Molecular Evolution of the *Escherichia coli* Chromosome. IV. Sequence Comparisons

Roger Milkman and Melissa McKane Bridges

Department of Biological Sciences, The University of Iowa, Iowa City, Iowa 52242–1324

Manuscript received June 11, 1992
Accepted for publication November 19, 1992

ABSTRACT

DNA sequences have been compared in a 4,400-bp region for *Escherichia coli* K12 and 36 ECOR strains. Discontinuities in degree of similarity, previously inferred, are confirmed in detail. Three *clonal frames* are described on the basis of the present local high-resolution data, as well as previous analyses of restriction fragment length polymorphism (RFLP) and of multilocus enzyme electrophoresis (MLEE) covering small regions more widely dispersed on the chromosome. These three approaches show important consistency. The data illustrate the fact that, in the limited context of intraspecific genomic sequence variation, clonality and homology are synonymous. Two estimable quantitative properties are defined: recency of common ancestry (the reciprocal of the $\log_{10}$ of the number of generations since the most recent common ancestor), and the number of nucleotide pairs over which a given recency of common ancestry applies. In principle, these parameters are measures of the *degree* and *physical extent* of homology. The small size of apparent recombinational replacements, together with the observation that they occasionally occur in discontinuous series, raises the question of whether they result from the superimposition of replacements of much larger size (as expected from an elementary interpretation of conjugation and transduction in experimental *E. coli* systems) or via an alternative mechanism. Length polymorphisms of several sorts are described.

COMMON ancestry is a relationship that can be quantified in *degree* (recency of common ancestry) and in the *extent* of the compared objects over which this degree is uniform. The chromosome of *Escherichia coli* is a good illustration of this statement. First, it is one dimensional, so that the extents of homologs are simply expressed. Second, it can be characterized in exhaustive detail by nucleotide sequencing. Third, divergence time from the most recent common ancestor can in principle be estimated from a sequence difference. Fourth, the comparison of corresponding chromosomal regions among some 37 strains of *E. coli* reveals clear spatial discontinuities in the degrees of difference; the differences appear to be neutral for the most part and so are suitable for the estimation of divergence times (MILKMAN and STOLTZFUS 1988; MILKMAN and BRIDGES 1990). Thus, we propose to define recency of common ancestry as the reciprocal of the $\log_{10}$ of the divergence time in generations. For example, if a 1% sequence difference in translated DNA in *E. coli* leads to the estimate of 50,000,000 generations of divergence time, then the recency of common ancestry estimated from this sequence difference is 0.13.

The intraspecific nucleotide sequence differences studied to date in *E. coli* (MILKMAN and BRIDGES 1990; BISERČIĆ *et al.* 1991; DYKHUIZEN and GREEN 1986, 1991; DYKUISZEN and HARTL 1988; NELSON and SELANDER 1992; NELSON, WHITTHAM and SELANDER 1991) center on the common laboratory reference strain K12 and the 72 ECOR strains of recent natural origin (OCHMAN and SELANDER 1984). These studies indicate that, with few exceptions, the chromosomal regions compared are similar enough for their recency of common ancestry, as defined above, to be estimated: they have not diverged to the level of neutral substitutional equilibrium. And so, in the terms of the discontinuous *segmental clonality* previously described (MILKMAN and STOLTZFUS 1988), corresponding stretches of DNA were assigned to hierarchical clonal levels (MILKMAN and BRIDGES 1990), which reflect their observed similarity and inferred recency of common ancestry. In this sense, a low clonal level reflects a high degree of homology.

Sequence types, clonal levels and clonal frames: The term *sequence type* has a single quantitative definition. The comparative sequence data reported here make desirable a small change in this definition. Sequence types are thus redefined here to differ from one another by at least 1% of their nucleotides, as opposed to the previous 0.5% (MILKMAN and BRIDGES 1990). Members of a Level II clone now share a single sequence type; a Level III clone can include more than one sequence type; and a Level IV clone an even broader range of sequence types. Finally, in a given genome, a predominant sequence type in which other sequence types are embedded is referred to as a *(Level II) clonal frame.* A Level III clonal frame would be

less uniform (MILKMAN and BRIDGES 1990), comprising more than one sequence type.

We would now like to amplify this subject in terms of a table of comparison of a 4,400-bp stretch of DNA beginning beyond the end of the *trp* operon and including *tonB* as well as a number of open reading frames that have now been expressed (STOLTZFUS, LESLIE, and MILKMAN 1988; MILKMAN and BRIDGES 1990; VANBOGELEN *et al.* 1992). Thirty-seven strains are compared completely or with minor omissions; fragments of three others are also of some interest. The data are presented in Figure 1, and the discontinuities in degree of similarity are summarized in Figure 2. The symbols in Figure 2 represent sequence types.

