Review

Insights from genomic comparisons of genetically monomorphic bacterial pathogens

Mark Achtman*

Environmental Research Institute and Department of Microbiology, University College Cork, Cork, Ireland

Some of the most deadly bacterial diseases, including leprosy, anthrax and plague, are caused by bacterial lineages with extremely low levels of genetic diversity, the so-called ‘genetically monomorphic bacteria’. It has only become possible to analyse the population genetics of such bacteria since the recent advent of high-throughput comparative genomics. The genomes of genetically monomorphic lineages contain very few polymorphic sites, which often reflect unambiguous clonal genealogies. Some genetically monomorphic lineages have evolved in the last decades, e.g. antibiotic-resistant Staphylococcus aureus, whereas others have evolved over several millennia, e.g. the cause of plague, Yersinia pestis. Based on recent results, it is now possible to reconstruct the sources and the history of pandemic waves of plague by a combined analysis of phylogeographic signals in Y. pestis plus polymorphisms found in ancient DNA. Different from historical accounts based exclusively on human disease, Y. pestis evolved in China, or the vicinity, and has spread globally on multiple occasions. These routes of transmission can be reconstructed from the genealogy, most precisely for the most recent pandemic that was spread from Hong Kong in multiple independent waves in 1894.

Keywords: genetically monomorphic bacterial pathogen; plague; ancient DNA; Black Death; phylogeography; historical reconstruction

1. INTRODUCTION
Evolutionary studies on microbes span the range of genetic diversity between the two microbial domains of life that evolved billions of years ago [1] down to genetic changes within a lineage that accumulate during the infection of a single diseased human over several weeks [2–4]. Clearly, the predominant genetic mechanisms are likely to be distinct over the extremes of such an extensive range. However, it is often not generally appreciated that different dynamics are also expected between comparisons of microbial species, even if they are closely related, versus analyses within a single species, because inter-species comparisons reflect fixed substitutions in genetically isolated taxa, whereas intra-species analyses reflect quantitative changes in frequency owing to neutral processes, such as genetic drift [5]. Within a species, the population structures also differ between taxa where homologous recombination is frequent, such as Helicobacter pylori [6], and those where the primary driving force is mutation, such as Yersinia pestis [7].

In some microbial species, the combination of homologous recombination, horizontal gene transfer from other species and deletions results in an ‘open’ pan-genome. An open pan-genome increases in size as more genomes are analysed, accompanied by reduction of the number of coding sequences that are common to all isolates (core-genome) [8]. Open pan-genomes have been described in multiple bacterial species, including Streptococcus agalactiae [9], and have become the current paradigm for the adaptability of bacteria to their ecological environment. Escherichia coli also has an open pan-genome [10]: a recent comparison of 29 genomes revealed greater than 17 500 genes in the pan-genome versus only 2356 core genes [11]. The number of genes within the pan-genome of E. coli continues to increase by 360 new genes for each new genome that is analysed.

Although E. coli as a whole has an open pan-genome, it also includes discrete and genetically isolated clades, such as the O157:H7 lineage that infects cattle and occasionally humans, in whom it can induce haemorrhagic colitis and the haemolytic–uremic syndrome [12]. The population dynamics of this clade differs from those of the entire species, and is typical of a class of microbes that I have designated as ‘genetically monomorphic’ [13]. Multiple other causes of dramatically life-threatening diseases are also genetically monomorphic (table 1). Similar to O157:H7 E. coli, some of these taxa correspond to lineages within a species of greater diversity, e.g. the cause of typhoid fever, serovar Typhi of Salmonella enterica [21] or the cause of pandemic cholera, O1 or O139 Vibrio cholerae [22]. In contrast, other genetically monomorphic lineages have been accorded species status because of the distinctive diseases they cause, e.g. the causes of plague (Y. pestis [23]), anthrax

* m.achtman@ucc.ie

One contribution of 14 to a Discussion Meeting Issue ‘Immunity, infection, migration and human evolution’.
Mycobacterium leprae that a parental species has not been identified, such still other species also have similar properties, except [26].

Table 1. Single nucleotide polymorphisms (SNPs) in core genomes among genetically monomorphic lineages.

