**MINI-
REVIEW****Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination**

David M. Gordon

Tel: +61 2 6125 3552. Fax: +61 2 6125 5573. e-mail: David.Gordon@anu.edu.au

Division of Botany & Zoology, Australian National University, Canberra, ACT 0200, Australia

Keywords: population genetics, virulence factors, enteric bacteria, ecological structure**Overview**

This mini-review presents the results of recent studies examining the extent to which among-population variation accounts for the genetic diversity of *Escherichia coli* and other members of the *Enterobacteriaceae*. The available population genetic data for *E. coli* are discussed in relation to an issue perceived to be of public health importance – the coliform contamination of water and efforts to trace the source of such contamination. For *E. coli*, the data suggest that geographical structure or host specificity accounts for little of the observed genetic diversity. Furthermore, the evidence indicates that there is little temporal stability in the clonal composition of *E. coli* populations. Consequently, most of the assumptions implicit in any programme attempting to identify the source of coliform contamination that focuses on commensal isolates of *E. coli* appear to be invalid. Other species of enteric bacteria appear to exhibit greater host specificity and geographical structure. The results of recent studies concerning the distribution of virulence factors among the clonal lineages of *E. coli* are highlighted and discussed in regard to alternative strategies for tracing the source of coliform contamination.

Background

The application of population genetic concepts and methods has made substantial contributions to the field of microbiology. One of its first applications was to investigate the genetic structure of bacteria at the species level. The early work with *E. coli* led to the clone concept (Milkman, 1973; Selander & Levin, 1980). This concept describes the consequences of asexual reproduction (vertical gene transfer) coupled with low rates of recombination (lateral gene transfer) for the genetic structure of a species. Subsequently, more species were studied and it became apparent that the clone concept required modification (Istock *et al.*, 1992; Maynard Smith *et al.*, 1993; Guttman & Dykhuizen, 1994). The

results of these and other studies demonstrated that for species such as *Neisseria gonorrhoeae* lateral gene transfer plays a far more significant role than it does for species such as *E. coli* or *Salmonella enterica* (Spratt & Maiden, 1999). It is now accepted that considerable among-species variation in genetic structure occurs. As discussed by Spratt & Maiden (1999), the extent to which lateral gene transfer varies among species has significant applied implications for epidemiological studies of bacterial pathogens.

Less effort has been made to determine the degree to which genetic variation is partitioned among populations of a species. Earlier work with *E. coli* demonstrated that the same clone, or at least the same multi-locus genotype, could be isolated from widely separated geographical localities (Selander *et al.*, 1987). These results suggested that little spatial structure existed in species like *E. coli*. However, the frequency of studies concerned with among-population variation has increased in recent years. Most of these studies have concerned soil bacteria and they have, almost universally, demonstrated a significant degree of population differentiation (Haubold & Rainey, 1996; Wernegreen *et al.*, 1997; Silva *et al.*, 1998; Bouzar *et al.*, 1999).

The mini-review is organized in four sections. First, the rationale behind efforts to trace the source of bacterial contamination in the environment and the goals of these efforts are outlined, together with the assumptions that are implicit in these goals. The second section discusses the validity of these assumptions in light of studies concerning the population genetics of *E. coli* and other enteric species. The third section summarizes recent data concerning the distribution of virulence factors among the clonal lineages of *E. coli*. The implications of these data, as they concern tracing the source of environmental bacterial contamination, are discussed. The mini-review closes with suggestions of alternative approaches to tracing the source of environmental bacterial contamination.

Tracing the source of environmental *E. coli* contamination

As faecal contamination of water is considered a human health risk, there has always been a great deal of concern regarding the level of coliform bacteria counts in water. Many bodies of water throughout the world are considered to have counts above acceptable levels. The sources of these coliforms are thought to be faecal contamination from humans, domestic animals and wildlife, as well as runoff from agricultural land, inadequate septic systems or sewer overflow. Coliforms may themselves represent a health risk or indicate the presence of other potential human pathogens such as the hepatitis A virus.

