PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Review

Cite this article: Croucher NJ, Harris SR, Grad YH, Hanage WP. 2013 Bacterial genomes in epidemiology—present and future. Phil Trans R Soc B 368: 20120202. http://dx.doi.org/10.1098/rstb.2012.0202

One contribution of 18 to a Discussion Meeting Issue 'Next-generation molecular and evolutionary epidemiology of infectious disease'.

Subject Areas:
genomics, health and disease and epidemiology, microbiology, evolution

Keywords:
horizontal gene transfer, coalescent, next-generation sequencing, second-generation sequencing

Author for correspondence:
William P. Hanage
e-mail: whanage@hsph.harvard.edu

Bacterial genomes in epidemiology—present and future

Nicholas J. Croucher[1], Simon R. Harris[2], Yonatan H. Grad[1,3] and William P. Hanage[1]

[1]Department of Epidemiology, Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA, USA
[2]Pathogen Genomics, Wellcome Trust Sanger Insitute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[3]Division of Infectious Diseases, Brigham and Women's Hospital, Boston, MA, USA

Sequence data are well established in the reconstruction of the phylogenetic and demographic scenarios that have given rise to outbreaks of viral pathogens. The application of similar methods to bacteria has been hindered in the main by the lack of high-resolution nucleotide sequence data from quality samples. Developing and already available genomic methods have greatly increased the amount of data that can be used to characterize an isolate and its relationship to others. However, differences in sequencing platforms and data analysis mean that these enhanced data come with a cost in terms of portability: results from one laboratory may not be directly comparable with those from another. Moreover, genomic data for many bacteria bear the mark of a history including extensive recombination, which has the potential to greatly confound phylogenetic and coalescent analyses. Here, we discuss the exacting requirements of genomic epidemiology, and means by which the distorting signal of recombination can be minimized to permit the leverage of growing datasets of genomic data from bacterial pathogens.

1. Introduction

A single genome from one isolate of a species provides a snapshot of that organism at one point in time. It provides evidence on the loci present and their function, and the relationship of that species (or the loci within it) to other species. What it does not tell us is anything about the genetic diversity within the species, and how it is partitioned among lineages or has arisen over time through the combination of selective and demographic processes. We can only detect such variation by sequencing multiple genomes, and the result is the developing field of population genomics.

In the epidemiology of bacterial diseases there are several important questions that may be answered using such data: how much time has elapsed since the most recent common ancestor of a sample (or outbreak), have there been bottlenecks in transmission or with what rate is the population of an emerging pathogen growing? Genomic data also reveal complementary information on the distribution of variation in gene content, which may impact on both virulence and drug resistance. Necessary for both types of analysis is a means of assessing relationships between sequences using polymorphic sites, typically using phylogenetics or sequence clustering methods. However, in order to identify the similarities between strains that are a result of vertically inherited variation indicative of recent common descent, two other sources of polymorphisms must be addressed. The first is the 'noise' generated by inherent errors in sequencing and analysis, which will vary according to the technology and software used in research. The second is the 'confounding signal' of polymorphisms introduced through horizontal transfer of sequence mediated by both homologous and non-homologous recombination. We will consider how these can be best handled from the perspective of molecular epidemiology.

Royal Society Publishing
Informing the science of the future

2. Requirements for genomic data in epidemiology

The ideal data for epidemiology would be robust finished genomes with confident assignment of single nucleotide polymorphisms (SNPs), insertions and deletions throughout, including repetitive regions and plasmids, generated economically and quickly (this last point being particularly important for outbreak analysis). Unfortunately, this is not currently achievable. As technology stands, with the caveat that it can change quickly, finished genomes come at an unacceptably slow pace and at considerable cost. Recent advances in the molecular studies of bacterial genomes are the result of the ability to economically generate large sets of short sequence reads using second generation (Illumina, 454 and SOLiD) or third generation (currently, PacBio) technologies. These can be processed to give draft genome sequences as is most appropriate for the particular questions being studied.