<table>
<thead>
<tr>
<th>lineage</th>
<th>SNPs (genomes)</th>
<th>core genome size (MB)</th>
<th>SNPs per 100 kb</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycobacterium leprae</td>
<td>222 (7)</td>
<td>3.3</td>
<td>7</td>
<td>[14,15]</td>
</tr>
<tr>
<td>Burkholderia mallei</td>
<td>515 (7)</td>
<td>5.7</td>
<td>9</td>
<td>[16]</td>
</tr>
<tr>
<td>Bordetella pertussis</td>
<td>471 (7)</td>
<td>4.1</td>
<td>12</td>
<td>[17]</td>
</tr>
<tr>
<td>Yersinia pseudotuberculosis</td>
<td>1364 (17)</td>
<td>4.8</td>
<td>28</td>
<td>[7]</td>
</tr>
<tr>
<td>Salmonella enterica serovar Typhi</td>
<td>1964 (19)</td>
<td>4.4</td>
<td>44</td>
<td>[18]</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td>2798 (19)</td>
<td>5.5</td>
<td>51</td>
<td>[19]</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>9945 (22)</td>
<td>4.4</td>
<td>226</td>
<td>[20]</td>
</tr>
</tbody>
</table>

(Bacillus anthracis [24]), glands (Burkholderia mallei [25]) and tuberculosis (Mycobacterium tuberculosis [20]), although each of them corresponds to a lineage within a species of greater diversity (Yersinia pseudotuberculosis, Bacillus cereus, Burkholderia pseudomallei and the M. tuberculosis complex, respectively). Finally, still other species also have similar properties, except that a parental species has not been identified, such as the causes of leprosy (Mycobacterium leprae [14]) and Buruli ulcer (Mycobacterium ulcerans [26]).

All these genetically monomorphic lineages are associated with life-threatening disease, suggesting that there might be an evolutionary link between disease intensity and transmission [27]. Alternatively, their association with severe disease may simply reflect a discovery bias. Genetically monomorphic lineages were discovered because many medical microbiologists have dedicated themselves since the late-nineteenth century to the epidemiological tracking of pathogens of animals. Genetically monomorphic lineages are now also being identified among microbial pathogens of plants, where epidemiological tracking was previously not pursued with the same intensity. In contrast, scientists working with environmental microbes have only rarely attempted to catalogue the global diversity of individual lineages from the environment. As a result, it is unclear whether an association between genetically monomorphic lineages and disease reflects more than the biased interests of microbiologists. However, their very existence makes them particularly suitable for comparative genomic analyses, which have provided important insights into the population dynamics of bacteria in nature.

2. DARWINIAN SELECTION DURING ADAPTATION

Experimental studies have provided convincing evidence for the importance of Darwinian selection in the laboratory. Parallel lineages of E. coli that evolved after daily sub-cultivation for a total of 20 000 generations underwent convergent evolution to increased fitness through non-synonymous mutations within the same genes [28]. In contrast, neutral evolution associated with synonymous mutations was less frequent, indicating that selection for adaptation is the dominant evolutionary dynamic under these conditions. Similarly, after alternating cycles of static and aerated growth, Pseudomonas aeruginosa evolved adaptability as manifested by an ability to reversibly generate variants that were particularly suitable for each of the two environments [29]. Laboratory experiments often select for the so-called mutators, in which the general frequency of genetic variants is elevated [28,30]. Indeed, simply storing Salmonella enterica in agar tubes over decades selects for multiple independent mutations that result in increased fitness, at least partially through a combination of nutritional deficiencies plus the increased production of bacteriophages that can kill competitors [31]. Thus, observations from evolution in the laboratory strongly support the thesis that the predominant evolution within bacteria is due to selection for greater fitness. Based on these and other, similar observations, Darwinian selection in bacteria is thought to result from ‘periodic selection’, which consists of sequential cycles of purification of genetic diversity through the replacement of the existing population by monophyletic, fitter variants [32].

