

Molecular Evolution of the *Escherichia coli* Chromosome. III. Clonal Frames

Roger Milkman and Melissa McKane Bridges

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

Manuscript received May 3, 1990

Accepted for publication July 27, 1990

ABSTRACT

PCR fragments, 1500-bp, from 15 previously sequenced regions in the *Escherichia coli* chromosome have been compared by restriction analysis in a large set of wild (ECOR) strains. Prior published observations of segmental clonality are confirmed: each of several *sequence types* is shared by a number of strains. The rate of recombinational replacement and the average size of the replacements are estimated in a set of closely related strains in which a *clonal frame* is dotted with occasional stretches of DNA belonging to other clones. A clonal hierarchy is described. Some new comparative sequencing data are presented.

THE second paper of this series (MILKMAN and STOLTZFUS 1988) centered on early indications of the mosaic phylogeny of the chromosome of *Escherichia coli*. The elements of the mosaic are clonal segments, which are stretches of DNA descended from a single ancestral molecule and bounded by recombinational borders. Recent data suggest a simple quantitative relationship between the frequency of recombinational replacements and the frequency of retained nucleotide substitutions. The ratio of these frequencies is about 0.2, and the size of the replacements appears to be on the order of 1 kb. Thus, where the frequency of retained substitutions is about 5 per 10 kb, for example, we can expect the pattern illustrated in Figure 1.

The population dynamics of *E. coli* presents one relatively simple aspect: a clone that spreads throughout the species arises rarely. Its spread is powered by the selection of a widely favorable "motivating" allele. This selection is fast enough, relative to recombination, to bring along the entire chromosome to a frequency of a few percent before it is dotted perceptibly with segments of less closely related DNA ("replacements"). In the previous paper (MILKMAN and STOLTZFUS 1988) it was assumed that these recombinational replacements were on the order of 100 kb in length, and this led to a focus on their progressive shortening due to the superimposition of further replacements. The observations we now report show that the replacements are much smaller. Thus, for some time each chromosome of the clone consists of a *clonal frame* pocked by rare short replacements belonging to other clones. These replacements are also clonal segments. Also at this point, diversification of the ances-

tral sequence due to neutral substitution is minor. This picture contrasts with the case of organisms with frequent recombination of large chromosomal segments, in which almost all of the genome soon becomes randomly associated with the motivating allele, so that only a small adjacent clonal region hitchhikes to prominence.

The contrast is not permanent, however. Any clone must be defined from a point of origin in time. Its original genome is itself inevitably a mosaic of superimposed contributions from various ancestors (Figure 2a), due to recombination in the past. The size and distribution of the elements of the mosaic (themselves clonal segments) depends on the time of reference. If a rare broadly favorable allele now arises [and escapes loss due to random genetic drift until it has reached a "safe" number (CROW 1986; KIMURA 1983, p. 35)], it will initiate a clone. We can now redraw the linear mosaic as uniformly white (Figure 2b). This stretch of DNA will now multiply, hitchhiking with a favored allele, and become numerous enough to appear repeatedly in the samples we have taken from nature (MILKMAN 1973; MILKMAN and CRAWFORD 1983; MILKMAN and STOLTZFUS 1988). During this increase in numbers and range, and thereafter, the clonal frame will lose its uniformity in nucleotide sequence as neutral substitutions occur, and it will undergo changes in the linear extent of clonality, as recombination introduces new (and let us restrict our consideration to homologous) DNA from sources outside the clone.

The picture that emerges will be, initially, a white chromosome containing patches of various patterns, each pattern corresponding to a different clone. Eventually, the prevalence of the original DNA clone may be reduced to an equal share with numerous other

This paper is dedicated to the memory of IRVING P. CRAWFORD.

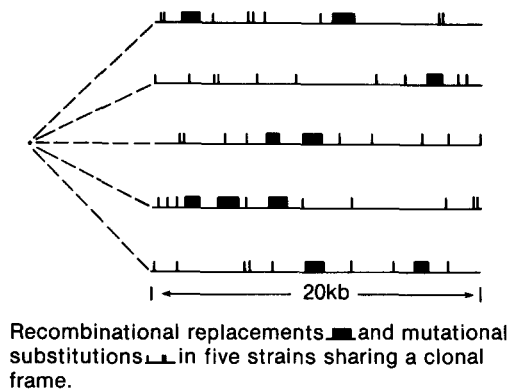


FIGURE 1.—Pattern expected from observations on 5 strains in the K12 level II clone.

patterns, as the chromosome becomes saturated with contribution upon contribution, from outside the clone and within it. Each contributed segment will alter the length of the existing segments it replaces entirely or in part, and it will be shortened or eventually replaced by subsequent contributions.

In wild strains of *E. coli* of diverse origin, both early comparative sequence data in and near the *trp* operon (MILKMAN and CRAWFORD 1983; MILKMAN and STOLTZFUS 1988) and extensive analyses of restriction fragment length polymorphisms (RFLPs) to be reported in this paper reveal a small number of *sequence types*, indicating the presence of a small number of large, broadly distributed clones. This in turn implies a class of extremely rare, new, broadly favorable alleles; if every broadly favorable allele arose frequently, many small clones would arise, and the recurrence of very similar sequences in independent isolates of diverse origin would not be observed (MILKMAN 1985). Finally, we notice a hierarchical distribution of clones of various ages, all contained in one grand clone.

The structure of the resulting variation thus depends upon the values of a small number of parameters. These are (1) the rate of mutation producing a motivating allele of general selective advantage that will bring along adjacent DNA, initially the entire chromosome; (2) the rate of spread of the motivating allele throughout the species, a function of its fitness and of *E. coli*'s vagility (rate of geographical relocation); (3) the rate of recombination; (4) the size of the recombinational replacement; and (5) the rate of neutral substitution, which also serves as the basis for a divergence clock that is useful until the members of a clone diverge beyond recognition. A real time equivalent may be obtained via an estimate of the average number of generations per year.

It appears that a clone may be able to spread worldwide in as few as 4 million generations. We now estimate, as will be described, a recombination rate of 2 replacements per 10^{12} nucleotides per generation, or about 1 per 10^5 genomes per generation (at about 4.7 million bp/genome). In 1 million generations,

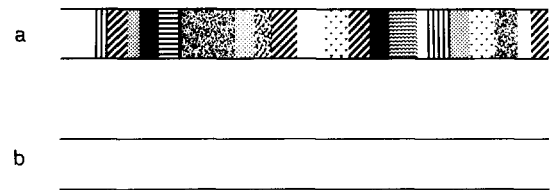


FIGURE 2.—(a) Stretch of chromosome showing its multiple origins dating from some arbitrary earlier time. Each pattern represents a contribution from a different ancestor. (b) The pattern is reset to reflect the initial absence of variation. As the clone forms, its descendants will be white, and replacements from extrachloneal sources will have different patterns.

then, an individual genome will acquire an average of 10 replacements. If they average 10^5 bp in length, roughly 20% of the genome will have been replaced, and 80% will never have been replaced (Figure 3a); if the average replacement size is 10^3 bp, only 0.2% will be extra-clonal (Figure 3b). The importance of the size of these replacements becomes clear after 10 million generations. At that point, if the average replacement is 10^5 bp, the chance of any site never having been replaced in a given chromosome is e^{-2} , or only 14%. The picture that appears is a linear quilt (Figure 4a), and it would be difficult to identify the ancestral DNA. But if the replacement averages 1000 bp, then 98% of the genome has remained unreplaced since the origin of the clone's motivating allele. The picture in this case is a *clonal frame* dotted with small replacements (Figure 4b). The two theoretical patterns will ultimately converge beyond distinction, but each is illustrated in cases to be described. In the case of a clone whose origin was about 40 million generations ago, an average replacement size of 10^5 bp would leave the average clonal frame composing about e^{-8} , or 0.03% of the genome. A 1000-bp replacement size would mean a clonal frame containing $e^{-0.08}$, or 92% of a given chromosome.

