

REVIEW

Bacterial epidemiology and biology - lessons from genome sequencing

Julian Parkhill¹ and Brendan W Wren^{2*}

Abstract

Next-generation sequencing has ushered in a new era of microbial genomics, enabling the detailed historical and geographical tracing of bacteria. This is helping to shape our understanding of bacterial evolution.

Whole-genome sequencing - a transformation in bacterial epidemiology

The application of whole-genome sequencing affords the opportunity to generate bacterial nucleic acid sequence data of extraordinary resolution, making it possible to identify single base changes within entire genomes. The development of second (and third) generation sequencing technology has largely been driven by the desire to assess human genetic variation rapidly by mapping genome-wide single-nucleotide polymorphisms (SNPs). Several recent studies have applied similar analyses at the whole-genome level to the much smaller genomes of bacteria, providing data of very fine-scale resolution and enabling the evolutionary history of multiple strains within a clonal lineage to be determined [1-8]. Not surprisingly, these studies have focused on bacterial pathogens because of their importance in disease.

Distinguishing individual bacterial lineages within a species, initially by phenotypic and subsequently by genotypic typing techniques, has been the cornerstone of infectious disease epidemiology, allowing the identification and tracking of the organisms responsible for infection and disease. On a wider scale, molecular typing is central to determining the population structure and understanding the evolution of bacterial pathogens. To date, sequence-based typing approaches, such as multi-locus sequence typing (MLST), have relied on variation within a few genes [9]. Although such techniques are

highly informative, they have limited resolution when applied to closely related isolates. Thus, they are often unsuitable for identifying fine-grained evolutionary events or for distinguishing clonal strains within a recent epidemic [10]. This situation has now been improved by the application of next-generation sequencing to bacterial collections of known origin and provenance. We review recent examples in which such studies have informed our understanding of the epidemiology and evolution of bacterial pathogens.

Phylohistory and phylogeography: from the 'Black Death' to hospital-acquired infections

Studies of the phylogeny of a species or of a clonal lineage within a species are highly dependent on the quantity and diversity of isolates sampled in the study. Recently, it has become possible to fully sequence significant numbers of isolates in strain collections, revealing new information on their temporal and spatial transmission dynamics.

