Review

Bacterial genomic epidemiology, from local outbreak characterization to species-history reconstruction

Stefano Gaiarsa[1,2], Leone De Marco[3,4], Francesco Comandatore[2,4], Piero Marone[1], Claudio Bandi[2], Davide Sassera[4]

[1]Struttura Complessa di Microbiologia e Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy, [2]Dipartimento di Scienze Veterinarie e Sanità Pubblica, Università degli Studi di Milano, Milan, Italy, [3]Scuola di Bioscienza e Medicina Veterinaria, Università di Camerino, Camerino, Italy, [4]Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy

Bacteriology has embraced the next-generation sequencing revolution, swiftly moving from the time of single genome sequencing to the age of genomic epidemiology. Hundreds and now even thousands of genomes are being sequenced for single bacterial species, allowing unprecedented levels of resolution and insight in the evolution and epidemic diffusion of the main bacterial pathogens. Here, we present a review of some of the most recent and groundbreaking studies in this field.

Keywords: Genomic epidemiology, Bacteria, Next-generation sequencing, *Mycobacterium tuberculosis*, *Salmonella* Typhi, Carbapenem-resistant Enterobacteriaceae, *Staphylococcus aureus*, *Streptococcus pneumoniae*

Introduction

It seems that lately all scientific articles presenting results based on next-generation sequencing start with slight variations of the same formula 'In recent years the advent of novel sequencing technologies has revolutionized the field of …'. This uniformity can teach us a couple of lessons. First of all that scientists do not apply their unquestionable creativity to the writing of introductions, but more importantly that maybe we are actually really facing a scientific revolution. These technologies allow to obtain unprecedented levels of resolution and standardization in genomic data at affordable prices and turnaround times.

Many researchers quickly understood the power of these novel sequencing approaches, generating wealth of data for a number of different biological systems, and designing novel methods to exploit these information. Among them, many bacteriologists fully embraced the revolution, understanding that the generation of high numbers of genomes from a single species, strain or clonal group would allow to reconstruct the history of a bacterium in time and space, to trace its movements and relevant evolutionary events, and to understand the success of specific strains. This genomic epidemiology approach has been applied to a plethora of bacteria, first and foremost to pathogens, with the final goal of obtaining novel strategies to effectively deal with diseases. In this review, we present a small collection of some of the most novel and groundbreaking results obtained in this field, to provide the reader with basic knowledge of the new advances and hopefully to inspire others to pursue innovative lines of research.

Correspondence to: Davide Sassera, Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Italy. Email: davide.sassera@unipv.it

Local Studies

Third-generation sequencing to tackle the plasmid issue

Multidrug-resistant (MDR) bacteria are currently considered a health problem of primary importance both in Europe and USA.[1] Enterobacteriaceae are among the most common MDR bacteria involved in nosocomial infection worldwide. These bacteria, *Escherichia coli* and *Klebsiella pneumoniae, in primis* are often commensals in healthy subjects, as part of the gut flora composition. When colonizing subjects with depressed or weakened immune system, they can turn into dangerous pathogens. These bacteria are able to rapidly develop antibiotic resistance, both by acquiring resistant factors from other bacteria and by developing favourable chromosome mutations. Among the

DOI 10.1080/20477724.2015.1103503

most feared are carbapenem-resistant Enterobacteriaceae (CRE), capable of surviving even treatments with the most recently discovered molecules.

Due to antimicrobial resistance and opportunistic pathogenicity, many of the fatal infections caused by CRE occur within the hospital intensive care units, hosting patients in precarious health conditions, where antibiotic pressure is continuous, often resulting in nosocomial outbreaks. Indeed, during 2013, MDR Enterobacteriaceae were responsible for over 9000 nosocomial infections in the USA, causing over 600 deaths.[2] Multiple genomic epidemiology studies have been performed on Enterobacteriaceae global evolution, in particular focusing on the most common antibiotic-resistant clones.[3–6] Additionally, the application of genomic approaches has allowed to characterize nosocomial outbreaks in detail, identifying dangerous clones and detecting transmission patterns.[7–9] These studies used the 'standard' next-generation technologies for genome sequencing (Illumina and 454). These now established technologies, sometimes now referred to as 'second-generation technologies',[10] output very large sets of short sequences, suitable to obtain draft genome assemblies. Such methods, however, do not allow to fully characterize the genomic structure, that is to completely detect genomic rearrangements and fully reconstruct complete plasmid sequences. In order to overcome these issues, we can turn to the latest advancements in DNA sequencing, the so called third-generation sequencing platforms.[10] Among them is the single molecule real time (SMRT) sequencing technology which allows to obtain very long sequences (> 10,000 nt)[11] and thus to perform genome assemblies that easily result in the full reconstruction of 'closed' chromosomes and plasmids.