Finally, the two patterns differ in their relation to genome-wide phenograms based on an analysis of allelic variation by enzyme electrophoresis and other methods (Figure 5). Clearly, Figure 4b illustrates a situation in which a close relationship among a group of strains in one region predicts a close relationship among most of them in any other region. In contrast, Figure 4a predicts variation in groupings from locus to locus. In both cases the segments belong to clones, but the entire chromosomes do not.

We report here a set of comparisons of restriction fragment patterns in 1500-bp PCR fragments amplified from 15 regions that have been sequenced previously (BACHMANN 1983; BELL *et al.* 1989; DUBOSE, DYKHUIZEN and HARTL 1988; TSE-DINH and WANG 1986; YANOFKY *et al.* 1981; *Nucleotide Sequences 1986/1987*). The dotted pattern (Figure 4b) best fits the younger clones. The linear quilt pattern appears in older clones. Also, the analyses of the data from

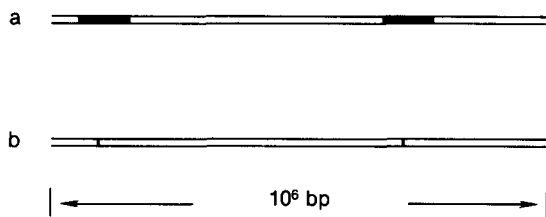


FIGURE 3.—A 3×10^6 -bp stretch of chromosome, assuming 5 replacements per 10^{12} bp per generation, 1 million generations after the origin of a clone. (a) Average replacement size is 10^5 bp; (b) average replacement is 10^3 bp.

some younger clones supports within an order of magnitude the earlier (MILKMAN and STOLTZFUS 1988) estimate of recombination rate.

MATERIALS AND METHODS

E. coli strains and chromosomal DNA: In addition to the standard reference strain K12, wild strains from the 72-strain ECOR collection (OCHMAN and SELANDER 1984) were studied. Each strain is derived from a single-cell isolate and has since been maintained in pure culture. Chromosomal DNA was isolated from liquid cultures by the method of SILHAVY, BERMAN and ENQUIST (1984). The sources of all the ECOR strains are given by OCHMAN and SELANDER (1988); a few of particular interest here are ECOR 4 (Human, Iowa), 16 (Leopard, Zoo), 19 (Celebes Black Ape, "freshly caught," Zoo), 21 (Steer, Bali), as well as 51 (Human infant, Massachusetts), 52 (Orangutan, Zoo), 54 (Human, Iowa) and 56 (Human, Sweden).

Regions studied: Some of the 15 regions were chosen for a wide chromosomal distribution; others to form a 12-kb continuous stretch. In addition, *fumB* was chosen because it is paralogous (evolved after gene duplication) with *fumA*. The regions are represented by the following symbols, with corresponding gene locus and map location given where appropriate: GN (*gnd*, 44); EF (*fumA*, 36); NR (*nirR*, 29); TA (*topA*, 28, about 8 kb upstream of the *trp* operon); ED (part of *trpE* and *trpD*, 28); CB, BA, FB, AL, LK, ON (*tonB*), BU, and GL, a continuous region from 3596 to about 15738 bp beyond the start of the *trp* operon; PH (*phoA*, 9); and MB (*fumB*, 93.5). In 23 of the 72 ECOR strains, a large insert called Atlas (Figure 6) is present between positions 8888 and 8889, splitting the AL fragment region in two. Atlas has been cloned from three strains (A. STOLTZFUS, unpublished). All 23 inserts are homologous at the right end; there is variation at the left. The AT-R PCR fragment (Table 4) straddles the right border of Atlas; its right half is the right half of AL. Physical distances are given in Table 5 with reference to *trpB*, which is at 28 min on the *E. coli* map (BACHMANN 1983).

PCR reactions: Reactions (SAIKI *et al.* 1988; SAMBROOK, FRITSCH and MANIATIS 1989) were generally carried out in 100 μ l volume and contained: 50 mM KCl, 10 mM Tris (pH 8.4), 1.5 mM MgCl₂, 100 μ g/ml gelatin, 1 μ M each primer, 200 μ M each dNTP, 2 units AmpliTaq, and 300 ng chromosomal DNA. The samples were overlaid with paraffin oil and run through 25 cycles as follows: 30 sec 94° (denature), 2 min 60° (anneal), and 3 min 72° (extend). For a few of the primer sets, it was necessary to anneal at 55° and/or add 5 more cycles. Typically, 24-nucleotide primers were chosen to generate 1500-bp fragments from translated DNA whose sequence was known. The primers were synthesized locally or by Oligos Etc. Inc., Guilford, Connecticut.

Restriction analysis: Typically 10 μ l of the 100- μ l PCR

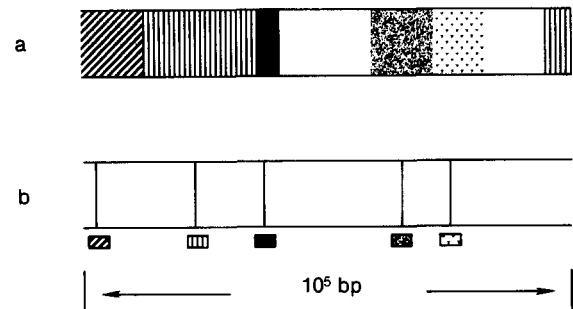


FIGURE 4.—A shorter (3×10^5 bp) stretch after a longer period of time (10 million generations), otherwise as in Figure 2.

product was digested without purification for 1 hr. The digests were now loaded on a 15-well 5% polyacrylamide vertical gel and run about 37 v hr cm⁻¹ to a standard dye position. Four to eight enzyme digests were made for each set of fragments. Fragment sizes were estimated and compared with the expectations from K12 sequences. Restriction site losses were noted; new sites were ordinarily interpreted as base substitutions, favoring synonymous changes (and transitions, given a choice). Subsequent sequencing, when performed, confirmed the interpretations almost invariably. Occasional deletions were obvious from a consistent local effect, as well as a reduction in total fragment size. Any variation in the primer regions themselves was of course obscured.

Sequencing: We have used several methods for direct sequencing of PCR products including that of MAXAM and GILBERT (GYLLENSTEN and EHRLICH 1988; INNIS *et al.* 1988; SAMBROOK, FRITSCH and MANIATIS 1989). We have now settled on the following method, for reasons of quality, speed and reliability over a variety of PCR fragments. PCR-amplified dsDNA is acrylamide-gel purified (to remove both genomic DNA and residual oligonucleotides, as well as any adventitious fragments) and subjected to a 15-cycle (typically) amplification with one primer only. The product is tested for quantity and authenticity on an acrylamide gel and sequenced following acrylamide gel purification, using the opposite amplification primer or an internal one, and the Sequenase T7 DNA Polymerase kit (U.S. Biochemical). A pooled sample of about 250 ng (0.5 pmol) of ssDNA is sequenced according to the manufacturer's protocols, except that labeling is carried out at 0° for 3 min and termination at 39° for 3 min [α -³⁵S]dATP and/or dCTP (Amersham) is used, and exposure to X-ray film is usually 1–3 days.

Terminology. The practice of expressing some sequence differences on a *per codon* basis (MILKMAN and STOLTZFUS 1988) has been abandoned in favor of a uniform *per nucleotide* basis. Also, the recombination rate is now stated in terms of replacements, rather than replacement borders (each replacement has two borders).