In an effort to manage the problem of elevated coliform counts, programmes are under way in many parts of the world to develop methods which will allow the source of the faecal contamination to be determined. For example, can elevated coliform counts be attributed to human-derived or animal-derived faecal contamination, or can the geographical source of the contamination be identified? Much of this work focuses on *E. coli*, as it is the dominant member of the aerobic flora of humans and the cause of a significant fraction of human bacterial disease (Siitonen, 1994). The techniques in use include examining antibiotic resistance profiles and genotyping bacteria using ribotyping or PCR-based methods such as rep-PCR (Parveen *et al.*, 1997, 1999; Dombek *et al.*, 2000). The aim of these methods is to obtain a fingerprint of the environmental isolate that can be then be classified as derived from a particular host group or locality.

Regardless of the bacterial species being monitored, the success of these efforts depends on several assumptions being valid. (1) The species shows geographical structure. That is, the clonal composition of populations differs among localities. (2) The species exhibits some degree of host specificity. That is, it is more likely that particular clones will be isolated from one host species, or group of species, than another. (3) The clonal composition of the species isolated from soil and water represents the clonal composition of the species in the host populations responsible for the faecal inputs to the environment. (4) The clonal composition of populations is stable through time. That is, that the same clones can be recovered from the same locality or host populations for 'extended' periods.

Geographical structure

Geographical structure appears to account for little of the genetic variation observed in *E. coli*. Spatial structure accounted for 2% of the genetic diversity observed in *E. coli* isolated from two populations of feral house mice 15 km apart (Gordon, 1997). Among *E. coli* isolated from mammals collected throughout Australia, locality effects explained 5% of the observed allelic diversity (Gordon & Lee, 1999). At an even larger scale, contrasting *E. coli* isolated from rodents living in Australia and Mexico revealed that continent differences explained about 10% of the observed diversity (Souza *et*

al., 1999). For *E. coli* from humans, variation between families within the same city explained 6% of the diversity, whilst 1% was attributable to variation between cities (Caugant *et al.*, 1984). In a collection of human *E. coli* isolates from four continents, spatial effects accounted for only 2% of the observed allelic diversity (Whittam *et al.*, 1983). The mobility of the human population probably accounts for the apparent lack of meaningful levels of geographical structure in *E. coli* isolated from humans. Isolates from non-domesticated animals appear to show slightly more geographical differentiation than isolates from humans and this probably reflects the more restricted distributions and movement patterns of most mammal species.

Geographical structure seems to explain more of the observed genetic variation in other enteric species than it does for *E. coli*. For isolates collected from Australian mammals, spatial effects accounted for 12% of the allelic variation in *Citrobacter freundii*, 17% in *Hafnia alvei* and 22% in *Klebsiella pneumoniae*, compared to 5% in *E. coli* (Gordon & Lee, 1999).

Host specificity

Although based on a very limited number of studies, there is little evidence that there are clones of *E. coli* adapted to particular host species or host groups. In a study of *E. coli* isolated from four orders and ten families of mammals in Australia, the taxonomic family of the host was found to explain only 6% of the genetic variation among clones (Gordon & Lee, 1999). Indeed the same multi-locus genotype was recovered from individuals of different taxonomic orders.

Other species of enteric bacteria, such as *Enterobacter cloacae*, *H. alvei* and *Klebsiella oxytoca* appear to exhibit a greater degree of host specificity (Fig. 1), although this is not always reflected in the genetic structure of the species (Gordon & FitzGibbon, 1999). For example, *K. oxytoca* is far more likely to be recovered from a bat (Vespertilionidae) than a member of any other family of Australian mammal (Fig. 1). Yet, host taxonomic family explained none of the observed genetic variation in *K. oxytoca* (Gordon & Lee, 1999).