Recently, Conlan and co-workers[12] exploited this technology to bring local genomic epidemiology to a higher level. Twenty strains of CRE were sampled in the NIH Clinical Center in the time span between 2011 and 2013 both from patients and from the hospital environment. Initial sequencing, performed with Illumina or 454 technologies was, as expected, not precise enough to characterize the genomic content on a structural basis. For this reason, genomes were sequenced a second time using SMRT technology and polished on two levels. Illumina MiSeq reads allowed to ensure the highest precision in base content, while OpGen physical maps[13] gave fundamental information on chromosome and plasmid structures. The effort allowed to obtain complete genomes at a very high level of accuracy, which were used together with epidemiological information to reconstruct the transmission routes of the pathogens. In five cases, the aid of complete genomes and plasmids was essential to detect the exact pattern of transmission. Among the 20 isolates analysed, four showed novel genomic features, never described before. A *Klebsiella oxytoca* strain carrying a plasmid with two copies of the bla_{KPC} gene, an isolate of *K. pneumoniae* with three plasmids each carrying a copy of the bla_{KPC} gene and two isolates with a chromosomal copy of the resistance gene. Is such an high-percentage of novel features (4/20) surprising? Are the bacteria isolated at the NIH Clinical Center so unique? The more parsimonious explanation is actually that these features are more common than we think, and that we had never seen them before because the quality of our sequencing was previously insufficient.

The approach by Conlan and colleagues also allowed to detect the movement of KPC plasmids between isolates of different species through the environment. A p55-like plasmid of *K. pneumoniae* isolated from a patient was the same found in *Citrobacter freundii* and *Enterobacter cloacae* sampled from the sink in the patient's room, differentiating only for two adaptive changes in the entire sequence. This suggested that the transmission happened in the sink, as a consequence of a transient presence of *K. pneumoniae* in the local biofilm. This finding underlines, once more, the importance of environmental controls in the hospital wards and suggests specific inspections that should be included in the safety protocols. Lastly, thanks to the finesse of the sequencing, it was possible to analyse a case of co-colonization in a 2011 patient who was found to carry both a carbapenem-resistant *K. pneumoniae* and a *E. cloacae.* Since the patient resulted not colonized at admission, and the two pathogens present two different copies of KPC plasmids, the authors concluded that there was no plasmid transfer and that both resistant strains were circulating in the institution prior of the patient's arrival.

Hubs of gene flow in Streptococcus pneumoniae

Streptococcus pneumoniae is a gram-positive bacterium that resides, usually as an innocuous commensal, in the nasopharynx of healthy carriers, with prevalence varying from 5 to 90%.[14,15] When *S. pneumoniae* colonizes immunocompromised individuals, children or elderly, it can act as an opportunistic pathogen, leading not only to pneumonia, but also to a variety of other important diseases, such as meningitis and febrile bacteraemia.[16] Multiple vaccines are currently available. For example, conjugate vaccine PCV7 was introduced in the United States in 2000, rapidly proving its effectiveness. Indeed, by 2003, the occurrence of infant pneumococcal disease in Massachusetts was 69% lower.[17] Nevertheless, *S. pneumoniae* remains a leading infectious disease worldwide, with higher impact in developing countries. Multiple studies have used genomics to study this important pathogen, starting with the first complete genome sequence presented in 2001.[18] Subsequent comparative works allowed to discover that this species exhibits a high frequency of genomic recombination, which in turn favours the rapid acquisition of novel genetic features. This is often followed by rapid diffusion of those characteristics that give a distinct selective advantage, such as traits conferring resistance to antibiotics and vaccines.[19–22] Such behaviour

was described even within the course of a single-patient chronic paediatric polyclonal infection.[23]

Recently a study presented the sequencing of over 3000 genomes of *S. pneumoniae* isolated during the course of 4 years from the population of a single refugee camp in Thailand.[24] This is, to our knowledge, the largest study of bacterial genomic epidemiology to date, structured so that it allows an unprecedented level of sampling density, providing novel insight in the genomic evolution of *S. pneumoniae* and of bacterial populations in general. The authors used the Bayesian approach BAPS[25,26] to investigate the structure of the bacterial population of the refugee camp, identifying 33 clusters, that could be divided in 183 secondary, mostly clonal, subclusters. An analysis focused on the capsule biosynthesis locus and detected a strong presence of non-typeable (NT) isolates, lacking the capsule, but also the impressive number of 191 'plausible capsule switching events', with numerous of these recombinations causing switches between the capsulated and non-capsulated states.

While nucleotide substitution rate did not differ between clusters, rate of recombination was variable. This ratio, intended as the ratio between recombination and mutation events (*r/m*), resulted to be very different among population clusters and significantly higher in NT isolates. The more striking result of this analysis was, however, the consistently higher rate of recombination of six genomic loci, specifically antigens and antibiotic-resistance genes. Among these 'recombination hotspots' were genes providing resistance to beta-lactams and cotrimoxazole. Interestingly, the recombination histories of genes providing resistance to the two antibiotics were different, and the authors found them to match the variable use of beta-lactams and cotrimoxazole, respectively, increasing and decreasing during the time of the study within the refugee camp area.

The extreme genome density of this study allowed also to tackle one important issue of recombination studies, the detection of the donor isolates. Indeed, the authors identified 443 blocks that were identical between the recipient and the donor, nine of which were the results of recombination between nine donors and one single recipient. These blocks were not uniformly distributed in the population, and the NT isolates resulted to be not only good recipients, but good donors as well. In summary, this study reported the presence of 'recombination hotspots' that can easily move among isolates, but also the presence of specific lineages that can act as 'hubs of gene flow'. Since these lineages are not necessarily those that present higher virulence, such as NT isolates in *S. pneumoniae*, these results provide novel insights that can help to change the way we look at the dynamics of bacterial populations.