RESULTS

For each locus the 1.5-kb PCR fragments amplified from K12 and the various ECOR strains are grouped by sequence similarity, as inferred from restriction analysis. For example, Table 1 lists the inferred differences among a subset of 14 TA fragments, chosen to represent the entire range of variation. The number of nucleotides screened is estimated as follows. First, for the set of enzymes used, the number of

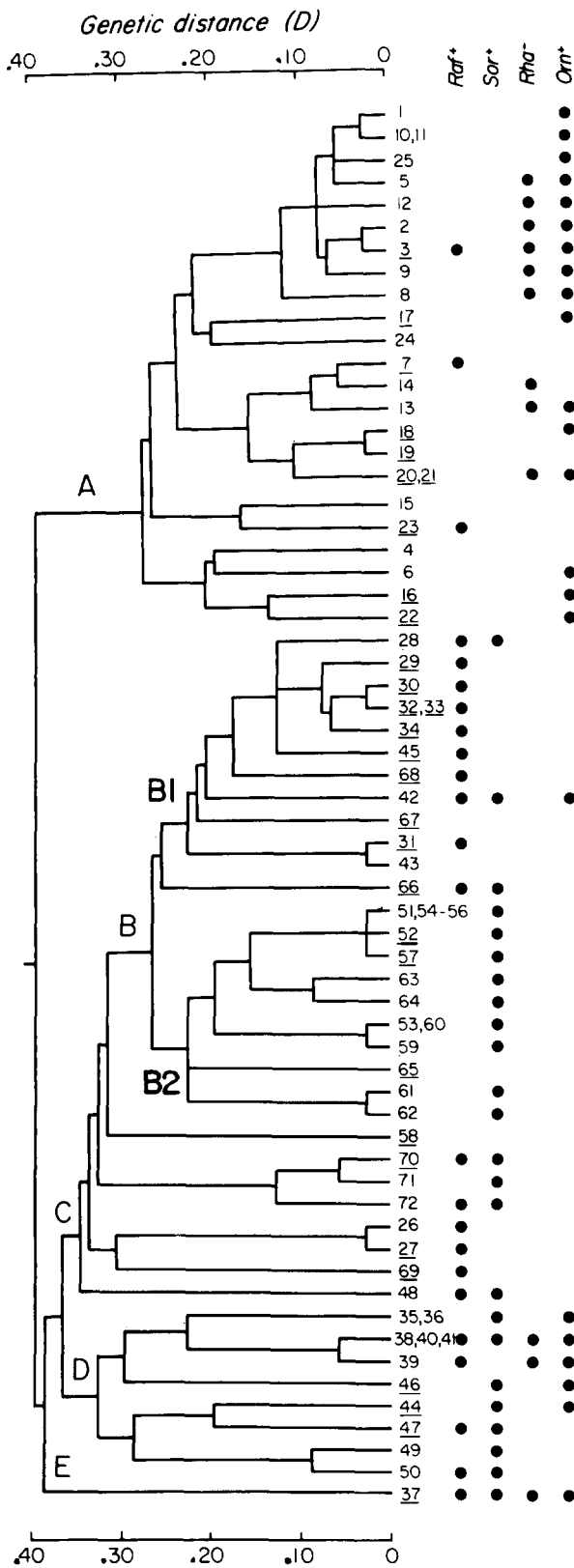


FIGURE 5.—Phenogram of the 72 ECOR strains based on electrophoretic morphs determined at 35 scattered loci (Dots refer, additionally, to 4 biotypes.) From SELANDER, CAUGANT, and WHITTAM (1987). Reproduced by permission.

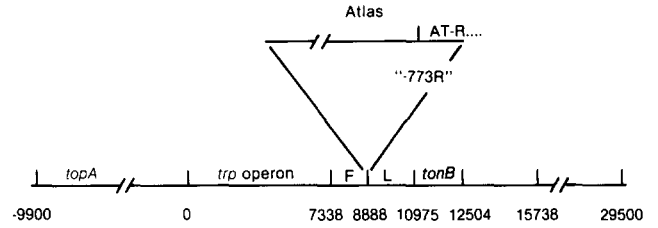


FIGURE 6.—General map of a 40-kb region including the *trp* operon. F = Fall, and L = Lika. These names are arbitrary and temporary; they do not correspond to genetic loci. Boundaries as indicated. Atlas, a region of varying length has been cloned in ECOR 36 (length 21 kb), in ECOR 49 (54 kb), and in ECOR 51 (47 kb). It is present in 23 of the ECOR strains (2, 5, 8, 9, 11, 12, 24, 35–41, 48–51, 61, 62, 64, 70, and 71), and ends on the right in a 121-np near duplicate of nucleotides 8678–8888. The *tonB* region includes the open reading frame P14, the *tonB* locus, and a short further stretch (POSTLE and GOOD 1985). We have cloned the entire 40-kb region, anticipating possible future investigations. See also STOLTZFUS, LESLIE and MILKMAN (1988).

restriction sites in K12 whose loss would be detectable is counted: for the TA example it is 40. Next, a steady state of variation is assumed; that is, the number of sites expected to be gained over a given period (*vs.* K12) is equal to those lost. Thus 40 is doubled to 80. At four nucleotides per restriction site, the total number of nucleotides screened here is 320, or about 20% of the total PCR fragment. Table 2 is a comparable display over the 12-kb continuous region. Note that ECOR 16 differs from K12 only in the LK and ON regions. This is evidence, at the level of restriction analysis, of a replacement in ECOR 16. The corresponding sequence data for LK in Table 3 show 11 or 12 nucleotide differences between K12 and ECOR 16, including those predicted by restriction analysis at positions 9929, 10004, and 10609. This sequence difference is not great enough to permit a close definition of the replacement's extent. Two moderately closely related strains are involved, as will be discussed.

Table 4 is a difference matrix for the 14-fragment TA subset. In this RFLP analysis, K12, ECOR 21, and ECOR 68 are identical, as are ECOR 16 and ECOR 28. The number of nucleotide differences inferred is given for each pair. To display the collective results of the large number of analyses performed, a linear arrangement is necessary, and this does not conform to the branched relationship of the inferred sequences. A further complication is that a particular PCR fragment is often a mosaic of clonal segments. Nevertheless, the display is informative in terms of identities among clonal segments, changes in similarity, and difference from K12. The linear array at the bottom of Table 4 shows identical segments separated by commas; single differences represented by semi-colons; multiple differences of 1% or less would be represented by a space (no examples in this set); and greater differences indicated by an asterisk.

TABLE 1

Sequence differences from K12 concluded from RFLP comparisons in the light of the known K12 sequence: TA PCR fragment

Position	Change	ECOR strain										Enzyme	
		68	71	28	35	4	37	40	65	52	50		47
1626	C-T								+	+			<i>AluI</i>
1671	C-T						+						<i>AluI</i>
1779	Loss							+				+	<i>MspI</i>
1784	Loss									+			<i>AluI</i>
1836	Loss					+							<i>BstUI</i>
1895	Loss											+	<i>MspI</i>
1960	G-A								+	+		+	<i>RsaI</i>
2031	Loss							+			+		<i>HinPI</i>
2124	G-C				+		+	+	+	+	+	+	<i>TaqI</i>
2239	Loss										+		<i>HinPI</i>
2259	Loss											+	<i>AluI</i>
2470	G-T										+		<i>HinPI</i>
2607	T-C					+	+	+	+	+	+	+	<i>HinPI</i>
2670	A-G				+			+	+	+	+	+	<i>MspI</i>
2742	C-G										+	+	<i>BstUI/HinPI</i>
2771	Loss			+		+	+					+	<i>AluI</i>
2782	Loss		+	+		+	+						<i>BstUI</i>

Table 5 contains similarly derived linear displays of the same subset of 14 strains for all regions studied. Table 6 displays the entire current data set of over 600 fragments, each subjected to 4–8 restriction digests. The overall order is based on differences from the K12 sequence type. It will be seen for CB, in Tables 5 and 6, that K12 is not in the leftmost group. This reflects the tentative inference that K12 has received a replacement and thus does not have its ancestral sequence type here. This inference is based on comparisons with fragments of related strains. On the other hand, ordering *within* the leftmost groups in PH, GL, ON, LK and NR merely reflects the one-site difference relationships.

Clonal levels: The main conclusion to be drawn from these data is that certain strains contain similar or identical sequence types in certain regions, and that these affinities are not constant. This conclusion may be detailed further as follows.

1. For any region, strains can be grouped by similarity into sets of various sizes. Strains that differ by fewer than 0.5% of their nucleotides in a region are said to share a *sequence type*. Sequence types are named for groups centered on reference strains, which include, respectively, K12, ECOR 15, ECOR 68, ECOR 71, ECOR 52, ECOR 65 and others, often infrequently represented. Note that a recombinational replacement may locally place a reference strain in a different sequence type, as apparently K12 in the trpCB region. Also, in some regions recombination makes a sequence chimeric, thus dividing its classification (as in portions marked *a* and *b*).

