clin Microbiol 53:3565-3573.

Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, Choi SY, Kim SH, da Silveira WD, Pickard DJ, Farrar JJ, Parkhill J, Dougan G, Thomson NR (2012). *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet 44:1056-1059.

Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40:987-993.

Holt KE, Thieu Nga TV, Thanh DP, Vinh H, Kim DW, Vu Tra MP, Campbell JI, Hoang NVM, Vinh NT, Minh PV, Thuy CT, Nga TTT, Thompson C, Dung TTN, Nhu NTK, Vinh PV, Tuyet PTN, Phuc HL, Lien NTN, Phu BD, Ai NTT, Tien NM, Dong N, Parry CM, Hien TT, Farrar JJ, Parkhill J, Dougan G, Thomson NR, Baker S (2013). Tracking the establishment of local endemic populations of an

emergent enteric pathogen. Proceedings of the National Academy of Sciences USA 110:17522-17527.

Hornsey M, Loman N, Wareham DW, Ellington MJ, Pallen MJ, Turton JF, Underwood A, Gaulton T, Thomas CP, Doumith M (2011). Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy. J Antimicrob Chemother 66:1499-1503.

Ioerger TR, Feng Y, Chen X, Dobos KM, Victor TC, Streicher EM, Warren RM, van Pittius NCG, Helden PD, Sacchettini JC (2010). The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa. BMC Genomics 11:1.

Jolley KA, Hill DMC, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MCJ (2012). Resolution of a Meningococcal Disease Outbreak from Whole-Genome Sequence Data with Rapid Web-Based Analysis Methods. J Clin Microbiol 50:3046-3053.

Ju W, Cao G, Rump L, Strain E, Luo Y, Timme R, Allard M, Zhao S, Brown E, Meng J (2012). Phylogenetic analysis of non-O157 Shiga toxin-producing *Escherichia coli* by whole genome sequencing. J Clin Microbiol.

Kanamori H, Parobek CM, Weber DJ, van Duin D, Rutala WA, Cairns BA, Juliano JJ (2016). Next-generation sequencing and comparative analysis of sequential outbreaks caused by multidrug-resistant *Acinetobacter baumannii* at a large academic burn center. Antimicrob Agents Chemother 60:1249-1257.

Kato-Maeda M, Ho C, Passarelli B, Banaei N, Grinsdale J, Flores L, Anderson J, Murray M, Rose G, Kawamura LM, Pourmand N, Tariq MA, Gagneux S, Hopewell PC (2013). Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS ONE 8:e58235.

Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S (2011). *Helicobacter pylori* genome evolution during human infection. Proceedings of the National Academy of Sciences USA 108:5033-5038.

Kinnevey PM, Shore AC, Mac Aogáin M, Creamer E, Brennan GI, Humphreys H, Rogers TR, O'Connell B, Coleman DC (2016). Enhanced Tracking of Nosocomial Transmission of Endemic Sequence Type 22 Methicillin-Resistant *Staphylococcus aureus* Type IV Isolates among Patients and Environmental Sites by Use of Whole-Genome Sequencing. J Clin Microbiol 54:445-448.

Knetsch CW, Connor TR, Mutreja A, van Dorp SM, Sanders IM, Browne HP, Harris D, Lipman L, Keessen EC, Corver J (2014). Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011. EuroSurveillance 19:30-41.

Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. New England Journal of Medicine 366:2267-2275.

Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ (2013). Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. New England Journal of Medicine 369:290-292.

Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP (2016). Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol 54:333-342.

Lee RS, Radomski N, Proulx JF, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA (2015). Reemergence and amplification of tuberculosis in the Canadian Arctic. Journal of Infectious Diseases 211:1905-1914.

Lévesque S, Plante PL, Mendis N, Cantin P, Marchand G, Charest H, Raymond F, Huot C, Goupil-Sormany I, Desbiens F (2014). Genomic characterization of a large outbreak of *Legionella pneumophila* serogroup 1 strains in Quebec City, 2012. PLoS ONE 9:e103852.

Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboy D, Pallen MJ (2010). High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. Journal of Hospital Infection 75:37-41.

Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R (2011). Identification of a salmonellosis outbreak by means of molecular sequencing. New England Journal of Medicine 364:981-982.