Yersinia pestis

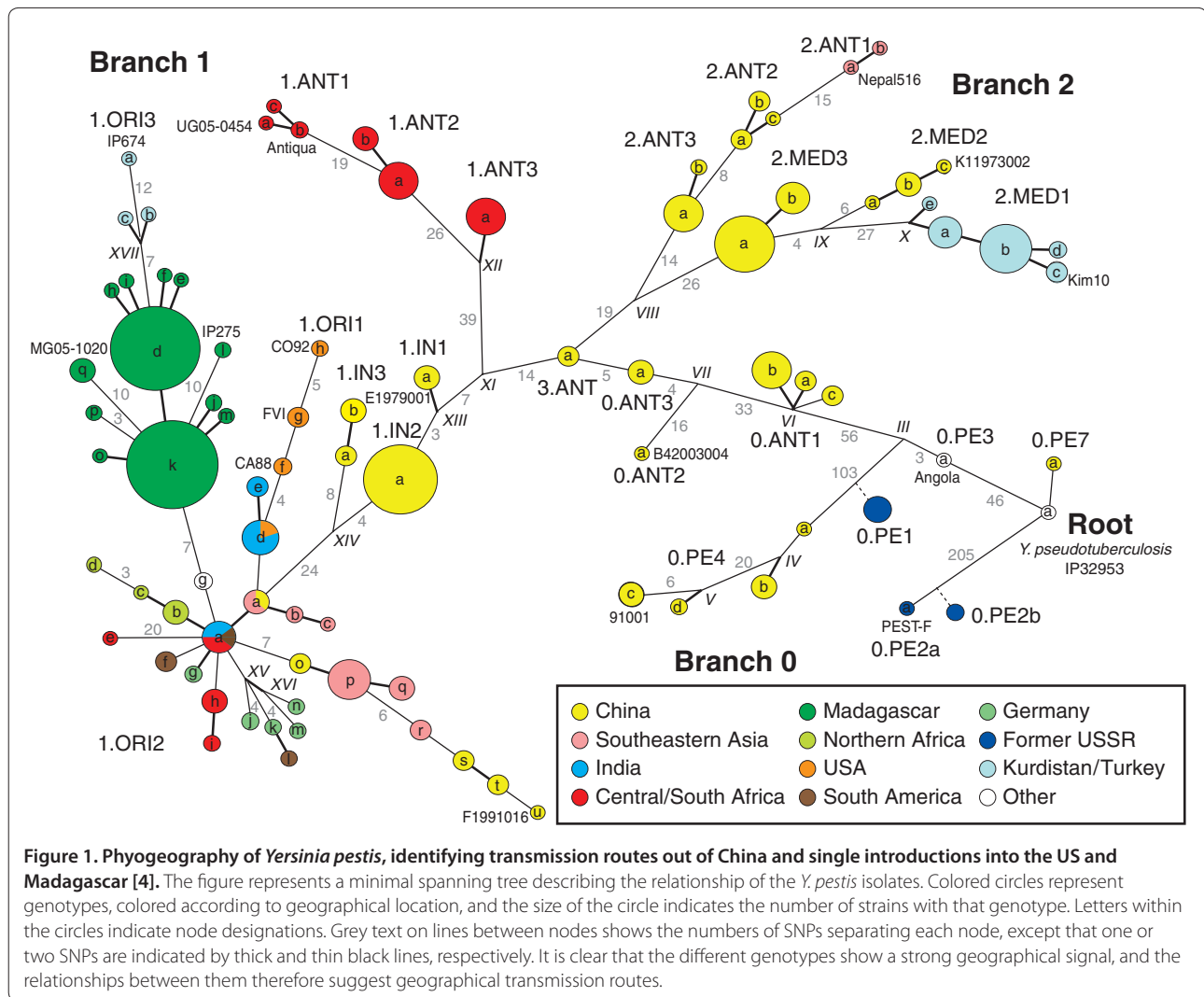
As a recent pathogen, the plague bacillus *Yersinia pestis* is genetically uniform. Thus, it is relatively simple to trace the source and to determine the global phylogenetic diversity of this bacillus using whole-genome SNP analysis. Such an analysis of a global collection of strains showed that *Y. pestis* originated in or around China, confirming the 'out of China' hypothesis whereby initial transmission was along East-West trade routes such as the Silk Road. Multiple radiations from China to Europe, Africa and Southeast Asia can be traced to country-specific SNP lineages [4]. Further fine analysis uncovered the phylogeography of *Y. pestis* in North America, where it radiated from a single introduction, and in Madagascar, where again it radiated from a single introduction. These and other radiations largely matched historical accounts of the disease and known trading routes [4] (Figure 1).

The causative agent of the 'Black Death' has been a matter of considerable debate [11]. Disease symptoms such as inflamed buboes led to the assumption that the 'Black Death' was a typical *Y. pestis* infection, but other etiologic agents, including hemorrhagic viruses, have been proposed on the basis of clinical and epidemiological

*Correspondence: brendan.wren@lshtm.ac.uk

²Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

Full list of author information is available at the end of the article



information. The SNP analysis used to trace the phylohistory of *Y. pestis* [4] is a development of an earlier SNP analysis [12] that has already been used to provide information on ancient bacterial DNA. Specifically, it had been used to analyze DNA extracted from human dental pulp taken from medieval mass graves in several European locations [11]. The analysis made it possible not only to amplify *Y. pestis* DNA taken from the victims but also to place these bacterial DNA sequences in the correct historical context through phylogenetic analysis. The results imply that at least two distinct ancient clones of *Y. pestis* were responsible for the ‘Black Death’, at least one of which appears to be extinct. These findings definitively show that *Y. pestis* was the causative agent of the ‘Black Death’.

Vibrio cholerae

Cholera, caused by *Vibrio cholerae*, is another ancient scourge that has recently been studied by rapid genome

sequencing. The current (seventh) cholera pandemic, recently manifest in Zimbabwe, Pakistan and Haiti, is caused by the El Tor biotype of serogroup O1 (El Tor O1) [13,14]. Two clinical *V. cholerae* isolates from the current outbreak in Haiti were fully sequenced in less than 24 hours using third-generation sequencing with the PacBio *RS* sequencing system [15]. Third-generation single-molecule real-time sequencing with this system involves direct observation of the DNA polymerase while it synthesizes a strand of DNA. It has advantages over second-generation sequencing in terms of increased read length and speed. Whole-genome sequence comparisons with reference strains confirmed that the Haitian cholera epidemic is clonal and El Tor O1, and suggested that the *V. cholerae* strain was introduced into Haiti by human activity from a distant geographic source [15]. Although this was a limited study, the data suggest that a South Asian variant of *V. cholerae* El Tor O1 has recently been accidentally introduced into Haiti.

