Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 1751–1755, March 1983
Population Biology

Multilocus genetic structure in natural populations of *Escherichia coli*

(enzyme polymorphism/electrophoretic types/linkage disequilibrium/neutral allele theory/*Shigella*)

THOMAS S. WHITTAM, HOWARD OCHMAN, AND ROBERT K. SELANDER

Department of Biology, University of Rochester, Rochester, New York 14627

Contributed by Robert K. Selander, December 8, 1982

ABSTRACT **A survey of allozyme variation at 12 enzyme loci in 1,705 clones of the genetic species *Escherichia coli* (including four species of *Shigella*) from natural populations revealed 302 unique allele combinations (electrophoretic types). Single-locus diversity estimates fall within the range predicted by the neutral allele theory of molecular evolution, but the combinations of alleles in electrophoretic types are highly nonrandom, as indicated by a test of association over all loci and by evidence of complex linkage disequilibria in several four-locus combinations. These linkage disequilibria reflect genetic differentiation of *E. coli* into three groups of strains. Because of restricted recombination, both the stochastic extinction of lines and selective differences between particular genetic combinations may have contributed to the evolution of subspecific structure in *E. coli*.**

Genetic studies of *Escherichia coli* have focused primarily on a small number of laboratory strains originally isolated from human hosts (1, 2). As a consequence, relatively little is known about the genetic structure of natural populations inhabiting the lower intestines of warm-blooded vertebrates. Serotyping (3–5) and DNA hybridization (6–9) have indicated that *E. coli* consists of many genetically distinct types, but only recently have quantitative estimates of allelic diversity at individual gene loci (10) been obtained for natural populations through the use of protein electrophoresis. Milkman (11, 12) obtained a mean genetic diversity of 0.23 for five enzyme loci in several hundred clones isolated from a wide variety of mammalian hosts, but this estimate was later revised upward to 0.47 by Selander and Levin (13), who assayed 20 enzyme loci in 109 clones. Genetic diversity in *E. coli* is two or three times greater than comparable estimates for eukaryotic species (14–16).

A second observation by Selander and Levin (13) contradicted the model of genetic structure proposed by Milkman (11, 12), in which populations were viewed as essentially panmictic, with rates of recombination sufficiently high to generate, *in situ*, most of the diversity of strains occurring in individual hosts. The repeated recovery of clones with identical multilocus allozyme profiles from unassociated hosts suggested that recombination is severely limited and that the diversity of types within an individual host results primarily from the continual immigration of new strains. This interpretation derives support from two sources: First, rates of phage-mediated and conjugative-plasmid transfer of genes in populations in chemostats are very low, being on the order of the mutation rate (17, 18). Second, the turnover rate of unrelated strains within the flora of an individual human host is high, with complete replacement of strains often occurring within a period of 2 weeks (19).

The implication of low rates of recombination in *E. coli* is that natural populations are mixtures of more or less independently evolving lines (strains). This has two evolutionary consequences: First, genetic drift may affect allele frequencies through the random extinction of lines, despite an enormous total population size (20–22). Moreover, frequent local extinction may generate large variances in the coefficients of linkage disequilibrium in a finite subdivided population (23). Second, with low rates of recombination, natural selection acting on variation at one genetic locus will cause allele frequency changes at other loci (24). Thus, neutral or slightly deleterious alleles can "hitchhike" with favorable mutations (25), and neutral alleles at two loci will tend to be in linkage disequilibrium through the action of selection on a third locus (26). The overall effect of the frequent extinction of lines, whether by random processes or periodic selection of clones of high fitness, is to reduce the effective population size and, hence, the amount of genetic variation carried by the population as a whole.

In an effort to understand the interaction between genetic drift and natural selection in determining the genetic structure of natural populations of *E. coli*, we have combined allozyme data from several studies into a single comprehensive analysis. Because *E. coli* and *Shigella* are considered to be one genetic species (7, 27, 28), we have also included a number of strains of *Shigella* in much of the analysis. We here address the following questions. (*i*) How polymorphic are natural populations of *E. coli* and *Shigella*, as revealed by protein electrophoresis? (*ii*) Do the observed values of genetic diversity fit the expectations of the neutral allele theory of molecular evolution (29)? (*iii*) What is the extent and degree of multilocus linkage disequilibrium? (*iv*) Is there evidence of subspecific structure within the species *E. coli*?