Primary versus secondary habitats

Most species of host-associated bacteria find themselves in an environment external to the host at some stage in their life cycle. Savageau (1983) has argued that the typical *E. coli* cell spends, on average, half its life in the external environment. The fate of clones moving from hosts to the external environment is poorly understood.

Linkage disequilibrium is a hallmark of a clonal genetic structure and simply means that alleles at one locus are non-randomly associated with alleles at other loci (Maynard Smith *et al.*, 1993). Virtually all studies of *E. coli* isolated from hosts have demonstrated linkage disequilibrium. Pupo & Richardson (1995) sampled *E. coli* from the inflow to a sewage treatment plant serving a population of about 16000 people and identified 159

Fig. 1. Host specificity in six species of enteric bacteria. The histograms depict the probability of isolating the bacterial species from a particular taxonomic family of hosts. Statistically significant host family effects were detected for all species except CIFR and KLPN (Gordon & FitzGibbon, 1999). CIFR, *Citrobacter freundii*; ENCL, *Enterobacter cloacae*; ESCO, *Escherichia coli*; HAAL, *Hafnia alvei*; KLOX, *Klebsiella oxytoca*; KLPN, *Klebsiella pneumoniae*.

haplotypes. Although the authors observed linkage disequilibrium for some pairwise locus comparisons, a multi-locus analysis of their data failed to reveal linkage disequilibrium among the 159 haplotypes (Gordon 1997). Whittam (1989) studied *E. coli* isolated from domestic birds and the litter on which the birds were raised. Only 10% of the 113 distinct clones isolated were recovered from both habitats. Further, the clones isolated from the environment were genetically distinct from those isolated from the host population. These studies suggest that the clonal composition of *E. coli* communities changes substantially during the transition from the host to the external environment (Whittam, 1989).

Temporal variation within populations

A quarter to a half of the *E. coli* clonal diversity observed in an individual host is a result of temporal variation in the number of clones present and their relative abundance (Whittam, 1989). The conventional explanation for this variation is that individuals are continually exposed to bacteria when they ingest food and water (Hartl & Dykhuizen, 1984). Although many of these strains fail to establish, or are present in the host for only a few days, some become resident members of the gut community and can persist for months, if not years (Caugant *et al.*, 1981).

Temporal changes in the genetic composition of *E. coli* communities occur at scales beyond that of an individual host. Significant variation over a 4 month period was observed in the composition of *E. coli* sampled from the inflow to a sewage treatment plant (Pupo & Richardson, 1995). No obvious environmental factor (pH, temperature) appeared to account for this variation. In another study, significant changes were observed over a 6 month period in the clonal composition of *E. coli* isolated from feral house mice (Gordon, 1997). Some clones were observed in every sample, others were

recovered intermittently, while others were only recovered over part of the study. Not all of this variation could be solely attributed to stochastic variation in the relative abundance of strains. A statistically significant decline (30%) in the frequency of colicinogenic isolates and an increase in the frequency of colicin-resistant isolates was observed over the 6 month sampling period (Gordon *et al.*, 1998).

Is *E. coli* the appropriate coliform to examine?

Geographical structure and host specificity are not absent in *E. coli*, and statistically significant levels of among-population variation have been observed. However, there appears to be minimal population differentiation, and typically only 5% of the observed genetic diversity can be attributed to among-locality or among-host group variation. This degree of population differentiation is inadequate if the goal is to unambiguously assign environmental isolates to their source populations (Cornuet *et al.*, 1999).

Most of the structure observed in *E. coli* populations appears to occur at the individual host level. Differences among individuals of the same species living in close proximity, such as members of the same human family, baboon troop or bird flock, can account for 25–60% of the observed diversity (Caugant *et al.*, 1984; Routman *et al.*, 1985; Whittam, 1989).