This theory is consistent with the epidemiological evidence [16].

Clostridium difficile

Genetic epidemiological studies are likely to prove particularly useful for tracing the routes of transmission and sources of infection for hospital-acquired infections. Whole-genome sequencing has been used to explore the phylogeny, horizontal gene transfer, recombination, and micro- and macroevolution of the major hospital-acquired pathogen *Clostridium difficile* [5,17,18]. This infection produces a wide range of symptoms from mild diarrhea to life-threatening pseudomembranous colitis, and characteristically occurs after treatment with broad-spectrum antibiotics. The hospital environment in which there are people undergoing antibiotic treatment provides a discrete ecosystem in which *C. difficile* persists and select virulent clones thrive. Consequently, *C. difficile* is the most frequent cause of nosocomial diarrhea worldwide [19,20].

Phylogenetic analysis demonstrates that *C. difficile* is a genetically diverse species, with estimates of the date of the most recent common ancestor (MRCA) varying from 1.1 to 85 million years before present. By contrast, the disease-causing isolates (PCR-ribotypes 017s, 027s, and 078s) have arisen from multiple lineages over very short evolutionary timescales. This suggests that virulence has evolved independently in several highly epidemic lineages [5], and contradicts the notion that a single lineage evolved to become pathogenic. For example, the MRCA of the very recently emerged PCR-ribotype 027 hypervirulent lineage was confirmed at approximately 30 years ago, consistent with the dates of the transcontinental spread of *C. difficile* [5,21,22]. This has implications for the emergence of *C. difficile* as a human pathogen. Although *C. difficile* appears to be an ancient species, it was recognized as a pathogen only 30 years ago, indicating that genetic modifications, changes in interactions between host and pathogen, and factors such as human activity, hospital design and antibiotic use might have contributed to the recent emergence of *C. difficile* as a major pathogen.

Methicillin-resistant *Staphylococcus aureus*

The first large-scale whole-genome analysis of a clonal lineage within a species was undertaken for an epidemic sequence type (ST239) of the notorious hospital pathogen methicillin-resistant *Staphylococcus aureus* (MRSA). The collection of 63 geographically diverse ST239 representatives over four decades revealed clear evidence of both geographical grouping and intercontinental transmission over time [6]. The resolution of the whole-genome SNP analysis was so fine that it revealed the microevolution of MRSA within a Thai hospital over a