MATERIALS AND METHODS

Our analysis is based on 1,705 clones (single-cell isolates from feces or urine of hosts), including 109 clones of *E. coli* isolated from various human and animal sources in North America (13), 550 clones collected from a human host in Massachusetts over an 11-month period (19), 268 clones from fecal samples and urine specimens collected from a human population in Göteborg, Sweden (30), and 655 clones from six human families in Massachusetts and New York. We also examined 123 clones of *Shigella*, representing *S. boydii*, *S. dysenteriae*, *S. sonnei*, and *S. flexneri*, isolated from human hosts in various parts of the world.

In each study, protein extracts from individual clones were subjected to standard starch-gel electrophoresis (13, 31). The following 12 enzymes were common to all the studies cited above: malate dehydrogenase (MDH), 6-phosphogluconate dehydrogenase (6PGD), β-galactosidase (βGA), adenylate kinase (AK), phenylalanylleucine peptidase (PE2), glutamic-oxaloacetic transaminase (GOT), isocitrate dehydrogenase (IDH), phosphoglucose isomerase (PGI), aconitase (ACO), mannose-6-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviation: 6PGD, 6-phosphogluconate dehydrogenase.

phosphate isomerase (MPI), glucose-6-phosphate dehydrogenase (G6PD), and alcohol dehydrogenase (ADH). All the enzymes assayed are known to be encoded by chromosomal genes (19).

For each enzyme, we distinguished a number of allozymes, which we equated with alleles; clones lacking activity for a particular enzyme were designated as "null" at that locus. Clones were classified by the combination of alleles observed at the 12 loci, each unique combination being designated as an electrophoretic type (19).

RESULTS

Enzyme Polymorphism. In the sample of 1,705 clones, we identified 302 electrophoretic types, of which 279 were *E. coli* and 23 were *Shigella*. The number of allelic states (allozymes and null alleles) among electrophoretic types averaged 9.3 per locus and ranged between 3 at *G6PD* and 23 at *βGA* for *E. coli*. *Shigella* had only 2.9 alleles per locus, including two monomorphic loci (*AK* and *G6PD*); 27 (77%) of the allozymes in *Shigella* were indistinguishable from those found in *E. coli*. Genetic diversity for each locus was calculated as $h = 1 - \Sigma x_i^2$, in which x_i is the frequency of the ith allelic state (10). Mean diversity over 12 loci was 0.52 for *E. coli* and 0.29 for *Shigella*.

We compared the levels of genetic diversity for single loci with those predicted under the hypothesis of strict neutrality of molecular polymorphism (32). According to this hypothesis, all genes mutate with the same probability and each new mutant is a novel allelic type; allele frequencies fluctuate through time and allele substitutions occur purely by chance through stochastic processes acting in a finite population, because there are no selective advantages of one allele over another. Ewens (32) developed a test of the neutral hypothesis in which the expected genetic diversity is simply a function of the sample size and the number of alleles observed. Using recent modifications of this test (33, 34), we compared the observed values of $\hat{F}$ $(= 1 - h)$ for 12 loci to the empirical significance levels for rejection of the neutral hypothesis (Fig. 1). All the values fall within the neutral range.