In infectious disease epidemiology, this will largely depend on whether the study is addressing 'local' or 'global' (or short and long term) questions [1]. Short term, local questions focus on very recently diverged isolates, such as are found in an outbreak, and as a result require methods with great resolving power (as an example, one recent study found isolates from a single outbreak differing at between 21 and 0 SNPs over more than 5 Mb of the genome [2]). Longer-term questions are concerned with deeper relationships, such as the probable origin of the Haiti outbreak strain of *Vibrio cholerae* [3], or whether antibiotic resistant strains on one continent are derived from those circulating on another [4]. In contrast with local epidemiology, these are normally considered to unfold over decades and at national or intercontinental scales.

3. High-resolution approaches for local epidemiology

For local or outbreak epidemiology, the crucial concern is the resolving power. In the pre-genomic era, this has been achieved by focusing on comparatively rapidly accumulating variation, for example, pulsed field gel electrophoresis (PFGE) [5–7]. For very uniform species, such as *Mycobacterium tuberculosis*, even more highly variable markers, such as the numbers of tandem repeats [8,9] or insertion sequence (IS) elements [10,11], are required. However, by virtue of their repetitive nature, such regions are frequently difficult to infer accurately using the relatively short reads produced by second generation sequencing technologies. Nevertheless, there is much excitement around sequencing platforms such as PacBio that promise to be able to span such regions with single reads (although these typically have higher per nucleotide error rates; for further discussion see [12,13]). There is considerable interest in mixed sequencing strategies that combine data from multiple platforms, and combine the strengths of each.

Even without access to the information in such repetitive regions, second generation sequencing technologies are able to detect differences between closely related strains with high sensitivity as they can identify SNP variation across a large fraction of the genome. This has potential as an alternative source of variation to discriminate between different lineages [14]. Such data make use of a much greater proportion of the genome than repeat sequences, and are far better suited to downstream analyses [15], such as the generation of phylogenies. However, very closely related isolates may differ at a scant handful of SNPs [2]. In the first instance, it is important to distinguish those strains that constitute a single outbreak from other co-circulating conspecific bacteria, which can be difficult when there is little extant variation to detect. For instance, a recent TB outbreak in British Columbia was found to be composed of two independent lineages spreading concurrently using genomic data, where analysis of tandem repeats had indicated the isolates were clonal [14]. Similarly, genomic analysis of a methicillin-resistant *S. aureus* outbreak in a UK hospital was able to discriminate between outbreak and non-outbreak strains that were genetically indistinguishable by conventional methods [16].

Analysing variation within outbreaks requires an even greater level of precision. For instance, the European *E. coli* O104 : H4 outbreak in 2011 included strains that could be distinguished by just 1 SNP [2]. Under such circumstances, it is necessary to identify polymorphisms with high confidence. Such analysis is often performed through 'read mapping': the alignment of large numbers of short sequence reads produced by high-throughput sequencing against a complete reference genome. SNPs can then be called when the consensus of the multiple reads spanning a particular nucleotide specifies a different base to that in the reference sequence. In practice, there is a great deal of variation in the choice of threshold conditions (e.g. minimum depth of coverage, proportion of reads that must agree to call a consensus base) that are used to call SNPs. These have to be optimized to minimize error rates, which must inevitably strike a balance between false-negatives that will deprive us of a genuine signal, or false-positives that exaggerate the amount of SNP variation. Possible sources of such false-positives and -negatives are summarized in table 1.

At present, comparison of independent analyses is complicated by differences in sequencing technology, quality of output data, software and parameter choices. Calling variants by comparison against a reference genome also introduces specific concerns around the choice of reference, of which there are a limited number available. Even if the genome is closely related to the sequenced isolates overall, divergent loci can cause problems. If they are sufficiently different from the sequence reads to preclude mapping in the region, then the method suffers from a loss of specificity through not being able to extract information from the whole genome. However, mobile genetic elements (MGEs) can cause the converse problem, as their ability to introduce paralogous sequence at multiple loci, in conjunction with their highly dynamic nature, can lead to reads being mapped against non-homologous sequence. This leads to a fall in specificity owing to the introduction of false-positive polymorphisms through misalignment.