MATERIALS AND METHODS

The new sequence data were obtained by standard dideoxy methods (SAMBROOK, FRITSCH and MANIATIS 1989) of two types, both designed to exploit PCR amplification of specific 0.7–1.6 kb stretches of genomic DNA. Our main method has been to use gel purified ds DNA made by PCR amplification as a template for ss DNA production in a 15-cycle (typically) amplification using one primer only. The ss DNA is acrylamide-gel purified and sequenced using the opposite amplification primer or an internal primer. Approximately 250 ng (0.5 pmol) of ss DNA is sequenced with the Sequenase T7 DNA Polymerase Kit (U. S. Biochemical Corp.) according to the manufacturer's protocols, except that labeling is carried out at 0° for 2 min and extension/termination at 41° for 3 min. Recently, we also have been sequencing the purified ds DNA fragment (50 fmol) with the ΔTaq Cycle-Sequencing Kit (U.S. Biochemical Corp.), using the manufacturer's "Cycled Labeling Step Protocol" except that 0.5 μl 0.1 M dithiothreitol is added to the samples before heating/loading, to eliminate ^{35}S degradation background, which we had observed. In both methods described, [α-^{35}S]dATP, -dGTP, -dCTP and -TTP (NEN Research Products) were used in the labeling step. ^{35}S-labeled DNA ladders were produced by constant-power electrophoresis in standard 4% or 6% (or buffer-gradient, 8%) acrylamide gels (SAMBROOK, FRITSCH and MANIATIS 1989); films were exposed for one to several days; no automatic procedures (reactions, reading) were used. In general, only one strand has been sequenced, since the comparison of many strains affords a considerable check on accuracy, but additional sequencing has been used to resolve some ambiguities. The standard reference sequences had of course been done on both strands.

The base position numbers we use are one less than the distance from the origin of the Genbank/EMBL sequence of the *trp* operon in K12 (ECOTGP, Accession Number J01714), and 7,337 greater than the Genbank/EMBL sequence ECOTRTOI (Accession Number X13583), which begins at ECOTGP position 7,339. Beyond this, our position 10,979 corresponds to the end (position 1,697) of the oppositely running Genbank/EMBL sequence ECOTONB (Accession Number K00431), whose position 189 corresponds to the end (position 12,487) of the region described here. Finally, the numbers correspond directly to those in STOLTZFUS, LESLIE and MILKMAN (1988) and in MILKMAN and BRIDGES (1990). Previous calibration to Genbank/EMBL numbering for the *trp* operon, which has changed, is now obsolete.

The 4,400-bp region described (ALO) is one part, essentially completed, of a 12-kb region being sequenced comparatively. [An adjacent 4,400-bp stretch, CAF, which runs from the middle of the *trp* operon to the present region; CAF's sequencing is about 95% complete. Finally, 3,145 bp of DNA, including *attB* (the attachment site of φ80), continues beyond *tonB*. It has been sequenced for K12, and several other strains. This last region is called BUG.]

Abbreviations and specialized terms.

Chromosomal regions: ALO = the approximately 4,400-bp region from positions 8,089–12,490 as defined above. It includes the PCR fragments AL, LK and ON (MILKMAN and BRIDGES 1990). CAF = an adjacent 4,400-bp stretch from positions 3,630–8,100 covering the PCR fragments CB, BA and FB.

Analytic techniques: RFLP = restriction fragment length polymorphism. MLEE = multilocus enzyme electrophoresis.

Clonal properties: Clonal segment = stretch of DNA sharing recent common ancestry with one or more stretches in a corresponding position. Clonal frame: when most of the DNA in a chromosome belongs to a single designated clone, this DNA constitutes a clonal frame. Clonal level: hierarchical term defined in principle by recency of common ancestry (thus clones can be nested), but in practice by degree of sequence uniformity.

Miscellaneous: Atlas = A set of lambdoid phages, complete or incomplete, inserted between positions 8,888 and 8,889 (as defined here) in *E. coli*. ECOR designates any of the 72 *E. coli* Reference strains of recent wild origin. Indel = observed length difference attributable to *in*sertion or *del*etion. PCR = polymerase chain reaction used to amplify a specific stretch of DNA from a genome or other source.

RESULTS

The body of data presented is best seen by referring to Figures 1 and 2 in conjunction with the text. Figure 1 differs from most published sequence comparisons in presenting the polymorphic sites vertically. The comparisons in this table are read horizontally. The sequence described is that of the strand continuous with the nontranscribed strand of the *trp* genes. The first four columns describe the comparisons. At the very left, the nature of the difference(s) is stated: transversion (#), C-T transition (iC), A-G transition (iA), mixed (M), deletion (Δ), insertion (>). In the next column is the *phase* (remainder after dividing the overall position number by 3). This is used to determine the *codon position* of the site (1, 2 or 3) in the respective translated regions, as shown in column 3; NT indicates that the base is not translated. Note that when the translated DNA sequence illustrated is that of the transcribed strand (*i.e.*, where transcription proceeds in the direction opposite to that of *trp*), the phase number is preceded by "R"; the transitions shown in column 1 still refer to the nontranscribed strand, which is not illustrated. The fourth column lists the *overall position of the site* numbered as described in MATERIALS AND METHODS. The next column gives the *most frequent ("consensus") base* for each site listed. The respective bases for K12 are in the column headed by K; hyphens indicate identity to the consen-

POLYMORPHIC SITES IN THE ALO REGION (8089 - 12490)