Liu Q, Via LE, Luo T, Liang L, Liu X, Wu S, Shen Q, Wei W, Ruan X, Yuan X, Zhang G, Barry CE, Gao Q (2015). Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. Scientific reports 5:17507.

Loman NJ, Constantinidou C (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic *Escherichia coli* O104:H4. JAMA 309:1502-1510.

Loman NJ, Gladstone RA, Constantinidou C, Tocheva AS, Jefferies JM, Faust SN, OΓÇÖConnor L, Chan J, Pallen MJ, Clarke SC (2013). Clonal expansion within pneumococcal serotype 6C after use of seven-valent vaccine. PLoS ONE 8:e64731.

Loman NJ, Pallen MJ (2015). Twenty years of bacterial genome sequencing. Nat Rev Micro advance online publication.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences USA 95:3140-3145.

Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, Kusebauch U, Vigl EE, Malfertheiner P, Megraud F, OΓÇÖSullivan N, Cipollini G, Coia V, Samadelli M, Engstrand L, Linz B, Moritz RL, Grimm R, Krause J, Nebel A, Moodley Y, Rattei T, Zink A (2016). The 5300-year-old Helicobacter pylori genome of the Iceman. Science 351:162-165.

Mathers AJ, Peirano G, Pitout JDD (2015). The Role of Epidemic Resistance Plasmids and International High-Risk Clones in the Spread of Multidrug-Resistant Enterobacteriaceae. Clinical Microbiology Reviews 28:565-591.

McAdam P, vander broek C, Lindsay D, Ward M, Hanson M, Gillies M, Watson M, Stevens J, Edwards G, Fitzgerald R (2014). Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. Genome Biology 15:504.

McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, Bargawi HJA, Spratt BG, Bentley SD, Parkhill J, Enright MC, Holmes A, Girvan EK, Godfrey PA, Feldgarden M, Kearns AM, Rambaut A, Robinson DA, Fitzgerald JR (2012). Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. Proceedings of the National Academy of Sciences USA 109:9107-9112.

McDonnell J, DALLMAN T, Atkin S, Turbitt DA, Connor TR, Grant KA, Thomson NR, Jenkins C (2013). Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of *Shigella sonnei* in the UK. Epidemiology & Infection 141:2568-2575.

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R (2008). Microbiology in the post-genomic era. Nat Rev Micro 6:419-430.

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W (2011). Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104: H4 outbreak by rapid next generation sequencing technology. PLoS ONE 6:e22751.

Mendum T, Schuenemann V, Roffey S, Taylor G, Wu H, Singh P, Tucker K, Hinds J, Cole S, Kierzek A, Nieselt K, Krause J, Stewart G (2014). *Mycobacterium leprae* genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. BMC Genomics 15:270.

Metzker ML (2010). Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.

Mortimer PP (2003). Five postulates for resolving outbreaks of infectious disease. Journal of Medical Microbiology 52:447-451.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462-465.

Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, Tomasz A (2007). Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. Proceedings of the National Academy of Sciences USA 104:9451-9456.

Niemann S, K+Âser CU, Gagneux S, Plinke C, Homolka S, Bignell H, Carter RJ, Cheetham RK, Cox A, Gormley NA, Kokko-Gonzales P, Murray LJ, Rigatti R, Smith VP, Arends FPM, Cox HS, Smith G, Archer JAC (2009). Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS ONE 4:e7407.

Nübel U, Nachtnebel M, Falkenhorst G, Benzler J, Hecht J, Kube M, Bröcker F, Moelling K, Bührer C, Gastmeier P, Piening B, Behnke M, Dehnert M, Layer F, Witte W, Eckmanns T (2013). MRSA transmission on a neonatal intensive care unit: Epidemiological and genome-based phylogenetic analyses. PLoS ONE 8:e54898.

Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E (2012). Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. Nature Genetics 44:1215-1223.

Onori R, Gaiarsa S, Comandatore F, Pongolini S, Brisse S, Colombo A, Cassani G, Marone P, Grossi P, Minoja G, Bandi C, Sassera D, Toniolo A (2015). Tracking nosocomial *Klebsiella pneumoniae* infections and outbreaks by whole-genome analysis: Small-scale Italian scenario within a single hospital. J Clin Microbiol 53:2861-2868.