2. There are five different sets whose strains are identical over all regions compared, namely, ECOR 35 and 36; 39 and 40; 51, 52, 54, and 56; 61 and 62;

and 70 and 71. The modest amount of sequencing to date has not revealed differences within *these* RFLP-identical sets. Note also that the restriction analysis of 51, 52, 54 and 56, which presently covers 55 PCR fragments, has screened about 16,500 bp without revealing a difference. Each of these sets is called a level I clone.

3. Among less closely related strains the degree of similarity is not constant over the regions studied. For example, some strains are identical or very similar over most, but not all, regions (K12, ECOR 4, 16, 19 and 21; likewise strains related to ECOR 52). K12, ECOR 4, 16, 19 and 21 thus share a clonal frame but do not have it in all regions: even K12 does not always have the K12 clonal frame. Sequencing often reveals a few nucleotide differences between such strains in regions where no restriction site differences have been observed. K12 and ECOR strains 4, 16, 19 and 21 compose a level II clone; the clonal relationship is not genome-wide, but is restricted to their clonal frames. Unlike a sequence type, whose definition is constant, a clonal frame's diversity is determined by its level. Thus the K12 level II clonal frame is not shared by ECOR 71, but the K12 level III clonal frame is.

4. Strains K12, 15, 68, and 71 are generally similar and occasionally identical. They (and their respective level I and II clone-mates) constitute a level III clone. ECOR 52 and 65 (and level I and II affiliates) constitute another level III clone. Some additional strains are expected to fall into each of these two level III clones.

5. The entire set of fragments studied appears to belong to one grand clone, with two exceptions. At the *gnd* locus, sequences from ECOR 4 and ECOR 16 were shown by DYKHUIZEN (SAWYER, DYKHUIZEN and

TABLE 2

Inferred nucleotide substitutions in a long continuous region

Fragment	Position	Sequence type							
		K12					68		
		K12	4	16	19	21	28	68	71
GL	<i>Bsu</i> UI ^a	-	-	-	-	-	+	+	+
	<i>Msp</i> I ^a	-	-	-	-	-	+	+	-
BU	13843	-	-	-	-	-	+	ND ^b	-
	13017	-	-	-	-	-	+	+	-
	12820	-	-	-	-	-	+	+	-
ON	12473	-	-	-	-	-	+	+	+
	11949	-	-	-	-	-	+	+	+
	11947	-	-	-	-	-	+	+	-
	11943	-	-	-	-	-	+	+	-
	11688	-	-	+	-	-	-	-	-
LK	10900	-	-	-	-	-	-	+	+
	10609	-	-	+	-	-	-	-	-
	10238	-	-	-	-	-	-	+	-
	10004	-	-	+	-	-	-	-	-
	9929	-	-	+	-	-	-	-	-
	9704	-	-	-	-	-	-	-	+
AL	9606	-	-	-	-	-	+	+	+
	9091	-	-	-	-	-	+	+	+
	9040	-	-	-	-	-	+	+	+
	8590	-	-	-	-	-	+	+	-
	8575	-	-	-	-	-	-	+	-
	8489	-	-	-	-	+	-	-	-
FB	7870	-	-	-	-	-	+	-	-
	7738	-	-	-	-	-	-	+	+
	7714	-	-	-	-	-	-	-	+
BA	6394	-	-	-	-	-	-	+	+
CB	4809	-	+	-	+	+	-	-	+
	4656	-	-	-	-	-	-	+	+
	4156	-	+	-	+	+	-	-	-
	4041	-	+	-	+	+	-	-	-
	4038	-	+	-	+	+	-	-	-
	4032	-	+	-	+	+	-	-	-
	3999	-	+	-	+	+	+	-	-
	3852	-	-	-	-	-	-	+	-
	3779	-	-	-	-	-	+	+	-
3621	-	-	-	-	-	-	+	-	

^a Unordered.^b ND = not done.

HARTL 1987) to differ from one another and from K12 as much as a *Salmonella typhimurium* strain differs from each of them. We obtained a GN fragment from ECOR 4 which was extremely different from the others, and attempts to amplify ECOR 16 failed, a result also consistent with a great sequence difference. Otherwise, no inferred sequence differences even approach the proportion expected at equilibrium, 0.167 per nucleotide.

As several lines of descent diverge from a common ancestor, they accumulate nucleotide substitutions independently of one another, and they accumulate recombinational replacements independently as well. Without introducing the time element, chromosomes may

be compared for individual substitutions and for replacements. This comparison ordinarily reveals which line has undergone the change. In the case of mutation, individual nucleotide differences are observed between each chromosome and the others, or between a set of several chromosomes and the rest. Recombination can produce an abrupt change over chromosomal distance: two chromosomes may be identical (or dotted with sporadic nucleotide differences) and then become different at a much higher proportion of sites. A return to the original relationship is observed further down the chromosome. A comparison between strains whose clonal frames are more distantly related (levels III and IV) will produce the converse result if one has donated a replacement to the other: chromosomes may appear to be quite different, then very similar, then different again.

The K12 level II clone: These features direct our approach to the questions of the relative frequency and the size of recombinational replacements. In order to maximize the changes of observing individual replacements, as opposed to superimposed multiple events, a set of very closely related strains should be studied first: not the ones that are always identical in studies to date, but the next category, those that reveal occasional abrupt changes in similarity. A set of five strains provides a good example of recombinational replacements. RFLP data on the level II clone strains K12, ECOR 4, ECOR 16, ECOR 19, and ECOR 21 are taken from 14 regions (excluding GN because of its extreme variation), each about 1.5 kb in length. The strains are identical by this criterion in most cases; they are clearly different in a few. Inspection of the data in Table 6 shows that all five strains are identical in 3 regions (BU, FB, BA); in 4 regions (GL, ON, AL, NR), one strain differs from the other four by one site, as indicated by a semicolon. In these 7 regions, all five strains are concluded to have no recombinational replacements. In 7 other regions, one strain (in MB, LK, ED, TA, EF) or an identical pair (in PH, CB) differ from the rest by 3-6 sites (data not shown). These differences are considered to reflect replacements; in the two cases of identical pairs, each pair's presumable ancestral replacement is counted twice. Finally, returning to TA, ECOR 16 differs from K12 by two nucleotides a dozen bp apart and within 40 bp of the primer at the end of the fragment. This is taken to be the beginning of a replacement that continues beyond the region amplified, rather than two random nucleotide substitutions, for two reasons. First, at this level nucleotide substitutions are rare, and second, the two differences from K12 are shared by several more distantly related strains. The direction of replacement in CB is uncertain. Comparisons with related strains suggests that a common ancestor of K12 and ECOR 16 received the replacement, but it is