Strain Numbers

Ch	Ph	Cp	N	Con	K	"RED" 08 19 04	"PURPLE" 29 45 58 28 68 72 70 27	"GREEN" 66 40 64 62 61 65 60 52 54 51	31 37	35	49 50	46 48	Amino acid changes
#	1	3	8095	C	-	------	--------	------ -	--	-	AA	---	
iC	2	1	8105	C	-	------	--------	TTTTTT -	--	-	--	T-T	
iA	1	3	8107	G	-	------	--------	------ -	--	A	--	---	
#	1	3	8119	C	-	------	--------	------ -	--	-	--	--A	
#	2	1	8120	C	A	AAAAAA	--A-----	------ -	--	-	--	---	l→i
#	0	2	8121	T	-	------	--------	------ -	G-	-	--	---	l→r
iC		NT	8167	C	T	TTTTTT	--TTTT--	------ -	--	-	--	T--	
iC	1	3	8248	C	-	------	--------	------ -	TT	-	--	---	
iC	2	1	8261	C	-	------	--------	---T-- -	--	-	--	---	q→STOP
#	1	3	8272	T	-	------	--------	AA---- -	--	-	--	---	
iA	1	3	8281	A	-	------	--------	--G-GG -	--	-	--	---	
iC	1	3	8290	C	-	------	--------	------ -	--	-	--	-T-	
iC	1	3	8293	T	C	CCCCCC	CCCC--CC	------ C	CC	-	--	C--	
iC	1	3	8320	C	-	------	--------	------ -	-T	-	--	---	
iC	1	3	8356	C	-	------	--------	----TT -	--	-	--	---	
iA	2	1	8363	G	-	------	--------	------ -	--	-	--	A--	v→i
iC	2	1	8369	C	-	------	-------T	------ -	--	-	--	---	
iA	1	3	8374	A	-	------	--------	------ -	--	-	-G	---	
iC	2	1	8387	C	-	------	--------	------ -	T-	-	--	---	r→c
iA	1	3	8434	G	-	------	A-------	----AA -	--	-	--	---	
iA	1	3	8440	G	-	------	--------	AAAAAA -	--	-	--	---	
iA	1	3	8446	G	-	------	--------	------ -	--	-	AA	---	
iC	1	3	8467	C	-	------	--------	------ -	--	T	--	---	
iC	1	3	8485	C	-	------	---TTT--	------ -	--	-	--	---	
iA	0	2	8490	G	-	--A---	--------	------ -	--	-	--	---	g→d
iC	1	3	8536	C	-	------	--------	TTTTTT -	--	$	TT	---	
#	1	3	8542	A	-	------	--------	------ -	T-	$	--	---	
#	1	3	8545	A	-	------	--------	CCCCCC -	--	$	CC	---	
iC	0	2	8577	C	-	------	---TTT--	------ -	--	$	--	---	a→v
iC	1	3	8578	T	-	------	---CCC-C	---CCC -	CC	$	CC	---	
iA	1	3	8587	A	-	------	---GGG-G	---GGG -	GG	$	GG	---	
iC	0	2	8589	C	-	------	---TTT-T	---TTT -	TT	$	TT	---	t→i
iC	1	3	8590	C	-	------	--------	---TTT -	--	$	TT	---	
#	1	3	8611	G	-	------	--------	---TTT -	--	$	--	---	e→d
iA	1	3	8617	G	-	------	--------	------ -	--	$	--	-AA	
#	2	1	8618	C	-	------	--------	------ -	--	$	--	-GG	q→e
#	1	3	8620	A	-	------	--------	CCCCCC -	--	$	CC	---	q→h
iA	1	3	8635	G	-	------	--------	AAAAAA -	--	$	AA	---	
#	2	1	8642	C	A	AAAAAA	--------	------ -	--	$	--	---	h→n
iC	1	3	8668	T	-	------	--------	CCCCCC -	--	$	CC	---	
iC	1	3	8677	C	-	------	---TTT-T	------ -	--	$	--	---	
iC	1	3	8689	T	-	------	CC----C-	------ C	--	$	--	---	
iA	1	3	8695	A	-	------	--------	------ -	--	$	GG	---	
#		NT	8714	C	-	------	AA----A-	------ A	--	$	--	---	
iA		NT	8721	A	-	------	GG----G-	------ G	GG	$	--	-G-	
Δ		NT	8744	T	-	------	--------	ΔΔΔΔΔΔ -	--	$	ΔΔ	---	
#		NT	8755	C	-	------	--------	------ -	--	$	--	G--	
iA		NT	8756	G	-	------	--------	AAAAAA -	--	$	JA	---	
#		NT	8762	G	-	------	TT----T-	------ T	--	$	--	---	
iC		NT	8765	C	-	------	TT----T-	------ T	--	$	--	---	
iA		NT	8767	G	A	-|--A-	--------	------ -	--	$	--	---	|:1-, 8:A
iC	1R	3	8812	G	-	------	--------	------ -	--	$	--	--A	
iC	1R	3	8827	G	-	-----	--------	------ -	AA	$	--	---	
#	1R	3	8839	A	-	-----	--------	----TT -	--	$	--	---	
iA	1R	3	8845	C	-	-----	TT-TTTT-	------ -	--	$	--	---	
iA	1R	3	8848	C	-	-----	--------	TTTTTT T	TT	$	TT	-T-	
iC	1R	3	8860	G	-	-----	--------	AAAAAA A	-	$	AA	- -	
#	1R	3	8863	G	-	-----	--------	--CC-- C	-	$	--	- -	
iC	1R	3	8869	G	-	-----	--------	AAAAAA A	-	$	--	- -	
#	1R	3	8884	C	-	-----	------G-	GGGGGG -	--	$	GG	G -	
>			8888-8889		-	-----@	-|------	|---@@@@	@-	@	@@	- -	|:51@,54-,56-; 70@,71@,72-
iC	1R	3	8893	G	-	------	--------	------A-	-A	-	-	---	
#	1R	3	8896	A	-	------	--------	------C-	--	-	-	---	
#	0R	1	8898	C	-	------	--------	------G-	--	-	-	---	v→l
iC	1R	3	8914	A	-	------	--------	------G-	--	-	-	---	