Parwati I, van Crevel R, van Soolingen D. (2010). Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect Dis 10:103–111.

Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH, Holden MTG, Morgan FJE, Ba X, Koop G, Harris SR, Maskell DJ, Peacock SJ, Herrtage ME, Parkhill J, Holmes MA (2015). Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nat Commun 6:6560.

Pérez-Lago L, Martínez Lirola M, Herranz M, Comas I, Bouza E, García-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. Clinical Microbiology and Infection 21:249.

Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E, García de Viedma D (2014). Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: Potential impact on the inference of tuberculosis transmission. Journal of Infectious Diseases 209:98-108.

Pinholt M, Larner-Svensson H, Littauer P, Moser CE, Pedersen M, Lemming LE, Ejlertsen T, Söndergaard TS, Holzknecht BJ, Justesen US, Dzajic E, Olsen SS, Nielsen JB, Worning P, Hammerum AM, Westh H, Jakobsen L (2015). Multiple hospital outbreaks of *vanA Enterococcus faecium* in Denmark, 2012−13, investigated by WGS, MLST and PFGE. J Antimicrob Chemother 70:2474-2482.

Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, Bowden R, Sarah Walker A, Peto TEA, Paul J, Llewelyn MJ (2014). Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* on an intensive care unit. Clin Infect Dis 58:609-618.

Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R, Xu J (2016). Population structure and minimum core genome typing of *Legionella pneumophila*. Scientific reports 6:21356.

Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, de Pinna E, Robinson E, Struthers K, Webber M, Catto A, Dallman T, Hawkey P, Loman N (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. Genome Biology 16:114.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ou+®draogo N, Afrough B, Bah A, Baum JHJ, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-S+ínchez +, Carter LL, Doerrbecker J, Enkirch T, Dorival IGa, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano Is, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D+, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner J,

Pollakis G, Hiscox JA, Matthews DA, Shea MKO, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, W+Âlfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Ke+»ta S, Rambaut A, Formenty P, G++nther S, Carroll MW (2016). Real-time, portable genome sequencing for Ebola surveillance. Nature advance online publication.

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren KG, Pedersen AG, Schubert M, Van Dam A, Kapel CMO, Nielsen HB, Brunak S, Avetisyan P, Epimakhov A, Khalyapin MV, Gnuni A, Kriiska A, Lasak I, Metspalu M, Moiseyev V, Gromov A, Pokutta D, Saag L, Varul L, Yepiskoposyan L, Sicheritz-Pontén T, Foley RA, Lahr MM, Nielsen R, Kristiansen K, Willerslev E (2015). Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. Cell 163:571-582.

Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch JD, Smith KL, Schupp JM, Solomon D, Keim P, Fraser CM (2002). Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 296:2028-2033.

Reuter S, Ellington MJ, Cartwright EP (2013a). Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Internal Medicine 173:1397-1404.

Reuter S, HARRISON TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Török ME (2013b). A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open 3.

Reuter S, Török ME, Holden MTG, Reynolds R, Raven KE, Blane B, Donker T, Bentley SD, Aanensen DM, Grundmann H, Feil EJ, Spratt BG, Parkhill J, Peacock SJ (2016). Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. Genome Research 26:263-270.

Rinder H, Mieskes KT, Löscher T (2001). Heteroresistance in *Mycobacterium tuberculosis*. The International Journal of Tuberculosis and Lung Disease 5:339-345.

Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S (2013). Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: A longitudinal molecular epidemiological study. PLoS Med 10:e1001387.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002). Genetic structure of human populations. Science 298:2381-2385.

Sánchez-Busó L, Comas I, Jorques G, González-Candelas F (2014). Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. Nature Genetics 46:1205-1211.

Sánchez-Busó L, Guiral S, Crespi S, Moya V, Camaró ML, Olmos P, Adrián F, Morera V, González Morán F, Vanaclocha H, González-Candelas F (2016). Genomic investigation of a legionellosis outbreak in a persistently colonized hotel. Frontiers in Microbiology 6:1556.

