THE ROYAL SOCIETY

Bacterial population genetics, evolution and epidemiology

Brian G. Spratt and Martin C. J. Maiden

Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK (brian.spratt@zoo.ox.ac.uk)

Asexual bacterial populations inevitably consist of an assemblage of distinct clonal lineages. However, bacterial populations are not entirely asexual since recombinational exchanges occur, mobilizing small genome segments among lineages and species. The relative contribution of recombination, as opposed to *de novo* mutation, in the generation of new bacterial genotypes varies among bacterial populations and, as this contribution increases, the clonality of a given population decreases. In consequence, a spectrum of possible population structures exists, with few bacterial species occupying the extremes of highly clonal and completely non-clonal, most containing both clonal and non-clonal elements. The analysis of collections of bacterial isolates, which accurately represent the natural population, by nucleotide sequence determination of multiple housekeeping loci provides data that can be used both to investigate the population structure of bacterial pathogens and for the molecular characterization of bacterial isolates. Understanding the population structure of a given pathogen is important since it impacts on the questions that can be addressed by, and the methods and samples required for, effective molecular epidemiological studies.

Keywords: population structure; horizontal gene transfer; linkage disequilibrium; non-clonal; nucleotide sequencing; phylogenetics

1. INTRODUCTION

Since pathogenic bacteria exist as populations, the members of which exhibit varying degrees of virulence, the integration of population genetic, evolutionary, and epidemiological studies can provide important insights into the origins and spread of bacterial diseases (Musser 1996). In addition to enhancing our understanding of the microbial causes of disease, these insights are helpful in designing effective public health interventions. In recent years a number of developments have improved our capacity to perform such integrated studies, perhaps the principal being the increasing speed and decreasing cost of automated nucleotide sequence determination. The generation of large volumes of sequence data, combined with the development of novel analytical techniques and conceptual advances, promise a better understanding of the complexities of the evolution of bacterial populations. Here we will review some of the insights into the population genetics of bacterial pathogens that have resulted from the use of nucleotide sequence data and examine their implications for epidemiology.

2. THE CLONAL MODEL

Bacterial population genetics, in common with many other features of prokaryotic biology, is at first sight deceptively simple. Haploid bacteria reproduce asexually by binary fission, mother cells giving rise to two daughter cells, each containing a chromosome identical to that of the mother cell. In the absence of sexual processes, chromosomal variation occurs by *de novo* mutations, which can spread only by being passed on to the descendants of the cells in which they arose, and new lineages emerge by the accumulation of such mutations over successive generations. This transmission of genetic information can be regarded as 'vertical', as it passes exclusively from mother to daughter cell, as opposed to 'horizontal' movement of genetic material between cells that do not necessarily share a recent ancestor, by sexual processes.

In the absence of the horizontal genetic exchange of chromosomal genes, a given mutation will be associated with the other mutations that have accumulated in the chromosome during the history of the lineage in which it arose. Consequently, the distribution of chromosomal polymorphisms within an asexual (clonal) bacterial population will be non-random, or in linkage disequilibrium. This contrasts with populations of sexual organisms where mutations are continually reassorted, resulting in linkage equilibrium, i.e. mutations at different sites occur in random combinations.

Asexual bacterial populations therefore exist as assemblages of independent lineages which are stable, with evolutionary change occuring only by *de novo* mutation, although diversity may also arise by other mechanisms, e.g. the gain and loss of plasmids or the movement of insertion sequences. Differences in the frequencies of particular lineages in the population will occur over time as a consequence of selection or stochastic events. When mutations that increase fitness arise, the lineages that contain them will increase in frequency, resulting in the loss of other lineages, and this process (periodic selection) reduces the genetic diversity within the population (Atwood *et al.* 1951; Levin 1981). Similarly, bacterial

populations are subject to rapid expansions and severe bottlenecks which can also reduce the diversity of clonal populations (Achtman 1997). In addition to linkage disequilibrium, clonal bacterial populations are therefore characterized by lower levels of sequence diversity than would be expected from the size, age and mutation rate of an idealized bacterial population (Levin 1981). While limited diversity is predicted under the clonal model, it is not the only reason why a bacterial population may be uniform, and lack of diversity should not, on its own, be used as a proof of clonality.

Clonal bacterial populations exhibiting linkage disequilibrium and subject to regular diversity reducing events are straightforward to study as identification of the limited number of lineages can be achieved by the characterization of the nucleotides present at a few variable sites. These data will also indicate the evolutionary history of a given isolate and establish its phylogenetic relationship to other members of the population. The phylogeny can then be used as a framework on to which other characteristics, such as pathogenicity, serology, host specificity and the presence of virulence genes, can be mapped (Selander *et al*. 1990*a*,*b*). In this way, important insights into the origins of the pathogenicity and host adaptation of the lineages within a clonal species can be obtained. The concept of the asexual clonal population is a powerful simplifying assumption for bacterial epidemiology, as the identification of a few variable sites quickly and unambiguously identifies disease-causing bacterial clones and establishes their relationship to each other. Unfortunately, the appeal of this simplification has been such that it has often been applied inappropriately.

Just as prokaryotes confounded the initial expectations of simplicity in their metabolism and cellular organization, and in the range of habitats they can exploit, bacterial population structures do not usually conform to the idealized clonal model described above. Indeed, far from being simpler than those of higher organisms, bacterial population structures are complex and not amenable to the theory and methods of population genetic analysis developed for either sexual or asexual eukaryotic organisms (Maynard Smith 1995). Even the concept of species is difficult in the prokaryote world, leading to various more or less unsatisfactory definitions of a bacterial species. As it is difficult to avoid using the term, we will use the word 'species' to refer to the assemblages of isolates defined as species by traditional microbiological criteria.

3. THE IMPACT OF SEXUAL PROCESSES

The complications of bacterial population biology arise from the fact that, while bacteria are largely asexual, they possess a number of mechanisms for the horizontal genetic exchange of chromosomal DNA. The three most important mechanisms, conjugation, transduction and transformation, are properly regarded as parasexual processes as none of them involves wholesale exchange of chromosomal genetic material. Rather, they permit the transfer of fragments of chromosomal DNA, and horizontal genetic exchanges lead to bacterial genomes being pocked by small chromosomal replacements from other lineages, a process that has been referred to as localized sex (Maynard Smith *et al*. 1991). Plasmids, prophages, transposons and insertion sequences can also be transferred horizontally, providing mechanisms for mobilizing DNA among distantly related bacteria.