#	1R	3	8917	C	-	------	--------	------G-	--	-	-	---	
#	1R	3	8926	A	-	------	--------	------C-	--	-	-	---	
#	0R	1	8928	C	-	------	--------	------G-	--	-	-	---	v→l
#	1R	3	8929	C	-	------	--------	------A-	--	-	-	---	
iA	1R	3	8932	C	-	------	--------	------T-	--	-	-	---	
iA	1R	3	8935	C	-	------	--------	------T-	--	-	-	---	
#	1R	3	8941	G	-	------	--------	------C-	--	-	-	---	
#	1R	3	8944	A	-	------	--------	------C-	--	-	-	---	
iC	2R	2	8945	G	-	------	--------	------A-	--	-	-	---	a→v
#	1R	3	8947	T	-	------	--------	------G-	--	-	-	---	
MC	1R	3	8953	G	-	------	--------	----AAT-	--	-	-	---	
iC	1R	3	8956	A	-	------	--------	GGGGGG-G	--	-	G	-G-	
iA	1R	3	8959	T	-	------	--------	------C-	--	-	-	---	
iC	1R	3	8965	A	-	------	-----G--	--------	--	-	-	---	
iC	1R	3	8968	G	-	------	--------	------A-	--	-	-	---	
iC	1R	3	8971	A	-	------	--------	------G-	--	-	-	---	
iC	1R	3	8974	G	-	------	--------	------A-	--	-	-	---	
#	1R	3	8977	A	-	------	--------	------T-	--	-	-	---	
iC	0R	1	8979	G	-	------	--------	------A-	--	-	-	---	
#	1R	3	8980	C	-	------	--------	------A-	--	-	-	---	
MA	1R	3	8983	C	T	TTTTTT	--TTT-T-	------A-	--	-	-	---	
iC	2R	2	8984	G	-	------	--------	------A-	--	-	-	---	a→l*
#	0R	1	8985	C	-	------	--------	------G-	--	-	-	---	a→l*
#	1R	3	8986	C	-	------	--------	------A-	--	-	-	---	
iA	2R	2	8987	T	-	------	--------	------C-	--	-	-	---	e→g
iA	0R	1	8991	T	-	------	--------	------C-	--	-	-	---	k→e
iC	1R	3	8992	G	-	------	--------	------A-	--	-	-	---	
iC	1R	3	8995	A	-	------	--------	------G-	--	-	-	---	h→t*
#	2R	2	8996	T	-	------	--------	------G-	--	-	-	---	h→t*
#	0R	1	8997	G	-	------	--------	------T-	--	-	-	---	h→t*
iC	1R	3	8998	A	-	------	--------	------G-	--	-	-	---	
iC	1R	3	9001	G	-	------	--------	------A-	--	-	-	---	
iC	1R	3	9004	A	-	------	--------	------G-	--	-	-	---	
#	1R	3	9007	T	-	------	--------	------G-	--	-	-	---	g→d*
iA	2R	2	9008	C	-	------	--------	------T-	--	-	-	---	g→d*
iA	0R	1	9015	C	-	------	--------	------T-	--	-	-	---	d→n
iC	1R	3	9019	G	-	------	--------	------A-	--	-	-	---	
MA	1R	3	9040	T	-	------	GGGGGGGG	------C-	--	-	-	---	
#	1R	3	9043	C	-	------	--------	AAAAAAAA	--	-	A	-A-	
iA	0R	1	9048	C	-	------	--------	------T-	--	-	-	---	v→i
iC	1R	3	9049	G	-	------	--------	--------	-A	-	-	---	
iC	1R	3	9064	G	-	------	--------	------A-	--	-	-	---	
iA	0R	1	9069	T	-	------	--------	------C-	--	-	-	---	s→g
#	1R	3	9070	G	-	------	--------	------T-	--	-	-	---	
#	0R	1	9072	C	-	------	--------	------A-	--	-	-	---	a→s

```
iC 1R 3   9076  A  -  ------  --------  ------G-  --  -  -  ---
iA 1R 3   9088  C  -  ------  --------  ------T-  --  -  -  ---
iA 1R 3   9091  C  -  ------  --------  ------T-  --  -  -  ---
iA 1R 3   9106  C  T  TTTTTT  --------  --------  -T  -  -  ---
iA 1R 3   9109  C  -  ------  --------  ------T-  --  -  -  ---
iC 0R 1   9111  G  -  ------  --------  ------A-  --  -  -  ---
iC 1R 3   9115  A  -  ------  GGGGGG--  ------G-  --  -  -  ---
iC 1R 3   9118  A  -  ------  --------  ------G-  --  -  -  ---
iC 1R 3   9121  G  -  ------  --------  ------A-  --  -  -  ---
#  1R 3   9133  G  -  ------  --------  ------C-  --  -  -  ---
#  1R 3   9139  C  -  ------  --------  ------A-  --  -  -  ---
iA 0R 1   9141  C  -  ------  --------  ------T-  --  -  -  ---  a→t
#  1R 3   9142  C  -  ------  --------  ------G-  --  -  -  ---
#  2R 2   9143  C  -  ------  --------  ------G-  --  -  -  ---  r→a*
#  0R 1   9144  G  -  ------  --------  ------C-  --  -  -  ---  r→a*
#  1R 3   9145  G  -  ------  --------  ------T-  --  -  -  ---
#  0R 1   9150  C  -  ------  --------  ------A-  --  -  -  ---  g→c
iC 1R 3   9151  G  -  ------  --------  ------A-  --  -  -  ---
