The genetic structure of enteric bacteria from Australian mammals

David M. Gordon and Joannah Lee†

Author for correspondence: David M. Gordon. Tel: +61 2 6249 3552. Fax: +61 2 6249 5573. e-mail: David.Gordon@anu.edu.au

A total of 246 isolates representing five species of the family Enterobacteriaceae, taken from a variety of Australian mammal species, were characterized using multi-locus enzyme electrophoresis. Genome diversity estimates varied significantly among species, with the Klebsiella pneumoniae sample exhibiting the lowest diversity and the Citrobacter freundii sample the highest. Multi-locus linkage disequilibrium estimates revealed that alleles were non-randomly associated in all five species samples, but the magnitude of the estimates differed significantly among species. Escherichia coli had the lowest linkage disequilibrium estimate and Klebsiella oxytoca the largest. Molecular analysis of variance was used to determine the extent to which population structure explained the observed genetic variation in a species. Two population levels were defined: the taxonomic family of the host from which the isolate was collected and the geographical locality where the host was collected. The amount of explained variation varied from 0% for K. oxytoca to 22% for K. pneumoniae. Host locality explained a significant amount of the genetic variation in the C. freundii (12%), E. coli (5%), Hafnia alvei (17%) and K. pneumoniae (22%) samples. Host family explained a significant fraction of the variation in E. coli (6%), H. alvei (7%) and K. pneumoniae (20%). Estimates of effective population size for all five species, based on the probability that two randomly chosen isolates will be identical, failed to reveal any relationship between the effective population size and the genetic diversity of a species.

Keywords: Enterobacteriaceae, genetic structure, population structure, clonality

INTRODUCTION

Over recent years there has been a growing body of theoretical and empirical evidence that has led to the view that the clone concept as a framework for understanding the population genetics and evolution of bacteria requires modification. Although studies of electrophoretic variation at multiple enzyme loci (MLEE) has provided much support for the clonal paradigm, especially for Escherichia coli (Milkman, 1973; Selander & Levin, 1980; Caugant et al., 1981, 1984; Whittam et al., 1983; Pupo & Richardson, 1995; Gordon, 1997), the results of these studies have also highlighted some of the problems with the concept.

Maynard Smith (1991) maintains that two observations are common in the majority of MLEE studies: (1) bacterial populations exhibit extensive genetic diversity; and (2) populations are ‘clonal’, that is, the same haplotype is recovered repeatedly. He argues that these two observations can only be reconciled in the simplest model of clonal structure by assuming unreasonably small effective population sizes. Problems with the clone concept have also been revealed by the results of DNA sequence studies of a variety of E. coli genes (Guttman, 1997). Many of these sequence studies have provided evidence of recombination, sometimes extensive recombination (Dykhuizen & Green, 1986; Dubose et al., 1988; Biseric et al., 1991; Milkman & Bridges, 1993; Guttman & Dykhuizen, 1994; Nelson & Selander, 1994).

Thus, Maynard Smith (1991) argues, the observations of extensive variability of bacterial populations and repeated isolation of particular clones cannot be explained...
by a simple model of clonal structure. Guttman (1997) also reasons that the clone concept in its basic form cannot account for the extensive linkage disequilibrium observed in many MLEE studies in the face of the frequent recombination suggested by some of the DNA sequence studies. These apparently contradictory observations can be reconciled if one assumes that bacterial species exhibit ecological structure—that is, different strains within a species are adapted to particular ecological niches.

The existence of ecological structure in bacterial species would have significant consequences for two processes that are important determinants of genetic structure—periodic selection and recombination. Periodic selection is the name given to the phenomenon whereby a mutation occurs in a clone that results in the clone replacing every other clone in the population, thereby purging all genetic variation (Atwood et al., 1951; Levin, 1981). Ecological structure would limit the extent to which genetic variation is reduced by a periodic selection event (Maynard Smith, 1991). If a species consists of multiple ecological populations then it is unlikely that a beneficial mutation would occur that confers a fitness advantage in all situations. A periodic selection event would purge genetic variation within an ecological population, but not in the species as a whole (Maynard Smith, 1991; Cohan, 1994). Ecological structure would also limit the extent to which recombination might randomize allele associations. If strains are adapted to particular niches then it is possible for there to be extensive recombination within ecological populations, but not between populations. Under such a scenario, simulation studies suggest that linkage disequilibrium would be maintained even when recombination occurs at much greater rate than mutation (Guttman, 1997).