Homologous recombination requires as little as 70% nucleotide sequence identity between the recipient and donor bacteria (Lorenz & Wackernagel 1994), although recombinational exchanges are far more likely to occur between closely related DNA sequences, since the efficiency of RecA-mediated recombination decreases sharply with increasing sequence divergence. A consequence of this promiscuity is that bacteria can recruit variation from other members of the same species and, at lower frequency, from related species. Evidence for localized sex has been found in many bacterial species, some of which are naturally transformable. In species that are not naturally transformable, such as the enteric bacteria, it is probable that phage-mediated transduction is the most important mechanism of localized genetic exchange.

As it is unlinked to reproduction, the frequency of recombinational replacement in bacterial populations could vary from very low (or zero) to very high. Localized sex disrupts clonal population structures by providing a means of reassorting genetic variation, thereby enabling mutations to escape the lineage in which they arose. Under the strictly clonal model, diversity-reducing events result in the loss of all of the nucleotide variation present in the lineages which become extinct; however, the introduction of sexual processes results in mutations spreading independently of the lineage in which they arose, moderating the power of diversity-reducing events to purge the variation in a bacterial population. Conversely, in other circumstances, horizontal genetic exchange can be a cohesive evolutionary force, by enabling a variant allele to spread horizontally in a population, replacing other variants and reducing diversity.

4. THE SPECTRUM OF BACTERIAL POPULATION STRUCTURES

Differences in the ratio of genetic change caused by recombination relative to *de novo* mutation leads to a spectrum of population structures, from the extremes of strictly clonal, where effectively no recombination has occurred in the evolutionary history of the species, to non-clonal, or 'panmictic', where recombinational exchanges are sufficiently frequent to randomize the alleles in the population and to prevent the emergence of stable clones. We contend that the extremes of strictly clonal and non-clonal population structures are rarely found in bacterial species, and that most bacterial populations occupy a middle ground where recombination is highly significant in the evolution of the population, but is not sufficiently frequent to prevent the emergence of clonal lineages. Further, it is increasingly clear that bacterial populations are complex and a single bacterial species may not always be adequately described as being highly clonal, weakly clonal or non-clonal. For example, populations may be non-clonal in the longer term, although the emergence of unstable clones may be an important feature of the population. A mixture of non-clonal and clonal elements within populations of recombinogenic bacterial pathogens may often be related

to differences in their ecology and epidemiology. For example, the impact of recombination on population structure will be very different in situations in which different lineages rarely meet, compared to those in which mixing of lineages is frequent.

5. PHYLOGENETIC MODELS AND BACTERIAL POPULATIONS

Bacterial population structures, shaped by the processes outlined above, can be described with appropriate combinations of three phylogenetic models: bifurcating trees, bundles and nets (Maynard Smith 1989). The bifurcating tree, for some time an important working hypothesis for the reconstruction of phylogenetic history, assumes that each member of a population is related by a series of direct antecedents. Under this model, analysis of the traits of the current members of a population will enable at least a partial reconstruction of the series of variants that led to it. Although an adequate model for clonal populations, a bifurcating tree is inappropriate if any degree of horizontal genetic transfer is introduced, as a given member of the population may have genetic material acquired from a variety of other members of the species, or even from members of related species.

To accommodate horizontal genetic exchange it is necessary to postulate a network, connecting each member of a population to all of the members of the population that have contributed to its genome. In practice, even with a small amount of horizontal genetic exchange such a network will be highly complex over evolutionary timescales, and can only be reconstructed for small parts of the genome (e.g. individual genes or gene fragments) over a relatively short period of time. Neither trees nor networks may be appropriate models if a microbial population has experienced a selective sweep, rapid population growth or very high levels of recombination, all of which remove phylogenetic signal and result in a bundle or 'star phylogeny'. The term 'epidemic clonal' has been used to describe a situation of this type, where a particularly effective lineage within a basically non-clonal bacterial population arises and rapidly spreads, so that, in the short term, a large number of related organisms come to predominate the population (Maynard Smith *et al*. 1993). This phenomenon is particularly clear where the emerging lineage has increased capacity to cause disease, as the analysis of isolates obtained exclusively from disease can result in a large amplification of the significance of the epidemic clone as a consequence of sampling bias (see below). Such organisms clearly represent the descendants of a founding bacterium, but their rapid diversification by frequent recombinational replacements results in their appearance as a bundle of closely related genotypes whose relationships to the bulk of the recombining non-clonal population cannot be established. It is possible for all of these three types of structure to be present in a single bacterial species, as illustrated in figure 1.

6. ANALYSING BACTERIAL POPULATIONS

(a) *Sampling bacterial populations and genes*

To understand the population biology of a given bacterial species it is essential to establish the relative contribution of each of the above processes to the overall population structure. A number of analytical approaches will achieve this but, before they can be applied, it is essential to acquire appropriate data. As in all population studies, this involves obtaining a representative collection of the organisms that make up the population and characterizing them accurately and appropriately. There are a number of pitfalls specific to the study of bacterial pathogens which have to be negotiated with care, with population sampling principal among them. For example, a sample containing one *Escherichia coli* isolate from the faeces of 100 different mammals is very different from a sample containing one isolate from 100 different people, or 100 isolates from one individual. If recombinational exchanges occur relative frequently in *E. coli*, they are more likely to be observed in the latter two samples than the former. Similarly, isolates of bacterial pathogens are usually collected from cases of disease and most strain collections are heavily biased towards isolates that are particularly virulent, often neglecting the less pathogenic isolates which frequently comprise the majority of the population. In most cases, analysis of the small fraction of isolates that are from disease will underestimate the diversity of the population as a whole and will overestimate the extent of clonality in the population.

Many years of serological typing, together with the needs of vaccine development, led to the use of surface antigens as the principal means of characterizing bacterial pathogens. In almost every case these genes are subject to strong diversifying selection by the host immune system and they evolve rapidly, since non-synonymous mutations and recombinational exchanges that alter antigenicity are positively selected. Only rarely will these markers provide reliable information on the phylogenetic history of the species and they will give highly misleading information about the underlying rates of recombination and *de novo* mutation in bacterial populations.

For studies of population structure it is necessary to index variation that is neutral. The best means of achieving this is to sequence housekeeping genes, or fragments of them, ideally using genes encoding cytoplasmic enzymes involved in central metabolism. The nucleotide changes in these genes are constrained by the essential biochemical functions of the proteins which they encode, but are not normally exposed to other selective pressures, and most of the variation that is observed is neutral and slowly evolving.