#  2R 2   9152  A  -  ------  --------  ------C-  --  -  -  ---  i→s
iC 1R 3   9160  G  -  ------  --------  ------A-  --  -  -  ---
iC 1R 3   9163  G  -  ------  --------  ------A-  --  -  -  ---
#  0R 1   9171  T  -  ------  --------  ------A-  --  -  -  ---  t→s
iA 1R 3   9172  C  -  ------  --------  TTTTTTTT  --  -  -  ---
#  1R 3   9175  T  -  ------  --------  --------  --  A  -  ---
iA 1R 3   9181  T  -  ------  --------  --------  --  -  C  CCC
iA 1R 3   9184  T  -  -----C  --------  --------  --  -  -  ---
iC 1R 3   9199  G  -  ------  --------  --------  --  -  A  AAA
#  1R 3   9223  C  A  AAAAAA  --------  --------  --  -  -  ---
iC 0R 1   9225  G  -  ------  AAAAAAA-  --------  --  -  -  ---
#  1R 3   9235  A  C  CCCCCC  --------  --------  --  -  C  CCC
iC 0R 1   9264  G  -  ------  --------  --AAAA-A  --  -  -  ---
#  1R 3   9295  T  -  ------  --------  --------  -A  -  -  ---
#  1R 3   9313  A  -  ------  --------  --------  --  C  C  CCC
#  1R 3   9334  G  -  ------  --------  --------  -T  -  -  ---
#  1R 3   9337  T  -  ------  --------  GGGGGGGG  --  -  -  ---
#  0R 1   9339  C  -  ------  --------  -------A  --  -  -  ---  a→s
iA 1R 3   9346  C  -  ------  --------  ----TT--  --  -  -  ---
#  2R 2   9350  A  -  ------  --------  --------  --  T  T  TTT  f→y
iC 1R 3   9364  G  -  ------  AAAAAAAA  --------  --  -  -  ---
#  1R 3   9370  A  -  ------  --------  --------  -T  -  -  ---
iA 0R 1   9375  T  -  ------  --------  CCCCCCCC  --  -  -  ---  t→a
iC    NT  9452  T  -  ------  --------  CCCCCCCC  --  C  -  ---
iA    NT  9460  A  -  ------  --------  GGGGGGGG  --  -  -  ---
iC    NT  9479  T  -  ------  --------  CCCCCCCC  --  C  -  ---
iA    NT  9482  A  -  ------  --------  --------  --  -  -  G-G
iC    NT  9491  C  -  ------  --------  --------  --  T  -  ---
iA    NT  9496  G  -  -----A  --------  --------  --  -  -  ---
iA    NT  9497  A  -  ------  --------  GGGGGGGG  --  -  -  ---
iA    NT  9504  A  -  ------  --------  --------  --  G  -  ---
iC    NT  9506  C  -  -T----  --------  --------  --  -  -  ---
#     NT  9507  C  -  ------  AAAAAAAA  AAA-----  A-  -  A  AAA
iC    NT  9527  T  -  ------  C-------  --------  --  -  -  ---
iA    NT  9536  G  -  ------  --------  --A-----  --  -  -  ---
iC    NT  9539  T  -  ------  --------  --------  --  C  -  ---
#     NT  9542  A  -  ------  --------  --------  --  T  -  ---
iA    NT  9543  G  -  ------  --------  AAAAAAAA  --  -  -  A--
iC    NT  9556  T  -  ------  --------  --------  CC  -  -  ---
iC    NT  9573  T  -  ------  --------  --------  --  C  -  ---
iC    NT  9574  T  -  ------  --------  --------  --  C  -  ---
>     NT >9575     -  ------  --------  --T-----  --  -  -  ---
#     NT  9577  A  -  ------  --------  --TTTTTT  --  T  -  ---
Å     NT  9581  T  -  ------  --------  ----ÅÅ--  --  -  -  ---
MA    NT  9586  G  -  ------  TTTTTTTT  -TJ-----  --  A  -  ---
iC    NT  9596  T  -  ------  --------  --------  --  C  -  ---
```

```
iC?   NT  9602  C  -  ?-----  --------  --------  --  -  -  ---  ?: T?
iC    NT  9604  C  -  ------  --------  --------  -T  -  -  ---
MAC   NT  9606  C  G  GGGGGG  AAAAAAAA  --------  TT  -  ?  -TA  ?: T/C?
iC    NT  9607  T  -  ------  --------  --------  --  -  -  --C
#     NT  9616  G  -  ------  --------  ----TT--  --  -  -  ---
MC    NT  9618  C  -  ------  --------  AAAAAAAA  --  T  -  A--
iA    NT  9622  G  A  AAAAAA  --------  --------  --  -  -  --A
iA    NT  9626  G  -  ------  --------  AAA-----  --  -  -  ---
iC    NT  9631  C  T  TTTTTT  --------  --------  --  -  -  ---
iA    NT  9632  G  -  ------  --------  --J-----  --  A  -  ---
MC    NT  9633  C  -  ------  --------  --------  --  G  -  --T