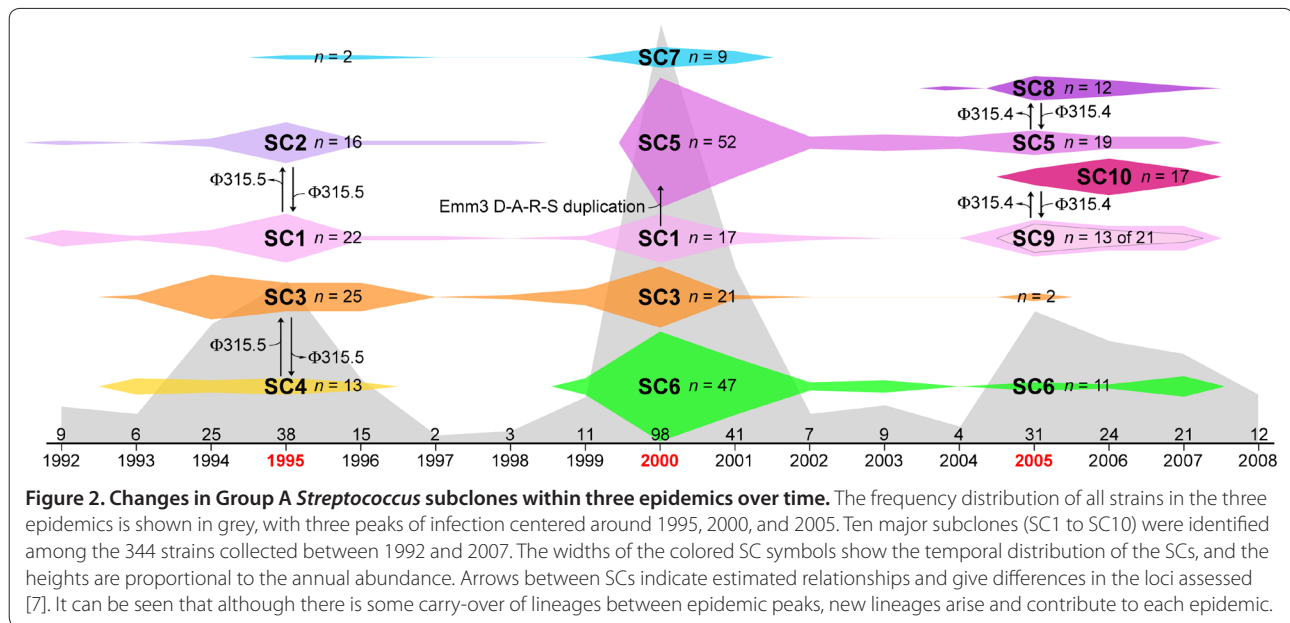
7-month period. It was even able to determine when new strains were introduced from the community, or when they were acquired by person-to-person transmission within the hospital environment. Thus, very high-resolution phylogenetic analyses enable detailed epidemiological reconstructions at the global (that is, spread by modern transport) and local (that is, within hospital) levels, suggesting that such analyses are likely to be clinically applicable in the near future.

Streptococcus pneumoniae

A similar whole-genome analysis of 240 *Streptococcus pneumoniae* isolates of the antibiotic-resistant PMEN1 clonal lineage (which is often serotype 23F) tracked the phylogeny of this recombinogenic lineage by SNP analysis, after accounting for the confounding effect of extensive recombination [3]. The phylogeny of the PMEN1 clone confirms that it arose around 40 years ago. Analysis of genetic markers, together with the provenance of the strains, suggests that the introduction of the heptavalent PCV7 glycoconjugate vaccine in the US (designed against 7 of the 90 pneumococcal serotypes, including 23F) effectively led to the depletion of the resident 23F population. However, this selective pressure opened the niche to non-vaccine serotypes such as 19A. These studies strongly suggest that although some of these non-vaccine isolates were from the same lineage as the 23F serotype, they had acquired a new non-vaccine capsule type before the introduction of the vaccine. This demonstrates that, despite the remarkable adaptability of recombinogenic bacteria such as the pneumococcus, strong vaccine pressure can remove the population that expresses vaccine-type capsules before it can switch its capsule genes. The emergence of vaccine-escape strains such as the 19A serotype thus has important implications for the introduction of partial species coverage vaccines. Such interventions require vigilance and the genomic surveillance of all bacterial clones. Overall, this study shows the surprisingly rapid evolution of a recombinogenic bacterial pathogen that can be linked to clinical interventions such as antibiotic usage and the introduction of vaccines.

Group A *Streptococcus*

The Group A *Streptococcus* (GAS) is a Gram-positive, human pathogen responsible for diseases ranging from pharyngitis to life-threatening invasive disease. It has been dubbed the 'flesh eating bug' by the popular media. The whole-genome analysis of 344 serotype M3 GAS strains from three epidemics (over a 16-year period) in Ontario, Canada, was undertaken using a combination of sequencing and SNP analysis [7]. This revealed a relationship among the 344 invasive strains responsible for the three epidemics, which comprised a dynamic



mixture of distinct clonally related complexes (Figure 2). The resulting evolutionary genetic framework facilitated an assessment of the phylogeographic features of these epidemics. The data confirm that each epidemic is composed of strains that are genetically distinct from those involved in the preceding epidemic, rather than re-emerging genetically identical organisms. Further investigations should provide an enhanced understanding of the genomic relationships between the serotype M3 strains that cause pharyngitis and invasive infections, thereby potentially forewarning of lethal GAS infections.