Global Trends

The explosion of Salmonella Typhi H58

Salmonella enterica is a widely studied pathogenic agent, of such importance that has caused multiple health and economic crises worldwide.[27] The three *S. enterica* clusters of greater importance for human health are serovar Typhi (or *S.* Typhi), serovar Typhimurium (or *S.* Typhimurium) and serovar Enteritidis (or *S.* Enteritidis). *S.* Typhi is the causative agent of typhoid fever, a disease particularly diffused in Asia, and endemic in the Indian subcontinent.[28] It was estimated that, just during 2010, typhoid fever affected 26.9 million people worldwide, with fatality rates ranging from 1 to 30%.[29] *S.* Typhi is able to infect human blood and intestine, producing an array of symptoms including nausea, vomiting, fever and death. A portion of colonized people usually remain asymptomatic for long periods of time (up to years), shedding the bacterium into their stool and urine,[27] with the effect of sustaining the pathogen transmission.

Blantyre is a ~1.3 million people district, localized in the south of Malawi. In this area, between 1998 and 2010, several cases of bloodstream infection (BSI) caused by nontyphoidal serovars of *Salmonella* were reported, while the typhoidal *S.* Typhi serovar was rare.[30] Starting from 2011, an increase in the number of BSI caused by MDR *S.* Typhi has been reported. In particular, the *S.* Typhi lineage H58 resulted to be predominant in the infected population. Feasey and colleagues[30] used a genomic epidemiology approach to reconstruct the origin of the emergence of *S.* Typhi H58 in this area: they collected epidemiological data from the Blantyre district (covering the period 1998–2011) and sequenced the genome of 112 *S.* Typhi strains, isolated there from 2004 to 2011. Whole-genome analysis was performed using a genome mapping approach, and the obtained SNPs were subjected to maximum-likelihood phylogenetic analysis. Merging the information from the resulting tree and the epidemiological data, the authors were able to reconstruct that the increase of typhoid fever reported from 2011 was due to the rapid diffusion of a monophyletic *S.* Typhi lineage, the aforementioned H58-haplotype. H58 resulted to be associated with MDR with a much higher frequency (89.3%) than the other circulating *S.* Typhi types (21.4%).

Strong phylo-geographical clusters were described within the H58 lineage, indicating it to be endemic in the areas included in the study. H58 isolates, collected in the same areas, resulted to be clustered on the phylogenetic tree, independently of the date of isolation. This geographical clusterization is consistent with the existence of reservoirs. Furthermore, genomes of the H58 strain result to be very conserved within the lineages, in contrast with the other *S.* Typhi monophyla. Indeed, the H58 tree branch lengths are shorter than the other *S.* Typhi lineages. These data can be explained hypothesizing that a strong purifying selective pressure affects the H58 lineage, and/or that frequent genomic recombinations occurred among the H58 strains.

Due to the undeniable importance of this haplotype, a second study was performed to describe its emergence at the global level. Wong and colleagues[31] considered the impressive collection of 1,832 *S.* Typhi isolates collected

in the period 1905–2014, from 63 countries spanning 6 continents. Whilst the most ancient isolate included in the study was collected in 1905, the first *S.* Typhi H58 isolate was from 1992, indicating a very recent origin of this lineage. Since 1992, the H58 haplotype represents ~40% of all the isolates collected each year, a remarkable explosive diffusion. Indeed, H58 genomes differ by a mean of only six SNPs, with 93% of them having less than five isolate-specific SNPs. These data show that H58 isolates are very closely related, consistent with the hypothesis of an impressive recent clonal expansion. It must be noted that the number of isolates obtained before 1992 is limited ($n = 50$), and this may skew the perception of the H58 diffusion, nevertheless the result is remarkable. The authors then inferred the date of the H58 origin to be between 1985 and 1992. After 1993, they observed a drastic increase of the H58 effective population size. Furthermore, on the basis of the phylogenetic reconstruction, the authors traced the major geographical transfers of the *S.* Typhi H58 haplotype: the strain originated in India, and through independent events reached Southeast Asia, Fuji, Western Asia, East Africa and Malawi. In Africa, it then invaded Malawi a second time through East Africa and then diffused from Malawi to South Africa.

Novel insights into the genomic evolution of Staphylococcus aureus

Staphylococcus aureus is among the most important antibiotic-resistant pathogens worldwide. Methicillin-resistant strains (MRSA), in particular, are spread in all continents and can be up to 70% of all *S. aureus* isolates in the most affected countries.[32] The first report of MRSA was an hospital-acquired infection in 1960,[33] but the pathogen has since then developed endemic status and can be transmitted outside of the nosocomial environment (the first cases were reported in the mid-1990).[34] The terms HA-MRSA (health care-associated MRSA) and CA-MRSA (community-associated MRSA) reflect this distinction. And if this was not enough, LA-MRSA (livestock-associated) is the zoonotic variant, which is common in farms.[35] Resistance to methicillin is encoded in the staphylococcal cassette chromosome *mec* (SCC*mec*) which contains the *mecA* gene. Several variants of the cassette have been discovered and found to be able to transmit and move between strains. The spread of MRSA strains has been the focus of a number of high-profile studies that used genomic approaches to investigate their diffusion and evolution, starting from the pioneering study of Harris,[36] already discussed in previous reviews (e.g. [37]).