#     NT  9636  G  -  ------  --------  --------  --  T  -  ---
iA    NT  9640  G  -  ------  --------  --------  --  A  -  ---
#     NT  9643  A  -  ------  --------  --------  --  C  -  ---
iC    NT  9652  C  -  ------  --T-----  --------  --  -  -  ---
                        [  GAP: 9672-9708]
iC 2  3   9809  C  -  ------  --------  ---TTTTT-  --  -  -  ---
iA 0  1   9840  G  -  ------  --------  AA------  --  -  -  ---  v→i
iC 1  2   9901  C  -  ------  --------  ---------  -T  -  -  ---  a→v
#  2  3   9911  G  -  ------  --------  ---------  --  T  -  ---
iA 0  1   9927  G  -  ------  -------A  ---------  --  -  -  ---  v→m
iA 2  3   9929  G  -  ----AA  --------  AAAAAAAAA  --  -  -  ---
iC?2  3   9932  C  -  ------  --?-----  ---------  --  -  -  ---  ?: 69T?
iA 0  1   9957  G  -  ----|-  --------  ---------  --  -  -  ---  d→n; |: 15-,16A
#  2  3   9980  G  -  ------  --------  ---------  --  T  -  ---
iA 2  3  10001  G  -  ------  --------  ---------  --  -  -  -A-
iA 2  3  10004  G  -  ----AA  --------  ---------  --  -  -  ---
#  2  3  10016  G  -  ----TT  --------  ---------  --  -  -  ---
iA 2  3  10034  A  -  ----GG  --------  ---------  --  -  -  ---
iA 2  3  10058  A  -  ------  --------  GGGGGGGGG  --  -  -  ---
#  2  3  10097  A  -  -|----  --------  ---------  --  -  -  ---  |: 1T; 8-
iA 2  3  10109  G  -  ------  --------  ------A--  --  -  -  ---
#  2  3  10112  G  -  ------  --------  ---------  --  T  -  T--
iA 3  1  10119  A  -  ------  --------  --G------  --  -  -  ---  i→v
iC 2  3  10124  T  -  ------  --------  --C------  --  -  -  ---
iA 2  3  10142  G  -  ------  --------  ----AA---  --  -  -  ---
#  2  3  10145  A  -  ----CC  CCC?CCCC  ---------  --  -  C  ---
iA 2  3  10202  G  -  ------  --------  ---------  -A  -  -  ---
iC 2  3  10229  C  -  ------  --------  ---------  -T  -  -  ---
#  2  3  10232  T  -  ------  --------  ---------  A-  -  -  ---
iA 2  3  10235  A  -  ----GG  --------  ---------  --  -  -  ---
iC 0  1  10236  C  -  ------  --------  ---------  --  -  T  ---
iC 2  3  10241  C  -  ------  --------  TTTTTTTTT  --  T  -  TT-
iA 0  1  10245  G  -  -----A  --------  ---------  --  -  -  ---
iC 0  1  10257  C  T  TTT---  --------  ---------  --  T  T  -?-
iA 0  1  10278  G  -  ----AA  AAAAAAAA  ---------  -A  A  A  ---  v→i
#  2  3  10292  G  -  ------  --------  ---------  C-  -  -  ---  m→i
#  2  3  10295  T  -  ------  --------  ---------  --  -  A  ---
iC 2  3  10301  T  -  ------  --------  ---------  --  -  C  ---
#  2  3  10307  G  -  ------  --------  ---------  --  -  C  ---
MC 2  3  10313  T  -  ------  --------  AAAAAAAAA  --  A  C  -A-
iA 2  3  10325  G  -  ------  --------  ------A--  --  -  -  ---
#  2  3  10343  A  -  ------  --------  --T------  --  -  -  ---
iA 2  3  10367  A  -  ------  --------  ---------  --  -  G  ---
iC 0  1  10377  C  -  ------  --------  --T------  --  -  -  ---
#  2  3  10427  A  -  ------  ------T-  ---------  --  -  -  ---
#  2  3  10457  A  -  ------  TTTTTTTT  ---------  --  -  -  ---
iA 0  1  10461  A  -  ------  --------  GGGGGGGGG  -G  G  G  GGG  i→v
iC 0  1  10464  T  -  ------  --------  --C------  --  -  C  CCC
iC 2  3  10484  C  -  ------  --------  ---------  -T  -  -  ---
iC    N  10526  C  -  ------  --------  -------T-  --  -  -  ---
iA    N  10533  G  -  ------  --------  --------A  --  -  -  ---
iC 1  3  10585  T  -  ------  --------  ---------  --  C  -  ---
iA 1  3  10591  G  -  ------  --------  ---------  A-  -  -  ---
```

```
iC 1  3  10606  T  -  ------  ---C----  ---------  --  -  -  ---
iA 1  3  10609  G  -  ----AA  --------  ---------  --  -  -  ---
#  1  3  10624  G  -  ------  --------  CC-CCCCC-  --  C  -  ---
#  1  3  10651  C  -  ------  --------  AAAAAAAA-  -A  -  A  - -
iC 0  2  10701  T  -  -----C  --------  ---------  --  C  -  - -  f→s