TABLE 3
Sequences compared in LK fragment

Position	Strain											
	K	1	4, 12, 19	8	21	16	15	37	69, 72	51, 52	50	
(Start 11065)												
10956	G	-	-	-	A	-	-	-	-	-	-	-
10930	T	-	-	-	-	-	-	-	-	C	-	-
10903	C	-	-	-	-	-	-	-	T	-	-	-
10894	G	-	-	-	-	-	-	-	A	-	-	-
10877	C	-	-	-	-	-	-	-	-	T	-	-
10867	G	-	-	-	-	-	-	-	-	C	-	-
10798	A	-	-	-	-	-	-	-	-	T	-	-
10786	T	-	-	-	-	-	-	-	-	C	-	-
10774	T	A	A	A	A	A	-	A	A	-	A	A
10729	G	-	-	-	-	-	-	-	-	A	-	-
10701	T	-	-	-	-	-	-	-	-	-	-	C
10651	C	-	-	-	-	-	-	-	-	A	-	A
10624	G	-	-	-	-	-	-	-	-	C	-	-
10609	G	-	-	-	-	A	A	-	-	-	-	-
10591	G	-	-	-	-	-	-	A	-	-	-	-
10464	T	-	-	-	-	-	-	-	-	-	-	C
10461	A	-	-	-	-	-	-	-	-	G	-	G
10457	A	-	-	-	-	-	-	-	T	-	-	-
10367	A	-	-	-	-	-	-	-	-	-	-	G
10313	T	-	-	-	-	-	-	-	-	A	-	C
10307	G	-	-	-	-	-	-	-	-	-	-	C
10301	T	-	-	-	-	-	-	-	-	-	-	C
10295	T	-	-	-	-	-	-	-	-	-	-	A
10292	G	-	-	-	-	-	-	C	-	-	-	-
10278	G	-	-	-	-	A	A	-	A	-	-	A
10257	T	-	-	-	-	C	C	C	C	C	-	-
10241	C	-	-	-	-	-	-	-	-	T	-	-
10236	C	-	-	-	-	-	-	-	-	-	-	T
10235	A	-	-	-	-	G	G	-	-	-	-	-
10232	T	-	-	-	-	-	-	A	-	-	-	-
10145	A	-	-	-	-	C	C	-	C	-	-	C
10097	A	T	-	-	-	-	-	-	-	-	-	-
10061	T	-	-	A	-	-	-	A	G	-	-	A
10058	A	-	-	-	-	-	-	-	-	G	-	-
10034	A	-	-	-	-	G	G	-	-	-	-	-
10016	G	-	-	-	-	T	T	-	-	-	-	-
10004	G	-	-	-	-	A	A	-	-	-	-	-
9957	G	-	-	-	-	A	-	-	-	-	-	-
9929	G	-	-	-	-	A	A	-	-	A	-	-
9840	G	-	-	-	-	-	-	-	-	A	-	-
(Stop 9708)												

Position numbers start with the beginning of the *trp* operon. Only polymorphic positions are listed, and only differences from K12 are indicated explicitly. Strain ECOR 16 may have a replacement about 400 bp long, extending from about 9900 to 10300. This would be supported by the finding of a matching sequence that is part of one that differs from K12 over a greater length. The difference at 10774 is attributed to a substitution in K12, and that at 10609 to a substitution in ECOR 16. The range of replacement sizes, while centering near 1 kb, may turn out to be fairly broad.

possible that a common ancestor of 4, 19 and 21 was the recipient.

These cases add up to 10 recombinational replacements in 5 strains, in 14 regions averaging 1500 bp, making a total of 105 kb. At the same time, from a smaller amount of sequence data, we estimate that there are roughly 50 nucleotide substitutions in these 105 kb. These strains correspond to K12, 39A, 191F, 201C and 213K, respectively, which are members of the "K12-like" group sequenced by MILKMAN and CRAWFORD (1983) and found to have an average

pairwise difference, roughly, of 1 in 1000 nucleotide pairs in *trpBA*. Thus each line averages 0.5 *substitutions* per 1000 bp, a value consistent with sequences of these strains in LK (Table 3), in which 3 substitutions were found in some 6700 bp, as well as one apparent replacement. Thus, whatever the absolute rates may be, the ratio of retained replacements to retained substitutions seems to be about 0.2, as illustrated in Figure 1.

The estimate of the average replacement size, about 1000 bp, is based on a variety of RFLP data including

TABLE 4
Difference matrix: TA PCR fragment

	71	16 28	35	4	37	40	65	52	50	47
K12,21,68	1	2	2	4	5	5	5	6	7	9
71		1	3	3	4	6	6	7	8	10
16,28			4	2	3	7	7	8	9	9
35				6	5	3	3	4	5	7
4					3	7	7	8	9	9
37						6	6	7	8	8
40							4	5	4	6
65								1	6	6
52									7	7
50										8

Linear display:

TA: K,21,68;71;16,28 * 35 * 4 * 37 * 40 * 65;52 * 50 * 47

those given for strains K12 through ECOR 21 in Table 2, as well as considerable sequence data (mostly not shown). It justifies the use of the fragment size, 1500 bp, as the number of bases screened by the analysis of each fragment: were the replacement size much larger, the number of bases screened would be larger as well. As to the actual proportion of unchanged DNA, the clonal frame, it is estimated at e^{-m} where m is the average proportion of DNA replaced since the origin of the clone, $10 \times 1000/105,000$ bp,

where 1000 is the average replacement size, and 105,000 is the total number of bp screened. Thus the proportion of the DNA remaining in this level II clonal frame is $e^{-0.1} = 0.90$.

A final feature of segmental clonality is that in any chromosomal region we observe rather few sequence types, instead of a vast array of continuously varying combinations. Among the ECOR strains examined, the *effective number*, $(1/\sum p_i^2)$, where p_i is the frequency of a given sequence type, averages 5.2 over the 14 regions aside from the unusually variable GN region. (If all sequence types were equally frequent, their actual number, n_a , would equal n_e . This is the standard definition for effective number of alleles, and for some other sets of alternatives. This effective number governs expectations but does not exclude rarities.) The fact that this number is small is of obvious importance in the successful interpretation of the abrupt changes in sequence similarity observed. Specifically, when recombination results in the replacement of a stretch of DNA, the new DNA will ordinarily belong to a recognizable sequence type. Most important, if a complete set of available sequence types is analyzed in a given region, almost all recombination events will be between two of the sequence types; therefore, they will result in abrupt stretches of similarity between two ordinarily different samples. We seem, in other

TABLE 5
RFLP Affinities of fragments from 14 strains in 15 regions

Region	DNA Sources
MB	K,16,21;28 68,71 47 4 37 * 50 * 35 * 40 65 52
1.5 mb	
PH	K;4,28;16 * 37 * 68 * 21;71 * 65;52 * 35 * 40 * 50 [47]
640 kb	
GL	4;K,16,21,37;71 50 28,68 * 47 52,65 * 35 40
BU	K,4,16,21;71 * 28,68 * 50 * 37 * 65 52 * 40 * [35*,47]
ON	16;K,4,21,37 71 28,68 * 52;65 40 35 * 47 * 50
LK	37;K,4,21,28 68 71 * 16 * 52,65;40 * 50 * 47 * 35
AL	K,4,16;21 72 28,68 * 52;65 * 47 (Atlas intervenes in 35,37,40,50)
FB	K,4,16,21;28 68 71 * 37 47 50 * 52D * 65 35 40
BA	K,4,16,21,28 40b 68 71 * 37 * 52;65 35,40a * 47 * 50
CB	4,21 * 71 * K,16 * 28 * 68 * 47 * 50 * 37 * 35;40 * 52,65
ED	K,4,21 * 16,28,68,71 * 37 * 40 * 35 * 52 * 65 * 47 [50]
-4.5 kb	
TA	K,21,68;71;16,28 * 35 * 4 * 37 * 40 * 65;52 * 50 * 47
-15 kb	
NR	16,28;K,4,21;58;37 * 71;52,65 * 47 * 35 40 * 50
-80 kb	
EF	K,4,21 28 52 * 37 * 16 * 40 * 50 35,47 * 71;65 * 68
-320 kb	
GN	K * 21 * 71 * 52,68 * 28 * 40 * 50 * 35 * 65 ** 4 [16,37,47]
-640 kb	

Unless distance from *trpBA* is given, fragments are contiguous or overlap slightly. In gnd, ** indicates that ECOR 4 is especially different from the others, as was already known (see text).

Symbols: D = (small) deletion present. , = no difference. ; = 1 site difference. [space] = >1 site but ≤1% sites differ. * = >1% difference. a,b = chimeric portions. [] = PCR has failed, so far.

* 35 and 36 have a 4.7 kb insertion which appears to be within Burton. Occasionally for no known reason a strain cannot be amplified. Source strains: K-4-16-21-28-35-37-40-47-50-52-65-68 and 71 (72 for Ally).