Holden and colleagues[38] used genomics to reconstruct the evolution of EMRSA-15, a strain belonging to sequence type 22, which is considered the most rapidly spreading and tenacious *S. aureus* in Europe, currently invading other continents. Genomes were obtained from 193 ST22 strains of *S. aureus* isolated from 1990 to 2009 and a SNP-based phylogeny was reconstructed. A molecular clock analysis allowed to distinguish and date different clades with variable virulence levels, corresponding to different stages in the epidemic diffusion. Genomic variability among the clades was analysed both at the SNP level and the gene content level, with the aim of correlating genomic changes with fitness and virulence.

The study concluded that sequence type 22-A (ST22-A) was the first of this lineage to obtain resistance to methicillin, from the primitive community-associated methicillin-sensitive ST22, and dated this event to before 1977. This led to a health care-associated MRSA epidemic that spread in England in the 1980s (ST22-A1). In the mid 1980s, a sublineage acquired resistance to fluoroquinolones, and EMRSA-15 (also called ST22-A2) was born. The authors performed a bayesian analysis and detected a considerable difference in population size between ST22-A1 and ST22-A2, putative consequence of a fitness boost which caused the worldwide diffusion of the latter strain. Lastly, genomic traits were correlated with antimicrobial resistance profiles, thus highlighting the potential of genome sequencing as a diagnostic tool. Appearance on the tree of genetic variants responsible for antimicrobial resistance was found to agree with the variations of antibiotic prescription policy-making in the different regions during the years. Indeed, EMRSA-15 spread through the UK when fluoroquinolones where highly used, while a subsequent spread in Germany was a consequence of the development of yet another resistance, to clindamycin, which was heavily used in that country. Recently, other similar works reconstructed and dated the origin of other MRSA epidemic clones. Genomic variants were mapped on the trees and correlated with phenotypic changes. For example, Planet and coworkers[39] worked on USA300 and USA300-LV clones, while Stinear and colleagues[40] on CA-MRSA ST93. Stegger *et al.*[41] studied CC80 and Baines *et al.*[42] focused their efforts on HA-MRSA ST239.

In addition to these phylogeny-based works, the genomics of *S. aureus* has been used to investigate variations in genome structure. Indeed recombinations, transmission of plasmids and pathogenicity islands represent big adaptive steps in the history of MRSA, as they do for CRE and *S. pneumoniae*. Recently, Méric and coworkers[43] studied the evolution and genomic flow between two species that share the same niche: *S. aureus* and *Staphylococcus epidermidis*. These species are indeed both common commensals on the human skin and in the nasal pharynx. The authors selected and sequenced 324 isolates from archives and databases, in order to represent the global diversity of the two species, choosing among different genomic variants, location and sources of isolation. Shared genes and alleles were searched between each pair of isolates, and the two species resulted to share a maximum of nine core genome alleles between them, thus suggesting that genomic recombination is very rare between individuals of the two species. On the contrary, mobile elements were highly shared, in particular genes associated with

the SaPIn1 pathogenicity island, metal detoxification and the methicillin-resistance island *SCCmec*. The authors use these interesting results to discuss the concept of evolution in relation with the host as a niche, that is two strains or species that share the same host, also share the same selective pressure. Recombination can be driven by direct contact between donor and receiver, but also by evolution of niches, as genomic material can be shared in these enclosed environments.

In this review, we report multiple examples of how exchange of genomic material can involve the accessory parts of the genomes, but also homologous recombinations in core genome loci. The latter is a common thread in global genomic epidemiology as it is commonly found in most bacterial species involved in nosocomial infections. This was recently pointed out by Croucher and Klugman[44] who observed that large recombinations (even bigger than one megabase) seem to be an evolutive weapon that pathogens use to rapidly gain fitness and survive in the hospital environment. The two authors compare recombined bacteria to the 'hopeful monsters' of the Cambrian period, citing the use by Stephen Gould of the term introduced earlier by Richard Goldschmidt.[45] The presence of recombinations should be taken into account because they need to be removed from the genomic alignment in order to obtain resolved and correct phylogenies, which represent the real evolutive history of the pathogen.

Historical Perspectives

Mycobacterium tuberculosis through history

Mycobacterium tuberculosis is an obligate aerobic pathogenic bacterium and the causative agent of tuberculosis (TB).[46] The presence of mycolitic acids in *M. tuberculosis* coating confers the bacterium resistance to weak disinfectants and dehydration and prevents the effective activity of hydrophobic antibiotics. Additionally, it allows the bacterium to grow inside of macrophages, effectively hiding it from the host's immune system.[47] All these characteristics contribute to the ease with which it is transmitted, despite its extremely slow replication time compared to other bacteria. Tuberculosis is a disease that accompanied human populations since antiquity, it has been prevalent worldwide and, if left untreated, causes death in 50% of cases. In the last two centuries, progress has been made both in diagnosis and treatment with the advent of screening programs, antibiotics and vaccines, relegating the emergency to third world countries. Nevertheless, deaths are increasing after an almost 40 years decline,[46] and the emergence of multiple antibiotic-resistant strains[48] makes *M. tuberculosis* one of the most important re-emerging bacterial pathogens to date, and the leading bacterial killer worldwide with 1.3 million deaths a year. The so called *M. tuberculosis* Beijing family is a heterogeneous group of strains, among which hypervirulent subtypes stand out, equipped with multiple antibiotic resistances and the ability to cause disease outbreaks.[49] The whole family, considered the predominant genotype in East Asia and still currently spreading, can be accounted for more than a quarter of the total tuberculosis cases worldwide. Despite previous epidemiology studies showed high genetic similarity even among strains isolated in different geographic areas,[49] pathobiological characteristics appeared heterogeneous.[50] The increasing availability of standard genotyping leads to the identification of several Beijing sublineages.[51] This approach, however, proved itself limited for fully understanding the diversity of this family due to the insufficient amount of nucleotide variation detected by this technique. Once again, genomic epidemiology can come to our rescue.