iA 1  3  10729  G  -  ------  --------  AA-A--AA-  --  A  -  - -
#  1  3  10732  G  -  ------  --------  --C-------  --  -  -  - C
iC 1  3  10738  C  -  ------  --------  ---T-----  --  -  -  - -
iC 1  3  10741  C  -  ------  --------  --------T  --  -  -  - -
#  1  3  10759  T  -  ------  --------  ---------  -A  -  -  ---
#  1  3  10774  A  T  ------  --------  TTTTTTTTT  -T  T  -  --T
iC 1  3  10786  T  -  ------  --------  CCCCCCCCC  --  C  -  -CC
iA 1  3  10789  A  -  ------  -G---G-G  GGGGGGGGG  -G  G  -  GGG
#  1  3  10798  A  -  ------  --------  TTTTTTTTT  --  T  -  -TT  e→d
iC 1  3  10813  C  -  ------  --------  ---------T  --  T  -  ---
M# 0  2  10833  C  -  ------  ---G----  ---------  -A  -  -  ---  p→r,q
iC 0  2  10866  C  -  ------  ----T---  ---------  --  -  -  ---  t→m
MA 1  3  10867  G  -  ------  --------  CCACCCCCC  --  C  -  ---
#  2  1  10868  C  -  ------  --------  --G-------  --  -  -  ---  l→v
iC 2  1  10877  C  -  ------  --------  TT-TTTTTT  --  T  -  ---  p→s
iA 1  3  10879  G  -  ------  --------  ---------  --  -  -  A--
iC 0  2  10887  C  -  ------  ----T---  ---------  --  -  -  ---  s→l
iA 1  3  10894  G  -  ------  AAAA-A-A  ---------  --  -  -  ---
iC 1  3  10903  C  -  ------  TTTT-T--  ---------  --  -  -  ---
iC 1  3  10930  T  -  ------  --------  CC-C---CC  --  C  -  ---
#  1  3  10933  G  -  ------  --------  ---------  --  -  -  --T
#  0  2  10956  G  -  --?---  --------  ---------  --  -  -  ---  21: T? w→l?
iC 1  3  11005  C  -  ---A--  --------  ------ --  --  -  -  -T-
iC     NT 11093  T  -  ---A--  --------  CC----- --  --  -  -  ---
iC     NT 11095  C  -  ---A--  --------  ------ --  T-  -  -  ---
>      NT 11095>    -  ---A--  --------  ----AA --  --  -  -  ---
iA     NT 11098  G  -  ---A-A  --------  ------ --  --  -  -  ---
  JAM  NT 11101-47- -  ---A--  --------  JJJ--- --  --  -  -  ---
iC 0  3  11253  T  -  ---A--  --------  ------ -C  --  -  -  ---
iA 0  3  11319  G  A  AAAAAA  ----A-A-  ------ --  A-  -  -  ---
iC 0  3  11328  T  -  ------  --------  ------ --  --  C  -  ---
iC 0  3  11334  C  -  ------  --------  TTT-TT --  --  -  -  ---
iC 0  3  11340  C  -  ------  --------  TTTTTT TT  --  -  -  ---
iC 0  3  11349  T  -  ------  --------  CCCCCC CC  --  C  -  CC-
iA 1  1  11356  G  -  ------  --------  ------ --  --  A  -  ---  v→i
iA 0  3  11379  G  -  ------  --------  ------ --  --  A  -  A--
iA 0  3  11442  G  -  ------  --------  AAAAAA AA  --  -  -  ---
iC 0  3  11445  C  -  ------  --------  ------ --  --  -  -  T--
MA 0  3  11478  A  -  ------  --------  GGGGGG GG  --  C  C  CC-
#  0  3  11490  A  -  ------  --------  ------ --  --  C  -  --C
iC 0  3  11493  C  T  TTT---  --------  ------ --  --  -  -  ---
#  0  3  11502  C  -  ----GG  --------  ------ --  --  -  -  ---
iA 0  3  11511  G  -  ------  --------  AAAAAA AA  --  -  -  ---
iC 0  3  11547  T  -  ------  --------  CCCCCC CC  --  -  -  ---
iC 0  3  11571  T  -  ------  --------  CCCCCC -C  --  -  -  ---
iA     NT 11583  A  G  GGGGGG  -G-----G  ------ --  G-  -  -  G--
iA     NT 11586  A  -  ------  G-GGG---  ------ --  --  -  -  ---
iA     NT 11590  G  -  ------  ----A---  ------ --  --  -  -  ---
#      NT 11594  C  -  ------  --------  ---?-- --  --  -  -  ---  ?: 65A?
#      NT 11595  C  -  ------  ----G---  ------ --  --  -  -  ---
#      NT 11616  C  -  ----AA  --------  AAAAAA AA  --  A  -  AA-
iC 1R 3  11635  G  -  ------  --------  AAAAAA AA  --  A  -  ---
iA 1R 3  11653  C  -  ------  -----T--  ------ --  --  -  -  ---
iC 2R 2  11690  G  -  --A---  --------  ------ --  --  -  -  ---  p→l
iA 1R 3  11725  C  -  ------  --------  ------ -T  --  T  -  ---