TABLE 6
RFLP data grouped by similarity

Region	DNA Sources
MB	K, 1, 2, 5, 9, 11, 12, 47; 4 * 6 * 37 * 31 * 49, * 35, * 39, 65 51, 52, * 46 16, 17, 19, 21, 24, 44 50 36 40 54, 56, 30; 27, 28, 45, 67, 60, 61, 69; 15, 29, 68, 70, 62, 64; 71, 72; 58 66
PH	27; K, 4, * 37 * 29, 45, * 19, 21, * 65, 66; 51, * 35, * 39, * 49, 28, 67; 16 68, 69 70, 71; 52, 56, 61, 36 40 50 6, 72 62
GL	4, 15; K, 1, 9, 16, 19, 49, 27, 28, * 47 51, 52, 54 * 61, 64; * 35, 39, * 44 21, 24, 37; 29, 30, 50 58, 68 56, 60, 62 66 36 40 31, 67, 70, 71, 72 65; 69
BU	K, 1, 2, 4, 5, 6, * 27, 28, 15, * 31 * 37 * 65 * 51, 52, * 39, * 44 * 58 8, 11, 12, 16, 29, 30, 50 54, 60, 40, 17, 19, 21, 24; 68; 45, 61, 62, 46 70, 71, 72 69 64
ON	21; 31; K, 1, 2, 17D 27, 69, 28, 29, * 61, 62; 51, 39, 35, * 47 * 49, 4, 5, 8, 11, 70, 71, 30, 58, 52, 54, 56, 40 36 50 12, 15, 16, 72 67, 68 60, 64; 65, 19, 24, 37; 6 66
LK	45; K, 1, 2, 4, 68; 27, * 44 * 6, * 40; 51, 52, * 49, * 47 * 35, 5, 8, 9, 11, 12, 69, 70, 15, 54, 56, 65, 50 36 19, 21, 28, 37, 71, 72; 16, 66; 61, 62; 67; 31; 46 29, 30 24 60
AL	21; K, 1, 4, 27; 28, 58 * 31 * 44 * 46 * 47 * 65; 52, 6, 15, 16, 45, 68; 54, 56, 17, 19, 72, 29, 30 60, AT-R ^a 2, 5, 8, 9, 70, 51; 35, * 49, * 38, 39, * 64 11, 12, 37, 71 61, 36 50 40, 41 48; 24 62
FB	K, 1, 4, 27, 30, * 67 * 37 47 49, * 51D, * 60D, * 61, 65 35, 39, * 66D 6, 15, 29, 70, 50 52D, 64D 62 36 40 16, 19, 45, 71, 54D, 21, 24, 68, 72 56D 58; 17; 69 28
BA	K, 1, 4, 15, 27, 29, * 37 * 51, 52, 35, * 47 * 49, 16, 17, 19, 30, 68, 54, 56, 36, 50 21, 24, 28, 70, 71, 60, 61, 39a, 58, 67; 6; 45, 72 62, 64; 40a 69; 39b, 40b 65; 66
CB	17 4, 19, 24 * 67, * 29 K, 1, * 28 * 68 * 47 * 49, * 37 * 35, * 51, 52, * 61, * 46 21, 30 70, 5, 6, 9, 50 36; 54, 56, 62 45; 27, 71, 11, 12, 39, 60, 64, 69; 23 72 15, 16, 40 65, 66
ED	K, 1, 2, 6 * 15 * 16, 28, 30, * 37 * 39, * 35, * 51, 52, 65 * 44, * 66 4, 17, 45, 68, 70, 40 36 54, 56; 21; 19 71, 72; 27; 60, 61, 58, 69 62, 64
TA	31; K, 1, 2, 9, 69 * 58 * 35, * 4 * 37 * 39 6; * 65 61, 51, * 49, * 46 47; 11, 12, 19, 21, 36, 40 15; 62 52, 50 44 68; 27, 29, 30 60, 17 54 45, 70, 71, 72; 64 56; 16, 28, 67 66
NR	24; 9, 15, 16, 27, 28, 45, 69; 70, * 47 * 35, 40 * 47 * 49; 29, 30, 58, 67, 72; 71; 51, 52, 36 50 K, 4, 6, 8, 11, 12, 56, 60, 64, 19, 21, 31; 68; 37 65, 66
EF	K, 1, 4 28 51, 52, * 37 * 15a, * 39, * 49, * 46 * 35, * 29, 30; * 68, * 44 17, 19 54, 56, 16; 40 50 36, 70, 71, 69, 21, 24, 60, 61, 58; 45 47 72; 27a, 15b, 31 62; 66 67 65 27b
GN	K * 21 * 29 * 70, * 51, 52, * 19 * 28 * 40 * 50 * 35, * 65 * 69 * * 4 71 68 36

^a AT-R comprises 773 base pairs in Atlas, a lambdaoid-phage-like insert in 23 of the 72 ECOR strains, and the right-hand 811 bases of AL. The affinities indicated by corresponding positions refer to the DNA shared by the AL and AT-R PCR fragments. Thus AL 19 = AT-R 2; AL 72 = AT-R 70, and AL 60 = AT-R 51. The AT-R fragments from ECOR 35 on are not especially similar to any in the AL group in this shared 811-bp section.

^b Variation in GN among some 20 additional strains is not yet grouped.

The 773-bp region exclusively in Atlas is complicated by length polymorphisms and will not be described here. Other symbols and explanatory notes are presented in Table 5.

words, to be dealing with an essentially closed system. Any infrequent exceptions (namely, rare sequence types present in isolated stretches only) will result in

a loss of similarity between two sequences not being accompanied by a new similarity to a different known sequence.

TABLE 7
Clonal levels

Clonal level	Estimated age (gen.)	Example strains	Clonal frame	
			Uniformity	Proportion (%)
I	4×10^6	51,52,54,56	0.0001?	100
II	4.5×10^7	K,4,16,19,21	0.001	90
III	4×10^8	K,15,68,71	0.01	87
IV	2×10^9	K,35,47,50,52	0.04	>99

Notes: Recall that any replacement that has occurred before the clonal ancestor is part of the clonal frame; it may become obvious at the next higher (=older) level. The percentage of replacement is the proportion of the genome replaced, assuming a replacement size averaging 1000 bp.

To summarize the clonal hierarchy, the strains belong to one grand clone which is subdivided into progressively more recent clones. The age of a clone is defined by the time since its most recent common ancestor. The age of a clone determines the uniformity and the extent of its clonal frame. Sequence type, in contrast, has a constant definition (p. 13). Clonal level, as described in Table 7, is a pragmatic classification. Age is estimated as divergence time in generations, based on differences seen in the pairwise comparison of clone members. The clonal frame is the DNA derived from the clonal ancestor; its uniformity is stated in terms of maximum pairwise divergence of clone members; its proportion of the genome is the part of the chromosome that has not been replaced by extra-clonal recombination. At levels III and IV intra-clonal replacement is also discernible. The grand (level IV) clone includes individual strains and sets which are generally quite different but occasionally are very similar or even identical. Such occasional affinities are seen in Table 6 and in a variety of

sequence comparisons made to date, including that of ECOR 47 and 50 in BA (MILKMAN and STOLTZFUS 1988).

A tentative estimate of replacement rate, r_r : There is no reason to assume that any substantial proportion of replacements is not retained; but of course there is good reason to believe that most nucleotide substitutions are not retained. In coding sequences, which predominate in our samples, we use the proportion 1/9 for retention on the assumption that at one of three codon positions, one of the three substitution pathways is neutral. To the extent that this is accurate, and that the nucleotide substitution rate is 3×10^{-10} per nucleotide per generation, then the rate of recombinational replacement, r_r , is about $0.6 \times 1/9 \times 10^{-10}$ or roughly 5×10^{-12} per nucleotide per generation. That is, the chance of a replacement *beginning* at a given nucleotide in a given generation is about 5×10^{-12} . Finally, since K12-like strains are abundant in independent collections from various parts of the world, this value confirms a relationship between spread rate (rise in frequency) and recombination rate such that the entire clonal chromosome rises to high frequency before extensive modification takes place. Thus, while the most recent clones have evidently been observed soon after reaching high frequency, broader groups in the hierarchy evince greater clonal age, and a greater time since rising to high frequency. Figures 2, 3 and 4 summarize the process.