A novel aspect of this study was the use of information generated from whole-genome analysis to select representative strains for comparative transcriptome and mass spectroscopy SNP analyses [7,23]. A comparative study of transcriptome expression in a subset of strains revealed that closely related strains, which are differentiated by apparently modest genetic changes, can have significantly divergent transcriptomes. Therefore, subtle genetic changes could have more significant phenotypic consequences than previously appreciated.

Deriving evolutionary information from mutation rates and phylogenetic analysis

One advantage of whole-genome SNP analysis is that it could be used to identify non-synonymous mutations that are more likely to be influenced by evolutionary selective pressure than synonymous mutations or mutations in intergenic DNA. The identification of genes that have undergone non-synonymous mutations should provide clues to the evolutionary pressures experienced over time by a bacterial species or clonal lineages within a species. Traditionally, the selective forces acting on a

bacterial genome were investigated by calculating the ratio of non-synonymous to synonymous substitutions (dN/dS) for a given species comparison [24]. A ratio significantly less than 1 suggests strong purifying or stabilizing selection, whereas a ratio close to 1 suggests a neutral selection pressure, and a ratio greater than 1 indicates diversifying selection. For very closely related genomes (for example, those within clonal lineages), however, dN/dS can be close to 1 simply because there has been insufficient time for selection to act [25]. This, combined with the very small number of mutations in individual genes within these lineages, actually makes the identification of genes under selection very difficult. Nevertheless, the increasing number of strains within bacterial collections that are currently being sequenced does allow some inferences relating to genetic selection.

The whole-genome sequencing of 19 isolates of *Salmonella enterica* serovar Typhi (*S. Typhi*), a human-restricted bacterial pathogen that causes typhoid fever, confirmed the very limited genetic variation within this species. The mean dN/dS of each isolate in comparison with the last common ancestor was approximately 0.66, suggesting that there has been a weak trend towards stabilizing selection since the occurrence of the MRCA) of *S. Typhi* [8]. Detailed analysis of the SNPs showed little evidence of diversifying selection, antigenic variation or recombination between isolates. Only 38% of genes had any sequence variation at all, and the occurrence of variants in almost all of these genes appeared to match random expectation. Nevertheless, evidence for selection could be found by looking for SNPs that occurred independently on different branches of the tree (that is, for homoplasy, or convergent evolution). Examples of

such non-synonymous mutations include those in *gyrA* that are responsible for resistance to quinolone-based antibiotics. No genes besides *gyrA* contained multiple homoplasic SNPs, and very few genes had an excess of non-synonymous SNPs, showing that unlike the strong adaptive selection for mutations conferring antibiotic resistance, there is little evidence for selective pressure for antigenic variation driven by immune selection. This is consistent with the previous assertion that most of the variants in the Typhi genome accumulate by genetic drift [26]. The adaptive mutations evident in the *gyrA* gene highlight the strong selective pressure on the *S. Typhi* genome associated with antibiotic use in the human population. This is not particularly surprising as the fitness advantage associated with increased antibiotic resistance is very strong. The paucity of similar evidence for other adaptive mutations suggests that *S. Typhi* is under relatively little selective pressure from its human host, consistent with its long-term carriage within a protected niche, the gall-bladder.

The technique of identifying genes with a significant excess of SNPs has also been used in the Group A *Streptococcus* data set [7]: just under 5% of the variable genes were found to have a statistical excess of variants. These included several surface proteins and virulence factors, along with a regulator, *ropB*, that controls multiple virulence genes, indicating that selection by the host is a significant factor in the recent evolution of this organism.

For *C. difficile*, dN/dS was calculated for different clonal lineages. The data from deeply diverging lineages provided evidence of strong purifying selection. For example, the average dN/dS between the divergent lineage 078 and the other strains tested is ≈ 0.08 [5]. By contrast, for recently diverged lineages, such as the 027 ribotype, dN/dS is very close to 1, consistent with the delayed action of purifying selection [25] and again making it difficult to identify genes that are under selection. The extensive full-genome data in this study did, however, permit the identification of genes that had significantly increased rates of non-synonymous nucleotide polymorphism, thereby providing clues about the operation of selective forces in the host.