The existence of among-species variation in genetic structure has been an additional piece of evidence that has lead to the realization that the clone concept requires revision. MLEE studies of an increasing number of species have revealed that significant linkage disequilibrium, the hallmark of a clonal genetic structure, is not always observed (Maynard Smith et al., 1993; Gordon et al., 1995). Although these studies suggest that there is considerable variation among species in genetic structure, the reasons for this variation are unknown. This is in part due to the life-history diversity of the species examined—free living, commensal, and facultative or obligate pathogens. The manner in which the characterized isolates were obtained has varied among studies, and these isolates may represent clones collected from single populations, clinical isolates, or subsamples of strains collected in an *ad hoc* fashion.

Recent work has demonstrated that a number of species of enteric bacteria are non-randomly distributed with respect to either the taxonomic family of the host from which they were isolated or the geographical locality where the host was collected (Gordon & FitzGibbon, 1999). Such an observation demonstrates the existence of population structure and suggests that these species may exhibit adaptations specific to particular types of primary (host environment) and secondary (external environment) habitats (sensu Savageau, 1983).

There were two main goals to this study: (1) to determine if species of bacteria with similar life-histories, in this case commensals of the gastro-intestinal tract of mammals, differed in their genetic structure, and (2) to assess the extent to which host and geographical effects accounted for the observed genetic diversity of a species. To these ends, strains of five species of the family *Enterobacteriaceae* isolated from mammals in Australia were characterized using MLEE.

**METHODS**

**Bacterial isolation and identification.** Faecal swabs were obtained from a variety of mammalian hosts representing 79 species from 16 families collected from over 75 localities throughout Australia (Gordon & FitzGibbon, 1999). Primary isolation of the strains was performed by streaking the swabs for single colonies on a MacConkey agar plate (Power & McCuen, 1988). The plates were then examined and a single representative of each colony morphology was selected. These colonies were purified on a second MacConkey agar plate, and then another MacConkey agar plate. Following overnight incubation in LB broth (Power & McCuen, 1988) at 37 °C, the isolates were stored in glycerol at −70 °C.

Colonies morphologically consistent with *E. coli* were then tested for growth on minimal lactose and minimal citrate agar plates (Power & McCuen, 1988) and for indole production. The Lac<sup>+</sup> Ind<sup>−</sup> Cit<sup>−</sup> isolates were considered to be *E. coli* (Ewing, 1986). The biochemical characteristics of the non-*E. coli* isolates were determined using BBL Crystal Enteric/Nonfermenter and BBL Crystal Rapid Stool kits (Becton Dickinson). The isolates were identified from their biochemical profiles using the BBL Crystal System Electronic code book. Further details on strain isolation and identification are presented by Gordon & FitzGibbon (1999).

The species investigated in this project were: *Citrobacter freundii* (n = 32), *Escherichia coli* (n = 59), *Hafnia alvi* (n = 52), *Klebsiella oxytoca* (n = 52) and *Klebsiella pneumoniae* (n = 51). Except for *E. coli*, essentially all of the isolates available were analysed. In the case of *E. coli* over 400 isolates were at hand, so a subset was chosen, at random, for characterization.

**MLEE characterization.** Protein extracts were obtained by first growing each isolate overnight in 12 ml Terrific Broth (Power & McCuen, 1988) at 37 °C with shaking at 200 r.p.m. The cells were pelleted and washed in 1 ml extraction buffer (0.15 M potassium phosphate, pH 7.0). The cells were then pelleted again, after discarding the supernatant, frozen in liquid nitrogen. Following the addition of 150 μl lysis buffer (per litre: 100 mg NADP and 0.5 ml 2-mercaptoethanol) the cells were sonicated using three 5 s bursts. The cellular debris was removed by centrifugation at 14,260 g for 20 min and the supernatant transferred to new tubes for storage at −70 °C. All steps in the extraction procedure were carried out at 4 °C.

The MLEE assays were run on Titan III cellulose acetate gels (Helena Laboratories). A total of 25 enzyme systems were tested using three isolates of each species. The 13 enzyme systems used in this study were selected because they gave reliable results across all species and no attention was paid as to whether the loci were polymorphic for every species. The following enzyme systems were used: ADH (alcohol dehydrogenase, EC 1.1.1.1), AK (adenylate kinase EC 2.7.4.3), G6PD (glucose-6-phosphate dehydrogenase, EC
1.1.1.49), IDH (isocitrate dehydrogenase, EC 1.1.1.42),
MDH (malate dehydrogenase, EC 1.1.1.37), ME (malic
enzyme, EC 1.1.1.40), MPI (mannose phosphate isomerase,
EC 5.3.1.8), MR (menadione reductase, EC 1.6.99.2), PDG
(6-phosphogluconate dehydrogenase, EC 1.1.1.44), PGI
(phosphoglucone isomerase, EC 5.3.1.9), PGM (phospho-
glucomutase, EC 5.4.2.2), SDH (shikimic acid dehydro-
genase, EC 1.1.1.25) and SOD (superoxide dismutase, EC
1.15.1.1). ADH, MDH, PGD, PGI, PGM and SOD were run
in 0.01 M Tris/glycine (TG) buffer, pH 8.5, all others were run
in 0.1 M Tris/EDTA/maleate/MgCl₂ (TEM) buffer, pH 7.4
(Hebert & Beaton, 1989) All strains were run at least twice
against strains giving similar band mobility in order to confirm
their genotype.