Although allele frequency distributions for single loci conform to expectations of the neutral hypothesis, the observed mean genetic diversity (0.52) is much less than that expected on the basis of the order of magnitude estimates of mutation rate and effective population size made by Milkman (12) for *E. coli*. But,

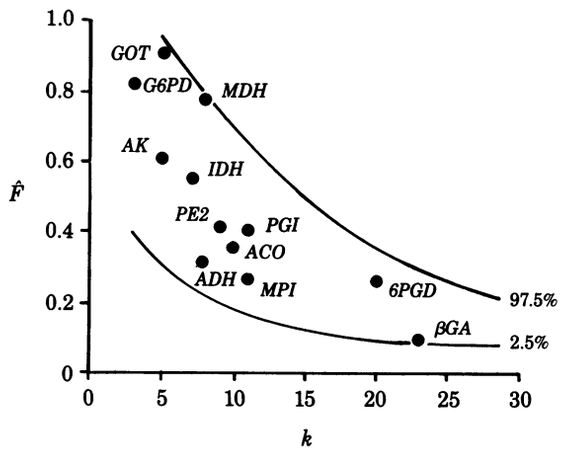

FIG. 1. The test statistic $\hat{F}$ $(= 1 - h)$ for 12 loci presented in Table 1 as a function of the number of allelic states, k. Lines indicate empirical significance levels (2.5%, 97.5%) generated from computer simulations for a sample of 300 proteins, as given in appendix C of Ewens (33).

as noted above, genetic diversity can be substantially reduced when recombination rates are very low and line extinction rates are high (20). Under these conditions, both natural selection of lines and genetic drift should generate linkage disequilibrium among loci, and allele associations are expected to persist for long periods of time. In the following sections, we assess the degree of multilocus disequilibrium manifested over all loci and in several combinations of four loci.

Overall Assessment of Multilocus Structure. To test the null hypothesis of independent occurrence of alleles in strains (linkage equilibrium), we used the method of Brown *et al.* (35) for detecting overall deviations from random assortment of alleles at many loci. Comparing each of the 302 electrophoretic types with itself and with every other type in turn, we recorded the number of loci, from 0 to 12, that mismatched (i.e., the number of loci having different alleles). We then calculated the three central moments of this empirical distribution (method B of ref. 35). The central moments of the expected distribution of the number of mismatches were calculated from single-locus genetic diversities by equations 3–5 in ref. 35.

Both the empirical and the expected moments are presented in Table 1. The first row shows that the observed variance is inflated by associations among alleles at different loci. The observed third and fourth moments also differ from the expected moments: the observed distribution of mismatches is more left-skewed and leptokurtic than expected.

Although this test demonstrated strong multilocus disequilibrium, it provided no information about which particular alleles are associated within electrophoretic types. The next step in our analysis was to measure degrees of association among specific combinations of alleles.

Four-Locus Combinations. From a principal components analysis, we chose subsets of alleles at different loci whose occurrence tended to be correlated in electrophoretic types. For the principal components analysis, each electrophoretic type was represented by a binary code indicating the presence or absence of each allele. The first two principal axes (Fig. 2) explain 10% of the total variance among types; we used those alleles with high loadings on these axes to analyze four-locus associations.

The relative frequencies of combinations of four alleles for three separate comparisons (Table 2) are shown in Table 3. Each of the 16 possible allele combinations within a comparison corresponds to a cell in a 2^4 contingency table in which a dimension represents the presence or absence of a designated allele. The observed frequency of a combination can be compared with the frequency expected if alleles occurred independently in electrophoretic types by examining the relative deviation given in parentheses in Table 3.

Comparison I involves four alleles with high negative loadings on factor 1, and thus represents the cluster of electrophoretic types on the left side of Fig. 2. The four-allele combination of ADH^1, ACO^7, PGI^4, and IDH^2 (combination number 16) is represented in approximately 12% of the 302 types, but it would

Table 1. Estimates of multilocus genetic parameters for 302 electrophoretic types of *E. coli* and *Shigella*

Moment	$M(i)$	$\mu(i)$	$X(i)$
$i = 2$	3.218*	2.292	0.404
$i = 3$	−2.786	−0.151	17.415
$i = 4$	32.431	15.187	1.135

* Exceeds upper 95% confidence limit ($L = 2.655$) calculated from equation 23 in ref. 33. $M(i)$ = observed central moment; $\mu(i)$ = expected central moment; $X(i) = [M(i)/\mu(i)] - 1$ = measure of multilocus structure.