Similar phenomena have also been observed outside the laboratory during long-term infections of single humans. For example, 19 non-synonymous point mutations and 10 indels accumulated over 12 weeks of (unsuccessful) antibiotic therapy of a patient suffering from Staphylococcus aureus endocarditis, as did six synonymous point mutations [3]. Many of these mutations were in genes that are linked to antibiotic resistance or encode transcriptional global regulators. Similarly, Ps. aeruginosa infection of the lungs of patients with cystic fibrosis is accompanied by the sequential accumulation of non-synonymous mutations over several years in parallel to changes in growth rate, reduced expression of lipopolysaccharide, mucoidy and the accumulation of mutators [2].

Darwinian selection has also been observed in natural bacterial populations. The extensive use of antibiotics for treating bacterial diseases of humans and domesticated animals, and as a growth supplement in animal husbandry, has resulted in a very dramatic increase in the frequencies of antibiotic-resistant pathogens. Over a 20 year period, multiple independent mutations that reduce susceptibility to fluoroquinolones have occurred at the same nucleotide(s) in the gyrA gene within all the lineages of serovar Typhi [21]. Similarly, genetic islands encoding resistance to meticillin and other antibiotics have been imported repeatedly in recent decades within one discrete lineage of S. aureus [33]. Immune evasion has resulted in repeated imports from other neisseriae into Neisseria meningitidis of antigenic variants of the tbpB gene, which encodes an immunogenic transferrin-binding outer membrane protein [34,35]. Another example of immune selection may be provided by Streptococcus pneumoniae, where
a vaccine against acapsular polysaccharide seems to have selected for variant lineages that express a serologically distinct polysaccharide [36]. Some of the examples presented above also exemplify that Darwinian selection for one trait often only leads to transient selection in nature. Many of the selected variants are apparently less fit than their parents. Despite an increase in the frequency of mutant gyrA alleles in Typhi, each progressive node within the genealogy also included isolates that lacked these mutations [21]. Twenty years of selection byfluoroquinolone treatment have apparently not yet been able to select epistatic mutations that can ameliorate this reduced fitness, and the basal genealogy of Typhi remains susceptible to fluoroquinolones. Similarly, the variant tcpB genes imported into N. meningitidis were repeatedly lost from the population during successive bottlenecks caused by geographical spread [35]. Antibiotic resistance in M. tuberculosis is predominantly associated with particular lineages, in which it has smaller effects on fitness [37]. The progressive changes in phenotype within P. aeruginosa during infection of the cystic fibrosis lung may represent source–sink dynamics: the bacteria infecting the sink (the lung) are apparently usually incapable of further transmission to other patients. In those rare cases in which further transmission does occur, continued evolution is neutral after a few initial years of adaptation to a human host [38]. Thus, the lessons taught by experimental evolution may be largely restricted to a short phase of adaptation to a novel environment, and are not applicable to the full dynamics of evolution in existing natural populations.

3. PURIFYING SELECTION IN NATURAL POPULATIONS

Instead of periodic selection of fitter variants, the population dynamics of genetically monomorphic pathogens largely seem to represent neutral evolution in the form of mild purifying selection. Genomic comparisons of approximately 20 genomes from each of Typhi, Y. pestis and M. tuberculosis have revealed comparable rates of accumulation of synonymous and non-synonymous mutations [7,18,20]. Instead of convergent evolution through repeated non-synonymous mutations in the same genes, their genealogies show an extremely low, almost negligible level of homoplasies, except for a few genes involved in antibiotic resistance, as described above. T cell epitopes in P. aeruginosa are less variable than the remainder of the genome [20], again suggesting purifying rather than diversifying (Darwinian) selection. Gene acquisition by horizontal genetic exchange is absent or exceedingly rare, except for the occasional acquisition of bacteriophages and plasmids, which seem to act as selfish DNAs. Similar conclusions also apply to long-term persistence and transmission of P. aeruginosa in the lungs of patients with cystic fibrosis after the initial period of adaptation [38]. If Darwinian selection occurs within genetically monomorphic pathogens, then it is both rare and of only limited extent in comparison with the sequential, neutral accumulation of single nucleotide polymorphisms (SNPs) that is associated with clonal diversification.