While it is important that all analyses are tailored to the species and collection under consideration, there are advantages to applying a consistent set of standards. Making processed genomic data more easily comparable would greatly enhance the portability of those data, meaning that results from one laboratory may be confidently compared with those from others. This has been a perennial issue in epidemiology [17–19]. Before Sanger sequencing became commonplace and widely available, epidemiologists were forced to use methods that characterize genomic diversity

Table 1. Summary of the processes that generate variation in genomic alignments in epidemiology. It should be noted that of these, point mutations inherited clonally should form the basis of any phylogenetic analysis.

	processes
true variation	point mutation
	insertions and deletions (tandem repeats/homopolymers, etc)
	homologous recombination
	mobile genetic elements
false-positives	SNPs generated *in vitro*
	mapping errors (frequently associated with repeat sequences)
	systematic sequencing errors
	'ectopic' mapping—reads mapped to paralogous rather than homologous loci in reference.
	poor filters
false-negatives	failure to assemble repetitive regions
	systematic sequencing errors
	poor filters
	distantly related reference sequence, such that variants cannot be called throughout the genome

by recording the mobility of proteins or DNA fragments separated by electrophoresis (such as PFGE or multilocus enzyme electrophoresis). The data from such methods have long been recognized as flawed because of the difficulty of comparing results among laboratories (despite attempts to standardize methods [7]). In response, the multi-locus sequence typing (MLST) approach was developed to detect sequence variation at a small fraction of the genome—typically seven loci—and use it to characterize and define the strain [1,19]: alleles at the MLST loci can usually be unambiguously determined by Sanger sequencing and compared with online databases. However, MLST is limited by the fact that it uses such a small fraction of the genome; hence the availability of an equivalent system for genome sequences would be a great improvement.

4. Genomes in long term or global epidemiology

The experience of the MLST databases demonstrates that community contributions to a common resource can generate large and valuable datasets [20]. And while the resulting samples are flawed (being voluntary contributions rather than the results of systematic surveillance, and hence biased for population genetic purposes) they are far better than none at all. However, the optimal form of the processed genomic data that could be made available is not clear. Where no appropriate reference sequence is available for mapping, or the samples being analysed are too diverse for such a technique to produce precise results, comparisons of de novo draft genome assemblies may be more appropriate. However, such sequences are inevitably imperfect, and their usefulness is further limited by the difficulty of multiple genome alignment. Hence, studies of diverse collections of bacteria have often resorted to identifying and aligning groups of orthologous genes, resulting in data that resemble whole genome MLST schemes [21] that nevertheless show potential for use even in local outbreak resolution [22].

Given that methods for such large-scale analysis exist, the growing use of genome sequencing to inform infection control in individual hospitals [16] and communities [14] could thus be combined into a semi-systematic dataset that would constitute a broader dataset for following the more 'global' concerns of epidemiology. Such an approach has had success in taking local surveillance data on antibiotic resistance, and making it more globally relevant through international programs such as WHOnet, designed to standardize and improve practice in measuring resistance in clinical laboratories [23–25] but with potential for additional functions including outbreak detection [26]. It is reasonable to think a similar approach could be beneficial for genomics.

5. The importance of context

Global datasets can also be used to inform local epidemiology. Following an outbreak of disease, investigators frequently wish to know where it originated. Comparisons of outbreak genomes with others that have already been sequenced can shed light on this, but such comparisons require a larger sample so that the outbreak can be placed in context. Following the cholera epidemic that arose in Haiti following the 2010 earthquake [27], rumours suggested that peacekeepers from Nepal, a region of endemic cholera transmission, travelling as part of the United Nations relief operation could have inadvertently introduced the outbreak strain. Genome sequencing confirmed a close relationship between the Haiti outbreak strain and recent *V. cholerae* isolates from Nepal, consistent with this hypothesis [28]. It should be noted that existing samples of diversity are not such that we can definitively link the outbreak to Nepal through sequence data alone—although strong evidence supports a South Asian origin, closely related genomes have also been sampled from Cameroon [29]. However, the combination of traditional and genomic epidemiology suggests a probable Nepalese origin [3,30].