Table 8 is a matrix of identities (upper right) and close similarities (non-zero differences no greater than 1%, lower left). Notice that changes frequently involve a shift from one identity group to another (Tables 5 and 6). Where a majority of the fragments are iden-

TABLE 8
Identities and similarities

	K	4	16	19	21	28	35	37	40	47	50	52	65	68	71
K	—	10	8	11	12	4	0	4	0	0	0	0	0	1	0
4	2	—	5	9	9	4	0	1	0	0	0	0	0	0	0
16	3	4	—	6	5	6	0	1	0	0	0	0	0	1	1
19	1	3	3	—	11	3	0	4	0	0	0	0	0	1	0
21	1	2	3	3	—	3	0	2	0	0	0	0	0	1	1
28	6	5	4	7	6	—	0	0	0	0	0	0	0	4	0
35	1	1	1	0	0	1	—	0	1	1	0	0	0	0	0
37	1	3	2	1	2	2	0	—	0	0	0	0	0	0	0
40	0	0	0	0	0	0	6	0	—	0	0	0	0	0	0
47	0	1	0	1	0	0	1	2	0	—	0	0	0	0	0
50	0	0	0	1	0	1	1	1	0	3	—	0	0	0	0
52	1	0	0	2	1	1	1	0	2	0	0	—	4	1	0
65	2	2	2	1	1	1	6	1	4	0	0	4	—	0	0
68	8	8	8	8	7	8	1	4	1	1	1	1	2	—	3
71	8	7	7	10	8	8	0	2	1	0	0	2	3	6	—

Upper right: number of identities (regions where the strains are identical); lower left: number of nonidentities differing by 1% of their nucleotides or less. The total number of segments surveyed is 14. The following are tentatively designated sequence types: K12, 35, 37, 40, 47, 50, 52, 65, 68 and 71. Types K12, 68 and 71 are relatively closely related, as are 52 and 65.

Genomes containing a given sequence type in preponderance are said to have a clonal frame of that sequence type. Thus, K12, 4, 16 and 21 each share a K12 clonal frame.

tical, as between K12 and ECOR 4, the two strains share a level II clonal frame; where a majority are either identical or closely similar, as between K12 and ECOR 68, the two strains share a level III clonal frame. Infrequent similarities reflect replacements.

DISCUSSION

ECOR 1 is likely to be placed in the K12 level II clone after further sequencing. The two nucleotide differences from K12 in LK (Table 3) support this; the absence of apparent replacements in the 13 PCR fragments examined to date is weak evidence for level I: if they average about 2 replacements in 20 kb, one of 7 level II clone members should show none ($e^{-2} = 0.14$).

Southern hybridizations and limited sequence comparisons suggest that the Atlas insertions represent several distinct species of lambdoid prophages (and their cryptic derivatives) sharing a unique attachment site. An Atlas phage has been isolated from one of the ECOR strains (ECOR 2) and propagated on *E. coli* strain C (A. STOLTZFUS, personal communication). Some sets of otherwise identical strains (by the present RFLP criteria) are not uniform with respect to the presence of Atlas. RFLP analysis (AL and AT-R, Table 6) shows that ECOR 51 (which contains Atlas) is identical, in the adjacent region, to ECOR 52, ECOR 54, and ECOR 56 (which do not contain Atlas). Similarly, ECOR 70 and ECOR 71 are identical at the right end of Atlas and identical, in the adjacent region, to ECOR 72 (which does not contain Atlas). These observations imply that Atlas is a relatively transient resident.

The recombination rate r was previously estimated from sequence-type discontinuities (MILKMAN and STOLTZFUS 1988) to be about 4×10^{-13} discontinuities (replacement borders) per nucleotide per generation. Since then, the emphasis has shifted from overlapping replacements (which shorten a clonal segment) to isolated replacements (which interrupt a clonal frame). Also, J. W. DRAKE (personal communication) now estimates the nucleotide substitution rate, k , upon which other calculations are based (MILKMAN and STOLTZFUS 1988), at about 3×10^{-10} , not 2×10^{-10} .

The cumulative recombination rate, r_c , is the product of the recombinational replacement rate per nucleotide per generation, r_r , and the number of generations, g , since divergence. For a pair of compared sequences, divergence time is calculated as

$$g = (1/4k) \ln [X_c/(X_c - X_0)] \\ = (1/4k) \ln [0.167/(0.167 - X_0)],$$

where k is the nucleotide substitution rate per nucleotide per generation, X_c is the proportion of nucleotide differences expected at equilibrium and X_0 is the

observed proportion of nucleotide differences. For the K12 level II clone, whose members' average substitution frequency of 0.0005 must be doubled to give pairwise divergence, the divergence time comes to 15×10^6 generations. The proportion of nucleotide differences expected at equilibrium in translated DNA is taken to be $1/3 \times 1/2 = 1/6$, or 0.167, assuming that an average 1/3 of all nucleotides have an average of one neutral alternative, so that half of these nucleotides will differ in any pairwise comparison at equilibrium (MILKMAN and STOLTZFUS, 1988; note that $0.5/\text{codon} = 0.167/\text{nucleotide}$).

Actually, the observations of CAIRNS, OVERBAUGH and MILLER (1988) and HALL (1988, 1990) give us some reason to question the applicability in nature of the estimate of k in laboratory conditions (nutrient broth, exponential growth, one strain). This uncertainty makes it worthwhile to state the value of r_r/k , which is 0.02.

We have noticed that replacements occasionally appear to occur in series. For example, restriction analysis shows that ECOR 37 and K12 are usually very different. However, they are identical in GL (sequencing of 450 bp reveals 2 differences), very different in BU, and identical in ON, in LK (6 differences in Table 3), and in the shared right half of AL and AT-R. Also they differ by only two restriction sites in NR. Since ECOR 37 is rarely like the other ECOR strains compared here (and is the most different from the others in SELANDER, CAUGANT, and WHITTAM'S (1987) comparison, Figure 5), it seems unlikely that these similarities are independent of one another. One possibility is that the entry into a cell of a large (*ca.* 100 kb) DNA fragment is followed by a number of separate incorporations. This possibility can be investigated by examining a set of PCR fragments at intervals and comparing the replacement patterns with those of more widely spaced sets to see if replacements tend to be clustered.

The case of K12 and ECOR 37 also illustrates the determinability of the direction of DNA transfer. In this case, K12 was the donor, because here K12 and ECOR 37 both resemble close relatives of K12. ECOR 4, ECOR 28, ECOR 68, and ECOR 70 are still similar to K12, *yet clearly not as similar as ECOR 37 is*. Had ECOR 37 been the donor, K12 would have become quite different from these strains in the replacement region, and ECOR 37 would have remained so. (Note that a resemblance to ECOR 21, for example, is of no interest, because a common ancestor of K12 and ECOR 21 could have received DNA from ECOR 37.) Moreover, the sequence comparisons tell us something about the divergence between K12 and the presumed set of replacements that ECOR 37 received. If the difference is 8 nucleotides in 1800 (6 differences in Table 3, plus 2 differences in 450 bp in GL, not

shown), then the replacement DNA diverged from K12 DNA about 67 million generations ago. The recombinational event could have taken place any time between then (from K12's direct ancestor) and a few years ago (from a collateral relative of K12). In BU, K12 and ECOR 37 differ by 20 nucleotides of 1100, and this suggests a divergence time of 290 million generations. Treating the 1800 bp in LK and GL as coming from one source and the 1100 bp in BU as coming from one source, the probability that the sources are the same is much less than 0.001 ($\chi^2 = 13$), assuming equal local substitution rates. Further evidence as to whether the discontinuous replacements in LK and GL are really derived from a single DNA entrant will likely emerge from additional sequencing.

The RFLP procedure is quite simple when fragments are identical in one place and quite different in another, or when they change from one identity group to another. Definition of replacements requires further information when, as in the case of K12 and ECOR 71, the clonal frames appear to differ by an average of only 1.1% (MILKMAN and STOLTZFUS 1988). Here, local differences may be due to random fluctuations in substitution rate, or they may indeed reflect replacements. Sequencing provides much greater detail and is likely to provide a definitive answer in many cases, not merely by adding detail, but by demonstrating affinity to a different known sequence type. Note that some measurable degree of sequence divergence is always required to estimate the replacement rate in terms of the time standard (substitution rate) and from there to generations and real time. An extensive program of sequencing is now in progress and should produce a firmer estimate of r , as well as detailed features of the individual replacements.

