Molecular Evolution of the *Escherichia coli* Chromosome. II. Clonal Segments

Roger Milkman and Arlin Stoltzfus

Department of Biology, The University of Iowa, Iowa City, Iowa 52242

Manuscript received March 12, 1988

Revised copy accepted June 17, 1988

ABSTRACT

Remarkable sequence similarities in the trp region among *Escherichia coli* strains of diverse natural origins imply the existence of worldwide clones of very recent origin. This in turn implies a low rate of fixation of new universally favorable alleles, which carry adjacent stretches of chromosome to high frequency. These clonal segments begin as entire chromosomes; recombination shortens them progressively by substituting less closely related homologous DNA. The rate of this recombination, comprising the introduction of a homologous chromosomal fragment to a cell and the replacement of part of the original chromosome, is estimated from observations.

The term "clone" may be applied to a group of organisms or cells descended without recombination from a single ancestor; to a similarly descended group of discrete genomic components like mitochondrial DNA (Cann, Stonerking and Wilson 1987; Wilson *et al.* 1985); to certain Y chromosomes or segments thereof having no X homologs (Bishop *et al.* 1985); and equally to a group of homologous single nucleotides derived from a single template, since recombination cannot take place within them.

The view has been held for some time (Kubitschek 1974; Maruyama and Kimura 1980; Levin 1981) that the species *Escherichia coli* consists of a set of widely distributed clones, rather than the vast array of allelic combinations expected in a frequently recombining (panmictic) species. The mechanism proposed to account for this clonal species structure is periodic selection (Atwood, Schneider and Ryan 1951), in which a rare generally favorable mutant allele rises in frequency, carrying the rest of the genome along as a passive hitchhiker. Smaller genomic elements are known with pertinent properties: β-globin-gene frameworks, defined by sequence, and related haplotypes, defined by restriction sites, appear to be polymorphic clones in humans (Orkin and Kazazian 1984). The frameworks have been investigated in several primates as well, with phylogenetic implications (Savatier *et al.*, 1987). There is now evidence that the clonal portion of the genome decreases in length due to recombination even as its numbers increase by selection.

Evidence favoring a clonal species structure comes from the observation of remarkable nucleotide sequence similarity in the trpCBA region of *E. coli* strains of diverse origin (Milkman and Crawford 1983), and from the observed recurrence of specific genomic types, consisting of electrophoretic alleles at 20 or at 35 scattered loci, among diverse *E. coli* strains (Selander and Levin 1980; Selander, Caugant and Whittam 1987). Both observations point to the descent of significant portions of this entire ancient species from single common ancestors that lived relatively recently.

If the clones which *E. coli* comprises were indeed genome-wide, then recombination would have to be negligible during the spread of a clone following a favorable mutation. But in fact, recent evidence (Dykhuizen and Green 1986; D. Dykhuizen, personal communication) favors the view that the clones are only segments of the chromosome. This possibility had been suggested by Hartl and Dykhuizen (1984). Using 9 of the same strains previously compared at trpCBA (Milkman and Crawford 1983), Dykhuizen and Green sequenced some 800 bp in the gnd locus, 16 min (about 1/6 of the chromosome) away from trp. The gnd results were strikingly different from the trpCBA observations: the degree of variation was far greater, and the phylogeny deduced was totally different. For example, strain 191F, which was identical to the standard K12 strain in trpCBA (Milkman and Crawford 1983), turned out to be the most different from K12 in gnd. Indeed, the degree of difference between these two strains is close to the postulated equilibrium level, which will be described. We have now found localized evidence of recombination in and near the trp operon (Stoltzfus, Leslie and Milkman, 1988; and see below).

The present paper describes an initial effort to identify and evaluate a set of rate processes which may result in the origin of clonality in the genome in *E. coli*, beginning with the entire chromosome and continuing with the progressive reduction in length and diversification of the clonal segment. This effort en-
and stop codons expected) will be different between approximately equal to the rate of fixation of new and TTC for TTT) for every amino acid with two or more codons. On the basis of this observation, we make the working assumption that on the average two homologous sequences. Note, too, that clones can arise within clones. A given clone may be defined by its neutral diversity, which reflects its age. Also, the size of a clone (number of individuals) must be distinguished from its segmental length (number of nucleotide pairs). Older clones are larger in number than their contained younger subclones, but these younger clones will actually have a greater segmental length.