Before the introduction of automated nucleotide sequence determination, multilocus enzyme electrophoresis (MLEE) was the method of choice for analysing bacterial population structures (Selander *et al*. 1986). This is an indirect approach, the genetic variation present at multiple housekeeping loci being inferred by measuring differences in the electrophoretic mobilities of their gene products on starch gels. The disadvantages of MLEE compared with nucleotide sequencing, apart from technical complexity, are that only a minority of mutations are detected (those that alter the electrophoretic mobility of the enzyme), and enzymes with the same electrophoretic mobility can be encoded by very different gene sequences. There are numerous other characterization methods used in medical microbiology but virtually all of

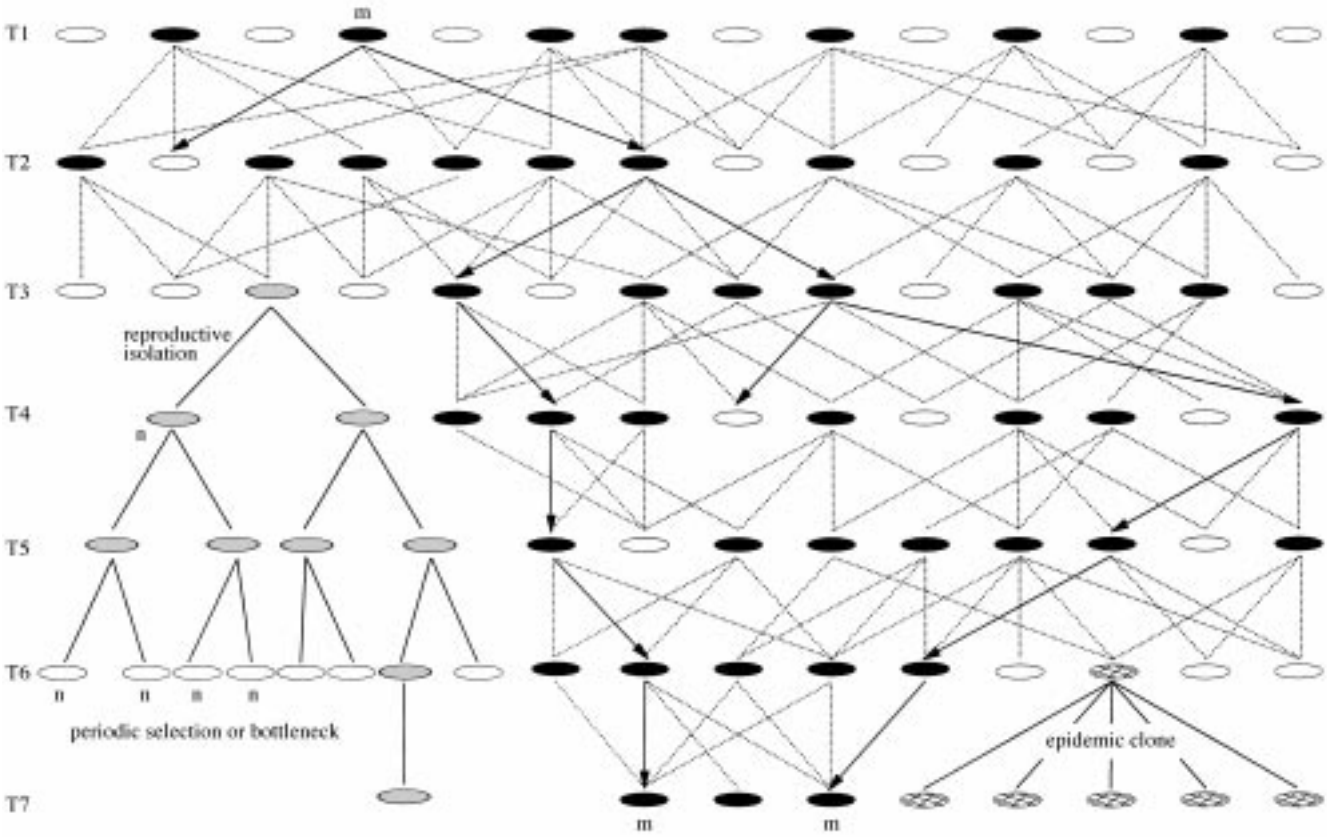

Figure 1. Illustration of three phylogenetic processes operating within a single bacterial species. The figure represents a bacterial species sampled at seven time points (T1–T7). The black and grey ovals represent bacterial cells that donate genetic material to subsequent generations and the white ovals members of the populations that do not, which in reality greatly outnumber the number that do. In the time period represented by T1–T3, the population is non-clonal with horizontal genetical exchange, represented by broken lines, reassorting variation such that a given member of the population at T3 may have genetic material which originated in a number of members of previous populations. Horizontal genetic exchange enables a mutation (m) that arises in the T1 population to spread by a number of routes through the population, shown by the solid black arrows; it will not be lost if any individual cell in which it occurs leaves no descendants. In the population at T3 the grey cell gives rise to a subpopulation that is reproductively isolated from the rest of the population. This could be due to genetic factors, e.g. the appearance of a non-transformable lineage in a transformable species, resulting in an isolated set of clonal lineages (as shown). A mutation (n) arising in a member of this clonal subpopulation will be passed on only to its direct descendants, and may be lost if periodic selection or a population bottleneck amplifies lineages that do not possess the mutation, as shown between T6 and T7. Alternatively, reproductive isolation may be ecological, e.g. a lineage may become adapted to a new niche, with genetic exchange occurring within the niche-adapted descendants, but rarely with isolates of the bulk of the population. The hatched cells represent the emergence in T6 of a lineage with an increased capacity to cause disease (an epidemic clone). The cells at T7 are isolates recovered from disease and this lineage becomes highly overrepresented within this biased sample of the population. Analysis of isolates from T7 may therefore show linkage disequilibrium due to the sampling bias introduced by the recent emergence of the epidemic clone, whereas the population from T5 may show disequilibrium due to the presence of isolates from two ecologically isolated niches, and the population at T3 will show linkage equilibrium.

them are of no use for population genetic and evolutionary studies. Indeed, many are unsuitable for the routine characterization of isolates for which they were designed (Achtman 1996).

While there are general rules about what constitutes an appropriate choice of loci and isolates, these decisions are ultimately dependent on the biology of the particular species under investigation. For example, gene order differs among bacterial species and proximity of a housekeeping gene to an antigen-encoding gene may distort the phylogenetic signal from that locus in one species but not in another, a problem that can largely be avoided as the genome sequences of many of the bacterial pathogens become available. The strategy for obtaining an appropriate sample of the bacterial population is also species specific: for an obligate human pathogen, a collection of disease-causing isolates from global sources will sample the entire population, but for a pathogen that is normally commensal, rarely causing disease, a representative sample of the whole population must mainly comprise isolates from healthy carriers.