Perhaps the most significant problem with using *E. coli* to trace coliform contamination is that there appears to be substantial changes in *E. coli* community composition during the transition from host to external environment. Indeed, the very limited evidence available suggests that there is little similarity between the dominant *E. coli* community of the host population and the community in the environment where the faeces of that host population accumulate. Such differences were observed

in a largely closed system, where the bird population was restricted to a single area and there was little opportunity for outside contamination of the litter on which the birds were raised (Whittam, 1989). Therefore, it is likely that changes that are even more significant will be observed in environments that are more complex. For example, when comparing the composition of the inputs to a sewage system and the bacterial community entering the treatment plant after having travelled through kilometres of sewage pipe.

Assigning an environmental isolate to a particular source population not only requires the presence of geographical structure, but that the clonal composition of populations is stable over significant time scales. This assumption appears to be unwarranted. The evidence indicates that there is little temporal stability in the clonal composition of an *E. coli* population. The temporal variation seems to occur at every level, the individual host, host population and locality. Changes in the clonal composition of populations appear to occur within time frames measured in weeks, rather than months or years.

***E. coli* genetic structure and virulence factors**

Population geneticists have determined that the species *E. coli* consists of four major clusters of clones (Selander *et al.*, 1987; Herzer *et al.*, 1990). These clusters have been designated A, B1, B2, and D (Fig. 2). Most *E. coli* clones, at least those isolated in Europe and North America, can be assigned to one of the four clusters. Non-assignable strains appear to be rare (Whittam *et al.*, 1983; Goullet & Picard, 1986).

The distribution of many virulence traits is non-random with respect to the four *E. coli* genetic clusters (Fig. 2). Group B2, and to a lesser extent group D, have a higher prevalence of extra-intestinal virulence factors than do strains in the other genetic clusters (Picard *et al.*, 1999; Johnson & Stell, 2000; Johnson *et al.*, 2001). This is not to state that virulence factors are absent from the other *E. coli* clusters, but they occur at much lower frequencies compared to clones belonging to the B2 or D clusters.

E. coli isolates can also be 'classified' as commensal (non-disease causing) or as those that are capable of causing disease in their host. The majority of commensal *E. coli* clones, those isolated from the faeces of healthy humans, belong to the clusters A and B1 (Picard *et al.*, 1999). Although there is a lack of studies characterizing *E. coli* isolated from non-domesticated animals, the available evidence suggests that most commensal isolates from domestic animals are also members of the A or B1 clusters (Goullet & Picard, 1986). Isolates capable of causing disease may be further categorized as those responsible for extra-intestinal diseases, such as pyelonephritis or neonatal meningitis, and those occasioning intestinal diseases, such as haemorrhagic colitis. Isolates taken from patients with extra-intestinal disease are grossly overrepresented among the B2 cluster

of *E. coli* (Picard *et al.*, 1999). However, it has been estimated that B2 strains represent only about 5% of the strains isolated from the faeces of healthy humans (Picard *et al.*, 1999). Cluster D strains also seem to be under-represented in the faeces of asymptomatic humans (Picard *et al.*, 1999).

The situation for strains accountable for intestinal infection is somewhat different. Whilst all the intestinal disease-causing strains are highly clonal, they can occur among any of the four major clusters of *E. coli* (Fig. 2) (Pupo *et al.*, 1997; Reid *et al.*, 2000). Recent studies have suggested that clones causing intestinal disease have evolved several times in different lineages of *E. coli* and this explains why these clonal groups are present in all four clusters (Reid *et al.*, 2000).

Sampling considerations

Determining the source of a particular genotype requires that strains be isolated from potential source populations and this means sampling hosts. Most standard microbiological methods of sampling the bacterial community of a host are only capable of detecting those strains that are common [$\sim 10^7$ cells (g faeces) $^{-1}$]. There is evidence to suggest that there are strains of *E. coli* persisting in the gut at frequencies well below conventional detection levels. The VTEC strain O157:H7 is thought to persist at low frequencies in cattle (Wells *et al.*, 1991) and strains capable of causing urinary tract infections have been found in humans at low frequencies (Johnson *et al.*, 1998).