The genealogies of Typhi, Y. pestis and M. tuberculosis, are clonal and lack all traces of recent homologous recombination. They seem to have been genetically isolated over the majority of their ancestry, although their genomes show traces of recombination that occurred before these lineages arose [39,40]. Searching for signs of diversifying selection is relatively straightforward in such clonal genealogies, but is technically much more difficult with microbes that recombine more frequently because each recombination event can introduce multiple clustered polymorphisms whose importance for fitness is unclear. Of the few comparative genomic analyses of recombining monomorphic lineages that are currently available, none has revealed particularly dramatic, genome-wide signals of Darwinian selection [36,41–43], except for further examples of the acquisition of bacteriophages, mutations or imports leading to antibiotic resistance, and changes in the frequency of serological variants which can evade vaccination pressures. Similarly, none of the comparative genomic analyses has provided clear evidence supporting the Darwinian evolution of fitter variants and/or periodic selection, which would have been expected from laboratory experiments. It seems that the population dynamics in nature differ from our preconceptions, at least for pathogens that cause disease. Possibly these differences are dictated by the fact that each transmission event between humans or other animals represents a bottleneck, and that geographical transmission to distinct areas imposes additional bottlenecks on the genetic diversity of pathogens.

4. MUTATION CLOCK RATES

How old are genetically monomorphic lineages? Until recently, the time since the most recent common ancestor (TMRCA) was generally calculated on the basis of a universal bacterial clock rate [44]. However, the universality of this clock rate has now been discredited for inter-species comparisons [6,45] and an intra-species clock rate would be expected to differ from an inter-species clock rate in any event. Recent data shows that the intra-species clock rate of different bacterial lineages varies by several orders of magnitude [6], ranging from $9 \times 10^{-9}$ to $3 \times 10^{-3}$ mutations per nucleotide per year (table 2). The time span for which these clock rates were calculated extends only to the last few hundred years. Theoretical considerations and data from viruses indicate that short-term clock rates are faster than clock rates over longer time periods [50,51]. I only know of one attempt to perform such a comparison within bacteria, namely for H. pylori [6]. Paired samples from individual humans or family members that had diverged over a time period of up to 78 years yielded clock rate estimates of $1.4 \times 10^{-6}$ to $4.5 \times 10^{-6}$. By contrast, long-term clock rates that were calibrated by archaeological dating of human migrations over millennia were $3 \times 10^{-7}$, which is slower by 5–17-fold than the short-term rates.

Some of the clock rates in genetically monomorphic pathogens are fast enough that real time evolution can be measured within those lineages by genomic

Phil. Trans. R. Soc. B (2012)
comparisons of dated samples (table 2). The resulting estimates of TMRCA coincide with the epidemiological history for two lineages of *S. aureus* that have emerged in recent decades [42,48]. Over longer time periods, additional information will be needed for reliable estimates, such as can be obtained from phylogeographic analyses and ancient DNA.

5. PHYLLOGEographic patterns and historical reconstructions of Plague epidemics

Phylogeographic patterns of a handful of housekeeping genes from *H. pylori* correlate with ancient human migrations, and have been used to reconstruct patterns of spread of these bacteria over the last 60 000 years [52,53]. Phylogeographic patterns are also apparent for many other bacterial pathogens, ranging from relatively recent epidemic waves caused by serogroup A *N. meningitidis* [35,54] and O1 V *cholerae* [55] to global transmissions that have been attributed to old human migrations and/or trade for *M. tuberculosis* [56], *M. leprae* [14] and *B. anthracis* [57]. However, possibly the most detailed historical reconstruction of historical patterns of spread currently available is for plague.

Historical records document three pandemic waves of plague. The Justinianic Pandemic reached Alexandria from Northeast Africa in 541, and caused recurrent epidemic waves through most of Europe over the following 200 years [58]. In 1346, the Black Death was imported to Europe from Central Asia, where it continued to cause epidemics until the early eighteenth century [59]. The third pandemic began in central China in the mid-nineteenth century and spread globally in 1894 from Hong Kong via marine shipping [60].