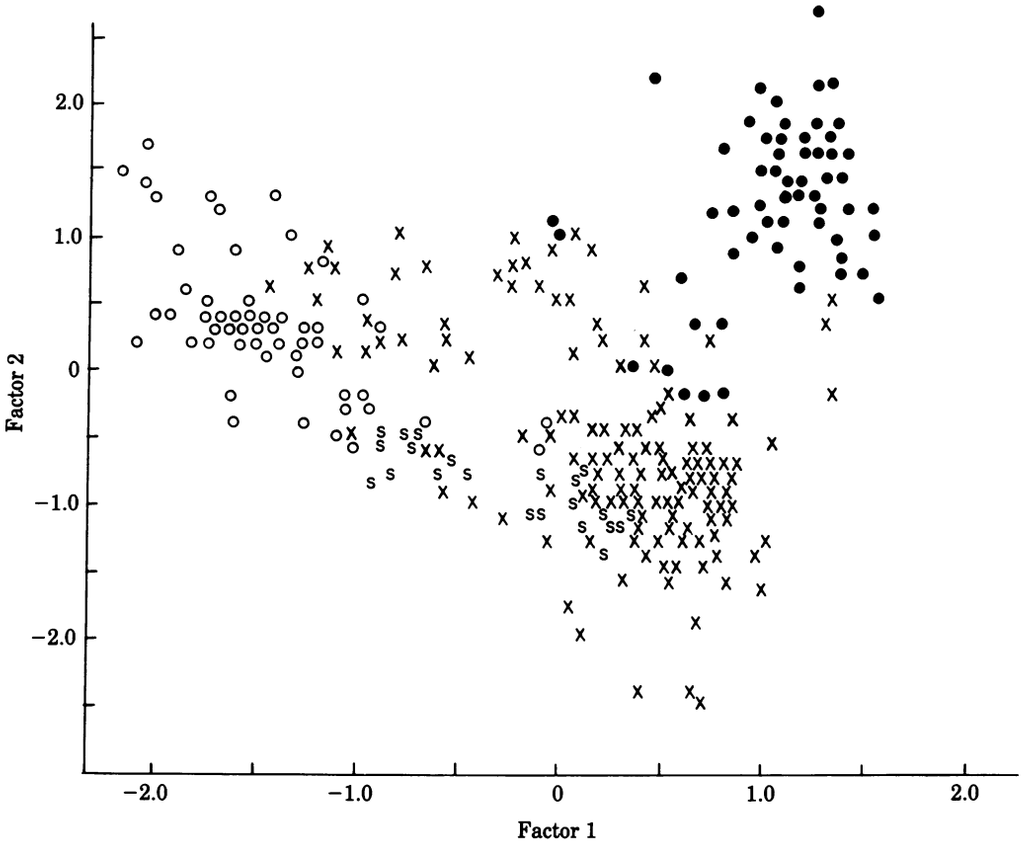

FIG. 2. Factor scores of electrophoretic types of *E. coli* ($n = 279$) and *Shigella* species ($n = 23$) for the first two principal axes. Group assignments of strains were determined by discriminant analysis (see text). ○, Group I; X, group II; ●, group III; and S, *Shigella*.

be expected in less than 1% of the sample if the alleles were independently assorted. The proportion (45%) of the types lacking these four alleles (number 1) is also much larger than expected.

The second comparison involves a subset of alternative alleles at the same four loci that were included in comparison I. These alleles have high positive loadings on factor 1 and therefore tend to occur in the two clusters of electrophoretic types on the right side of Fig. 2. Approximately 17% of all types have the four-allele combination of ADH^6, ACO^6, PGI^7, and IDH^5 (number 16), a proportion well in excess of that predicted from allele frequencies alone. The tendency for the relative deviations to be positive for three-allele combinations (numbers 8, 12, 14, 15) and negative for two-allele combinations further illustrates the strong associations among these four alleles.

The third comparison involves four alleles with high positive loadings on factor 2. Their presence defines the cluster of types near the upper right corner of Fig. 2. The frequency of the four-allele combination of MPI^3, $PE2^4$, AK^5, and βGA^7 is 5%, which is substantially greater than expected.