A similar approach for identifying genes that are under selection (that is, searching for homoplasic SNPs within the background of random SNPs that delineate the tree) was used in the MRSA ST239 study. Again, there was a relatively small number of homoplasic SNPs (less than 1%) in the core genome, but around 30% of these could be directly linked to the evolution of resistance to antibiotics currently in use (for example, quinolones, rifampicin, mupirocin and trimethoprim). These findings confirm that antibiotic use is a major driving force in the evolution of MRSA, and that this technique can detect

recent selection [6]. Nevertheless, the majority of homoplasies were found in genes for which no reason for selective pressure could be clearly identified. Understanding how and why these mutations are selected could provide novel information on the emergence of multi-drug-resistant clonal lineages.

The ability to distinguish vertically acquired substitutions from horizontally acquired sequences is crucial to reconstructing phylogenies for recombinogenic organisms. This was recently attempted for the *S. pneumoniae* data set, in which 88% of the variants were estimated to have been introduced by recombination. Despite this, the relative likelihood that a polymorphism was introduced through recombination rather than by point mutation (r/m ratio) was estimated to be 7.2, less than the previously calculated value of approximately 66 from MLST data [3]. By removing recombination events from the phylogeny, the number of homoplasic sites was reduced by 97%, and the apparent rate of SNP accumulation was much more consistent within the tree, thus considerably strengthening the core phylogeny and the inferences that could be made from it.

The large majority of the SNPs identified in these studies of recent clones are effectively neutral over these timescales, and thus they can be used to give a good estimate of the current mutation rate. Curiously, this estimate for *S. pneumoniae* and *S. aureus* is roughly 1,000 times greater than that estimated from the synonymous substitution rate between deeper bacterial lineages (such as *Escherichia coli* and *Salmonella*) [27]. This apparent discrepancy can be reconciled by the fact that synonymous sites, commonly assumed to be neutral, are in fact under selection over longer time periods because of pressures such as G+C content and codon usage.

Limitations and future perspectives

Advances in sequencing technologies have enabled the whole-genome phylogenies of multiple clonal isolates to be determined readily, but as with all microbial epidemiological investigations, the size and composition of the strain collection investigated is crucial for subsequent biological interpretations. This is particularly relevant for bacterial pathogens that reside in multiple niches: isolates are frequently collected only from patients. This might bias the data sets and provide an incomplete picture of the true diversity of a bacterial species, which might have alternative niches and reservoirs to humans. The lack of representation of the true diversity results in gaps in our evolutionary knowledge of a given pathogen. For example, there appears to be huge evolutionary distances between current epidemic clonal lineages of *C. difficile*. Nevertheless, because so much data can now be derived from bacterial isolates, it is now more important than ever to engage appropriately with

clinicians and strain collectors and to acquire accurate strain provenance.

Current sequencing approaches are limited for more divergent bacterial pathogens. As more representatives of clonal lineages within a bacterial species are sequenced and as sequencing technology continues to improve, however, more diverse and panmictic bacterial populations will be sequenced. This might be particularly revealing for organisms such as *Helicobacter pylori*, which is resident in the stomachs of half of the human population. MLST analysis of multiple *H. pylori* strains has been used to re-construct and confirm human population expansion and migration in Africa, Europe and the Pacific [28]. The higher resolution of whole-genome sequencing promises significant progress towards tracing human pre-history.

Notwithstanding these limitations, molecular epidemiology has clearly come of age for clonal bacterial pathogens. The major advantage of whole-genome sequencing is its power to discriminate between isolates, enabling the generation of robust phylogenies. This provides greater confidence in identifying the origins of infections and the routes of transmission, as demonstrated by the monitoring of patient-to-patient transfer of bacteria within a hospital [6,29] or within a community [2]. Such finely tuned transmission tracking will be vital in determining whether factors such as patient-to-patient transmission are important in the spread of the disease. In the future, this may facilitate the practice of proactive infectious disease surveillance to truncate or avert epidemics.