**MATERIALS AND METHODS**

Most materials and biochemical methods are described in Stoltzfus, Leslie and Milkman (1988). DNA to be sequenced was cloned in plasmids pUC8, pUC18, pUN121 (kindly given to Steven Clegg by Bjorn Nilsson, Royal Institute of Technology, Stockholm) and pIA15 (a modification of pBR322 containing part of the trpCBA region). Initially, clones were obtained by transduction to plasmids (Milkman and Crawford 1983). More recently, desired clones were identified by colony hybridization. The chromosomal regions investigated are primarily in the trpCBA region of the trp operon. The nearby tonB region and the intervening stretch, temporarily divided into two arbitrary regions called Fall and Lika, have also been studied.

Mutation rate, neutral divergence, and time: Divergence, in the form of neutral nucleotide substitutions, is taken as a simple rate process, as noted above. For estimating the number of generations, \( g \), that have elapsed since the most recent common ancestor of two strains, an integrated
**TABLE 1**

Polymorphic sites in four strains in trpBA region

<table>
<thead>
<tr>
<th>Place</th>
<th>K12 47 50 51</th>
<th>Place</th>
<th>K12 47 50 51</th>
<th>Place</th>
<th>K12 47 50 51</th>
<th>Place</th>
<th>K12 47 50 51</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 5114 T C C 3 5699 C T C 3 6430 T A A</td>
<td>3 5141 T C C 3 5717 G A A 3 6435 G A A</td>
<td>3 5186 A G A 3 5729 C C T 3 6436 T C C C</td>
<td>3 5189 G G G 3 5738 T A A 3 6442 A A G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5207 T C C 3 5744 C T T 3 6445 A A G G</td>
<td>3 5210 A G G 3 5750 T C C 3 6446 C T C C</td>
<td>3 5213 A C C 3 5757 A G G 3 6454 A A G</td>
<td>3 5216 C G C 3 5777 G C C 3 6466 G G G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5219 T T 3 5792 C C C 3 6478 A G G</td>
<td>3 5237 C G G 3 5780 G A A 3 6496 T C C C</td>
<td>3 5243 T C C C 3 5783 A A A 3 6502 G T T T</td>
<td>3 5276 T C C 3 5786 T C C 3 6517 T C C C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5288 G C C C 1 5823 C T T 2 6528 A G A</td>
<td>3 5291 G T T T 3 5831 T C C 1 6554 C A A</td>
<td>3 5297 G T T T 3 5843 C C C 3 6562 A A G G</td>
<td>3 5300 G T T T 3 5848 G A A 3 6601 A A C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5324 C G T T 3 5891 G G G 3 6604 T G G</td>
<td>3 5328 C G T T 3 5896 G G G 3 6616 G T T T</td>
<td>3 5333 C A C C 3 5912 C C C 3 6643 C T T</td>
<td>3 5345 T A A 3 5915 C C C 3 6646 C T T T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5402 C C T 3 5921 A A A 3 6661 C T T T</td>
<td>3 5405 T C A 3 5936 T C C 3 6664 C T T T</td>
<td>3 5444 T C C 3 5945 A * G 3 6673 A G G</td>
<td>3 5453 T C C C 3 5951 C * G 3 6682 T C C C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5456 T G G 3 5960 T * G 3 6685 C T T</td>
<td>3 5461 G A A 3 5990 T A A 3 6688 C T T T</td>
<td>3 5462 C T T 3 5999 T C C 1 6689 C T T T</td>
<td>3 5471 T C C 3 6008 G A A 3 6709 T C C C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5534 G A A 3 6062 A G G 3 6745 A A G</td>
<td>3 5549 T C C C 1 6078 T C C 3 6796 T C C</td>
<td>3 5555 C T T 3 6107 C C C 3 6797 G A A</td>
<td>3 5579 C T T 3 6113 T A A 3 6817 C T T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5615 G C C 3 6819 T C C 3 6892 G A A</td>
<td>3 5624 T C C 3 6835 A C C 3 6900 T A A</td>
<td>3 5630 G A A 3 6849 T A A A 3 6910 G A A</td>
<td>3 5630 C G G 3 6849 C T T T 3 6916 T A A 3 5648 C G G 3 6849 C T T T 3 6916 T A A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5654 G A A 3 6841 C C C 3 6963 C T T</td>
<td>3 5654 G A A 3 6841 C C C 3 6963 C C C</td>
<td>3 5690 C G G 3 6847 A A A 3 6968 C T T</td>
<td>3 5690 C G G 3 6847 A A A 3 6968 C T T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Position of base in codon is given first (this region is totally translated), followed by its position in the trp operon (subtract 203 to match sequence in Nucleotide Sequences 1986/1987). Only sites polymorphic for these four strains are listed. '*' indicates agreement with K12. ** indicates no datum.