Given the relevance of this family for public health globally, several studies focused on reconstructing the origin and spread of *M. tuberculosis* Beijing strains.[52–56] Merker *et al*.[57] focused on the biogeographical structure of strains belonging to the Beijing family, with a in-depth analysis on the association between sublineages and antibiotic resistance. The authors assembled a huge data set, comprising almost five thousand genotyped isolates plus 110 whole-genome sequences, the biggest and broadest collection of Beijing family strains to date, both in terms of sheer size and variety of geographical origins. Initially, a minimum-spanning tree (MStree) was constructed using genotyping data, grouping the genomic diversity into 6 major clonal complexes (CCs) and 3 distant branches which were collectively designated as basal sublineage 7 (BL7). CC1 through CC5 were classified as typical/modern Beijing while CC6 and BL7 comprised typical ancestral Beijing variants. The shape of the Mstree and the mean allelic richness confirmed this hypothesis, suggesting that CC1, CC2 and CC5 are in a state of population expansion, while CC6 and BL7 are more ancient and/or in a situation of milder expansion.

The authors then used the information for all the five thousand genotyped isolates to estimate past expansions and time to the most recent common ancestor (TMRCA). CC6 and BL7 were confirmed once again as the oldest sublineages with, respectively, a TMRCA of ~6,000 and 5,000 years, while CC5 resulted the youngest with a TMRCA of ~1500 years. For all sublineages, an estimate of the time elapsed since the beginning of the latest expansion was computed. This analysis showed a much recent timeframe, roughly 200 years ago, for CC1, CC2 and CC5, compared to the estimate detected for the more ancient lineanges, CC6 and BL7, which dated back to the middle age. Once again, genome information, available for 110 isolates, made it possible to obtain a more sensitive estimate of population changes in the recent past, by means of a Bayesian skyline plot. Two significant population growth phases were detected in conjunction with the Industrial Revolution and the First World War. The only decrease in population size was observed contemporary with the spreading of anti-tuberculosis drug usage, while

a slight expansion coincides with the advent of the HIV epidemics and the first tuberculosis outbreaks in the former Soviet Union and the United States in the 1990s. A subset of roughly one thousand clinical isolates with known drug resistance profiles was analysed to shed light on possible association between the identified CCs and antibiotic resistance, resulting in CC2 having the highest proportions of MDR strains. It is important to note that CC1, while having a similar proportion of MDR strains to the ancient lineages, showed a high clustering rate (95%) meaning that almost all isolates shared a single resistant haplotype, in contrast with CC6 and BL7 (42% and 57%, respectively). Additionally, strains from CC1 (central Asian outbreak) and CC2 (Russian-European outbreak) showed a higher similarity between them than with other clonal complexes, supporting the MDR outbreak hypothesis and a recent specific expansion of these two clonal complexes.

It is interesting to note that *M. tuberculosis* population growths are closely related to historical events and human migrations. Genotype data allowed to estimate the start of the last expansion events for the recent sublineages to 200–250 years ago, roughly around the time of the industrial revolution and matching known episodes of Chinese immigration towards Pacific islands, Americas and Russia. It is both interesting and not surprising that Beijing family strains experienced increases in population size in conjunction with the Industrial revolution and with the First World War. This is consistent both with Chinese immigration episodes, as noted above, and with the deprivations caused by war conditions and the co-mortality due to the influenza pandemics of that time.[58] The only decrease in population size detected coincides with the advent of antibiotics and mass vaccinations around the 1960s. It has to be kept in mind that the expansion of the two sublineages more associated with antibiotic resistance, CC1 and CC2, predates this event. This indicates that MDR is not the reason of the success of these clonal complexes but just the consequence of public health policies implemented on an already growing bacterial population. Finally, the most recent increasing trend is consistent with the onset of the global HIV epidemic and follows it closely.

The strange case of the amphibious Mycobacterium

Albeit recent studies such as the one described above are shedding light in the history of *M. tuberculosis*, there are still plenty of dark patches that need to be illuminated. It is clear that the co-evolution of the bacterium with humans started with the shift from the hunter/gatherer behaviour to the onset of agriculture and animal husbandry, especially cattle.[59] Until recent years, the most accredited theory had a zoonotic transfer of *Mycobacterium bovis* following animal domestication during the Neolithic.[60] However, recent comparative genomic analyses lean towards the opposite theory. That is, strains adapted to bovines and other animals may have originated from human strains.[61,62] For what concerns the Americas, given that strains currently present in the area are closely related to the European ones, the consensus is that the pathogen was brought by colonizers and settlers after the Columbian discovery.[63] This, however, is not consistent with several evidences of skeletal samples dated before 1492 with obvious signs of the disease. If tubercolosis was not brought on caravels, who or what carried it to the New World? And how can we explain the genomic similarity between American and European strains of *M. tuberculosis*? Bos *et al*.[64] try to give us an answer.