In contrast to most typing methods, whole-genome sequencing also facilitates the direct identification of gene losses and gene gains that can play a role in the evolution of a bacterial species or clonal lineage within a species. Such information has frequently identified the emergence of antibiotic resistance within populations, which is often associated with increased antibiotic usage. Nevertheless, other more subtle selective forces are also likely to be important in the emergence of bacterial pathogens, and our current knowledge in this area is lacking. High-throughput sequencing holds the promise of mapping more subtle associations between phenotype and genotype [30,31]. The next few years will see an increase in the biological interpretation of such mutations using high-throughput *in vitro* assays and the selected testing of representative isolates in animal infection studies. This should further our understanding of the host, ecological, environmental and human forces that are important in the evolution of bacterial pathogens and enable further appropriate interventions to be made. Another area in which we lack knowledge is how and why some bacterial lineages appear to diminish, or even become extinct. For example, the *S. pneumoniae* lineage BM4200 is a multidrug-resistant serotype 23F isolate

from 1978, but despite its similarity to the PMEN1 isolates, it is now seldom found [32]. As genome sequencing, SNP detection and geospatial information become more accessible, these methods will continue to transform the way molecular epidemiology is used to study populations of bacterial pathogens.

Concluding remarks

We are now in a new era of high-throughput, sequence-based microbiology that will have important implications for health service providers working with infectious diseases. Rather than merely identifying a particular bacterium by culture, whole-genome sequencing will provide a better understanding of its origin and disease potential. As global positioning systems become more accessible and merge with molecular epidemiology, more accurate geospatial information on the origins of strains and outbreaks will become available. Just as the DNA fingerprinting of human microsatellites changed our lives through diverse applications from paternity testing to crime screen investigations, next-generation sequencing means that molecular epidemiology is set to be revolutionized for clonal bacterial pathogens. The next few years promise a voyage of discovery in terms of the attribution of sources and transmission tracking of bacteria, the understanding of how and why epidemic clones emerge or disappear, and ultimately the management and treatment of infectious diseases.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We wish to acknowledge research funding from The Wellcome Trust.

Author details

¹The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ²Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Published: 24 October 2011

References

1. Shea PR, Beres SB, Flores AR, Ewbank AL, Gonzalez-Lugo JH, Martagon-Rosado AJ, Martinez-Gutierrez JC, Rehman HA, Serrano-Gonzalez M, Fittipaldi N, Ayers SD, Webb P, Willey BM, Low DE, Musser JM: **Distinct signatures of diversifying selection revealed by genome analysis of respiratory tract and invasive bacterial populations.** *Proc Natl Acad Sci U S A* 2011, **108**:5039-5044.
2. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodtkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *N Engl J Med* 2011, **364**:730-739.
3. 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: **Rapid pneumococcal evolution in response to clinical interventions.** *Science* 2011, **331**:430-434.
4. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang

- R, Carniel E, Achtman M: ***Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity.** *Nat Genet* 2010, **42**:1140-1143.
5. He M, Sebahia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HM, Quail MA, Rance R, Brooks K, Churcher C, Harris D, Bentley SD, Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, van Tonder A, Walker D, Wren BW, Dougan G, Parkhill J: **Evolutionary dynamics of *Clostridium difficile* over short and long time scales.** *Proc Natl Acad Sci U S A* 2010, **107**:7527-7532.
 6. Beres SB, Feil EJ, Holden MT, 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: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327**:469-474.
 7. Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, McGeer A, Willey BM, Green K, Tyrrell GJ, Goldman TD, Feldgarden M, Birren BW, Fofanov Y, Boos J, Wheaton WD, Honisch C, Musser JM: **Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics.** *Proc Natl Acad Sci U S A* 2010, **107**:4371-4376.
 8. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Iecsek C, Achtman M, Dougan G: **High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi.** *Nat Genet* 2008, **40**:987-993.
 9. 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: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A* 1998, **95**:3140-3145.
 10. Achtman M: **Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens.** *Annu Rev Microbiol* 2008, **62**:53-70.
 11. Haensch S, Bianucci R, Signoli M, Rajerison M, Schultz M, Kacki S, Vermunt M, Weston DA, Hurst D, Achtman M, Carniel E, Bramanti B: **Distinct clones of *Yersinia pestis* caused the black death.** *PLoS Pathog* 2010, **6**:e1001134.
 12. Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francois V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P: **Microevolution and history of the plague bacillus, *Yersinia pestis*.** *Proc Natl Acad Sci U S A* 2004, **101**:17837-17842.
 13. Cho YJ, Yi H, Lee JH, Kim DW, Chun J: **Genomic evolution of *Vibrio cholerae*.** *Curr Opin Microbiol* 2010, **13**:646-651.
 14. Butler D: **Cholera tightens grip on Haiti.** *Nature* 2010, **468**:483-484.
 15. 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: **The origin of the Haitian cholera outbreak strain.** *N Engl J Med* 2011, **364**:33-42.
 16. **Final Report of the Independent Panel of Experts on the Cholera Outbreak in Haiti** [http://www.un.org/News/dh/infocus/haiti/UN-cholera-report-final.pdf]
 17. Stabler RA, Valiente E, Dawson LF, He M, Parkhill J, Wren BW: **In-depth genetic analysis of *Clostridium difficile* PCR-ribotype 027 strains reveals high genome fluidity including point mutations and inversions.** *Gut Microbes* 2010, **1**:269-276.
 18. Stabler RA, He M, Dawson L, Martin M, Valiente E, Corton C, Lawley TD, Sebahia M, Quail MA, Rose G, Gerding DN, Gibert M, Popoff MR, Parkhill J, Dougan G, Wren BW: **Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium.** *Genome Biol* 2009, **10**:R102.
 19. Kuijper EJ, van Dissel JT, Wilcox MH: ***Clostridium difficile*: changing epidemiology and new treatment options.** *Curr Opin Infect Dis* 2007, **20**:376-383.
 20. Bartlett JG: ***Clostridium difficile*: history of its role as an enteric pathogen and the current state of knowledge about the organism.** *Clin Infect Dis* 1994, **18** (Suppl 4):S265-272.
 21. Goorhuis A, Van der Kooij T, Vaessen N, Dekker FW, Van den Berg R, Harmanus C, van den Hof S, Notermans DW, Kuijper EJ: **Spread and epidemiology of *Clostridium difficile* polymerase chain reaction ribotype 027/toxinotype III in The Netherlands.** *Clin Infect Dis* 2007, **45**:695-703.
 22. Loo VG, Poirier L, Miller MA, Oughton M, Libman MD, Michaud S, Bourgault AM, Nguyen T, Frenette C, Kelly M, Vibien A, Brassard P, Fenn S, Dewar K, Hudson TJ, Horn R, René P, Monczak Y, Dascal A: **A predominantly clonal multi-institutional outbreak of *Clostridium difficile*-associated diarrhea with high morbidity and mortality.** *N Engl J Med* 2005, **353**:2442-2449.
 23. Honisch C, Chen Y, Mortimer C, Arnold C, Schmidt O, van den Boom D, Cantor CR, Shah HN, Gharbia SE: **Automated comparative sequence analysis by base-specific cleavage and mass spectrometry for nucleic acid-based microbial typing.** *Proc Natl Acad Sci U S A* 2007, **104**:10649-10654.
 24. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends Ecol Evol* 2000, **15**:496-503.
 25. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ: **Comparisons of dN/dS are time dependent for closely related bacterial genomes.** *J Theor Biol* 2006, **239**:226-235.
 26. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, Achtman M: **Evolutionary history of *Salmonella* typhi.** *Science* 2006, **314**:1301-1304.
 27. Ochman H, Elwyn S, Moran NA: **Calibrating bacterial evolution.** *Proc Natl Acad Sci U S A* 1999, **96**:12638-12643.
 28. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu JY, Maady A, Bernhoft S, Thiberge JM, Phuanukoonnon S, Jobb G, Siba P, Graham DY, Marshall BJ, Achtman M: **The peopling of the Pacific from a bacterial perspective.** *Science* 2009, **323**:527-530.
 29. Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboym D, Pallen MJ: **High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak.** *J Hosp Infect* 2010, **75**:37-41.
 30. Bille E, Ure R, Gray SJ, Kaczmarek EB, McCarthy ND, Nassif X, Maiden MC, Tinsley CR: **Association of a bacteriophage with meningococcal disease in young adults.** *PLoS One* 2008, **3**:e3885.
 31. Falush D, Bowden R: **Genome-wide association mapping in bacteria?** *Trends Microbiol* 2006, **14**:353-355.
 32. Buu-Hoi A, Horodniceanu T: **Conjugative transfer of multiple antibiotic resistance markers in *Streptococcus pneumoniae*.** *J Bacteriol* 1980, **143**:313-320.

doi:10.1186/gb-2011-12-10-230

Cite this article as: Parkhill J, Wren BW: **Bacterial epidemiology and biology - lessons from genome sequencing.** *Genome Biology* 2011, **12**:230.