Another example is the large *E. coli* O104 : H4 outbreak that took place in northern Europe in the summer of 2011 [31]. Following the rapid determination of a genome sequence for the outbreak strain [32], the closest known

relative for which genome sequence was available was found to be an O104 : H4 isolate from the Central African Republic [33]. The outbreak was linked to a sprout farm in Lower Saxony and specifically a shipment of fenugreek seeds originating from Egypt [34] (another smaller outbreak in France was traced back to the same shipment [35]). The link to the African isolate could be taken to be consistent with the hypothesis that contamination occurred in Egypt, but we should be very wary of drawing links between samples from such different locations, and from a region (North Africa) from which very few samples are available. We do not have a sufficient sample of the diversity of O104 : H4 strains to be confident that such phylogeographical conclusions would not be spurious. Other isolates with similar PFGE profiles had been reported [36], but had not been sequenced prior to the outbreak. We lack the necessary context that would be provided by a better sample of *E. coli* diversity in Africa, or indeed anywhere else. Genomes from such a small number of isolates may be valuable in defining the gene content of the outbreak strain and the molecular roots of virulence [37], but we should be cautious of simple epidemiological conclusions.

6. Recombination

Inevitably, any effort to track the movement of bacteria using sequence data risks being confounded by the movement of sequence between bacteria. It was long thought that recombination, with the exception of plasmid transfer, did not happen in bacteria. However, this is not the case. Unlike sexually producing eukaryotes in which recombination typically occurs through the reciprocal exchange of genetic material at meiosis, in bacteria genetic material can be transferred non-reciprocally between members of the same or different named species. In some cases, usually described as horizontal gene transfer (HGT), recombination can insert genes with functions that were not originally present in the recipient strain. In others, relatively short tracts of DNA homologous to those that they replace are acquired in a process similar to gene conversion.

The contribution of recombination to the evolution of many bacteria is now well recognized [18], as is the potential for recombination to confound phylogenetic reconstruction or even render it impossible beyond very short linkage differences [38]. Because a single recombination event can introduce multiple polymorphisms, it will confound attempts to accurately measure the rate with which substitutions arise by mutation. If recombination occurs at an approximately homogenous rate relative to substitutions, and each event increases a taxon's divergence from the common ancestor, then using samples collected over time it may be possible to deduce a combined rate of recombination and substitution over the course of a lineage's evolution. This will result in a measurably evolving population [39], in which the rate of divergence is a statistical property of how genetic distance between two isolates relates to the time since they shared a common ancestor, rather than a mechanistic statement about mutation rates. However, recombination has the potential to disrupt or destroy any correlation between time and divergence, and it is clear that the frequency of recombination can be highly irregular [40] and the number of polymorphisms that may be imported [41] can be highly variable. This has the potential to obscure any correlation between divergence and time since a common ancestor [42]. Furthermore, recombination between two lineages in the dataset will also produce spurious similarities between them, making them appear more closely related than they are in reality [43], and thereby scramble phylogenetic signal. The resulting incorrect tree topology will further bias estimates of clock rates.

Indeed, a classic study of sequences of housekeeping loci in six bacterial pathogens found that in four of them, the phylogeny of one locus was no more similar to that of others than would be expected by chance [38]. It was concluded that the rate of homologous recombination was such that it had obliterated all phylogenetic signal beyond the relatively short term. An attractive and powerful way of handling this problem is to somehow identify those sites in an alignment that have been introduced by recombination and remove them. What is left will then reflect the 'clonal frame'—being those sites at which observed variation is the result of vertical inheritance and reflects recent common descent. These may then be used to estimate phylogenies free of the distorting influence of recombination.