There is another dimension to this sampling problem, at least when it concerns *E. coli*. The dominant strains isolated from the faeces of asymptomatic hosts are mostly from genetic clusters A and B1. These strains represent only a portion of the diversity to be found in *E. coli* and are less likely to be of clinical significance. Cluster B2 and D strains, those mostly likely to be the cause of extra-intestinal disease, are rarely isolated as the dominant members of the *E. coli* flora of a host. Where do the B2 and D strains typically persist? Do they occur at low frequencies in the gut of their hosts? If this is the case then conventional sampling regimes are not detecting the clones of potential clinical significance.

Alternative approaches for tracing the source of coliform contamination

Perhaps PCR-based detection of virulence factors would provide a better indication of the potential health significance of coliforms present in the environment and might be more suitable for tracing the source of environmental contamination. There are many known or suspected virulence factors present in *E. coli*. Primers have been designed to detect 30 virulence alleles associated with extra-intestinal infection using five PCR reactions (Johnson & Stell, 2000). Such an approach could easily be extended to include the virulence factors known to be associated with clones responsible for

Fig. 2. Genetic structure of *E. coli* and the distribution of virulence factors in each of the four major clusters of strains (A, B1, B2, D). The dendrogram was constructed using subsets of the allozyme data available for the ECOR collection (Selander *et al.*, 1987) and the allozyme data presented by Pupo *et al.* (1997). Solid circles denote urinary tract isolates. Triangles denote strains responsible for intestinal disease (solid triangles, *E. coli* isolates; open triangles, *Shigella* isolates). The frequency histograms present the fraction of ECOR strains in a cluster that were positive for a particular virulence determinant using a PCR-based assay. The data for these histograms were taken from Johnson *et al.* (2001). The virulence factors presented are (from left to right): *papAH*, *sfafoc*, *iha*, *hlyA*, *focG*, *kpsMT II*, *K1*, *iutA*, *fyuA*, *iroN* and *malX*.

intestinal infections in humans. The detection of 40 or more 'virulence factors' has the ability to produce a complex fingerprint, to lead to good among-strain discrimination and to identify those strains of potential clinical significance. Furthermore, the virulence factor screening of the ECOR collection suggests that the median number of virulence factors in strains of animal origin is half that found in strains isolated from human faeces (Johnson *et al.*, 2001).

Such a PCR assay could be applied directly to faecal material. Serial dilution of the faecal sample followed by PCR of the diluted samples would allow the relative frequency of a particular virulence profile to be determined. Such a sampling approach would not allow the profile of individual isolates to be determined, rather a 'host profile' would be produced. The advantage of such an approach is that it would allow the detection of low-frequency strains. The use of such an assay would of course depend on knowing the frequency of individual virulence factors and their 'linkage' relationships in *E. coli*, especially the dominant commensal strains of clusters A and B1. This knowledge would enable one to calculate the probability of obtaining a particular virulence profile. If the sample had a virulence profile that suggested the presence of a high-risk strain, and the probability of obtaining this profile by chance was low, then this would suggest that a virulent strain was present in the sample.

'Virulence finger-printing' has the advantage of detecting clones of potential clinical significance. However, it may have no advantages over conventional fingerprinting techniques given that *E. coli* appears to exhibit little spatial structure or host specificity and populations show a high degree of temporal variation in their clonal composition. Cluster B2 and D strains, those most likely to cause disease, represent a fraction of the diversity to be found in *E. coli*. Furthermore, it is possible that these strains exhibit a greater degree of among-population differentiation. An additional advantage to focusing on clinically significant strains is that it may eliminate much of the 'noise' introduced by examining the far more abundant commensal strains that are of little clinical significance.