7. METHODS FOR DEFINING STRUCTURE WITHIN BACTERIAL POPULATIONS

Bacterial population structure can be determined by a number of approaches. These can be grouped into methods that analyse levels of linkage disequilibrium among alleles, which to date have been applied almost exclusively to MLEE data, phylogenetic inference from nucleotide sequence data and studies of particular lineages as they diverge.

(a) *Analysis of linkage disequilibrium*

A number of methods are available to detect whether the alleles in a bacterial population depart from linkage equilibrium (Brown *et al*. 1980; Maynard Smith *et al*. 1993; Sved 1968; Whittam *et al*. 1983). One of the limitations of these approaches is that linkage disequilibrium between alleles can be present in populations in which recombination is frequent. For example, a named species may include two distinct populations (cryptic species), with recombination common within each population, but very rare between populations. Analysis of a sample of the whole species will detect significant levels of linkage disequilibrium between alleles, leading to an erroneous conclusion about the extent of recombination and the population structure (Maynard Smith *et al*. 1993). Linkage disequilibrium can also arise in a weakly clonal population as a consequence of poor sampling. This problem can be highly significant in those pathogens where most isolates rarely cause disease but the isolates in the data set are predominantly from disease. Sampling problems of this type may in some cases be detected by showing that linkage disequilibrium disappears, or is greatly reduced, if only one example of each lineage, rather than every isolate, is included in the analysis (Maynard Smith *et al*. 1993). Finally, most bacterial species occupy a number of different ecological niches, and while recombination may be relatively frequent within individuals occupying the same niche, it will be less frequent among isolates occupying different niches. Such population subdivisions can introduce linkage disequilibrium into samples of the population containing isolates from the different niches (Maynard Smith *et al*. 1991; Reeves 1992).

Linkage equilibrium is less likely to be caused by sampling artefacts, and provided the sample of the population is sufficiently large and sufficiently variable to provide a statistically robust test of linkage relationships, the absence of any association between the alleles at different loci implies that recombination must be frequent, and that distinct lineages share a common gene pool.

(b) *Analysis of nucleotide sequences*

Nucleotide sequence data from multiple housekeeping genes in an appropriately sampled population can be used in a variety of analyses. The simplest of these is to establish the alleles present at each locus and to use a clustering algorithm to determine the relationships among strains from the matrix of pairwise differences between their allelic profiles (multilocus sequence typing (MLST), Maiden *et al*. (1998)). While this is very effective in establishing that isolates are identical or closely related, the approach will not, unless the population is strictly clonal, provide much information about the relationships between more distantly related isolates.

Additional phylogenetic information can be recovered if the nucleotide sequences themselves are analysed, and a range of algorithms are available for this purpose. A problem with most of these algorithms is that they assume a bifurcating tree-like phylogeny and will construct such a phylogeny even if there is little or no evidence for one in the data. Using these approaches, evidence for a history of recombination can sometimes be inferred from the non-congruence between the trees constructed for different housekeeping genes, although low levels of sequence diversity and recombinational exchanges within the data set can result in poorly supported trees, confounding efforts to establish non-congruence. Alternative approaches, such as split decomposition, which do not assume a tree can be used to visualize the relationships among sequence data without recourse to a bifurcating tree (Bandelt & Dress 1992; Dopazo *et al*. 1993). In this procedure a history of recombinational exchanges can be uncovered by the appearance of a network of relationships between the sequences (figure 2).

Recombination within DNA sequences results in different parts of a gene having different evolutionary histories. This results in mosaic genes in which one portion of a gene may be identical in two isolates, whereas the other parts differ at many sites. A number of procedures are available to detect this non-random distribution of polymorphic sites within two sequences (Maynard Smith 1992), or a set of sequences (Sawyer 1989). Other methods look for evidence of recombination as runs of nucleotides, or the number of nucleotides, that are inconsistent with the optimal tree produced by a tree-building algorithm. The homoplasy test (Maynard Smith & Smith 1998), for example, constructs a maximum parsimony tree and evaluates whether the number of sites that are inconsistent with this tree is significantly more than would be expected from the occurrence of the same substitution independently in different branches of the tree (homoplasies). Tests for linkage disequilibrium can also be applied to variable sites within a gene and in some cases have shown random assortment of these sites within the population (Feil *et al*. 1996*b*; Suerbaum *et al*. 1998).

(c) *Analysis of the molecular events during clonal diversification*

Analysis of levels of linkage disequilibrium, non-congruence between gene trees, the absence of tree-like structure and the detection of mosaic genes all provide evidence for recombination, but they are relatively blunt tools that give little information about the relative role of recombination compared to mutation in the accumulation of variation within housekeeping genes in different bacterial populations. This information might be obtained from a knowledge of the rates of diversification of the lineages within different species if it could be assumed that rates of mutation among bacteria are constant, such that increased rates of diversification reflect increased rates of recombination. The problem with this approach is that mutation rates may not be constant among species, and the ages of bacterial clones are rarely known with any certainty. However, as will be detailed in the next section, such studies have been possible for a number of bacterial species including *Streptococcus pneumoniae*, *Neisseria meningitidis* and *E. coli*. The diversification of clones within a matter of years, or a few decades, is a hallmark of frequent recombinational exchanges in the population, since point mutations within housekeeping genes are very unlikely to be observed within this time-scale. This is particularly true when diversification is measured by MLEE, as a change in the electrophoretic mobility of a housekeeping enzyme has been estimated to require an average of 26 nucleotide changes (Boyd *et al*. 1994). Using MLST data this approach