Table 2. Four-locus allele combinations used in analyzing linkage disequilibrium (Table 3)

Comparison	Symbol A	Symbol B	Symbol C	Symbol D
I	ADH^1	ACO^7	PGI^4	IDH^2
	(0.219)	(0.334)	(0.450)	(0.235)
II	ADH^6	ACO^6	PGI^7	IDH^5
	(0.427)	(0.483)	(0.447)	(0.702)
III	MPI^3	$PE2^4$	AK^5	βGA^7
	(0.414)	(0.149)	(0.179)	(0.136)

The relative frequencies of the alleles in the sample of 302 electrophoretic types are in parentheses.

Table 3. Relative combination frequencies and their relative deviations (in parentheses) from products of allele frequencies in 302 electrophoretic types of *E. coli* and *Shigella*

Number	Allele combination A	B	C	D	Comparison I	II	III
1	1	1	1	1	0.447 (+0.228)	0.212 (+0.097)	0.490 (+0.136)
2	1	1	1	2	0.007 (−0.060)	0.113 (+0.064)	0.026 (−0.030)
3	1	1	2	1	0.129 (−0.050)	0.010 (−0.083)	0.010 (−0.067)
4	1	1	2	2	0.033 (−0.022)	0.026 (−0.013)	0.013 (+0.001)
5	1	2	1	1	0.073 (−0.037)	0.020 (−0.087)	0.000 (−0.062)
6	1	2	1	2	0.000 (−0.034)	0.089 (−0.018)	0.010 (+0.000)
7	1	2	2	1	0.033 (−0.057)	0.007 (−0.080)	0.033 (+0.019)
8	1	2	2	2	0.060 (+0.032)	0.096 (+0.059)	0.003 (+0.001)
9	2	1	1	1	0.013 (−0.048)	0.003 (−0.083)	0.258 (+0.008)
10	2	1	1	2	0.003 (−0.016)	0.050 (+0.014)	0.013 (−0.026)
11	2	1	2	1	0.017 (−0.033)	0.003 (−0.066)	0.030 (−0.024)
12	2	1	2	2	0.017 (+0.002)	0.099 (+0.096)	0.010 (+0.001)
13	2	2	1	1	0.007 (−0.024)	0.007 (−0.073)	0.013 (−0.031)
14	2	2	1	2	0.000 (−0.009)	0.060 (+0.026)	0.010 (+0.003)
15	2	2	2	1	0.046 (+0.021)	0.036 (−0.029)	0.030 (+0.020)
16	2	2	2	2	0.116 (+0.108)	0.169 (+0.142)	0.050 (+0.048)

In each of the 16 possible combinations of four alleles, 1 = absence and 2 = presence of a designated allele at a locus. (The absence of an allele means that a nondesignated allelic state was observed at that locus.) The designated alleles in each comparison and their relative frequencies are indicated in Table 2.

Table 4. Likelihood ratio χ^2s from tests of multilocus disequilibrium in three comparisons, partitioned by level of interaction

Source	df	Comparison I	Comparison II	Comparison III
Two-locus effects	6	293.32***	176.74***	168.23***
AB	1	20.17***	1.14	2.50
AC	1	15.50***	39.92***	3.60
AD	1	6.29*	12.58***	0.6
BC	1	7.27**	17.22***	59.64***
BD	1	12.81***	13.32***	10.74**
CD	1	44.76***	6.82**	5.94*
Three-locus effects	4	6.58	34.63***	16.34**
ABC	1	1.27	3.27	1.79
ABD	1	1.43	10.57**	1.76
ACD	1	1.97	2.37	0.62
BCD	1	1.92	0.76	12.07***
Four-locus effect, *ABCD*	1	0.01	0.16	8.65**
Total	11	299.91***	211.52***	193.22***

*$P < 0.05$.
**$P < 0.01$.
***$P < 0.001$.