**TABLE 2**

Sequence differences between ECOR 71 and K12

<table>
<thead>
<tr>
<th>Region</th>
<th>Substitutions/No. of codons</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>trpE</td>
<td>4/91</td>
<td>4</td>
</tr>
<tr>
<td>trpC</td>
<td>5/169</td>
<td>3</td>
</tr>
<tr>
<td>trpBA</td>
<td>10/501</td>
<td>2</td>
</tr>
<tr>
<td>Fall</td>
<td>22/625</td>
<td>4</td>
</tr>
<tr>
<td>Lika</td>
<td>10/183</td>
<td>5</td>
</tr>
<tr>
<td>tonB</td>
<td>5/120</td>
<td>4</td>
</tr>
</tbody>
</table>

Homogeneity χ² = 5.8; d.f. = 5; 0.25 < P < 0.50.

**RESULTS**

The sequence data upon which this analysis is based is excerpted in Table 1.

**Estimation of r**: Preliminary estimation of the basic outcrossing chromosomal recombination rate, r, rests on the observation that the sequences of several strains in the trpCBA region fall into a small number of types, as previously indicated, and that occasionally, a sequence changes type abruptly. The most striking example (Table 1), involves a change in sequence similarities between ECOR 47 and 50. These strains differ between positions 5101 and (arbitrarily) 5731 by 30 bases in 210 codons, a ratio of 0.14. Between 5731 and (arbitrarily) 6871, the strains differ by 6 bases in 380 codons, or 0.016. This suggests an exchange involving an ancestor of one of the strains. Less compelling is the subsequent difference of 4 bases in 32 codons (0.12).

Between positions 6409 and 6691, ECOR strain 51 differs from K12 by 29 bases in 94 codons (0.31) and by 5 in the next 95 codons (0.05). Again, an exchange is suggested, especially since there is no indication of a strong regional difference in the level of variation among the other three strains.

In contrast, K12 and ECOR 71 sequences compared at intervals over some 11 kb indicate no evidence of recombination since their separation: the two sequences differ by an amount, 0.033 base/codon on average, that is fairly uniform over the entire region.
Favorable mutations are fixed, or reach a very high frequency, so rarely in E. coli that one can cause the spread of a clone worldwide. The clone begins with the entire genome but continues as a shrinking segment.

Early in the process, when the clonal region hitchhiking with the favorable mutation is still quite long, all individuals are still identical, or nearly so, to their recent common ancestor, and therefore to one another.

Recombination with distantly related individuals reduces the length of the clonal segment. Also, neutral mutations accumulate within the clonal segment.

Eventually the clonal segment accumulates so many neutral mutations that variation reaches equilibrium. Now the clonal segment is indistinguishable from flanking regions.

The observed independent variable is divergence between two compared homologous sequences. This converts to time in generations and in years, represented on the two lines above. The shortening of the clonal segment due to recombination is displayed on the bottom line. At the top, a short early interval in the spread of the clone is expanded. The corresponding divergence is again shown, as is the rise in frequency of the new allele (spread) on the top line.

FIGURE 1.—Sequence divergence and shortening of clonal segments related to time and increase in frequency, assuming recombination rate per nucleotide to be $4 \times 10^{-13}$ and nucleotide substitution rate to be $3.5 \times 10^{-10}$. The observed independent variable is divergence between two compared homologous sequences. This converts to time in generations and in years, represented on the two lines above. The shortening of the clonal segment due to recombination is displayed on the bottom line. At the top, a short early interval in the spread of the clone is expanded. The corresponding divergence is again shown, as is the rise in frequency of the new allele (spread) on the top line.

Favorable mutations are fixed, or reach a very high frequency, so rarely in E. coli that one can cause the spread of a clone worldwide. The clone begins with the entire genome but continues as a shrinking segment.

Early in the process, when the clonal region hitchhiking with the favorable mutation is still quite long, all individuals are still identical, or nearly so, to their recent common ancestor, and therefore to one another.