No enzyme activity was detected at particular loci in some
strains and these were designated as nulls. Nulls were not
included when determining the total number of alleles per
locus, but were included in the calculations of the genetic
distances between strains and diversity estimates. Some strains
also produced double bands at particular loci; in other words
they were diploid at these loci. The diploid alleles were treated
as a different allele when calculating the distance between
strains and genetic diversity, but not for calculations of the
number of alleles, unless they were unique alleles. Four of the
enzyme systems, ME, MR, PGI and SOD, displayed multiple
loci. The number of loci and allele mobility varied between
species. Loci displaying similar allele mobility were assumed
to be homologous.

Statistical analyses. The genetic diversity of a locus is
determined by two components: the number of alleles present
at that locus (richness), and the frequency of those alleles
(evenness). One way in which evenness can be quantified is to
determine the probability (p_i) that two isolates, selected at
random from a population, will have different alleles at locus
j (Nei, 1978).

\[ h_j = \frac{1 - \sum p_i^2}{n} \]

where \( p_i \) is the frequency of allele \( i \) at locus \( j \) and \( n \) is the
number of isolates. Genome diversity (H) is estimated as the
arithmetic mean of \( h_j \) for \( m \) loci.

\[ H = \frac{\sum h_j}{m} \]

The haplotype of a strain is determined by the combination of
alleles it possess at \( m \) loci. Diversity estimates were also
calculated using only the unique haplotypes, that is the
frequency with which each haplotype was recovered was not
considered (\( H_{ET} \), where ET denotes electrophoretic type).

The haplotype diversity (G) is the chance that two strains,
drawn at random, will differ at least one of \( m \) loci,

\[ G = \frac{1 - \sum g_i^2}{n} \]

where \( g_i \) is the frequency of the \( i \)th haplotype, and \( n \) is the
total number of isolates.

Among-species comparisons of genome diversity were based
on only the 12 loci common to all species. A resampling
approach was taken in order to investigate whether there was
a significant difference in genome diversity (\( H \)) among species.
More specifically, the probability of a random sample drawn
from one species having an \( H \) as large (or as small) as the \( H \)
calculated for a random sample drawn from another species
was determined. The procedure was as follows. For each
species, 10000 pseudo-samples equal in size to the original
sample were created by randomly sampling, with replacement,
isolates from the original data set. Genome diversity was then
calculated for each random sample, \( H_i \). The species were
compared in a pairwise fashion. Each of the 10000 \( H_i \) values
for the two species were arbitrarily paired and the differences
between the \( H_i \) values calculated. The number of times that an
\( H_i \) for the less diverse species exceeded the \( H_i \) of the more
diverse species was determined. Not all possible pairwise
species combinations were tested. Among-species com-
parisons were made by first comparing the species with the
greatest difference between their observed \( H \) values. If this
difference was significant then the species pair with the next
greatest difference was tested, and so on, until the difference
was found to be insignificant.

Estimates of linkage disequilibrium utilized the multi-locus
index developed by Brown et al. (1980). \( K \) is the total number
of loci at which two individuals differ; there are \( n(n - 1)/2 \)
such pairs in a sample of \( n \) individuals. The mean difference
between two isolates is \( K = \Sigma h_j \). In the absence of linkage
disequilibrium, the expected variance of \( K \) is \( V_e = \Sigma h_j(1 - \Sigma h_j) \),
\( V_e \) is the observed variance of \( K \). In the absence of linkage
disequilibrium the observed variance \( (V_o) \) is expected to equal
the expected variance \( (V_e) \) and therefore the ratio \( V_o/V_e \)
should have a value of 1. To test if the observed value of \( V_o/V_e \)
is significantly greater than 1, a randomization approach is
taken. Samples the same size as the original data set were
generated by randomly sampling alleles, without replacement,
according to their respective frequencies at each locus (Souza
et al., 1993). For each random sample \( V_o \) and \( V_e \) were
calculated and the probability of observing a \( V_o/V_e \) ratio as
extreme as that for the original data was determined.

The value of \( V_o/V_e \) depends on the number of loci (Maynard
Smith, 1991; Souza et al., 1993) and to a lesser extent on allele
frequencies (Brown et al., 1981). For this reason, among-
species comparisons were based on the eight polymorphic
loci common to all species. To test whether there was a significant
difference in \( V_o/V_e \) ratios among species the same approach
was used as for comparing genomic diversity among species
(Souza et al., 1993).