A causal attribution of the third pandemic to *Y. pestis* is indisputable owing to extensive microbiological investigations [60,61]. The Black Death was also caused by *Y. pestis*, as documented by ancient DNA studies that demonstrated the existence of specific DNA sequences in the dental pulp [62–66] and/or the F1 capsular protein in the bones [66–68] of skeletons from dated mass graves distributed throughout Europe. Similar results from ancient DNA analyses have also been obtained for the Justinianic Pandemic. However, additional confirmation of the validity of these observations would be desirable because genotypes from skeletons from the Black Death and the Justinianic Pandemic were indistinguishable in those studies [62,69], whereas that genotype should not yet have existed at the time of the Black Death [7], and different, older genotypes were found by a later study [66,70].

Based on historical records of human diseases, historians have generally concluded that plague, and by inference *Y. pestis*, originated in Africa [59]. However, the primary hosts of *Y. pestis* are various rodents, within many of which it causes sylvatic disease. These rodent foci are scattered throughout the world, most extensively through a large portion of the former Soviet Union, Mongolia and China [71]. Fleas on infected rats have been implicated in the transmission of plague to humans in India and Madagascar in the third pandemic. However, rats are only a secondary host, and are therefore not necessarily relevant to the evolution or history of *Y. pestis*. In turn, historical records of human infections can only reflect those occasions when *Y. pestis* spread from its sylvatic environment, and accidentally caused epidemicisease in humans. The dynamics of such occasions is unknown although attempts have been made to associate them with climactic changes [72].

Microbiological analyses of *Y. pestis* have yielded a bewildering variety of designations that differ with the country of origin of the microbiologists [73]. A common language has now been implemented based on the population structure identified by genomic analyses on a globally representative sample [7,74]. Initial results with 17 genomes were recently published [7] and manuscripts on results with additional genomes are being submitted. The genealogy of *Y. pestis* is clonal, rooted in its ancestral species, *Y. pseudotuberculosis* [23]. Most nucleotides in the core genome are identical, and only very few SNPs distinguish individual isolates. However, because the frequency of homoplastic mutations is negligible, the genealogy is unambiguous and defines a root branch, branch 0, as well as three more recent branches, branches 1 to 3 (figure 1) [7]. Each branch contains multiple populations of closely related genotypes, many of which show geographical specificity in their distribution. The most recent date for the evolution of each of the branches and populations was estimated on the basis of the clock rate of diversity that accumulated over 70 years of sampling in Madagascar [7]. Figure 1 shows that these estimates predate the beginning of the third pandemic which was caused by the 1.ORI populations called 1.ORI1, 1.ORI2 and 1.ORI3. Similarly, genotypes recently identified in skeletons from the Black Death [66,70] map to a location on the genealogy at the beginning of branch 1 (figure 1). By inference from its historical dates, the Justinianic Pandemic might have been

---

**Table 2. Mutation clock rates (per nucleotide per year) in different taxa.**

<table>
<thead>
<tr>
<th>taxon</th>
<th>clock rate</th>
<th>age (years)</th>
<th>sampling period (years)</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Campylobacter jejuni</em></td>
<td>$\mu_S = 2.8 \times 10^{-5}$</td>
<td>400</td>
<td>3</td>
<td>[46]</td>
</tr>
<tr>
<td>pandemic <em>V. cholerae</em></td>
<td>$\mu_S = 1.1 \times 10^{-5}$</td>
<td>130</td>
<td>34</td>
<td>[47]</td>
</tr>
<tr>
<td><em>S. aureus</em> ST239</td>
<td>$\mu = 3.3 \times 10^{-6}$</td>
<td>45</td>
<td>21</td>
<td>[42]</td>
</tr>
<tr>
<td><em>S. aureus</em> ST225</td>
<td>$\mu = 2.0 \times 10^{-6}$</td>
<td>25</td>
<td>14</td>
<td>[48]</td>
</tr>
<tr>
<td><em>Buchnera aphidicola</em></td>
<td>$\mu_S = 2.2 \times 10^{-7}$</td>
<td>&lt;135</td>
<td>extant</td>
<td>[49]</td>
</tr>
<tr>
<td>endemic <em>Y. pestis</em></td>
<td>$\mu = 8.6 \times 10^{-9}$</td>
<td>100</td>
<td>70</td>
<td>[7]</td>
</tr>
</tbody>
</table>

---

*Phil. Trans. R. Soc. B* (2012)
associated with population 0.PE3 (figure 1). (0.PE4 is less likely because it cannot infect larger mammals).