Progresses in protocols for isolation of ancient DNA made it possible to collect three *M. tubercolosis* genomes from skeletal samples with signs of tuberculosis infection collected in Peru and dated back to roughly one thousand years ago. They completed a the data set by adding 259 modern *M. tuberculosis* complex (MTBC) genomes, 14 animal isolates and an additional ancient genome originated from an eighteen century Hungarian mummy. An alignment of 22,480 variable positions was the input for a phylogenetic analysis. The resulting tree came with a surprise: the ancient Peruvian samples did not cluster with other human isolates; they were closer to the animal strains, in particular to the modern *Mycobacterium pinnipedii* sample. As the latin suggests, *M. pinnipedii* has been isolated from seals and sea lions. Bayesian dating analysis, using radiocarbon dates as tip calibration, was implemented for dating the most recent common ancestor (MRCA). Using a relaxed molecular clock model, the MRCA was dated between 4000 and 4500 years ago. Since the Bering land bridge closed some 15,000 years ago, 10000 years before the estimated time of the MRCA, the researchers discarded a human migration hypothesis for the appearance of TB in the New World. The remaining, unexpected hypothesis is the amphibious one: seals contracted the disease on the coasts of Africa and carried it to South America where populations living in the seaside contracted by exploiting the marine mammals. This is consistent with similar cases in literature for other pathogens.[65] The later eradication of this strain and the almost total substitution with European-like lineages could be accounted by a spread-after-contact of the latter following favourable conditions.[66]

Conclusions: From Epidemiology to Diagnostics

The examples presented here testify how the use of genome sequencing in bacteriology for epidemiological purposes is now widespread. Favoured by the wealth of data that are being generated and by the continued advances in sequencing technologies, the next step will be to branch into microbiological diagnostics.[67–69] Efforts are being made in multiple directions, two of the most promising being direct sequencing from clinical samples and the use of genomic data to predict phenotypic characteristics.

The utility of direct sequencing of clinical samples for diagnosis purposes is obvious, and particularly important

for bacteria that are difficult, slow or impossible to culture. A possible approach, when investigating a single pathogen, is the use of specific baits to sequence the target bacterium starting from a mixed clinical sample, an approach that was used for example to detect *M. tubercolosis* from sputum.[70] In many clinical cases, it is, however, impossible to know a priori which bacterium is causing a pathological state. In these situations, a whole metagenomic sequencing approach could allow to identify the causative agent/s quickly and without bias. Ad-hoc bioinformatic methods are being developed to tackle the problematic issue of analysing the complex metagenomes that can be obtained from clinical samples, with the goal of correctly sorting and identifying bacterial populations.[71–73]

Some phenotypic characteristics of a pathogen can be readily inferred from its genome sequence. This is the case for example of antibiotic resistance traits that are determined by the acquisition of one single gene. Many other phenotypes are, however, multi-factorial, making such correlations more complex. These difficulties have not discouraged pioneering projects that use genome-wide association studies to link genotype to phenotype, in an effort to obtain diagnostic informations from the now quick and cheap whole-genome sequences. Laabei and colleagues[74] applied such an approach to MRSA, developing a model that can predict with a high degree of accuracy the toxicity of an isolate based on the sequence of signature sites. Another study integrated a genomic approach with gene expression analysis, performed with RNA-seq, to determine antibiotic resistance profiles in *E. coli*.[75] These are not the only examples, as other approaches are being tried and validated on multiple pathogens.[76,77]

Multiple issues will need to be addressed to allow the transition from genomic epidemiology to genomic diagnostics in bacteriology, and one that needs the most concerted effort is standardization of genomes, related metadata and bioinformatic analysis. Indeed, the full potentiality of genomics will only be exploited in clinical microbiology when all the wealth of data generated worldwide will be fully compatible, allowing real time evaluation and comparison of the characteristics of sequenced isolates, in a single global database. This will in turn allow to fully correlate genotype with phenotype, to optimize diagnostic and therapeutic approaches and to monitor movement of dangerous strains worldwide. Efforts towards this goal are already being made from the setting of standards for genome qualities[78] to the establishment of platforms using standarized bioinformatic protocols for genome analyses.[79,80]

It is only fitting to conclude this review how it started: Next-generation sequencing is revolutionizing bacteriology… and the best is yet to come.

Conflict of Interest

The authors declare no conflict of interest.