The extent to which the different species were influenced by
population-level effects was examined. Two population levels
were defined: the taxonomic group of the host from which an
isolate was recovered and the geographical locality where the
host was collected. Host level was chosen to be the taxonomic
family. The geographical level was defined as the state or
territory in which the host was collected. As a consequence of
the way hosts were collected, the majority of the samples were
collected from a single region within each state. For example,
most of the Northern Territory samples were collected from a
small area of Kakadu National Park. Therefore, the majority of
the localities within a state represent a single climatic
region.

Molecular analysis of variance (AMOVA) (Excoffier et al.,
1992) was used to determine how much of the allozyme
variation among strains of a species was explained by a
population effect. AMOVA is built on the principles of classic
analysis of variance. However, the data set analysed with
AMOVA is the matrix of pairwise genetic distances (the
number of loci at which two strains differ). The estimated
variance components are expressed as the proportion of the
total variance explained by the population level effect
\( (\phi_0) \), \( \phi_1 \) statistics have an interpretation equivalent to the
classic single-locus \( F \) statistics of Wright (1943), which are used for
assessing the extent of population differentiation. Although
AMOVA is based on analysis of variance, $F$ statistics cannot be used to assess significance levels. Instead a randomization approach is used, whereby strains are randomly assigned to populations and $\phi_{st}$ is recalculated. This procedure is repeated a number of times and the probability of obtaining a $\phi_{st}$ value as extreme as that observed is determined. The AMOVA analysis was done using the software WINAMOVA 1.5 (© L. Excoffier, 1993).

RESULTS

Allozyme variation

The allelic diversity at a locus varied within and between species (Table 1). For example in C. freundii, depending on the locus, from 1 to 6 alleles were identified and the diversity ($h_j$) ranged from 0 to 0.71 (Table 1). Whilst depending on the species, MDH exhibited from 1 to 5 alleles and diversities ($h_j$) ranging between 0 and 0.54 (Table 1). Correlation analysis of the number of alleles and allelic diversity at a locus was performed on all pairwise species combinations. In no case was any statistically significant correlation observed ($P > 0.05$). In other words, a low-diversity locus in one species was not necessarily a low-diversity locus in another, or a high-diversity locus in one species a high one in another. Estimates of genome diversity ($H_{ET}$) ranged from 0.28 for K. pneumoniae to 0.41 for K. oxytoca when haplotype frequency was not considered (Table 2). In order to compare genome diversity ($H$) among species a

Table 1. Allelic diversity estimates of each locus for five species of enteric bacteria

<table>
<thead>
<tr>
<th>Locus</th>
<th>E. coli</th>
<th>C. freundii</th>
<th>H. alvei</th>
<th>K. pneumoniae</th>
<th>K. oxytoca</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH</td>
<td>0.42 (5)</td>
<td>–</td>
<td>0.54 (4)</td>
<td>0.08 (2)</td>
<td>0.18 (3)</td>
</tr>
<tr>
<td>AK</td>
<td>0.36 (3)</td>
<td>0.37 (3)</td>
<td>0 (1)</td>
<td>0.08 (1)</td>
<td>0.43 (2)</td>
</tr>
<tr>
<td>G6PD</td>
<td>0.00 (1)</td>
<td>0.33 (4)</td>
<td>0.04 (2)</td>
<td>0.18 (3)</td>
<td>0.11 (2)</td>
</tr>
<tr>
<td>IDH</td>
<td>0.35 (4)</td>
<td>0.32 (4)</td>
<td>0 (1)</td>
<td>0.33 (4)</td>
<td>0.04 (2)</td>
</tr>
<tr>
<td>MDH</td>
<td>0.19 (3)</td>
<td>0 (1)</td>
<td>0.30 (4)</td>
<td>0.54 (5)</td>
<td>0.49 (5)</td>
</tr>
<tr>
<td>ME-1</td>
<td>0.03 (2)</td>
<td>0 (1)</td>
<td>0.64 (4)</td>
<td>0.15 (4)</td>
<td>0.51 (4)</td>
</tr>
<tr>
<td>ME-2</td>
<td>0.16 (3)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MPI</td>
<td>0.50 (5)</td>
<td>0.68 (5)</td>
<td>0.61 (5)</td>
<td>0.08 (2)</td>
<td>0.11 (2)</td>
</tr>
<tr>
<td>MR-1</td>
<td>0.22 (4)</td>
<td>0.28 (4)</td>
<td>0.58 (4)</td>
<td>0.14 (1)</td>
<td>0.50 (2)</td>
</tr>
<tr>
<td>MR-2</td>
<td>0.59 (2)</td>
<td>–</td>
<td>0.49 (2)</td>
<td>0.11 (2)</td>
<td>0.35 (3)</td>
</tr>
<tr>
<td>MR-3</td>
<td>0.18 (1)</td>
<td>0.22 (1)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MR-4</td>
<td>0.21 (2)</td>
<td>–</td>
<td>–</td>
<td>0.11 (1)</td>
<td>0.11 (3)</td>
</tr>
<tr>
<td>PGD</td>
<td>0.48 (4)</td>
<td>0.57 (4)</td>
<td>0.47 (5)</td>
<td>0.62 (4)</td>
<td>0.59 (5)</td>
</tr>
<tr>
<td>PGI-1</td>
<td>–</td>
<td>–</td>
<td>0.74 (5)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PGI-2</td>
<td>0.27 (4)</td>
<td>0.33 (5)</td>
<td>0.49 (2)</td>
<td>0.18 (5)</td>
<td>0.30 (3)</td>
</tr>
<tr>
<td>PGM</td>
<td>0.07 (2)</td>
<td>0.56 (5)</td>
<td>0.27 (3)</td>
<td>0.11 (2)</td>
<td>0.36 (3)</td>
</tr>
<tr>
<td>SDH</td>
<td>0.42 (4)</td>
<td>0.71 (6)</td>
<td>0.71 (5)</td>
<td>0.64 (4)</td>
<td>0.56 (6)</td>
</tr>
<tr>
<td>SOD-1</td>
<td>–</td>
<td>–</td>
<td>0.04 (1)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SOD-2</td>
<td>0.19 (3)</td>
<td>0.54 (4)</td>
<td>0.04 (1)</td>
<td>0.11 (2)</td>
<td>0.45 (2)</td>
</tr>
</tbody>
</table>