Other enterics, such as *C. freundii* or *H. alvei*, are at least as genetically diverse as *E. coli* and appear, generally, to exhibit a greater degree of host or spatial structure (Gordon & Lee, 1999). They also exhibit a greater degree of host specificity (Gordon & FitzGibbon, 1999). Although they are much less frequently responsible for human disease than *E. coli*, perhaps they might be more appropriate for tracing the source of coliform contamination. Whilst there are no data available, these species may show more temporal stability in their host populations and undergo less dramatic changes in their clonal community composition during the transition from the host to external environment. They are also easily isolated using selective plating methods.

This mini-review has highlighted the results of some of the recent studies concerning the population genetics of

bacteria. It has endeavored to discuss the results of these studies in the context of a problem perceived to be of potential significance to human health. There are many other aspects of the population genetics and ecology of bacteria that are of equal applied significance. There is every indication that bacteria will continue to be important pathogens of animals and plants, and much evidence to suggest that their significance as disease agents is increasing. The control of infectious diseases depends on, among other aspects, a sound understanding of their population ecology and genetics.

References

- Bouzar, H., Jones, J. B., Stall, R. E., Louws, F. J., Schneider, M., Rademaker, J. L. W., de Bruijn, F. J. & Jackson, L. E. (1999). Multiphasic analysis of xanthomonads causing bacterial spot disease on tomato and pepper in the Caribbean and Central America: evidence for common lineages within and between countries. *Phytopathology* **89**, 328–335.
- Caugant, D. A., Levin, B. R. & Selander, R. K. (1981). Genetic diversity and temporal variation in the *E. coli* populations of a human host. *Genetics* **98**, 467–490.
- Caugant, D. A., Levin, B. R. & Selander, R. K. (1984). Distribution of multilocus genotypes of *Escherichia coli* within and between host families. *J Hyg* **92**, 377–384.
- Cornuet, J., Piry, S., Luikart, G., Estoup, A. & Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.
- Dombek, P. E., Johnson, L. K., Zimmerley, S. T. & Sadowsky, M. J. (2000). Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. *Appl Environ Microbiol* **66**, 2572–2577.
- Gordon, D. M. (1997). The genetic structure of *Escherichia coli* populations in feral house mice. *Microbiology* **143**, 2039–2046.
- Gordon, D. M. & FitzGibbon, F. (1999). The distribution of enteric bacteria among Australian mammals. *Microbiology* **145**, 2663–2671.
- Gordon, D. M. & Lee, J. (1999). The genetic structure of enteric bacteria from Australian mammals. *Microbiology* **145**, 2673–2682.
- Gordon, D. M., Riley, M. A. & Pinou, T. (1998). Temporal changes in the frequency of colicinogeny in *Escherichia coli* from house mice. *Microbiology* **144**, 2233–2240.
- Goullet, P. & Picard, B. (1986). Comparative esterase electrophoretic polymorphism of *Escherichia coli* isolates obtained from animal and human sources. *J Gen Microbiol* **132**, 1843–1851.
- Guttman, D. S. & Dykuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383.
- Hartl, D. L. & Dykuizen, D. E. (1984). The population genetics of *Escherichia coli*. *Annu Rev Genet* **18**, 31–68.
- Haubold, B. & Rainey, P. B. (1996). Genetic and ecotypic structure of a fluorescent *Pseudomonas* population. *Mol Ecol* **5**, 747–761.
- Herzer, P. J., Inouye, S., Inouye, M. & Whittam, T. S. (1990). Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *J Bacteriol* **172**, 6175–6181.
- Istock, C. A., Duncan, K. E., Ferguson, N. & Zhou, X. (1992). Sexuality in a natural population of bacteria – *Bacillus subtilis* challenges the clonal paradigm. *Mol Ecol* **1**, 95–103.