ClonalFrame, a statistical model for the identification and removal of recombinant sequence from MLST data, was described in an influential paper attempting to deal with the problems outlined above [44]. This software uses a Bayesian approach to fit a sophisticated model of evolution to a set of sequences using a Markov chain Monte Carlo (MCMC) method. Recombination events are identified as segments of the alignment containing an atypically high density of polymorphisms, similar in principle to Sawyer's Runs Test [45]. However, as the model contains a large number of parameters, the algorithm is very slow with large whole genome datasets. Hence simpler, faster methods that take a similar approach to identifying SNP-dense regions have recently been developed and applied to large genome datasets of two lineages of *S. pneumoniae* [40,42].

As all of these methods rely on identifying regions with a high density of SNPs, they are appropriate for the analysis of closely related isolates. As isolates in such collections share a recent common ancestor, there is a relatively low background level of SNPs caused by point mutation, and there has been little time for selection to skew these SNPs away from a uniform distribution around the chromosome. This makes high densities of SNPs acquired through sequence import easy to identify. However, such methods need to be modified for the analysis of the more distantly related isolates that need to be compared to answer more long term, global epidemiological questions. Such collections encompass the diversity typically expected between donor and recipient in recombination events, hence the density of polymorphisms distinguishing taxa is not necessarily expected to be higher than average in regions corresponding to recently occurring recombinations. Hence, it is no longer possible to identify recombinations through spatial arrangements of SNPs alone.

Algorithms that address this problem by identifying recombinations as regions of unexpectedly high similarity between recipients and potential donors have been applied in order to detect sequence exchange between branches of a phylogeny. This is achieved by the program ClonalOrigin [46] through post-processing of the output of ClonalFrame, and as such is dependent on a successful ClonalFrame analysis. Localized regions of sequence where the same mutations

are observed to occur in parallel on different branches of the tree suggest there is likely to have been an exchange of DNA in the history of the studied collection. Such an approach has been used to trace the history of a diverse collection of *Chlamydia trachomatis* isolates [47].

An alternative approach has been implemented within a rapidly converging Bayesian framework by the program NextGenBRAT [48], which initially defines populations characterized by distinct allele frequencies in the data, and then seeks evidence in each sequence for runs of polymorphisms (i.e. alleles) that are characteristic of populations other than the presumed ancestral one. All methods that look for exchange of sequence occurring in the history of a set of isolates provide useful information on the pairings involved in, and directionality of, recombination. Yet the caveat remains that the power to detect these recombinations is reliant on isolates closely related to the donor and recipient being present in the collection. Hence, as with determining the geographical source of an isolate, identifying the bacterial source of DNA is limited by the original sampling.

Finally, we should recall that while recombination is usually considered a nuisance that impedes straightforward analysis, it is just as much a record of the evolutionary history of a genome as mutation, albeit being more difficult to interpret. Recombinant sequence may be of special interest to epidemiologists because it confers attributes such as drug resistance, or because it can reflect contact between the donor and the recipient strain. However, as the interpretation of horizontally transferred regions is complicated by the huge diversity of MGEs [49] and the highly variable, mosaic form of homologous recombination [41], exploiting rather than removing the signal of recombination will be a challenge.

7. Concluding remarks

The potential for genomic data to improve and inform the epidemiology of bacterial pathogens is vast, but whether it is realized will depend on the issues of context and consistency discussed above.

Our ability to place a genome into its proper evolutionary context, understanding its probable origins, and the processes that have given rise to it (and any associated properties such as virulence) all depend on sampling. At present, we have an extremely narrow and biased sample of the relevant diversity. As such, efforts to improve sampling must be supported. While this is an important goal, another one linked to it is care in the interpretation of results. We will only be able to accurately define ancestry with larger, better samples containing bacteria that share a recent common ancestor with the isolate of interest.