–, No alleles detected at this locus for the species.
* The number of alleles at a locus (does not include nulls) is shown in parentheses.

Table 2. Summary statistics of genetic diversity and structure for five species of enteric bacteria

<table>
<thead>
<tr>
<th>Species</th>
<th>n</th>
<th>m</th>
<th>Alleles</th>
<th>$G$</th>
<th>$H$</th>
<th>$H_{ET}$</th>
<th>$V_o/V_e$</th>
<th>$P &gt; V_o/V_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. freundii</td>
<td>32</td>
<td>13</td>
<td>3-4</td>
<td>0.98</td>
<td>0.38</td>
<td>0.39</td>
<td>2.59</td>
<td>$P &lt; 0.002$</td>
</tr>
<tr>
<td>H. alvei</td>
<td>52</td>
<td>16</td>
<td>3-1</td>
<td>0.99</td>
<td>0.37</td>
<td>0.39</td>
<td>2.80</td>
<td>$P &lt; 0.008$</td>
</tr>
<tr>
<td>K. oxytoca</td>
<td>52</td>
<td>15</td>
<td>2-5</td>
<td>0.95</td>
<td>0.34</td>
<td>0.41</td>
<td>3.06</td>
<td>$P &lt; 0.002$</td>
</tr>
<tr>
<td>K. pneumoniae</td>
<td>51</td>
<td>15</td>
<td>2-8</td>
<td>0.96</td>
<td>0.23</td>
<td>0.28</td>
<td>2.91</td>
<td>$P &lt; 0.005$</td>
</tr>
<tr>
<td>E. coli</td>
<td>59</td>
<td>17</td>
<td>3-1</td>
<td>0.99</td>
<td>0.27</td>
<td>0.30</td>
<td>1.56</td>
<td>$P &lt; 0.001$</td>
</tr>
</tbody>
</table>

Definitions: $n$, number of isolates examined; $m$, number of loci scored; alleles, mean number of alleles per locus; $G$, haplotype diversity; $H$, mean genome diversity (based on all isolates); $H_{ET}$, mean genome diversity (based on electrophoretic types), $V_o/V_e$, multi-locus linkage disequilibrium estimate; $P > V_o/V_e$, probability that alleles are randomly associated.
bootstrap approach was taken using only the 12 loci common to all species (Table 3). The frequency histogram for the 10000 pseudo-samples for each species is presented in Fig. 1. The results of the randomization showed that the diversity of *K. pneumoniae* was significantly less than that of *H. alvei* (*P < 0.011*) and therefore *K. oxytoca* and *C. freundii* also. The diversity of *E. coli* was also significantly less than that of *H. alvei* (*P < 0.001*), *K. oxytoca* and *C. freundii*. *H. alvei*, *K. oxytoca* and *C. freundii* were not significantly different from each other. Basing the comparisons on all of the loci resulted in similar conclusions.