For each comparison, we tested for association among alleles by methods for multiway contingency tables (36–38). (To avoid computational problems, we assumed that zero counts are a result of finite sample size and added 0.5 to each cell count.) Table 4 summarizes the components of linkage disequilibria resulting from overall tests of two-locus and higher-order effects and of partial association among alleles. The likelihood ratio statistics of total disequilibrium shown in the last row of Table 4 indicate highly significant deviations from linkage equilibrium within each comparison, a result in agreement with the earlier overall assessment of multilocus disequilibrium. In two cases, both the total two-locus and total three-locus disequilibria are significant. The tests of partial association (adjusted tests of ref. 38) specify the degree of association for particular alleles and reveal the components of the total disequilibrium. The results indicate that, although much of the total disequilibrium occurs between pairs of alleles, the pairwise disequilibria are not independent.

One measure of the nonindependence of the pairwise components is the difference between the total two-locus disequilibria (293.32 for comparison I) and the sum of the pairwise components (106.80 for comparison I). This difference, which results from the statistical dependence of the two-locus tests (38), is substantial for each of the comparisons. Furthermore, in comparisons II and III, there are significant higher-order interactions among the loci.

DISCUSSION

Our analysis clearly demonstrates that the multilocus allele combinations observed in natural populations of *E. coli* are not random samples from a single gene pool. Nonrandom associations of alleles were apparent in both an overall assessment of 12 loci and in three specific four-locus combinations. Moreover, the multilocus structure appears to be complex in that pairwise associations between alleles are not independent.

The simplest hypothesis to account for the multilocus structure in *E. coli* is that the observed disequilibria result from genetic drift. The potential role of drift in generating linkage disequilibrium is well known (39–42), and recent theoretical studies by Ohta (23, 43) have shown that large variances in disequilibrium coefficients can result from drift in finite, subdivided populations. Moreover, the conditions under which drift may become important—low recombination rates, frequent local extinction of lines, and limited migration—appear to be met in natural populations of *E. coli*.

At this point, we cannot reject the hypothesis of drift-generated linkage disequilibrium, especially in light of the fit of the allele frequency spectra for individual loci to neutral expectations. But the following observations suggest that some form of natural selection is also affecting the frequencies of certain genetic combinations.

Evidence for Subspecific Structure. The projection of the 302 electrophoretic types on the first two principal axes revealed three overlapping clusters of strains (Fig. 2), which represent the nonrandom associations of alleles that were analyzed in the four-locus comparisons (Table 4). Using a random subsample of 20–30 strains from each group, we classified the remaining strains into one of three groups (I, II, and III) by a discriminant function analysis of the data, recoded as follows. For each locus (variable), we ranked the allozymes by mobility, pooled nulls with the most common allozyme, and transformed the variables onto a common scale by ranging (44). The discriminant function provided a means of classifying intermediate electrophoretic types into one of the three groups with 95% confidence. The 23 electrophoretic types of *Shigella* fall into two clusters (Fig. 2) and lie largely between the *E. coli* groups I and II.

We further quantified the degrees of genetic differentiation among the groups by partitioning the total genetic diversity into within- and between-group components (45). For the 12 loci studied, 19% of the total genetic diversity in allele frequencies is attributable to genetic differences between the groups ($F_{ST} = 0.190$). This relative measure of differentiation is more than 10 times larger than that recorded in a comparison of populations of *E. coli* in Sweden and the United States (30) and is equivalent to that of geographic subspecies in eukaryotes (10). Moreover, the estimated minimal net codon difference among the three groups (0.147 per locus) is an order of magnitude greater than comparable estimates reported for local populations of animal and plant species (10).

To further examine the subspecific structure of *E. coli*, we represented the center of each group of strains by the hypothetical modal electrophoretic type, that is, the combination of alleles with the highest frequencies among types in the group. The modal type of group II differs from the modal types of groups I and III at five and four loci, respectively, whereas the modal types of groups I and III differ at seven loci. This comparison suggests that group II is the ancestral group from which I and III evolved. We also compared these hypothetical combinations with those of actual strains isolated from hosts. The first and third most commonly isolated types (discounting occurrences of clones of the same type in associated hosts), which were found in 28 and 16 host individuals, respectively, differ at only one locus from the modal type of group III (Fig. 1). The second most commonly recovered strain, which was isolated from 17 hosts, differs at three loci from the modal type of group II. The fourth most common strain, which was collected from 12 hosts, is identical to the modal type of group I. This strain is also indistinguishable in allozyme profile from laboratory strain K-12, which was isolated over 50 years ago (46). These common strains were found in both Sweden and North America and were not limited to a particular type of host (animal or human).