Sandegren L, Groenheit R, Koivula T, Ghebremichael S, Advani A, Castro E, Pennhag A, Hoffner S, Mazurek J, Pawlowski A, Kan B, Bruchfeld J, Melefors +, K+ñllenius G (2011). Genomic Stability over 9 Years of an Isoniazid Resistant *Mycobacterium tuberculosis* Outbreak Strain in Sweden. PLoS ONE 6:e16647.

Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann KH, Krahn I, Krause L, Krömeke H, Kruse O, Mussgnug JH, Neuweger H, Niehaus K, Pühler A, Runte KJ, Szczepanowski

R, Tauch A, Tilker A, Viehöver P, Goesmann A (2008a). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. Journal of Biotechnology 136:77-90.

Schlüter A, Krause L, Szczepanowski R, Goesmann A, Pühler A (2008b). Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. Journal of Biotechnology 136:65-76.

Schmid D, Allerberger F, Huhulescu S, Pietzka A, Amar C, Kleta S, Prager R, Preussel K, Aichinger E, Mellmann A (2014). Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011 – 2013. Clinical Microbiology and Infection 20:431-436.

Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jäger G, Bos KI, Herbig A, Economou C, Benjak A, Busso P, Nebel A, Boldsen JL, Kjellström A, Wu H, Stewart GR, Taylor GM, Bauer P, Lee OYC, Wu HHT, Minnikin DE, Besra GS, Tucker K, Roffey S, Sow SO, Cole ST, Nieselt K, Krause J (2013). Genome-wide comparison of medieval and modern *Mycobacterium leprae*. Science 341:179-183.

Schürch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, Smith NH, van Soolingen D (2010). The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. Infection, Genetics and Evolution 10:108-114.

Senn L, Clerc O, Zanetti G, Basset P, ProdΓÇÖhom G, Gordon NC, Sheppard AE, Crook DW, James R, Thorpe HA, Feil EJ, Blanc DS (2016). The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant Staphylococcus aureus. mBio 7.

Seth-Smith HMB, Harris SR, Skilton RJ, Radebe FM, Golparian D, Shipitsyna E, Duy PT, Scott P, Cutcliffe LT, O'Neill C, Parmar S, Pitt R, Baker S, Ison CA, Marsh P, Jalal H, Lewis DA, Unemo M, Clarke IN, Parkhill J, Thomson NR (2013). Whole-genome sequences of *Chlamydia trachomatis* directly from clinical samples without culture. Genome Research 23:855-866.

Shah MA, Mutreja A, Thimson N, Baker S, Parkhill J, Dougan G, Bokhari H, Wren BW (2014). Genomic epidemiology of *Vibrio cholerae* O1 associated with floods, Pakistan, 2010. Emerging Infectious Diseases 20:13-20.

Smit PW, Vasankari T, Aaltonen H, Haanperä M, Casali N, Marttila H, Marttila J, Ojanen P, Ruohola A, Ruutu P, Drobniewski F, Lyytikäinen O, Soini H (2015). Enhanced tuberculosis outbreak investigation using whole genome sequencing and IGRA. Eur Respir J 45:276-279.

Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program, Henderson DK, Palmore TN, Segre JA (2012). Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Science Translational Medicine 4:148ra116.

Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, Pallen MJ, Ala'Aldeen A (2013). Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. EuroSurveillance 18:20611.

Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer AM, Droz S, Butz C, Borrell S, Längle C, Feldmann J, Furrer H, Mordasini C, Helbling P, Rieder HL, 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. Journal of Infectious Diseases 211:1306-1316.

Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry CE, Mei J, Gao Q (2012). Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. Journal of Infectious Diseases 206:1724-1733.

Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, Boxrud D (2015). Characterization of foodborne outbreaks of *Salmonella enterica* serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. J Clin Microbiol 53:3334-3340.

Török ME, Peacock SJ (2012). Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory–pipe dream or reality? J Antimicrob Chemother 67:2307-2308.

Török ME, Reuter S, Bryant J, Köser CU, Stinchcombe SV, Nazareth B, Ellington MJ, Bentley SD, Smith GP, Parkhill J, Peacock SJ (2013). Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. J Clin Microbiol 51:611-614.

Underwood A, Jones G, Mentasti M, Fry N, Harrison T (2013a). Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiology 13:302.

Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J (2013b). Public health value of next-generation DNA sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. J Clin Microbiol 51:232-237.