### Linkage disequilibrium

The linkage disequilibrium estimate \( V_o/V_e \) ranged from 1.56 for *E. coli* to 3.06 for *K. oxytoca*, and for every species \( V_o/V_e \) was significantly greater than 1 (Table 2). Among-species comparisons of \( V_o/V_e \) were based on the eight polymorphic loci common to all species (Table 3). Because the magnitude of \( V_o/V_e \) depends on the number of loci considered, estimates based on only eight loci are smaller than those based on all loci, but were still significantly greater than 1 (Table 3).

Again a randomization approach was taken to determine whether \( V_o/V_e \) was significantly different among species. The frequency histogram for the 10000 pseudo-samples is presented in Fig. 1. The degree of linkage disequilibrium displayed by *K. pneumoniae* was not significantly different from any of the other species. This was because the variation in \( V_o/V_e \) displayed by *K. pneumoniae* was much greater than for the other species. The level of linkage disequilibrium displayed by *E. coli* was significantly smaller than that of *H. alvei* (*P < 0.009*) and hence *C. freundii* and *K. oxytoca* also. The levels of *C. freundii* and *K. oxytoca* were not significantly different from each other. The degree of linkage disequilibrium of *H. alvei* is significantly smaller than that of *K. oxytoca* (*P < 0.043*), but not that of *C. freundii*. Again, basing the comparisons on all of the loci would have resulted in similar conclusions, since the ranking remains the same for both data sets (except for *K. pneumoniae* and *C. freundii* switching positions).

### Population structure

As a result of the fact that bacterial isolates are not available for every possible host family/locality combination, a fully nested analysis of the genetic data was not possible. Rather, AMOVA was performed by partitioning the data with respect to host locality, host family, and the host locality/family combinations available.

Neither host locality nor host family was found to explain any of the observed genetic variation in *K. oxytoca* (Table 4). By contrast, each of these factors accounted for about 20% of the variation observed in *K. pneumoniae*.

### Table 3. Mean genome diversity (\( H \)) and linkage disequilibrium estimates (\( V_o/V_e \)) for five species of enteric bacteria based on the loci common to all species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Diversity (( H )) (12 loci)</th>
<th>( V_o/V_e ) (8 loci)</th>
<th>( P &gt; V_o/V_e )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>K. pneumoniae</em></td>
<td>0.26</td>
<td>1.61</td>
<td><em>P &lt; 0.001</em></td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>0.26</td>
<td>1.13</td>
<td><em>P &lt; 0.029</em></td>
</tr>
<tr>
<td><em>H. alvei</em></td>
<td>0.35</td>
<td>1.58</td>
<td><em>P &lt; 0.001</em></td>
</tr>
<tr>
<td><em>K. oxytoca</em></td>
<td>0.37</td>
<td>2.03</td>
<td><em>P &lt; 0.001</em></td>
</tr>
<tr>
<td><em>C. freundii</em></td>
<td>0.39</td>
<td>1.96</td>
<td><em>P &lt; 0.001</em></td>
</tr>
</tbody>
</table>

**Fig. 1.** Frequency distribution of genome diversity (\( H \)) and linkage disequilibrium estimates (\( V_o/V_e \)) derived from 10000 random resamplings of the original MLEE data sets for five species of enteric bacteria.
pneumoniae. Population structure explained a significant, but small, amount of the variation in E. coli. For C. freundii and H. alvei geographical effects explained more of the observed variation than did the family of host from which the strains were isolated.

DISCUSSION

Gordon & FitzGibbon (1999) demonstrated that each of the bacterial species considered in this study were significantly non-randomly distributed with respect to the family and/or geographical origin of the host from which the bacteria were isolated. That is, the probability of isolating a particular species of bacteria from a host depended on the taxonomic family of the host and its geographical locality. These same factors account for a significant proportion of the genetic diversity observed in four of the five species examined (Table 4). However, if host family or locality had an effect on the probability of occurrence for a species, this did not necessarily mean that the same factor(s) explained a significant amount of the genetic variation (Table 5). C. freundii was found to be non-randomly distributed with respect to host locality, and locality explained the greatest amount (12%) of the observed genetic diversity. By contrast, H. alvei was far more likely to be isolated from particular host families and its distribution was not influenced by locality. However, both host locality (17.5%) and host family (7.2%) explained a significant fraction of the observed diversity in this species. Host family, locality and family/locality combinations all explain about the same amount of genetic variation for K. pneumoniae. Although the probability of occurrence for this species depended on locality, it did not vary with respect to host family. The observation that neither host family nor locality explained any of the genetic variation observed in K. oxytoca does not accord with the observation that this species was far more likely to be recovered from particular host families. There are few other studies of enteric bacteria available for comparison. Whittam et al. (1983) reported that about 2% of the allelic diversity in E. coli isolated from humans living on three continents was explained by a geographical effect. In a study of E. coli isolated from two populations of feral house mice,