Typing of several hundred isolates from a global sample for most of the SNPs in the genealogy in figure 1 showed that \textit{Y. pestis} originated in China, or the vicinity \[7\]. This conclusion was reached because the basal populations on the genealogy are found in China, as are multiple other populations from all branches. An alternative source in Central Asia that subsequently migrated to China is unlikely because studies with other genetic markers \[75,76\] indicate that the diversity of isolates from Central Asia is lower than in China, not all populations on the genealogy are represented in Central Asia, and the oldest Central Asian populations are more recent than the oldest Chinese population. The most parsimonious interpretation of these observations is that sylvatic plague originated in China and spread to other global regions on multiple occasions.

The existence of 2.MED isolates in Kazakhstan and former Kurdistan would then reflect spread from China, possibly along the former Silk Road, along which most 2.MED isolates in China were found \[7\]. Similarly, the exclusive existence of 1.ANT isolates in East Africa implies that they must have been imported, possibly directly from China via the marine voyages of Zheng He in the early-fifteenth century \[7\]. Finally, if the Justinianic Pandemic were associated with 0.PE3, or a related population, it must have been transmitted from China to Northeast Africa prior to spreading throughout Europe.

The associations of individual populations with distinct waves of transmission are particularly clear for the third pandemic, where extensive historical literature documents the introduction of plague in the 1890s to India, Australia, Africa, the Americas and Europe. Each of these transmission routes (except

\begin{figure}
\begin{center}
\includegraphics{genealogy.png}
\end{center}
\caption{Minimal age of major nodes within the \textit{Yersinia pestis} genealogy. Dates (grey) are the most recent dates estimated for the nodes shown as grey circles. This genealogy contains three branches, branch 0, 1 and 2, as well as one leaf per genome sequenced. Branch 3 is not shown because no public genome is yet available but branches at the split between branches 1 and 2. Population assignments (1.ORI3, 0.ANT2, etc.) are indicated in colours, while the dates of the three plague pandemics are indicated within yellow-rounded rectangles. MRCA, most recent common ancestor. Adapted from Morelli \textit{et al.} \[7\].}
\end{figure}

\textit{Phil. Trans. R. Soc. B} (2012)
Australia) has left extant isolates that were assigned to geographically discrete populations or lineages (figure 2). Transmission from Hong Kong was thus associated with a very dramatic expansion of genetic diversity within the genealogy, a so-called polytomy, during which multiple independent mutations were fixed by independent chains of transmission.

6. SUMMARY

Genetically monomorphic pathogens have traditionally been difficult to analyse because their diversity is so limited, and a general overview of their properties was first published in 2008 [13]. Since then, there has been a veritable explosion of comparative genomic studies that have focused on such pathogens. Genetically monomorphic pathogens are particularly easy to analyse by comparative genomics simply because they are so monomorphic, much easier than more diverse taxa. Their population genetic patterns are often associated with clonal genealogies, which are particularly easy to interpret. And their recent origins allow easier reconstructions of their history than is possible with most other microbes. The recent recognition that some lineages have high mutation rates, almost as high as those of some RNA viruses, means that real time evolution can be used for epidemiological purposes in the near future because the technological challenges are now relatively minor.

The current microbiological literature emphasizes the importance of horizontal genetic exchange for the recent emergence of novel pathogens. A number of import events that were apparently selected by antibiotics and immune escape are described here. Yet such imports seem to be generally rare, and are often rapidly discarded, and an acquisition of enhanced virulence seems possibly even rarer! Experimental studies of evolution and long-term infections of humans have documented the power of Darwinian selection during initial adaptation to a novel environment. Such adaptation is likely to have preceded the emergence of the most recent common ancestors of natural populations of genetically monomorphic pathogens but most of their subsequent, ongoing microevolution seems to be neutral. If this interpretation is correct, the evolution of novel microbial pathogens is a very rare event, and increased levels of disease are more likely to reflect environmental changes than evolutionary adaptation.

M.A. was supported by grant 05/F1/B882 from the Science Foundation of Ireland.

REFERENCES