These results can be summarized as follows: Each of the four most commonly isolated and presumably, therefore, the most abundant strains in nature is very similar to or identical with one of the hypothetical combinations of alleles that locate the centers of the three groups of strains. This suggests that some electrophoretic types have persisted for periods of time sufficiently

long to have spread over large geographic areas and, through mutation, to have produced clusters of closely related strains. Many of the mutations may be neutral or nearly so, because the allele frequency distributions for individual loci are statistically indistinguishable from the expectation of strict neutrality. With low rates of recombination, the genetic combinations occurring in the persistent strains would also be represented in the related mutant strains, thus yielding the linkage disequilibrium we have observed.

The persistence of particular strains may have been promoted by the action of natural selection, which, by favoring certain genetic combinations in the past, may have played a role in the subspecific differentiation observed in *E. coli*. A potential for the action of selection on particular genetic combinations in natural populations of *E. coli* has recently been demonstrated for allozymes at the 6PGD locus by Hartl and Dykhuizen (47, 48). The selective effects of six alleles of 6PGD were estimated by transferring the alleles, by transduction, from wild strains onto the genetic background of the K-12 laboratory strain and then measuring growth rates in chemostat populations. These experiments suggested that all six alleles are normally neutral or nearly so on the K-12 genetic background and, by inference, in natural populations. However, in certain genetic and environmental backgrounds, functional differences between allozymes were expressed as selective differences between strains. This work strengthens the hypothesis that natural selection, in addition to stochastic factors, has had a significant part in the generation and maintenance of the multilocus genetic structure of natural populations of *E. coli*.

Most of the electrophoresis was performed by D. A. Caugant. This research was supported by grants from the National Science Foundation (DEB 78-23263) and the National Institutes of Health (GM 22126).