### Table 4. Amount of genetic variation explained by population effects in five species of enteric bacteria

<table>
<thead>
<tr>
<th>Species</th>
<th>Percentage among-population variation, ( \phi_{st} ) (probability of a greater ( \phi_{st} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Host locality</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>C. freundii</td>
<td>11.8 (P &lt; 0.001)</td>
</tr>
<tr>
<td>E. coli</td>
<td>4.5 (P &lt; 0.015)</td>
</tr>
<tr>
<td>H. alvei</td>
<td>17.5 (P &lt; 0.001)</td>
</tr>
<tr>
<td>K. oxytoca</td>
<td>0.3 (P &gt; 0.447)</td>
</tr>
<tr>
<td>K. pneumoniae</td>
<td>21.9 (P &lt; 0.001)</td>
</tr>
</tbody>
</table>

Table 5. Summary of the impact of host locality and host family on the occurrence and genetic structure of five species of enteric bacteria

<table>
<thead>
<tr>
<th>Species</th>
<th>Host locality</th>
<th>Host family</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probability of occurrence</td>
<td>Genetic Diversity</td>
</tr>
<tr>
<td>C. freundii</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>E. coli</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>H. alvei</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>K. oxytoca</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>K. pneumoniae</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*The probability of isolating the bacterial species from a particular host.
temperatures varying from 0 to 20 °C. In vitro studies with E. coli have demonstrated that this species adapts to the thermal environment in which it finds itself (Bennet & Lenski, 1993). It is therefore possible that these enteric bacteria are adapted to the specific internal and external environments to which they are exposed. Therefore, the observation that host and locality effects explain a significant amount of the genetic diversity in four of the five species examined may indicate the presence of multiple ecological populations within these species.

In addition to the possible existence of ecological structure, bacteria, particularly host-associated species, may exhibit population substructure. For example, Caugant et al. (1984) isolated E. coli from each member of five households. An AMOVA analysis of the E. coli they isolated from the human family members reveals that 74% of the total observed haplotype diversity is explained by differences among individuals within a family. Much of the observed difference among individuals probably arises as a consequence of two main factors. First, it is a consequence of stochastic immigration and extinction events (Hartl & Dykhuizen, 1984). That is, to a large extent chance may determine which particular haplotype from a pool of haplotypes a host is exposed to, and if this haplotype succeeds in establishing in the host. Second, experimental evidence suggests that once a haplotype has established in a host it is difficult for another haplotype to displace the resident strain (Hartl & Dykhuizen, 1984). Thus chance and limited migration will result in hosts having different haplotype communities.

At this time, there is insufficient data to conclude if the observed host family and/or host locality effects indicates the existence of ecological populations or simply reflects population substructure. It is possible that limited migration, rather than ecological adaptation, may explain the observation that host locality accounts for a significant amount of the observed genetic variation in four of the five enteric species examined in this study. Several families, many genera, and almost all species of Australian mammal are not represented in every region of Australia (Strahan, 1983). Furthermore, even in those groups that are present in multiple regions, there are many geographical or climatic features that will limit among-population movement (Clemens et al., 1989). Similarly, the host effects may arise through limited ‘migration’. When hosts from different families occur in the same geographical area they will usually use the habitat in a very different manner. For example, it is likely that bats, kangaroos and possums living in the same area will be exposed to different subsets of the enteric bacteria present in that habitat. Gordon & FitzGibbon (1999) showed that K. oxytoca was far more likely to be isolated from bats than most other families of mammal, yet host family explained none of the genetic diversity observed in this species (Table 4). The bats examined in this study were all insectivores, and K. oxytoca is commonly isolated from insects (Howard et al., 1985). Thus, host diet may explain why K. oxytoca is common in bats rather than the alternative expla-
nation that there are strains of *K. oxytoca* that adapted to living in the gut of bats.

Discriminating between structure due to the existence of ecological populations and population substructure is likely to be difficult for two reasons. First, ecological populations are defined on the basis of adaptive traits, and standard genetic analysis techniques are unlikely to detect these traits: techniques such as allozyme analysis are thought to detect primarily ‘nearly neutral’ variation (Maynard Smith, 1991; Selander et al., 1987), and the adaptive traits are likely to involve a relatively small fraction of the genome (Lenski et al., 1998). The second problem relates to the definition of an ecological population. Although an adaptation’s domain of competitive superiority is an elegant definition of an ecological population, in reality the extent of these domains may be quite restricted. For example, in relation to the primary habitat of enteric bacteria, two hosts of the same age, sex and species may well represent distinct environments. Diet and immune status are just two of the many host-related factors that may differ between individuals. These factors will not only influence the environment provided for a particular enteric species, but will also influence the composition of the entire gut community.