Recombination with distantly related individuals reduces the length of the clonal segment. Also, neutral mutations accumulate within the clonal segment.

Eventually the clonal segment accumulates so many neutral mutations that variation reaches equilibrium. Now the clonal segment is indistinguishable from flanking regions.

- Position of favorable mutation
- Sequence identical (or nearly so) to ancestral sequence
- Neutral differences, below equilibrium value, exist between present segment and ancestral segment
- Neutral differences between this and the clonal ancestral segment are at equilibrium level
- Borders of clonal segment

FIGURE 2.—Shortening and diversification of a clonal segment.
Shortening of clonal fragments below 100 kb and within one billion generations. Time of event stated in millions of generations; size proportioned to bar length = 100 kb.

Ten simulations, $r = 4 \times 10^{-13}$.

Taking the estimate of $g$ as $10^9$ generations, we estimate $r = 1/gL_A$ to be $4 \times 10^{-15}$, as explained below.

By comparison, *Drosophila melanogaster*’s basic recombination rate is about 1 in 70 million, or $1.4 \times 10^{-8}$ per nucleotide per generation, based on an average of one crossover per generation in the second chromosome (also the third), each having about 70 million base pairs (haploid number).

The length of any stretch (segment) of DNA will be called $l$. If it is a clonal segment, sharing a recent common ancestral template with other segments, the continuous length derived from that ancestral template (Table 2), excluding a large stretch of DNA present in ECOR 71 but not in K12 (STOLTZFUS, LESLIE and MILKMAN 1988). A divergence of 0.033 corresponds to 250 million generations. So far, it is mainly the sequences that have been separated from K12 and from one another on the order of a billion generations that occasionally show abrupt changes in similarity [differences are admittedly easier to detect in such cases]. Comparison of sequences in the *trp* and Fall regions totalling about 16,000 bp reveals evidence of 6 recombinational events, leading to a rough estimate of 2700 bp for $L_A$, average clonal segment length.

**Figure 3.**—Simulation of the shortening of ten clonal segments by random recombination.

**Figure 4.**—Nomogram relating recombination rate, $r$, pairwise diversity, $x$, and clonal segment length, $L$. Values for any two variables on the nomogram can be connected by a ruler, which will lie on the corresponding value for the third. Subscripts stand for average length ($A$) or minimum length in a given proportion of the clonal segments (decimal). The lengths other than $L_A$ can be calculated from $L_A$ (see text) or estimated graphically, the distances between given percentiles being constant. Nucleotide substitution rate taken as $2 \times 10^{-9}$, excluding a large stretch of DNA present taking the estimate of $g$ as $10^9$ generations, we estimate $r = 1/gL_A$ to be $4 \times 10^{-15}$, as explained below.
is \( L \). The mean length of a clonal segment in a group of individuals is \( L_A \). A basic recombination rate per nucleotide, \( r \), is defined with the following special properties: it is the rate of recombination between two adjacent nucleotides resulting from the apposition of, and the physical exchange between, two distinguishable chromosomes. Thus it is a compound process and one that takes place in nature. Since the frequency of apposition of distinguishable chromosomes (by conjugation or transduction) is itself a somewhat complex variable, \( r \) cannot be thought of as absolutely constant, but it will be taken as approximately so.

For a segment of any given length, \( l \), an expected cumulative recombination frequency, \( c \), is defined as the product of the basic recombination rate, \( r \), the pertinent number of generations, \( g \), and the length, \( l \). As \( c = rgL \), \( r = c/gL \), and so on. With regard to a set of clonal segments, \( g \) is used to represent the number of generations since the most recent common ancestor, in other words since the origin of the favorable mutation whose selection is responsible for the spread of the clone. By definition, \( rgL_A = 1 \), since \( L_A \) is the mean length of the segment after \( g \) generations, given a basic recombination rate, \( r \). When \( g \) has been estimated from Equation 1 and \( r \) has been estimated from observations to be described, \( L_A \) can be calculated as \((rg)^{-1}\), and this value can be tested empirically by sequence comparisons. The estimation of \( r \) is impeded by the difficulty of estimating difference level, and therefore divergence time, over relatively short segments. In any event, since \( r \) is taken to be a constant, \( g \) and \( L_A \) are inversely proportional, and so the length of an isoancestral segment decreases in proportion to the age of the clone. An error in the estimation of \( r \) will lead to an inverse error in the expected value of \( L_A \).