How these samples should be sequenced is another challenge in the form of the diverse sequencing platforms and approaches to analysis. Two samples of the same organism, sequenced by different methods, can yield drastically different results [2,37]. Hence, there are advantages to introducing a degree of consistency, in order to facilitate easier comparisons of results between publications or studies [17]. A lesson from the success of prior methods, notably MLST, is that consistency also leads to larger samples, and thereby allows data collected for answering local questions to be applied to global analyses. While making raw data available to other researchers is already required for publication in most cases, releasing processed forms would greatly enhance its practical availability to a large proportion of the research community. Yet there is little incentive for such public release of processed data at present, although this is a point that could be addressed by funding bodies or journals.

Developing a standard approach is likely to be highly problematic, however, with so many competing alternatives. Objective evaluations of different sequencing platforms [12] and analysis algorithms (e.g. the 'Assemblathon' [50]) are valuable, but bacteria with different genetic and epidemiological characteristics may merit different approaches. Any standard method would need to be under frequent review, as the quality and quantity of genome data continue to improve, and the sophistication and effectiveness of novel algorithms facilitates more detailed and accurate analysis. Otherwise, consistency may come at the cost of stagnation of analysis techniques and therefore suboptimal use of the available data.

Making the best use of the data is crucial for robust inference of relationships between isolates. This is crucial for constructing phylogenies, calculating clock rates and understanding population structure. Our attempts to clean our data of these sources of the confounding signal of recombination and the noise of sequencing and analysis errors are inevitably imperfect. This re-emphasizes the utility of applying consistent methods to data of consistent quality. In so doing, we generate alignments in which the effects of noise are uniformly distributed between the data. Once such filters have been applied, the output statistics will be limited in their application to the proportion of the genome that can be accurately inferred at a high level of confidence given these constraints.

Yet even with perfect sequences there are still limitations. It is notable that the way molecular epidemiologists dealt with recombination in the pre-genomic era was to focus attention on the very tips of the tree [51]. For all the greater resolution afforded by genomic data, this remains the case as we examine individual lineages. We, as yet, have no decent way of reconstructing the ancestral recombination graph for diverse genomic datasets; hence, while recombination can be satisfactorily dealt with in outbreak analyses, it remains a problem when approaching more global questions. In fact, this may not be possible. In a sample of 240 genomes from just one antibiotic resistant clone, 74 per cent of the genome had been replaced by recombination in at least one isolate [42]. The inescapable conclusion is that, given the rates of recombination that are believed to be characteristic of pneumococcus and some other species, the clonal frame will shrink as additional strains are added and dwindle to nothing. In other words, there will be no sites that have been solely inherited by vertical descent since the most recent common ancestor of any sample of reasonable size. This is not a barrier to the epidemiology of the pathogen, which focuses on relatively recent events, but it is a rebuke to the pretense that we will always be able to adequately resolve deep branching lineages in recombinogenic organisms.

N.J.C. is the recipient of an AXA postdoctoral fellowship. Y.H.G. is supported by a NERCE-BEID career development award and an ASTDA career development award. S.R.H. is funded by Wellcome Trust grant no. 098051. W.P.H. acknowledges funding from NIH/NIGMS GM088558-01 for the MIDAS Center for Communicable Disease Dynamics at HSPH. The authors would like to thank Jukka Corander for helpful discussions.