1. Glass, R. E. (1982) *Gene Function. E. coli and Its Heritable Elements* (Univ. of California Press, Berkeley).
2. Bachmann, B. J. & Low, K. B. (1980) *Microbiol. Rev.* **44**, 1–56.
3. Edwards, P. R. & Ewing, W. H. (1972) *Identification of Enterobacteriaceae* (Burgess, Minneapolis, MN), 3rd Ed.
4. Ørskov, I., Ørskov, F., Jann, B. & Jann, K. (1977) *Bacteriol. Rev.* **41**, 667–710.
5. Ørskov, F. & Ørskov, I. (1978) in *Methods in Microbiology*, eds. Bergan, T. & Norris, J. R. (Academic, New York), Vol. 11, pp. 1–77.
6. Brenner, D. J., Fanning, G. R., Johnson, K. E., Citarella, R. V. & Falkow, S. (1969) *J. Bacteriol.* **98**, 637–650.
7. Brenner, D. J., Fanning, G. R., Skerman, F. J. & Falkow, S. (1972) *J. Bacteriol.* **109**, 953–965.
8. Brenner, D. J., Fanning, G. R., Steigerwalt, A. G., Ørskov, I. & Ørskov, F. (1972) *Infect. Immun.* **6**, 308–315.
9. Brenner, D. J. (1977) *Prog. Clin. Pathol.* **7**, 71–117.
10. Nei, M. (1975) *Molecular Population Genetics and Evolution* (Elsevier, Amsterdam).
11. Milkman, R. (1973) *Science* **182**, 1024–1026.
12. Milkman, R. (1975) in *Isozymes*, ed. Markert, C. L. (Academic, New York), Vol. 4, pp. 273–285.
13. Selander, R. K. & Levin, B. R. (1980) *Science* **210**, 545–547.
14. Selander, R. K. (1976) in *Molecular Evolution*, ed. Ayala, F. J. (Sinauer, Sunderland, MA), pp. 21–45.
15. Brown, A. H. D. (1979) *Theor. Popul. Biol.* **15**, 1–42.
16. Nevo, E. (1978) *Theor. Popul. Biol.* **13**, 121–177.
17. Levin, B. R., Stewart, F. M. & Rice, V. A. (1979) *Plasmid* **2**, 247–260.
18. Levin, B. R., Stewart, F. M. & Chao, L. (1977) *Am. Nat.* **111**, 3–24.
19. Caugant, D. A., Levin, B. R. & Selander, R. K. (1981) *Genetics* **98**, 467–490.
20. Maruyama, T. & Kimura, M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6710–6714.
21. Wright, S. (1940) *Am. Nat.* **74**, 232–248.
22. Nei, M. (1976) *Trends Biochem. Sci.* **1**, N247–N248.
23. Ohta, T. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1940–1944.
24. Kubitschek, H. E. (1974) in *Evolution in the Microbial World*, eds. Carlile, M. J. & Skehel, J. J. (Cambridge Univ. Press, London), pp. 105–130.
25. Maynard Smith, J. & Haigh, J. (1974) *Genet. Res.* **23**, 23–35.
26. Thomson, G. (1977) *Genetics* **85**, 753–788.
27. Brenner, D. J., Fanning, G. R., Miklos, G. V. & Steigerwalt, A. G. (1973) *Int. J. Syst. Bacteriol.* **23**, 1–7.
28. Brenner, D. J. (1981) in *The Prokaryotes*, eds. Starr, M. P., Stolp, H., Truper, H. G., Balows, A. & Schleger, H. G. (Springer, Berlin), pp. 1105–1127.
29. Kimura, M. (1982) *Molecular Evolution, Protein Polymorphism and the Neutral Theory* (Jpn. Sci. Soc., Tokyo).
30. Caugant, D. A., Levin, B. R., Lidin-Janson, G., Whittam, T. S., Svanborg Eden, C. & Selander, R. K. (1983) *Prog. Allergy* **33**, 203–227.
31. Selander, R. K., Smith, M. H., Yang, S. Y., Johnson, W. E. & Gentry, J. B. (1971) *Stud. Genet.* **6**, 49–90.
32. Ewens, W. J. (1972) *Theor. Popul. Biol.* **3**, 87–112.
33. Ewens, W. J. (1979) *Population Genetics* (Springer, Berlin).
34. Watterson, G. A. (1978) *Genetics* **88**, 405–417.
35. Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980) *Genetics* **96**, 523–536.
36. Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975) *Discrete Multivariate Analysis. Theory and Practice* (MIT Press, Cambridge, MA).
37. Goodman, L. (1978) *Analyzing Qualitative Categorial Data. Log-Linear Models and Latent Structure Analysis* (Abt, Cambridge, MA).
38. Smouse, P. E. (1974) *Genetics* **76**, 555–565.
39. Kimura, M. & Ohta, T. (1971) *Theoretical Aspects of Population Genetics* (Princeton Univ. Press, Princeton, NJ).
40. Nei, M. & Li, W.-H. (1973) *Genetics* **75**, 213–219.
41. Hill, W. G. (1976) in *Population Genetics and Ecology*, eds. Karlin, S. & Nevo, E. (Academic, New York), pp. 339–376.
42. Hedrick, P., Jain, S. & Holden, L. (1978) *Evol. Biol.* **11**, 101–184.
43. Ohta, T. (1982) *Genetics* **101**, 139–155.
44. Sneath, P. H. A. & Sokal, R. R. (1973) *Numerical Taxonomy* (Freeman, San Francisco).
45. Nei, M. (1977) *Ann. Hum. Genet.* **41**, 225–233.
46. Lederberg, J. (1950) in *Genetics in the 20th Century*, ed. Dunn, C. (Macmillan, New York), pp. 263–289.
47. Dykhuizen, D. E. & Hartl, D. L. (1980) *Genetics* **96**, 801–817.
48. Hartl, D. L. & Dykhuizen, D. E. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 6344–6348.