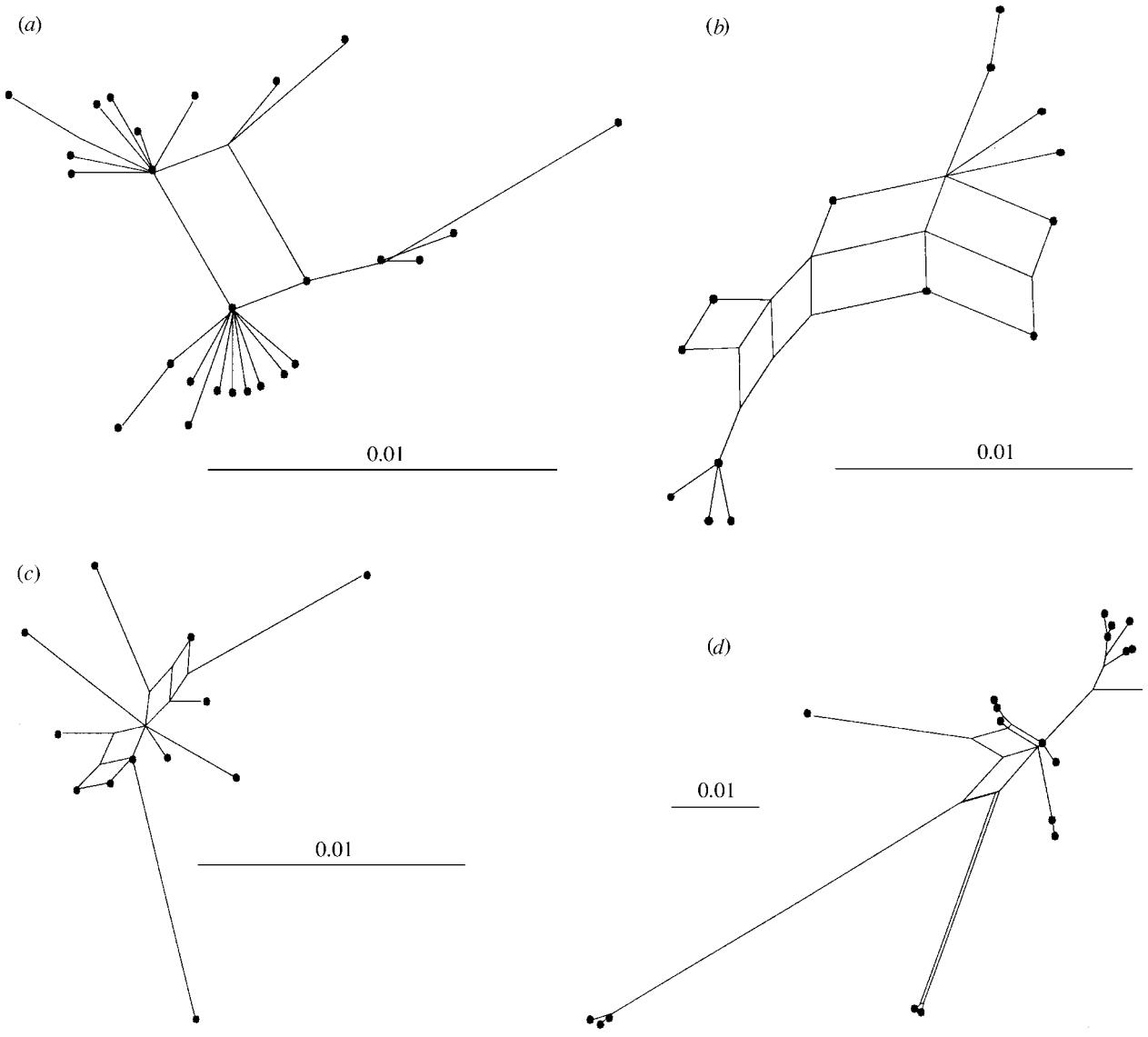

Figure 2. Split decomposition analysis of housekeeping genes from various bacterial species. These graphs illustrate the degree to which the sequences conform to a bifurcating tree-like phylogeny. Only the *mdh* (malate dehydrogenase gene) sequences from *Salmonella enterica* (*d*) show evidence of tree-like structures, and the *gdh* (glucose-6-phosphate dehydrogenase) gene of *Neisseria meningitidis* shows a networked structure (*b*), providing evidence of extensive recombination in the evolution of this gene in this species. There is also evidence for recombination being important in the *Escherichia coli mdh* gene (*c*), and recombination may have played a role in the evolution *S. enterica mdh* genes. There is little phylogenetic signal in the *Streptococcus pneumoniae gki* (glucose kinase) gene, but importantly, no evidence of a tree-like structure (*a*). All of these data would produce a tree when subjected to conventional phylogenetic analysis.

can be taken further to provide a relative measure of the recombination to mutation ratio in different populations, since the sequence variation that has occurred within housekeeping genes during the initial diversification of a clone can be analysed to obtain the frequency at which a nucleotide site has changed by recombination compared to mutation (Guttman & Dykhuizen 1994).

8. POPULATION STRUCTURES OF PATHOGENIC BACTERIA

The population structures of a number of pathogenic bacteria have been established. While these represent only a small proportion of the known bacterial pathogens, they do provide examples of several distinct bacterial population structures and, as more pathogens are studied by population genetic approaches, it is likely that an even wider range of population structures will be revealed.

(a) Salmonella enterica—*a clonal organism*

Salmonella enterica conforms, by and large, to the clonal model. Populations of this species exhibit high levels of linkage disequilibrium between alleles (Boyd *et al*. 1996), which do not appear to be a consequence of sampling artefacts, and the frequent association of a single antigenic profile with a single clone implies a low rate of horizontal transfer, even for genes that are expected to be under diversifying selection. Although nucleotide sequencing studies, particularly of antigen genes, show occasional examples of recombinational exchanges between *S. enterica* lineages, the dendrograms derived from the sequences of housekeeping genes are largely

congruent with each other and with those derived from MLEE data (Boyd *et al*. 1994; Nelson & Selander 1992; Nelson *et al*. 1991). *S. enterica* lineages also appear to be stable. For example, isolates of the lineage that causes typhoid fever, which is presumed to be hundreds or thousands of years old, recovered from worldwide sources, are largely uniform by MLEE and antigenic profile.

(b) Helicobacter pylori *and* N. gonorrhoeae—*non-clonal bacterial populations*

The population structures of *N. gonorrhoeae* and *H. pylori*, which are both naturally transformable, do not conform to the clonal model (Go *et al*. 1996; Salaun *et al*. 1998; Spratt *et al*. 1995). Neither of these species exhibit linkage disequilibrium between alleles. *H. pylori* populations are genetically diverse and, as expected for a recombining population, every isolate from epidemiologically unrelated individuals has a unique multilocus genotype (Go *et al*. 1996). There is also strong evidence for extensive localized recombination within *H. pylori* genes, as detected by the homoplasy test (Suerbaum *et al*. 1998), and by the near absence of linkage disequilibrium between variable nucleotide sites (Salaun *et al*. 1998). In the case of the gonococcus, analysis of the nucleotide sequences of housekeeping genes is not informative as they are too uniform to provide statistically significant evidence of recombination. This uniformity is probably a feature of the species being relatively young. Intriguingly, not all *N. gonorrhoeae* are non-clonal, as some isolates, the arginine-, hypoxanthine- and uracil-requiring (AHU) strains, have existed with little change of phenotype over a period of four decades (Kohl *et al*. 1986) and are a clonal lineage that can be accommodated within a non-clonal model (Gutjahr *et al*. 1997).