References

1. Enright MC, Spratt BG. 1999 Multilocus sequence typing. *Trends Microbiol.* **7**, 482–487. (doi:10.1016/S0966-842X(99)01609-1)
2. Grad YH *et al*. 2012 Genomic epidemiology of the *Escherichia coli* O104 : H4 outbreaks in Europe, 2011. *Proc. Natl Acad. Sci. USA.* **109**, 3065–3070. (doi:10.1073/pnas.1121491109)
3. Chin CS *et al*. 2011 The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42. (doi:10.1056/NEJMoa1012928)
4. McGee L *et al*. 2001 Nomenclature of Major Antimicrobial-Resistant Clones of *Streptococcus pneumoniae* defined by the Pneumococcal Molecular Epidemiology Network. *J. Clin. Microbiol.* **39**, 2565–2571. (doi:10.1128/JCM.39.7.2565-2571.2001)
5. Willshaw GA, Smith HR, Cheasty T, O'Brien SJ. 2001 Use of strain typing to provide evidence for specific interventions in the transmission of VTEC O157 infections. *Int. J. Food Microbiol.* **66**, 39–46. (doi:10.1016/S0168-1605(00)00511-0)
6. Gerner-Smidt P, Hise K, Kincaid J, Hunter S, Rolando S, Hyytia-Trees E, Ribot EM, Swaminathan B. 2006 Spring PulseNet USA: a five-year update. *Foodborne Pathog. Dis.* **3**, 9–19. (doi:10.1089/fpd.2006.3.9)
7. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV. 2001 PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* **7**, 382–389.
8. Crawford JT. 2003 Genotyping in contact investigations: a CDC perspective. *Int. J. Tuberc. Lung Dis.* **7**(12 Suppl. 3), S453–S457.
9. Keim P, Smith KL. 2002 Bacillus anthracis evolution and epidemiology. *Curr. Top. Microbiol. Immunol.* **271**, 21–32.
10. Achtman M. 2008 Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* **62**, 53–70. (doi:10.1146/annurev.micro.62.081307.162832)
11. van Embden JD *et al*. 1993 Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409.
12. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012 Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439. (doi:10.1038/nbt.2198)
13. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012 A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genom.* **13**, 341. (doi:10.1186/1471-2164-13-341)
14. Gardy JL *et al*. 2011 Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739. (doi:10.1056/NEJMoa1003176)
15. Keim P, Van Ert MN, Pearson T, Vogler AJ, Huynh LY, Wagner DM. 2004 Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infect. Genet. Evol.* **4**, 205–213. (doi:10.1016/j.meegid.2004.02.005)
16. Koser CU *et al*. 2012 Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275. (doi:10.1056/NEJMoa1109910)
17. Achtman M. 1996 A surfeit of YATMs? *J. Clin. Microbiol.* **34**, 1870.
18. Feil EJ, Spratt BG. 2001 Recombination and the population structures of bacterial pathogens. *Annu. Rev. Microbiol.* **55**, 561–590. (doi:10.1146/annurev.micro.55.1.561)
19. Maiden MC *et al*. 1998 Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145. (doi:10.1073/pnas.95.6.3140)
20. Maiden MC. 2006 Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–588. (doi:10.1146/annurev.micro.59.030804.121325)
21. Jolley KA, Maiden MC. 2010 BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595. (doi:10.1186/1471-2105-11-595)
22. Jolley KA, Hill DM, 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. (doi:10.1128/JCM.01312-12)
23. O'Brien TF, Stelling JM. 1995 WHONET: an information system for monitoring antimicrobial resistance. *Emerg. Infect. Dis.* **1**, 66. (doi:10.3201/eid0102.950209)
24. O'Brien TF, Stelling JM. 1996 WHONET: removing obstacles to the full use of information about antimicrobial resistance. *Diagn. Microbiol. Infect. Dis.* **25**, 162–168. (doi:10.1016/S0732-8893(96)00139-3)
25. Stelling JM, O'Brien TF. 1997 Surveillance of antimicrobial resistance: the WHONET program. *Clin. Infect. Dis.* **24**(Suppl. 1), S157–S168. (doi:10.1093/clinids/24.Supplement_1.S157)
26. Stelling J *et al*. 2010 Automated use of WHONET and SaTScan to detect outbreaks of Shigella spp. using antimicrobial resistance phenotypes. *Epidemiol. Infect.* **138**, 873–883. (doi:10.1017/S0950268809990884)
27. 2010 Cholera outbreak—Haiti, October 2010. *MMWR Morb. Mortal Weekly Rep.* **59**, 1411.