Urwin R, Maiden MCJ (2003). Multi-locus sequence typing: a tool for global epidemiology. Trends in Microbiology 11:479-487.

Vos M, Didelot X (2009). A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199-208.

Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, Enk J, Birdsell DN, Kuch M, Lumibao C, Poinar D, Pearson T, Fourment M, Golding B, Riehm JM, Earn DJD, DeWitte S, Rouillard JM, Grupe G, Wiechmann I, Bliska JB, Keim PS, Scholz HC, Holmes EC, Poinar H (2014). *Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis. The Lancet Infectious Diseases 14:319-326.

Walker MJ, Beatson SA (2012). Outsmarting Outbreaks. Science 338:1161-1162.

Walker TM, Monk P, Grace Smith E, Peto TEA (2013a). Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. Clinical Microbiology and Infection 19:796-802.

Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE (2013b). Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. The Lancet Infectious Diseases 13:137-146.

Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP, Del Ojo Elias C, Jeffery KJ, Bowler ICJW, Laurenson IF, Barrett A, Drobniewski F, McCarthy ND, Anderson LF, Abubakar I, Thomas HL, Monk P, Smith EG, Walker

AS, Crook DW, Peto TEA, Conlon CP (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.

Ward MJ, Gibbons CL, McAdam PR, van Bunnik BAD, Girvan EK, Edwards GF, Fitzgerald JR, Woolhouse MEJ (2014). Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of *Staphylococcus aureus* clonal complex 398. Appl Environ Microbiol 80:7275-7282.

WHO (2014). Global tubreculosis report, 2014.

Witney AA, Gould KA, Pope CF, Bolt F, Stoker NG, Cubbon MD, Bradley CR, Fraise A, Breathnach AS, Butcher PD, Planche TD, Hinds J (2014). Genome sequencing and characterization of an XDR ST111 serotype O12 hospital outbreak strain of *Pseudomonas aeruginosa*. Clinical Microbiology and Infectionn/a.

Worby CJ, Lipsitch M, Hanage WP (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Comput Biol 10:e1003549.

Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, White RG, Cohen T, Cobelens FG, Wood R, Moore DAJ, Abubakar I (2016). The transmission of *Mycobacterium tuberculosis* in high burden settings. The Lancet Infectious Diseases 16:227-238.

Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG (2012). Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proceedings of the National Academy of Sciences USA 109:4550-4555.

Zakour NLB, Venturini C, Beatson SA, Walker MJ (2012). Analysis of a *Streptococcus pyogenes* puerperal sepsis cluster by use of whole-genome sequencing. J Clin Microbiol 50:2224-2228.

Zhou Y, Gao H, Mihindukulasuriya K, Rosa P, Wylie K, Vishnivetskaya T, Podar M, Warner B, Tarr P, Nelson D, Fortenberry JD, Holland M, Burr S, Shannon W, Sodergren E, Weinstock G (2013). Biogeography of the ecosystems of the healthy human body. Genome Biology 14:R1.

Table 1. A summary of published works analyzing complete genome sequences of bacterial pathogens for the study of outbreaks and transmission chains.

Pathogen	Genome size (Mb)	Sequencing strategy	References
Acinetobacter baumannii	4.11	Illumina HiSeq 2000 2x100 bp Illumina MiSeq 2x150 bp, 2x250 bp Roche 454 GS FLX PacBio	(Lewis et al, 2010;Hornsey et al, 2011;Kanamori et al, 2016) (Fitzpatrick et al, 2016)
Bacillus anthracis	5.2	Sanger	(Read et al, 2002)
Campylobacter jejuni	1.64	Illumina HiSeq 2000 76 bp	(Cody et al, 2013)
Chlamydia trachomatis	≈1.0	Illumina GA Illumina GAII PE 2x37 bp	(Harris et al, 2012;Seth-Smith et al, 2013)
Clostridium difficile	4.0	Illumina GAII/GAIIx 2x51/100-108bp Illumina HiSeq2000 2x100bp; 2x54/108/76bp	(Didelot et al, 2012a;Eyre et al, 2012;Eyre et al, 2013a;Eyre et al, 2013b;He et al, 2013;Knetsch et al, 2014)
Enterobacter cloacae	5.31	Illumina MiSeq	(Reuter et al, 2013a)
Enterococcus faecium	2.9	Illumina MiSeq	(Reuter et al, 2013a;Pinholt et al, 2015)
Escherichia coli	≈5.2	Roche 454 GS Junior Roche 454 Titanium Illumina MiSeq Illumina Solexa Illumina HiSeq2000 2x101 Ilumina GAIIx Ion Torrent PGM	(Mellmann et al, 2011;Brzuszkiewicz et al, 2011;Ahmed et al, 2012;Ju et al, 2012;Grad et al, 2012;Grad et al, 2013;Underwood et al, 2013b;Shah et al, 2014;Holmes et al, 2015)
Helicobacter pylori	1.5-1.7	Roche 454	(Kennemann et al, 2011)
Klebsiella pneumoniae	5.6	Illumina Hi Seq 2000 Illumina MiSeq platform Roche 454 Titanium XLR	(Snitkin et al, 2012;Espedido et al, 2013;Onori et al, 2015)