Assuming that recombination events occur at random, by the time \( L_A \) has become relatively short the probability distribution of \( L \) is exponential:

\[
f(L) = rg \cdot e^{-rgL}.
\]

Integrating, \( F(L_T) = 1 - e^{-rgL_T} \); \( F(L_T) \) is the cumulative relative frequency of all values of \( L \) between 0 and \( L_T \). Rearranging, \( L_T = -\ln[1 - F(L_T)](rg)^{-1} \). Thus, with a basic recombination rate, \( r \), of \( 4 \times 10^{-10} \), after 60 million generations, in 90% [chosen as \( F(L_T) \)] of the individuals carrying it, a favorable mutation will be lodged in a clonal segment of 95,941 bp or less. Similarly, in 10% of the individuals the segment length will be 4390 or less. The ratio of lengths at these percentiles is of course always \( 0.10/\ln 0.90 = 21.85 \). \( L_A = (rg)^{-1} = 41,667 \) here. Also, \( 1 - F(L_A) \) is always \( e^{-1} = 36.79\% \). Note also that \( 1 - F(L_T) = e^{-rgL_T} \) is the probability that no recombination has taken place within a segment of length \( L_T \) in \( g \) generations. This is the proportion of cases in which the clonal segment length is \( L_T \) or greater.

**Graphic models:** A clonal segment experiences three basic microevolutionary processes which are placed on a common time scale in Figure 1: rise in frequency, shortening, and neutral sequence diversification. Values chosen for pertinent variables depend in considerable part on a substitution rate of \( \frac{3}{2} \times 10^{-10} \) per nucleotide per path. The finding in strains of diverse origin of 5 identical sequences and 5 more each differing by a single different base (Milkman and Crawford 1983 and unpublished data) is interpreted as due to sampling of a uniformly diverse set of "K12" sequences averaging about 1 nucleotide substitution per 350 codons. This divergence indicates that clones can rise to substantial frequency in some 20 million generations. Moreover, 5 "45E" sequences are known which show no differences at all in a total of 1700 codons, suggesting that the rise to prominence may take even less time. As a working illustration, a time of 2 million generations is used in Figure 1 (top).

The simultaneous shortening and diversification of a clonal segment is illustrated in Figure 2, which shows a single simulation based on random recombination starting after the length has declined below 100 kb and ending at 10 billion generations, at which time the neutral diversity within a set of clonal segments is expected to be 0.465 substitution per codon. This value is arbitrarily taken to be indistinguishable from the equilibrium value, 0.500, attributed to unrelated homologous segments.

The diverse random shortening processes among lines of descent are illustrated in Figure 3. These simulations are based on uniformly distributed recombination events which must leave the favorable mutation within the clonal segment. Only the clonal segment containing the favorable mutation is represented, though it is of course easy to envision that reintroduction of clonal stretches to a chromosome with subsequent recombinational events, or, on a larger scale, the retention of a clonal stretch beyond an introduced nonclonal portion.

In this regard, one can calculate the frequency, at any given site, of a clonal nucleotide (derived with or without mutation from the ancestral template) as a function of its distance from the original favorable mutation and from the latter’s frequency. If the favorable mutation becomes fixed, then a site \( l \) nucleotides away will have a frequency of \( e^{-rl} \) clone members, to a first approximation; the value of \( g \), again, is the number of generations from the origin of the favorable mutation to fixation. Surrounding the site considered, any clonal segment will decrease in length with time after fixation; thus the frequency of a clonal segment of given length will decrease, but the frequencies of given sites belonging to the clone (identical by descent) will remain essentially constant, though
some are no longer connected to the favorable allele.

Finally, a nomogram is presented in Figure 4 relating recombination rate, clonal segment length, and neutral sequence diversity, assuming a substitution rate, as before, of $3\times 10^{-10}$.

**DISCUSSION**

This estimate of $r$ can be applied to the aforementioned data on ECOR strain 71, which evidently has not recombined in an 11,000-bp region with K12 or any more distant strain. The mean segment length $L_{37}$ can be calculated as $(rg)^{-1}; (4 \times 10^{-13} \times 250 \times 10^9)^{-1} = 10,000$ bp, so the observation of no recombination over this distance is consistent with the estimate of $r$.