Eventually, the unusually variable GN region must be considered in any comprehensive description of recombination in *E. coli*, especially since other similar regions may exist. It has been suggested that adaptive antigenic polymorphism determined at the *rfb* locus, together with specialized transduction, may increase the genetic diversity nearby, including the *gnd* locus [SELANDER, CAUGANT and WHITTAM (1987), p. 1631]. While all recombinational replacements are of interest in this study, the possibility that recombination in the GN fragment does not share homogeneity with the other cases is good reason to consider it separately for the time being. SAWYER (1989) has inferred that "conversion events" in the *gnd* gene have a size range of 70–200 bp. These events may correspond to what we have called replacements, but they may instead be the remnants of overlapping replacements.

The amount of neutral polymorphism in a random

population sample, as influenced by recombination and selection, has been addressed theoretically by Hudson and collaborators. Most recently, a steady-state, stochastic finite-population model of the hitchhiking phenomenon has been developed by KAPLAN, HUDSON and LANGLEY (1989; see also references therein).

Recombination in nature is known for a broad array of eubacteria (LEVY and MILLER 1989), excluding the Cyanobacteria (for which the present approach seems well suited). The results of initial experiments on four *Bacillus subtilis* fragments from up to 33 strains with FREDERICK COHAN (Wesleyan University, Connecticut) and on two fragments in a variety of *Anabaena* and *Nostoc* strains with SUSAN BARNUM (Miami University, Ohio) show patterns comparable to those seen in *E. coli*.

Finally, clonal frames provide a conceptual link between segmental clonality and electrophoretic types (ETs). It is no accident that the ECOR strain numbers in identical sets (level I clones) according to RFLP analysis (*e.g.*, 35/36, or 70/71) are numbered consecutively or nearly so. The numbering scheme was devised on the basis of similarities established by SELANDER and colleagues (OCHMAN and SELANDER 1984) in the electrophoretic genotypes of the strains—the number of allozymes with identical mobilities. Where two strains have very similar ETs (Figure 5), they are likely to have diverged very recently, in terms both of individual nucleotide substitutions and of recombinational replacements. Their common clonal frame is evidence of a cellular clonality: their cell walls as well as their chromosomes are related by close physical ancestry (though the cell walls lack the distinctive features and precise replication of DNA). On the other hand, beyond a certain level of ET divergence, the clonal frames are obscured, as in Figure 4a, and so is the relationship of anything but the clonal segments themselves. Note that the K12 level II clone spans the entire "A" division of the phenogram. The minimal degree of ET similarity with phylogenetic correlations doubtless varies broadly among organisms, and with circumstances. The problem of genomic clonal mosaicism in a cladistic context has been discussed appealingly by WANNTORP (1983).

We are grateful for the participation of ARLIN STOLTZFUS in all aspects of this research. J. R. GUEST kindly alerted us to the existence of *fumB* and provided a pre-publication sequence. We thank ANNA CRANE, JODI FOX, ANNE HENNESSEY, KOMAL JAIPAUL, BLAKE SCHULTZ and JANIS VOETBERG for technical assistance and MITCHELL SOGIN for suggesting PCR. This work was supported in part by National Institutes of Health grant R01 GM 33518.

LITERATURE CITED

BACHMANN, B. J., 1990 Linkage map of *Escherichia coli* K-12, Edition 8. *Microbiol. Rev.* **54**: 130–197.

BELL, P. J., S. C. ANDREWS, M. N. SIVAK and J. R. GUEST,

- 1989 Nucleotide sequence of the FNR-regulated fumarase gene (*fumB*) of *Escherichia coli* K-12. *J. Bacteriol.* **171**: 3494-3503.
- CAIRNS, J., J. OVERBAUGH and S. MILLER, 1988 The origin of mutations. *Nature* **335**: 142-145.
- CROW, J. F., 1986 *Basic Concepts in Population, Quantitative, and Evolutionary Genetics*. Freeman, New York.
- DUBOSE, R. F., D. E. DYKHUIZEN and D. L. HARTL, 1988 Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **85**: 7036-7040.
- GYLLENSTEN, U. B., and H. A. EHRLICH, 1988 Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the *HLA-DQA* locus. *Proc. Natl. Acad. Sci. USA* **85**: 7652-7656.
- HALL, B. G., 1988 Adaptive evolution that requires multiple spontaneous mutations. I. Mutations involving an insertion sequence. *Genetics* **120**: 887-897.
- HALL, B. G., 1990 Spontaneous point mutations that occur more often when advantageous than when neutral. *Genetics* **126**: 5-16.
- INNS, M. A., K. B. MYAMBO, D. H. GELFAND and M. A. D. BROW, 1988 DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. Natl. Acad. Sci. USA* **85**: 9436-9440.
- KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking" effect revisited. *Genetics* **123**: 887-899.
- KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- LEVY, S. B., and R. V. MILLER (Editors), 1989 *Gene Transfer in the Environment*. McGraw-Hill, New York.
- MILKMAN, R., 1973 Electrophoretic variation in *Escherichia coli* strains from natural sources. *Science* **182**: 1024-1026.
- MILKMAN, R., 1985 Two elements of a unified theory of population genetics and molecular evolution, pp. 65-83 in *Population Genetics and Molecular Evolution*, edited by T. OHTA and K. AOKI. Japan Scientific Societies Press, Tokyo.
- MILKMAN, R., and I. P. CRAWFORD, 1983 Clustered third-base substitutions among wild strains of *Escherichia coli*. *Science* **221**: 378-380.
- MILKMAN, R., and A. STOLTZFUS, 1988 Molecular evolution of the *E. coli* chromosome. II. Clonal segments. *Genetics* **120**: 359-366.
- Nucleotide Sequences 1986/1987*, Volume V, 1987 Compiled by E. J. ATTENCIO *et al.* Academic Press, Orlando, Fla.
- OCHMAN, H., and R. K. SELANDER, 1984 Standard reference strains of *E. coli* from natural populations. *J. Bacteriol.* **157**: 690-693.
- POSTLE, K., and R. F. GOOD, 1985 A bidirectional rho-independent transcription terminator between the *E. coli tonB* gene and an opposing gene. *Cell* **41**: 577-585.
- SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS and H. A. EHRLICH, 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491.
- SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning: A Laboratory Manual*, Ed. 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- SAWYER, S. A., 1989 Statistical tests for detecting gene-conversion. *Mol. Biol. Evol.* **6**: 526-538.
- SAWYER, S., D. DYKHUIZEN, and D. HARTL, 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Natl. Acad. Sci. USA* **84**: 6225-6228.
- SELANDER, R. K., D. A. CAUGANT and T. S. WHITTAM, 1987 Genetic structure and variation in natural populations of *Escherichia coli*, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, edited by F. C. NEIDHARDT American Society for Microbiology, Washington, D.C.
- SILHAVY, T. J., M. L. BERMAN and L. W. ENQUIST, 1984 *Experiments With Gene Fusions*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
- STOLTZFUS, A., J. F. LESLIE and R. MILKMAN, 1988 Molecular evolution of the *E. coli* chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between *trp* and *tonB*. *Genetics* **120**: 345-358.
- TSE-DINH, Y.-C., and J. C. WANG, 1986 Complete nucleotide sequence of the *topA* gene encoding *Escherichia coli* DNA topoisomerase I. *J. Mol. Biol.* **191**: 321-331.
- WANNTORP, H.-E., 1983 Reticulated cladograms and the identification of hybrid taxa, pp. 81-88, in *Advances in Cladistics*, Vol. 2, edited by N. I. PLATNICK and V. A. FUNK. Columbia University Press, New York.
- YANOFSKY, C., T. PLATT, I. P. CRAWFORD, B. P. NICHOLS, G. E. CHRISTIE, H. HOROWITZ, M. VANCLEMPUT and A. M. WU, 1981 The complete nucleotide sequence of the tryptophan operon of *Escherichia coli*. *Nucleic Acids Res.* **9**: 6647-6668.

Communicating editor: W.-H. LI