This study demonstrates that the genetic diversities of the five species samples of enteric bacteria investigated are significantly different. The diversity of a species is determined by the mutation rate, the effective population size, the frequency and intensity of periodic selection and the rate of recombination. The only factor that might be rejected, on a priori grounds, as having a significant impact is among-species variation in the genome-level rate of mutation. Drake (1991) demonstrated that there is little variation in mutation rates to be found among DNA-based micro-organisms. However, none of the other factors appear to explain any of the among-species differences in diversity in a simple manner.

In the simplest model of an asexually reproducing organism, the probability that two isolates selected at random will be identical (F) is (Maynard Smith, 1991):

\[ F = \frac{1}{1 + 2N_e U} \]

where \( N_e \) is the effective population size and \( U \) is per-genome rate of recognizable mutations. \( U = \mu nm \), where \( \mu \) is the rate of an amino acid change, \( n \) is the number of possible allelic variants at a locus and \( m \) is the number of loci. As previously stated, there are no data to suggest that there are significant differences in \( \mu \) among the species examined. Nor are there any statistically significant differences among species in the number of alleles detected at the 12 loci common to all species. Hence, there should be no difference in \( U \) among species. Therefore, the effective population size of a species will be proportional to \( 1/F \). Estimates of \( F \), based on the 12 loci common to all species indicates that the rank order of \( N_e \) from largest to smallest is: *E. coli* (\( F^{-1} = 155 \)), *H. alvei* (58), *C. freundii* (41), *K. pneumoniae* (24), *K. oxytoca* (17). This suggests about a sixfold difference between *E. coli* and *K. pneumoniae*, two species with similar levels of diversity, whilst *K. oxytoca*, the species with the smallest estimated population size, is one of the most diverse species.

Alternatively, it could be assumed that effective population size will be proportional to real population size. Given this assumption, then *E. coli* should have the largest effective population size as it is isolated five times more frequently than any of the other species. Furthermore, the relative abundance of the other species suggests that they should have comparable effective population sizes, as they were recovered at similar frequencies (Gordon & FitzGibbon, 1999). However, *E. coli* is no more or less diverse than the less common species and the other five species, which were about equally common, had significantly different levels of diversity.

The diversity-purging effects of a periodic selection event will decline as the degree of structure, be it ecological structure or population substructure, within a species increases. This result occurs because while an adaptive mutation may result in the purging of variation within an ecological population, it will have little effect on the level of diversity within the species overall. The same factors that generate population substructure will also serve to lessen the impact of a periodic selection event. However, *K. pneumoniae* and *E. coli* exhibited comparable levels of diversity but structure accounted for very different amounts of the total variation in these two species. Furthermore host family or locality explained none of the observed variation in *K. oxytoca*, one of the most genetically diverse species. However, there may well be factors other than host or locality causing population structure in *K. oxytoca* that might explain this species genetic diversity. A sufficiently ‘novel’ adaptation might result in the purging of variation in the species as a whole, and differences among species in the time when such a major selective sweep occurred could explain the observed among-species differences in diversity. There are no data available to test this possibility.

Finally, the rate of within-species recombination will determine levels of diversity. Relatively high rates of within-species recombination will result in little loss of genetic variation following the appearance of an adaptive mutation, as the novel mutation can spread among existing genetic variants, thereby reducing the purging effect of the periodic selection event. There are no data available concerning the frequency of recombination in these species of bacteria. The magnitude of the linkage disequilibrium estimates is unlikely to provide a reliable indication of the frequency of recombination in these species, as the species exhibit differing degrees of population structure. Theoretical studies have shown that linkage disequilibrium can be present even in the face of frequent within-species recombination, provided...
that there is some sort of population structure present (Guttman, 1997).

This study, and that of Gordon & FitzGibbon (1999), has demonstrated that there are considerable differences among a group of closely related bacterial species with seemingly similar life histories, all of which were collected in a similar manner. The likelihood of isolating a particular bacterial species depended on the taxonomic group to which the host belonged and/or on where the host lived. There were significant differences among the species in their genetic diversity and multi-locus linkage disequilibrium estimates. The extent to which the taxonomic family of the host or host locality explained any of the observed variation varied considerably among species.

All of these results suggest that the relative importance of different epidemiological and genetic processes varies considerably among these species. However, virtually nothing is known about the ecology/epidemiology of these species. Whilst we have a good theoretical understanding of the processes shaping the genetic structure of a species, the results of this study illustrate that explaining the observed structure of a species in the context of the relative importance of these processes and how they interact will not be a straightforward undertaking.

ACKNOWLEDGEMENTS

This study could not have been undertaken without the assistance of a great many researchers throughout Australia and their help is greatly appreciated. Funding for this study was provided by the Australian Research Council.

REFERENCES


Received 10 March 1999; revised 21 June 1999; accepted 1 July 1999.