28. Hendriksen RS *et al*. 2011 Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* **2**, e00157-11. (doi:10.1128/mBio.00157-11)
29. Reimer AR *et al*. 2011 Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa. *Emerg. Infect. Dis.* **17**, 2113–2121. (doi:10.3201/eid1711.110794)
30. Frerichs RR, Keim PS, Barrais R, Piarroux R. 2012 Nepalese origin of cholera epidemic in Haiti. *Clin. Microbiol. Infect.* **18**, E158–E163. (doi:10.1111/j.1469-0691.2012.03841.x)
31. Frank C *et al*. 2011 Epidemic profile of Shiga-toxin-producing Escherichia coli O104 : H4 outbreak in Germany. *N. Engl. J. Med.* **365**, 1771–1780. (doi:10.1056/NEJMoa1106483)
32. Rohde H *et al*. 2011 Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104 : H4. *N. Engl. J. Med.* **365**, 718–724. (doi:10.1056/NEJMoa1107643)
33. Mossoro C, Glaziou P, Yassibanda S, Lan NT, Bekondi C, Minssart P, Bernier C, Le Bouguenec C, Germani Y. 2002 Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent *Escherichia coli* in adults infected with human immunodeficiency virus in Bangui, Central African Republic. *J. Clin. Microbiol.* **40**, 3086–3088. (doi:10.1128/JCM.40.8.3086-3088.2002)
34. Buchholz U *et al*. 2011 German outbreak of *Escherichia coli* O104:H4 associated with sprouts. *N. Engl. J. Med.* **365**, 1763–1770. (doi:10.1056/NEJMoa1106482)
35. Gault G *et al*. 2011 Outbreak of haemolytic uraemic syndrome and bloody diarrhoea due to *Escherichia coli* O104 : H4, south-west France, June 2011. *Euro Surveill.* **16**, pii19905.
36. Scheutz F, Nielsen EM, Frimodt-Moller J, Boisen N, Morabito S, Tozzoli R, Nataro JP, Caprioli A. 2011 Characteristics of the enteroaggregative Shiga toxin/verotoxin-producing Escherichia coli O104 : H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, May to June 2011. *Euro Surveill.* **16**, pii19889.
37. Rasko DA *et al*. 2011 Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717. (doi:10.1056/NEJMoa1106920)
38. Feil EJ *et al*. 2001 Recombination within natural populations of pathogenic bacteria: short term empirical estimates and long-term phylogenetic consequences. *Proc. Natl Acad. Sci. USA* **98**, 182–187. (doi:10.1073/pnas.98.1.182)
39. Gray RR, Pybus OG, Salemi M. 2011 Measuring the temporal structure in serially sampled phylogenies. *Methods Ecol. Evol.* **2**, 437–445. (doi:10.1111/j.2041-210X.2011.00102.x)
40. Golubchik T *et al*. 2012 Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat. Genet.* **44**, 352–355. (doi:10.1038/ng.1072)
41. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. 2012 A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog.* **8**, e1002745. (doi:10.1371/journal.ppat.1002745)
42. Croucher NJ *et al*. 2011 Rapid pneumococcal evolution in response to clinical interventions.

Science **331**, 430–434. (doi:10.1126/science.1198545)

43. Smith J. 1999 The detection and measurement of recombination from sequence data. *Genetics* **153**, 1021–1027.
44. Didelot X, Falush D. 2007 Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266. (doi:10.1534/genetics.106.063305)
45. Sawyer S. 1989 Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538.
46. Didelot X, Lawson D, Darling A, Falush D. 2010 Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–1449. (doi:10.1534/genetics.110.120121)
47. Harris SR *et al*. 2012 Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. *Nat. Genet*. **44**, 413–419. (doi:10.1038/ng.2214)
48. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012 Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6. (doi:10.1093/nar/gkr928)
49. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011 The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625. (doi:10.1101/gr.122705.111)
50. Earl D *et al*. 2011 Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* **21**, 2224–2241. (doi:10.1101/gr.126599.111)
51. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. 2004s eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186**, 1518–1530. (doi:10.1128/JB.186.5.1518-1530.2004)