(c) *Bacteria with weakly clonal population structures*

N. meningitidis, which causes meningitis and septicaemia worldwide, is an example of an organism with a weakly clonal population structure. This organism also illustrates the sampling problems associated with some bacterial pathogens. Asymptomatic nasopharyngeal carriage of *N. meningitidis* is common and only very occasionally do the bacteria invade the blood stream and cerebrospinal fluid to cause disease. Meningococcal populations are highly diverse but only a small number of lineages, representing a small fraction of the meningococcal population, are responsible for most cases of disease. Due to the concentration of attention on isolates from cases of disease, these lineages are greatly overrepresented in strain collections (Maiden & Feavers 1995). Consequently, MLEE studies of such highly biased samples of the population show strong linkage disequilibrium (Caugant *et al*. 1987*b*), although strong evidence supports a population structure which is, in the longer term, non-clonal.

This evidence includes the observations that levels of linkage disequilibrium are low when corrections for sample bias are made (Maynard Smith *et al*. 1993), and that meningococcal clones are unstable undergoing rapid diversification in their electrophoretic profiles (MLEE) (Caugant *et al*. 1987*a*), multilocus genotypes (MLST) (Maiden *et al*. 1998) and cell surface antigens (Maiden & Feavers 1995). Further, analysis of housekeeping genes shows non-congruence between gene trees and mosaic structure, and a total absence of tree-like structure using split decomposition (figure 2).

Some of the meningococcal hypervirulent lineages can be tentatively assigned ages based on their first appearance in strain collections (Maiden & Feavers 1995), enabling an analysis of clonal diversification. For example, members of the ET-5 hypervirulent lineage were first isolated in the mid-1970s, but by the mid-1980s had already diversified extensively to form a clonal complex containing many isolates that differed by MLEE and MLST at one or two loci (Caugant *et al*. 1986; Maiden *et al*. 1998). Lineages of serogroup A are thought to be somewhat older and more clonal, exhibiting evidence of point mutations as well as recombination events during their recent evolution (Morelli *et al*. 1997). Analysis of clonal diversification, using MLST data, has estimated that an individual nucleotide site in a meningococcal housekeeping gene is at least 80 times more likely to change by recombination than by point mutation (E. J. Feil and B. G. Spratt, unpublished data).

The meningococcal population therefore provides a good example of a basically non-clonal population where the relationships between isolates should be represented by a net rather than a tree. However, from this recombining population of isolates within the nasopharynges of carriers, isolates with an increased capacity to cause disease are believed to emerge at intervals and become highly overrepresented in samples of isolates from disease (epidemic clones). These hypervirulent lineages rapidly diversify, predominantly by the accumulation of recombinational replacements, and eventually will no longer be distinguishable from the diverse collection of genotypes that constitute the meningococcal population.

Meningococci coexist in the human nasopharynx with a number of closely related commensal *Neisseria* species, and another striking feature of this species is the frequency with which recombinational replacements from commensal *Neisseria* species, which differ in sequence by as much as 25%, are observed (Feil *et al*. 1996*b*). These interspecies recombinational replacements are found in genes encoding cell surface antigens (Vazquez *et al*. 1995), and antibiotic targets (Bowler *et al*. 1994; Maiden 1998), where rare interspecies recombinational exchanges may be strongly selected as they introduce adaptively useful variation, but they are also encountered within most housekeeping genes where strong selection is unlikely (Feil *et al*. 1996*a*; Zhou *et al*. 1997).

The population structure of *S. pneumoniae* has not yet been well studied, but clonal diversification can be usefully applied in this species, as an upper limit for the age of antibiotic-resistant clones can be confidentally assigned. There are several clones of multiply-antibiotic-resistant *S. pneumoniae* which cannot pre-date the introduction of antibiotics, and which are probably less than 20 years old (Crook & Spratt 1998). Most of the isolates of each of these very young clones have identical alleles at all of the seven loci used for MLST of *S. pneumoniae* (Enright & Spratt 1998), but there are already single-locus variants of each of the clones (J. Zhou, M. C. Enright and B. G. Spratt, unpublished results). In almost all cases so far examined, the variant alleles within each of the multiresistant *S. pneumoniae* clones differ from the normal alleles at multiple nucleotide sites, and have

therefore arisen by recombination rather than by point mutation (J. Zhou, M. C. Enright and B. G. Spratt, unpublished results). Recombination is therefore the predominant mode of evolutionary change within *S. pneumoniae* housekeeping genes and the population structure of this naturally transformable species is likely to be strongly influenced by recombination.

A final example of an organism with a mixed population structure is provided by *E. coli*. Although analyses of levels of linkage disequilibrium suggest that *E. coli* populations are clonal (Whittam 1995), analysis of the diversification of closely related clones indicated that diversification in *E. coli* also occurs more frequently by recombination than *de novo* mutation (Guttman & Dykhuizen 1994). This result indicates that even when there is significant linkage disequilibrium between alleles, possibly introduced by niche specialization (Guttman 1997), recombination may still be an important mechanism for evolutionary change at neutral loci.

9. POPULATION STRUCTURE AND MOLECULAR EPIDEMIOLOGY

The characterization of isolates of bacterial pathogens is central to many aspects of bacterial epidemiology, which aims to answer two main types of question. The first are short-term epidemiological questions, e.g. are the isolates recovered from a localized outbreak of disease the same or different, or is relapse of disease after intervention due to treatment failure or reinfection? The second type of question concerns long-term or global epidemiology, e.g. how do strains causing disease in one geographical area relate to strains recovered worldwide?

A knowledge of the extent of recombination in bacterial pathogens is important since low levels of recombination result in highly clonal populations, where lineages persist with little variation over hundreds or thousands of years. At the other extreme, high rates of recombination lead to non-clonal populations in which lineages diversify so rapidly that the isolates recovered in one decade may be completely different from those recovered in the next. Clearly, the way in which these two types of populations evolve and the methods that should be used to study their molecular epidemiology will differ, as will the types of epidemiological questions that we can sensibly ask.

For highly clonal pathogens, such as *S. enterica,* MLEE provides little variation that can be used to study the global or local epidemiology of the infections caused by the disease-associated clones. Detailed epidemiological studies of typhoid fever need to make use of the microvariation that inevitably accumulates even within a highly clonal lineage. In this case, ribotyping and pulsed-field gel electrophoresis fingerprinting have been shown to detect variation that can be used to discriminate between isolates (Altwegg *et al*. 1989; Navarro *et al*. 1996). The nature of this variation is largely uncharacterized although some recent studies suggest that genetic exchange between multiple copies of the ribosomal RNA genes, or inversions of chromosomal segments mediated by homologies provided by repetitive sequences, may be involved. Such methods are not, however, useful for identifying the lineages themselves.