- Johnson, J. R. & Stell, A. L. (2000). Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. *J Infect Dis* **181**, 261–272.
- Johnson, J. R., Brown, J. J., Carlino, U. B. & Russo, T. A. (1998). Colonization with and acquisition of uropathogenic *Escherichia coli* as revealed by polymerase chain reaction-based detection. *J Infect Dis* **177**, 1120–1124.
- Johnson, J. R., Delavari, P., Kuskowski, M. & Stell, A. L. (2001). Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*. *J Infect Dis* **183**, 78–88.
- 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.
- Milkman, R. (1973). Electrophoretic variation in *Escherichia coli* from natural sources. *Science* **182**, 1024–1026.
- Parveen, S., Murphree, R. L., Edmiston, L., Kasper, C. W., Portier, K. M. & Tamplin, M. L. (1997). Association of multiple-antibiotic resistance profiles with point and nonpoint sources of *Escherichia coli* in Apalachicola Bay. *Appl Environ Microbiol* **63**, 2607–2612.
- Parveen, S., Portier, K. M., Robinson, K., Edmiston, L. & Tamplin, M. L. (1999). Discriminant analysis of ribotype profiles of *Escherichia coli* for differentiating human and nonhuman sources of fecal pollution. *Appl Environ Microbiol* **65**, 3142–3147.
- Picard, B., Sevali Garcia, J., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J. & Denamur, E. (1999). The link between phylogeny and virulence in *Escherichia coli* extra-intestinal infection. *Infect Immun* **67**, 546–553.
- Pupo, G. M. & Richardson, B. J. (1995). Biochemical genetics of a natural population of *Escherichia coli*: seasonal changes in alleles and haplotypes. *Microbiology* **141**, 1037–1044.
- Pupo, G. M., Karaolis, D. K. R., Lan, R. & Reeves, P. R. (1997). Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. *Infect Immun* **65**, 2685–2692.
- Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K. & Whittam, T. S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64–67.
- Routman, E., Miller, R. D., Philips-Conroy, J. & Hartl, D. L. (1985). Antibiotic resistance and population structure in *Escherichia coli* from free-ranging African yellow baboons. *Appl Environ Microbiol* **50**, 749–754.
- Savageau, M. A. (1983). *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *Am Nat* **122**, 732–744.
- Selander, R. K. & Levin, B. R. (1980). Genetic diversity and structure in *Escherichia coli*. *Science* **210**, 545–547.
- Selander, R. K., Caugant, D. A. & Whittam, T. S. (1987). Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, pp. 1625–1648. Edited by F. C. Neidhardt and others. Washington, DC: American Society for Microbiology.
- Siitonen, A. (1994). What makes *Escherichia coli* pathogenic. *Ann Med* **4**, 229–231.
- Silva, C., Eguarte, L. E. & Souza, V. (1998). Reticulated and epidemic population genetic structure of *Rhizobium etli* biovar *phaseoli* in a traditionally managed locality in Mexico. *Mol Ecol* **8**, 277–287.
- Souza, V., Rocha, M., Valera, A. & Eguarte, L. E. (1999). Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. *Appl Environ Microbiol* **65**, 3373–3385.
- Spratt, B. G. & Maiden, M. C. (1999). Bacterial population genetics, evolution and epidemiology. *Philos Trans R Soc Lond B Biol Sci* **354**, 701–710.
- Wells, J. G., Shipman, L. D., Greene, K. D. & 7 other authors (1991). Isolation of *Escherichia coli* serotype O157:H7 and other Shiga-like-toxin-producing *E. coli* from dairy cattle. *J Clin Microbiol* **29**, 985–989.
- Wernegreen, J. J., Harding, E. E. & Riley, M. A. (1997). *Rhizobium* gone native: unexpected plasmid stability of indigenous *Rhizobium leguminosarum*. *Proc Natl Acad Sci USA* **94**, 5483–5488.
- Whittam, T. S. (1989). Clonal dynamics of *Escherichia coli* in its natural habitat. *Antonie Leeuwenhoek* **55**, 23–32.
- Whittam, T. S., Ochman, H. & Selander, R. K. (1983). Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. *Mol Biol Evol* **1**, 67–83.