Legionella pneumophila	3.5	Illumina HiSeq 2x100 bp Illumina MiSeq 2x250 bp, 2x150bp SOLiD 5500XL SE 75bp	(Reuter et al, 2013a;Reuter et al, 2013b;Sánchez-Busó et al, 2014;Bartley et al, 2016)
Listeria monocytogenes	3	Roche 454 GS-FLX	(Gilmour et al, 2010;Schmid et al, 2014;Kwong et al, 2016)
Mycobacterium abscessus	5-5.2	Illumina HiSeq 2x75 bp	(Bryant et al, 2013b)
M. abscessus subsp. bolletii	≈ 5	Life Technologies SOLiD	(Davidson et al, 2013)
Mycobacterium canettii	≈ 4.5	HiSeq2000 MiSeq Illumina	(Blouin et al, 2014)
Mycobacterium tuberculosis	4.4	Illumina GAII PE 2x36bp; 2x50 Illumina GAIIx PE 2x76; 2x108 Illumina HiSeq PE 2x75 bp Illumina MiSeq 150 bp Roche 454 GS FLX 36bp	(Ioerger et al, 2010;Schürch et al, 2010;Gardy et al, 2011;Sandegren et al, 2011;Casali et al, 2012;Kato-Maeda et al, 2013;Bryant et al, 2013a;Roetzer et al, 2013;Köser et al, 2013;Török et al, 2013;Walker et al, 2013a;Walker et al, 2013b;Pérez-Lago et al, 2014;Coscollá et al, 2015)
Neisseria gonorrhoeae	2.1	Illumina HiSeq	(Grad et al, 2014)
Neisseria meningitidis	2.2	Illumina GAIIx PE 2x76 Illumina MiSeq	(Jolley et al, 2012;Reuter et al, 2013a;Bennett et al, 2012)
Pseudomonas aeruginosa	6.26	Ion Torrent	(Witney et al, 2014;Snyder et al, 2013)
Salmonella enterica	4.76	Illumina MiSeq Illumina HiSeq 2500 MinION Roche 454	(Holt et al, 2008;Lienau et al, 2011;Quick et al, 2015;Allard et al, 2013;Cao et al, 2013;Allard et al, 2012;Taylor et al, 2015;Bekal et al, 2016)
Salmonella Typhimurium	4.7	Illumina GA II system	(Okoro et al, 2012)
Shigella sonnei	5.06	Illumina GAII PE 2x54 bp Illumina MiSeq Illumina HiSeq2000	(Holt et al, 2012;Holt et al, 2013;McDonnell et al, 2013)