References

1 Nordmann P, Cuzon G, Naas T. The real threat of *Klebsiella pneumoniae* carbapenemase-producing bacteria. Lancet Infect Dis. 2009;9(4):228–236.
2 White House 2014 National Strategy for Combating Antibiotic-resistant Bacteria. 2014. Available from www.whitehouse.gov/sites/default/files/docs/carb_national_strategy.pdf
3 Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, *et al*. Global dissemination of a multidrug resistant *Escherichia coli* clone. Proc Natl Acad Sci USA. 2014;111(15):5694–5699.
4 Bohlin J, Brynildsrud OB, Sekse C, Snipen L. An evolutionary analysis of genome expansion and pathogenicity in *Escherichia coli*. BMC Genomics. 2014;15:882.
5 DeLeo FR, Chen L, Porcella SF, Martens CA, Kobayashi SD, Porter AR, *et al*. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*. Proc Natl Acad Sci U S A. 2014;111(13):4988–4993.
6 Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, *et al*. Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. Antimicrob Agents Chemother. 2015;59(1):389–396.
7 Snitkin ES, Zelazny AM, Thomas PJ, Stock F; NISC Comparative Sequencing Program Group, Henderson DK, *et al*. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med. 2012; 4(148):148ra116.
8 Stoesser N, Sheppard AE, Shakya M, Sthapit B, Thorson S, Giess A, *et al*. Dynamics of MDR Enterobacter cloacae outbreaks in a neonatal unit in Nepal: insights using wider sampling frames and next-generation sequencing. J Antimicrob Chemother. 2015;70(4):1008–1015.
9 Onori R, Gaiarsa S, Comandatore F, Pongolini S, Brisse S, Colombo A, *et al*. Tracking nosocomial *Klebsiella pneumoniae* infections and outbreaks by whole genome analysis: small-scale Italian scenario within a single hospital. J Clin Microbiol. 2015;53(9):2861–2868.
10 Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Gen. 2011;52(4):413–435.
11 Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 2003;299(5607):682–686.
12 Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, *et al*. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med. 2014;6(254):254ra126.
13 Schwartz DC, Li X, Hernandez Li, Ramnarain SP, Huff EJ, Wang YK. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. Science. 1993;262(5130):110–114.
14 Pneumococcal Disease. Available from: http://www.cdc.gov/vaccines/pubs/pinkbook/downloads/pneumo.pdf
15 Thummeepak R, Leerach N, Kunthalert D, Tangchaisuriya U, Thanwisai A, Sitthisak S. High prevalence of multi-drug resistant *Streptococcus pneumoniae* among healthy children in Thailand. J Infect Public Health. 2015; 8(3):274–281.
16 Baron S (editor). Medical microbiology. 4th ed. Galveston, TX: University of Texas Medical Branch at Galveston; 1996.
17 Hsu K, Pelton S, Karumuri S, Heisey-Grove D, Klein J, Massachusetts department of public health Epidemiologists. Population-based surveillance for childhood invasive pneumococcal disease in the era of conjugate vaccine. Pediatr Infect Dis J. 2005;24(1):17–23.
18 Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, *et al*. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science. 2001;293(5529):498–506.
19 Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in *Pneumococcus*. Science. 2009;324(5933):1454–1457.
20 Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, *et al*. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331(6016):430–434.
21 Donkor ES, Bishop CJ, Gould K, Hinds J, Antonio M, Wren B, *et al*. High levels of recombination among *Streptococcus pneumoniae* isolates from the Gambia. MBio. 2011;2(3):e00040–11.
22 Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, *et al*. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013;45(6):656–663.
23 Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, *et al*. Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PloS Pathog. 2010;6(9):e1001108.