For non-clonal species, asking long-term epidemiological questions—tracking the global spread of particular strains, for example—will not be possible. Fortunately, bacterial pathogens are unlikely to diversify so rapidly that short-term epidemiological questions cannot be addressed. In the case of gonococci, it is possible to obtain information about the transmission of gonorrhoea among sexual partners, or within core groups (O'Rourke *et al*. 1995). Similarly, with *H. pylori* it is possible to ask questions about the transmission of the organism within families (Suerbaum *et al*. 1998), or if recurrence of colonization following antibiotic treatment is due to recolonization by a different strain or treatment failure (Shortridge *et al*. 1997). Whether longer-term epidemiological questions can be addressed with these non-clonal species requires a better understanding of the rates of diversification of lineages.

Characterization of weakly clonal pathogens (e.g. *N. meningitidis*, *S. pneumoniae*) is more problematic, since clones diversify relatively rapidly by the accumulation of recombinational exchanges. The identity of isolates recovered from a localized outbreak, caused by the same strain, can be established using any number of sufficiently discriminatory methods (e.g. pulsed-field gel electrophoresis fingerprinting), since the common ancestor of these isolates occurred so recently that there has been insufficient time for any significant variation to accumulate. However, it becomes more difficult to show whether an outbreak strain belongs to one of the clones associated with disease by comparing it to reference isolates from other countries, as substantial diversification of these clones will have occurred since the outbreak strain and the other members of the clone had a common ancestor. Techniques that assay variation at many loci (e.g. MLEE or MLST) are ideal for recognizing these unstable clones since recombinational exchanges that alter the alleles at one or two of the loci do not prevent their recognition as members of the clone using clustering techniques. This problem becomes more acute as the clones become older, until their diversification is so great that their members can no longer be distinguished from the background population (figure 1).

10. CONCLUDING REMARKS

Bacterial population genetics is a late-developing but rapidly maturing science. Now that bacterial variation in large samples of the population can be measured at multiple loci directly by nucleotide sequence determination, it is possible to design studies which permit the models developed in this article to be rigorously tested and further developed. While informative in themselves, these models have a number of very practical applications. It is vital to be able to predict the likely effects of public health interventions on the biology of the pathogens against which they are directed. The emergence of resistance to a particular antibiotic and the appearance of escape mutants which avoid immunity induced by a given vaccine are evolutionary phenomena operating on bacterial populations. Understanding the structure and dynamics of these populations will permit the design of effective public health policies that minimize the possibilities for microbial evolution to neutralize them.

B.G.S. is a Wellcome Trust Principal Research Fellow. M.C.J.M. is a Wellcome Trust Senior Fellow in Biodiversity Research.

REFERENCES

Achtman, M. 1996 A surfeit of YATMs? *J. Clin. Microbiol.* **34**, 18–70.

Achtman, M. 1997 Microevolution and epidemic spread of serogroup A *Neisseria meningitidis*—a review. *Gene* **192**, 135–140.

Altwegg, M., Hickman-Brenner, F. W. & Farmer, J. J. D. 1989 Ribosomal RNA gene restriction patterns provide increased sensitivity for typing *Salmonella typhi* strains. *J. Infect. Dis.* **160**, 145–149.

Atwood, K. C., Schneider, L. K. & Ryan, F. J. 1951 Periodic selection in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **37**, 146–155.

Bandelt, H. J. & Dress, A. W. 1992 Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* **1**, 242–252.

Bowler, L. D., Zhang, Q. Y., Riou, J. Y. & Spratt, B. G. 1994 Interspecies recombination between the *penA* genes of *Neisseria meningitidis* and commensal *Neisseria* species during the emergence of penicillin resistance in *N. meningitidis*: natural events and laboratory simulation. *J. Bacteriol.* **176**, 333–337.

Boyd, E. F., Nelson, K., Wang, F. S., Whittam, T. S. & Selander, R. K. 1994 Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. *Proc. Natl Acad. Sci. USA* **91**, 1280–1284.

Boyd, E. F., Wang, F. S., Whittam, T. S. & Selander, R. K. 1996 Molecular genetic relationships of the salmonellae. *Appl. Environ. Microbiol.* **62**, 804–808.

Brown, A. H. D., Feldman, M. W. & Nevo, E. 1980 Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* **98**, 523–536.

Caugant, D. A., Froholm, L. O., Bovre, K., Holten, E., Frasch, C. E., Mocca, L. F., Zollinger, W. D. & Selander, R. K. 1986 Intercontinental spread of a genetically distinctive complex of clones of *Neisseria meningitidis* causing epidemic disease. *Proc. Natl Acad. Sci. USA* **83**, 4927–4931.

Caugant, D. A., Froholm, L. O., Bovre, K., Holten, E., Frasch, C. E., Mocca, L. F., Zollinger, W. D. & Selander, R. K. 1987*a* Intercontinental spread of *Neisseria meningitidis* clones of the ET-5 complex. *Antonie van Leeuwenhoek J. Microbiol.* **53**, 389–394.

Caugant, D. A., Mocca, L. F., Frasch, C. E., Froholm, L. O., Zollinger, W. D. & Selander, R. K. 1987*b* Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern. *J. Bacteriol.* **169**, 2781–2792.

Crook, D. W. & Spratt, B. G. 1998 Multidrug resistance in *Streptococcus pneumoniae*. *Br. Med. Bull.* **54**, 593–608.

Dopazo, J., Dress, A. & Von Haeseler, A. 1993 Split decomposition: a technique to analyze viral evolution. *Proc. Natl Acad. Sci. USA* **90**, 10 320–10 324.

Enright, M. & Spratt, B. G. 1998 A multilocus sequence typing scheme for *Streptococcus pneumoniae* identification of clones associated with serious invasive disease. *Microbiology* **144**, 3049–3060.

Feil, E., Carpenter, G. & Spratt, B. G. 1996*a* Electrophoretic variation in adenylate kinase of *Neisseria meningitidis* is due to inter- and intraspecies recombination. *Proc. Natl Acad. Sci. USA* **92**, 10 535–10 539.

Feil, E., Zhou, J., Maynard Smith, J. & Spratt, B. G. 1996*b* A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal *Neisseria* species: evidence for extensive interspecies recombination within *adk*. *J. Mol. Evol.* **43**, 631–640.

Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. 1996 Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J. Bacteriol.* **178**, 3934–3938.

Gutjahr, T. S., O'Rourke, M., Ison, C. A. & Spratt, B. G. 1997 Arginine-, hypoxanthine-, uracil-requiring isolates of *Neisseria gonorrhoeae* are a clonal lineage within a non-clonal population. *Microbiology* **143**, 633–640.

Guttman, D. S. 1997 Recombination and clonality in natural populations of *Escherichia coli*. *Trends Ecol. Evol.* **12**, 16–22.

Guttman, D. S. & Dykhuizen, D. E. 1994 Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383.

Kohl, P. K., Knapp, J. S., Hofmann, H., Gruender, K., Petzoldt, D., Tams, M. R. & Holmes, K. K. 1986 Epidemiological analysis of *Neisseria gonorrhoeae* in the Federal Republic of Germany by auxotyping and serological classification using monoclonal antibodies. *Genitourin. Med.* **62**, 145–150.

Levin, B. R. 1981 Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* **99**, 1–23.

Lorenz, M. G. & Wackernagel, W. 1994 Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* **58**, 563–602.

Maiden, M. C. J. 1998 Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. *Clin. Infect. Dis.* **27**, S12–S20.

Maiden, M. C. J. & Feavers, I. M. 1995 Population genetics and global epidemiology of the human pathogen *Neisseria meningitidis*. In *Population genetics of bacteria* (ed. S. Baumberg, J. P. W. Young, E. M. H. Wellington & J. R. Saunders), pp. 269–293. Cambridge University Press.

Maiden, M. C. J. (and 12 others) 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.

Maynard Smith, J. 1989 Trees, bundles or nets. *Trends Ecol. Evol.* **4**, 302–304.

Maynard Smith, J. 1992 Analysing the mosaic structure of genes. *J. Mol. Evol.* **34**, 126–129.

Maynard Smith, J. 1995 Do bacteria have population genetics? In *Population genetics of bacteria* (ed. S. Baumberg, J. P. W. Young, E. M. H. Wellington & J. R. Saunders), pp. 1–12. Cambridge University Press.

Maynard Smith, J. & Smith, N. H. 1998 Detecting recombination from gene trees. *Mol. Biol. Evol.* **15**, 590–599.

Maynard Smith, J., Dowson, C. G. & Spratt, B. G. 1991 Localized sex in bacteria. *Nature* **349**, 29–31.

Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. 1993 How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4384–4388.

Morelli, G., Malorny, B., Muller, K., Seiler, A., Wang, J. F., del Valle, J. & Achtman, M. 1997 Clonal descent and microevolution of *Neisseria meningitidis* during 30 years of epidemic spread. *Mol. Microbiol.* **25**, 1047–1064.

Musser, J. M. 1996 Molecular population genetic analysis of emerged bacterial pathogens: selected insights. *Emerg. Infect. Dis.* **2**, 1–17.

Navarro, F., Llovet, T., Echeita, M. A., Coll, P., Aladuena, A., Usera, M. A. & Prats, G. 1996 Molecular typing of *Salmonella enterica* serovar *typhi*. *J. Clin. Microbiol.* **4**, 2831–2834.

Nelson, K. & Selander, R. K. 1992 Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. *J. Bacteriol.* **174**, 6886–6895.

Nelson, K., Whittam, T. S. & Selander, R. K. 1991 Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **88**, 6667–6671.

O'Rourke, M., Ison, C. A., Renton, A. M. & Spratt, B. G. 1995 Opa-typing: a high resolution tool for studying the epidemiology of gonorrhoea. *Mol. Microbiol.* **17**, 865–875.

Reeves, P. R. 1992 Variation in O-antigens, niche-specific selection and bacterial populations. *FEMS Microbiol. Lett.* **79**, 509–516.

Salaun, L., Audibert, C., Le Lay, G., Burucoa, C., Fauchere, J. L. & Picard, B. 1998 Panmictic structure of *Helicobacter pylori* demonstrated by the comparative study of six genetic markers. *FEMS Microbiol. Lett.* **161**, 231–239.

Sawyer, S. 1989 Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N. & Whittam, T. S. 1986 Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl. Environ. Microbiol.* **51**, 837–884.

Selander, R. K., Beltran, P., Smith, N. H., Barker, R. M., Crichton, P. B., Old, D. C., Musser, J. M. & Whittam, T. S. 1990*a* Genetic population structure, clonal phylogeny, and pathogenicity of *Salmonella paratyphi* B. *Infect. Immun.* **58**, 1891–1901.

Selander, R. K., Beltran, P., Smith, N. H., Helmuth, R., Rubin, F. A., Kopecko, D. J., Ferris, K., Tall, B. D., Cravioto, A. & Musser, J. M. 1990*b* Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. *Infect. Immun.* **58**, 2262–2275.

Shortridge, V. D., Stone, G. G., Flamm, R. K., Beyer, J., Versalovic, J., Graham, D. W. & Tanaka, S. K. 1997 Molecular typing of *Helicobacter pylori* isolates from a multicenter U.S. clinical trial by *ureC* restriction fragment length polymorphism. *J. Clin. Microbiol.* **35**, 471–473.

Spratt, B. G., Smith, N. H., Zhou, J., O'Rourke, M. & Feil, E. 1995 The population genetics of the pathogenic *Neisseria*. In *Population genetics of bacteria* (ed. S. Baumberg, J. P. W. Young, E. M. H. Wellington & J. R. Saunders), pp. 143–160. Cambridge University Press.

Suerbaum, S., Maynard Smith, J., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. 1998 Free recombination within *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **95**, 12 619–12 624.

Sved, J. A. 1968 The stability of linked systems of loci with a small population size. *Genetics* **59**, 543–563.

Vazquez, J., Berron, S., O'Rourke, M., Carpenter, G., Feil, E., Smith, N. H. & Spratt, B. G. 1995 Interspecies recombination in nature: a meningococcus that has acquired a gonococcal PIB porin. *Mol. Microbiol.* **15**, 1001–1007.

Whittam, T. S. 1995 Genetic population structure and pathogenicity in enteric bacteria. In *Population genetics of bacteria* (ed. S. Baumberg, J. P. W. Young, E. M. H. Wellington & J. R. Saunders), pp. 217–245. Cambridge University Press.

Whittam, T. S., Ochman, H. & Selander, R. K. 1983 Multilocus genetic structure in natural populations of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **80**, 1751–1755.

Zhou, J., Bowler, L. D. & Spratt, B. G. 1997 Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. *Mol. Microbiol.* **23**, 799–812.