Staphylococcus aureus	2.8-3	Illumina MiSeq PE 2x150 bp Illumina GAIIx PE Illumina GAII SE 150 bp Illumina HiSeq2000 Roche 454 GS FLX	(Harris et al, 2010;Eyre et al, 2012;McAdam et al, 2012;Young et al, 2012;Köser et al, 2012;Holden et al, 2013;Nübel et al, 2013;Harris et al, 2013;Price et al, 2014;Azarian et al, 2015;Paterson et al, 2015;Senn et al, 2016;Kinnevey et al, 2016;Reuter et al, 2016)
Streptococcus pneumoniae	1.98 - 2.19	Illumina HiSeq 2000 2x75 bp Illumina PE 2x54bp Roche 454	(Croucher et al, 2011;Loman et al, 2013;Croucher et al, 2013;Chewapreecha et al, 2014)
Streptococcus pyogenes	1.85	Illumina HiSeq 2000 Illumina GA1s Roche 454	(Zakour et al, 2012;Fittipaldi et al, 2013)
Streptococcus suis	2.15	Roche 454 / GS 20 Solexa	(Holden et al, 2009)
Vibrio cholerae	≈ 4	Illumina HiSeq Illumina GAI Illumina GAIIx PacBio-RS Roche 454 GS FLX	(Mutreja et al, 2011;Hendriksen et al, 2011;Chin et al, 2011;Hasan et al, 2012;Shah et al, 2014;Schmid et al, 2014;Devault et al, 2014;Wagner et al, 2014;Knetsch et al, 2014)
Yersinia pestis	5.46	Illumina	(Cui et al, 2013;Wagner et al, 2014)

Figure 1. General workflow followed during high throughput sequencing (Metzker, 2010).

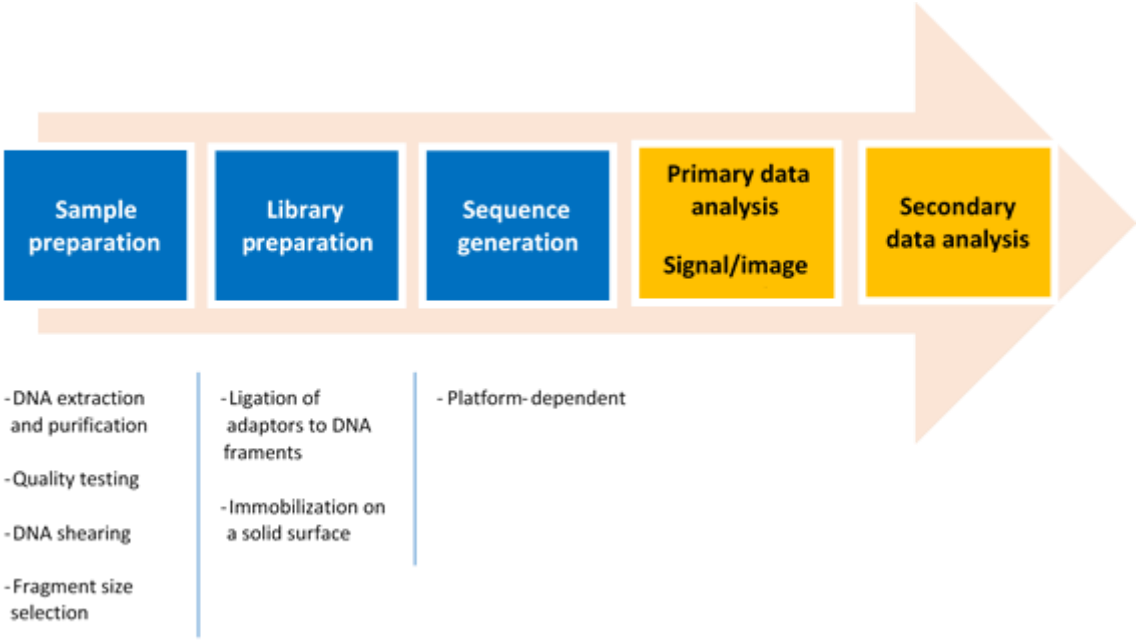

Figure 2. Estimates of evolutionary rate for different bacterial species and its relationship to the time elapsed since the most recent common ancestor of the isolates used to determine the rate. Sources of data: *H. pylori* (Kennemann et al, 2011), *C. difficile* (Didelot et al, 2012b), *Sh. sonnei* (Holt et al, 2012), *Y. pestis* rasmussen2015a (Rasmussen et al, 2015), *S. aureus* (Harris et al, 2010;Ward et al, 2014), *S. pneumoniae* (Croucher et al, 2011), *L. pneumophila* (Sánchez-Busó et al, 2014), *M. tuberculosis* (Comas et al, 2013), *M. leprae* (Schuenemann et al, 2013), *S. enterica* (Zhou et al, 2013).