Current evidence supports the following working hypothesis: about 1.5 billion generations ago, a universally favorable mutation in or near the trp operon arose and spread throughout the species in about 20 million generations, becoming essentially fixed and carrying the flanking DNA to fixation as well. Since that time, a number of universally favorable mutations have arisen in the general vicinity. Whether or not they (and closely adjacent DNA) have become fixed, they have not carried to fixation more distant sequences reaching into the trp operon—though these sequences have reached substantial frequencies. Thus, the failure of these sequences to reach fixation could be due either to the nonfixation of the motivating alleles, or to these alleles' being at a considerable distance from the trp operon. Next, 20 million generations ago, a favorable mutation carried the K12 sequence to a high frequency, but not to fixation. And finally, the 45E subclone was established most recently in a line related to K12. The lineage of strains, as Hartl and Dykhuisen (1984) suggested, has transitory significance, given their tendency to acquire a composite genome.

We turn now to the alleles whose selection must account for this clonal structure. Such alleles must be universally favorable (rather than locally). Also, they must arise so rarely that a single mutation in a particular region of the chromosome results in the formation of a vast clone without the effective competition of other new alleles. A fixation rate of $10^{-6}$ per generation would result in the establishment of a clone once every million generations on the average. In addition, some clones might become prominent without becoming fixed.

As noted, the K12 subclone's divergence time is about 20 million generations. For a single cell's descendants to reach the estimated species number of $10^{20} (10^{11} \text{ hosts averaging } 10^9 \text{ bacteria each})$ in 20 million generations, their fitness should be about $1.00000023$, since ideally $w^s = N$ (commonly expressed as $w = N$); therefore $s$, defined as $s = 1$, is $0.00000023$. We are also interested in the rate of fixation (or reaching a substantial allele frequency such as 0.10) of the favorable mutations whose selection creates the clone. The fixation rate is the number of loci per generation at which an allele reaches a frequency of 1 for the first time. It is not a measure of the time elapsed from mutation to fixation. The fixation rate of a given category of favorable mutations is the product of its mutation rate per individual, the number of individuals in the species (if haploid), and the individual probability of fixation of a new allele, which is $4N_s/N$ (Kimura 1983; Crow 1986). $N_s$ is the effective population size, as opposed to the actual population size, $N$. $N_s$ is used to standardize sampling effects, notably random genetic drift. The vast majority of new neutral alleles, and even new weakly advantageous alleles, are lost due to random genetic drift soon after their origin. Nevertheless, the small proportion of surviving favorable alleles can be far greater than that of neutral alleles. Some new favorable alleles that reach a "safe" frequency, where extinction is no longer likely, and are in a position to go steadily to fixation and thereby bring along a hitchhiking chromosomal segment. If $N_s = N$, the probability of the ultimate fixation of any new favorable mutation is $4s$, in the present example about $9 \times 10^{-6}$. A fixation rate of $10^{-6}$ would thus require a mutation rate per cell of $1 \times 10^{-21}$, far less than the single nucleotide substitution rate. On the other hand, the fixation probability could be as low as $9 \times 10^{-16}$, since it is quite possible that the effective population size in E. coli is as low as $10^{-10}$, on the grounds that each host animal's population is best regarded as nearly isolated, and on the basis of various empirical measurements (Maruyama and Kimura 1980; Ochman and Wilson 1987). This would permit a favorable mutation rate on the order of $1 \times 10^{-11}$ per cell, which is of course still quite low. Thus, although the fixation rate may indeed be low primarily due to the genetic structure of the species, it is worth considering the possibility that the mutations responsible for major clone formation are rare multiple events resulting in two or even three simultaneous amino acid substitutions (Milkman 1985). This possibility is not inaccessible to investigation. In any event, the product of specific mutation rate and fixation probability can be expected to be in E. coli a rate low enough to account for the apparent clonal structure of the species, provided, of course, that chromosomal recombination is sufficiently rare.

The evolution of a genome is more than the aggregate of changes in allele frequencies, and even two-locus classical models have not yet been enlightening. An approach to the understanding of clonal segments, particularly empirically, may be a step on a different path to the effective description of genome evolution.

We acknowledge the technical assistance of Robert Bebee, Helene Begin Donita, Teri Herbst and Laura Hutchings, the
participation of Michael Dardik and Michael Ciccolo, and the collaboration of Irving Crawford. Supported in part by the National Institutes of Health grant GM-33518 and funds from The University of Iowa.

LITERATURE CITED


Communicating editor: W.-H. Li