24 Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, *et al*. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46(3):305–309.
25 Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinf. 2008;9:539.
26 Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol. 2013;30(5):1224–1228.
27 Watson CH, Edmunds WJ. A review of typhoid fever transmission dynamic models and economic evaluations of vaccination. Vaccine. 2015;33(Suppl 3):C42–C54.
28 Kothari A, Pruthi A, Chugh TD. The burden of enteric fever. J Infect Dev Ctries. 2008;2(4):253–259.
29 Buckle GC, Walker CL, Black RE. Typhoid fever and paratyphoid fever: systematic review to estimate global morbidity and mortality for 2010. J Glob Health. 2012;2(1):010401.
30 Feasey NA, Gaskell K, Wong V, Msefula C, Selemani G, Kumwenda S, *et al*. Rapid emergence of multidrug resistant, H58-lineage *Salmonella typhi* in Blantyre, Malawi. PLoS Neglected Trop Dis. 2015 ;9(4):e0003748.
31 Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, *et al*. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella typhi* identifies inter- and intracontinental transmission events. Nat Genet. 2015;47(6):632–639.
32 MRSA infection rates by country. Available from: www.cddep.org/tool/mrsa_infection_rates_country.
33 Jevons MP, Coe AW, Parker MT. Methicillin resistance in *Staphylococci*. The Lancet. 1963;281(7287):904–907.
34 David MZ, Daum RS. Community-associated methicillin-resistant *Staphylococcus aureus*: epidemiology and clinical consequences of an emerging epidemic. Clin Microbiol Rev. 2010;23(3):616–687.
35 Hetem DJ, Bootsma MC, Troelstra A, Bonten MJ. Transmissibility of livestock-associated methicillin-resistant *Staphylococcus aureus*. Emerg Infect Dis. 2013;19(11):1797–1802.
36 Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469–474.
37 Parkhill J, Wren BW. Bacterial epidemiology and biology–lessons from genome sequencing. Genome Biol. 2011;12(10):230.
38 Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, *et al*. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. 2013;23(4):653–664.
39 Planet PJ, Diaz L, Kolokotronis SO, Narechania A, Reyes J, Xing G, *et al*. Parallel epidemics of community-associated methicillin-resistant *Staphylococcus aureus* USA300 in North and South America. J Infect Dis. 2015;212(12):1874–1882.
40 Stinear TP, Holt KE, Chua K, Stepnell J, Tuck KL, Coombs G, *et al*. Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant *Staphylococcus aureus*. Genome Biol Evol. 2014;6(2):366–378.
41 Stegger M, Wirth T, Andersen PS, Skov RL, De Grassi A, Simões PM, *et al*. Origin and evolution of European community-acquired methicillin-resistant *Staphylococcus aureus*. MBio. 2014;5(5):e01044–14.
42 Baines SL, Holt KE, Schultz MB, Seemann T, Howden BO, Jensen SO, *et al*. Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant *Staphylococcus aureus*. MBio. 2015;6(2):e00080.
43 Meric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L, *et al*. Ecological overlap and horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*. Genome Biol Evol. 2015;7(5):1313–1328.
44 Croucher NJ, Klugman KP. The emergence of bacterial "hopeful monsters". MBio. 2014;5(4):e01550–14.
45 Gould, SJ. The return of hopeful monsters. Nat Hist. 1977;86(6): 22–30.
46 World Health Organization. Global tuberculosis control: WHO report 2010. World Health Organization, Geneva; 2010.
47 Flynn JL, Chan J, Lin PL. Macrophages and control of granulomatous inflammation in tuberculosis. Mucosal Immunol. 2011;4(3):271–278.
48 Fonseca JD, Knight GM, McHugh TD. The complex evolution of antibiotic resistance in *Mycobacterium tuberculosis*. Int J Infect Dis. 2015;32:94–100.
49 Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol. 2002;10(1):45–52.
50 Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 2014;26(6):431–444.
51 Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, *et al*. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol. 2006;188(2):759–772.
52 Wang W, Hu Y, Mathema B, Jiang W, Kreiswirth B, Xu B. Recent transmission of W-Beijing family Mycobacterium tuberculosis in rural eastern China. Int J Tuberc Lung Dis. 2012;16(3):306–311.
53 Liu M, Jiang W, Liu Y, Zhang Y, Wei X, Wang W. Increased genetic diversity of the *Mycobacterium tuberculosis* W-Beijing genotype that predominates in eastern China. Infect Genet Evol. 2014;22:23–29.
54 Zanini F, Carugati M, Schiroli C, Lapadula G, Lombardi A, Codecasa L, *et al*. *Mycobacterium tuberculosis* Beijing family: analysis of the epidemiological and clinical factors associated with an emerging lineage in the urban area of Milan. Infect Genet Evol. 2014;25:14–19
55 Li D, Dong CB, Cui JY, Nakajima C, Zhang CL, Pan XL, *et al*. Dominant modern sublineages and a new modern sublineage of *Mycobacterium tuberculosis* Beijing family clinical isolates in Heilongjiang Province, China. China. Infect Genet Evol. 2014;27:294–299.
56 Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, *et al*. Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. Proceedings of the National Academy of Sciences. 2015;112(26):8136–8141.
57 Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, *et al*. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat Genet. 2015;47(3):242–249.
58 Drolet GJ. World War I and tuberculosis. A statistical summary and review. Am J Public Health Nations Health. 1945;35(7):689–697.
59 Hershkovitz I, Donoghue HD, Minnikin DE, May H, Lee OY, Feldman M, *et al*. Tuberculosis origin: The Neolithic scenario. Tuberculosis (Edinb). 2015;95(Suppl 1):S122–S126.
60 Cockburn TA. The evolution and eradication of infectious diseases. Perspect Biol Med. 1964;7:498–499.
61 Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, *et al*. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45(10):1176–1182.
62 Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, *et al*. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A. 2002;99(6):3684–3689.
63 Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, *et al*. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PloS Biol. 2008;6(12):e311.
64 Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, *et al*. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014;514(7523):494–497.
65 Patrucco R, Tello R, Bonavia D. Parasitological studies of coprolites of pre-hispanic peruvian populations. Curr Anthrop. 1983;24(3):393–394.
66 Herring DA, Sattenspiel L. Social contexts, syndemics, and infectious disease in northern Aboriginal populations. Am J Hum Biol. 2007; 19(2):190–202.
67 Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, *et al*. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012;8(8):e1002824.
68 Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect. 2013;19(9):803–813.
69 Fournier PE, Dubourg G, Raoult D. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med. 2014;6(11):114.
70 Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, *et al*. Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. J Clin Microbiol. 2015;53(7):2230–2237.
71 Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, *et al*. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 2014;24(7):1180–1192.
72 Hasan NA, Young BA, Minard-Smith AT, Saeed K, Li H, Heizer EM, *et al*. Microbial community profiling of human saliva using shotgun metagenomic sequencing. PLoS ONE. 2014;9(5):e97699.
73 Takeuchi F, Sekizuka T, Yamashita A, Ogasawara Y, Mizuta K, Kuroda M. MePIC, metagenomic pathogen identification for clinical specimens. Jpn J Infect Dis. 2014;67(1):62–65.

74 Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, *et al*. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014;24(5):839–849.
75 Suzuki S, Horinouchi T, Furusawa C. Prediction of antibiotic resistance by gene expression profiles. Nat Commun. 2014;5(5):5792.
76 Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, *et al*. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother. 2013;68(10):2234–2244.
77 Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, *et al*. Rapid antibiotic resistance predictions from genome sequence data for S. *aureus* and M. *tuberculosis*. BioRxiv. 2015:018564.
78 Bristow F, Adam J, Carriço JA, Courtot M, Dhillon B, Dooley D, *et al*. IRIDA: Canada's federated platform for genomic epidemiology. The 8th Meeting of the Global Microbial Identifier. May 11–13, 2015; Beijing, China.
79 Magalhães WC, Rodrigues MR, Silva D, Soares-Souza G, Iannini ML, Cerqueira GC, *et al*. DIVERGENOME: a bioinformatics platform to assist population genetics and genetic epidemiology Studies. Genet Epidemiol. 2012;36(4):360–367.
80 Land ML, Hyatt D, Jun SR, Kora GH, Hauser LJ, Lukjancenko O, *et al*. Quality scores for 32,000 genomes. Stand Genomic Sci. 2014;9:20.