iC 1R 3  11740  G  -  ------  --------  ------ --  --  -  -  A--
iA 1R 3  11743  C  -  ------  --------  TTTTTT --  --  -  -  ---
iA 1R 3  11749  C  -  ------  --------  ------ --  -T  -  -  ---
iC 1R 3  11752  G  -  ------  --------  ------ --  -A  -  -  ---

iC 1R 3  11758  G  -  ------  --------  ------ --  A-  -  -  ---
iC 1R 3  11761  G  -  ------  ---A----  ------ --  --  -  -  ---
iA 1R 3  11767  T  -  ------  --------  ------ --  -C  -  C  ---
iA 1R 3  11788  C  -  ------  --------  TTTTTT TT  --  T  -  ---
iC 1R 3  11794  G  A  AAA---  --------  ------ --  --  -  -  ---
iC 1R 3  11797  G  -  ------  --------  AAAAAA AA  -A  A  A  ---
iA 1R 3  11809  T  C  ------  --------  ------ --  --  -  -  ---
iA 0R 1  11814  C  -  ------  --------  --T--- --  --  -  ---  v→i
#  1R 3  11839  T  -  ------  ----A---  ------ --  --  -  ---
iA 1R 3  11842  T  -  ------  --------  ------ --  -C  -  ---
iA 1R 3  11917  T  -  ------  --------  ------ --  -C  -  -  ---
#  0R 1  11922  C  -  ------  --------  ------ --  --  -  -  ?--  ?: A?
iA 1R 3  11926  T  -  ------  --------  ------ --  --  -  -  ?-C  ?: C?
iA 1R 3  11929  C  T  TTTTTT  --------  ------ --  T-  -  -  ---
#  1R 3  11941  C  -  ------  -------G  ------ --  --  -  -  ---
iC 2R 2  11942  G  A  AAA---  AAAAAA--  ------ --  --  -  -  ---  p→l
iA 1R 3  11947  T  -  ------  ---CCCCC  CCCCCC CC  --  C  -  ---
#  1R 3  11950  G  C  CCCCCC  --------  ------ --  CC  -  C  CCC
#  2R 2  11978  G  -  ------  --------  ---T-- --  --  -  -  ---  p→q
iA 0R 1  12000  C  -  ------  ---TTTT-  ------ --  --  -  -  ---  v→i
#  2R 2  12011  G  -  ------  --------  ------ --  T-  -  -  ---  p→q
iA 1R 3  12016  C  -  ------  --------  ------ --  -T  -  -  --T
Δ        12020-2  -  ------  --------  ΔΔΔΔΔΔ ΔΔ  --  Δ  -  ---  Δq
#  1R 3  12052  A  -  ------  --------  ------ -T  --  -  -  ---
>        12057>    -  ------  --------  ------ --  --  -  >  ---  >CGGCTT=k,p
iC 1R 3  12076  G  -  ------  --------  ------ --  --  A  -  ---
iC 2R 2  12080  A  -  ------  GG?-----  ------ --  --  -  -  ---  v-a (69G?)
iA 1R 3  12082  C  -  ---T--  --------  ------ --  --  -  -  ---
#  1R 3  12085  T  -  ------  --------  AAAAAA AA  -A  A  -  ---
iA 1R 3  12097  C  -  ------  TTT-----  ------ --  --  -  -  ---
iC 1R 3  12103  G  -  ------  --------  AAAAAA AA  --  A  -  ---
iA 1R 3  12130  C  -  ------  --------  ------ --  --  -  -  -T-
iA 0R 1  12132  C  -  ------  --------  ------ --  -T  -  -  ---  e→k
MA 1R 3  12151  C  -  ------  --------  -T-GTT GT  --  T  -  ---
iA 1R 3  12154  C  -  ------  --------  ---T-- TT  --  T  -  ---
iC 1R 3  12175  G  -  ------  --------  AAA-AA -A  --  -  -  ---
#  1R 3  12181  A  -  ------  --------  ---C-- CC  --  C  C  CCC
#  1R 3  12184  A  -  ------  --------  ---T-- TT  --  T  -  ---
iA 0R 1  12189  C  T  TTTTTT  --------  ------ --  --  -  -  ---  v→t
MA 1R 3  12208  C  -  ------  TTTTTTTT  ---G-- GG  -T  -  -  TTT
iC 1R 3  12217  A  -  ------  --------  ---G-- -G  --  G  -  ---
#  2R 2  12254  T  -  ------  --------  ------ --  A-  -  -  ---  y→f
iC 0R 1  12261  G  -  ------  --------  AAAAAA -A  --  -  -  ---
iC 2R 2  12269  A  -  --G---  --------  ------ --  G-  -  -  ---  v→a
#  1R 3  12310  G  -  ------  --------  ------ T-  --  -  -  ---
iC 0R 1  12312  G  -  ------  --------  ------ --  A-  -  -  ---  p→s
iC 1R 3  12328  A  -  ------  --------  ------ --  --  -  -  G--
#  0R 1  12333  G  -  ------  --------  ---T-- --  --  -  -  ---  l→i
iC 2R 2  12341  G  -  ------  --------  ------ --  --  -  -  A--  s→l
iA     NT 12371  G  -  -----A  --------  ------ A-  --  -  -  ---
iC     NT 12373  T  -  ------  --------  CCCC-- CC  --  C  C  CCC
iC     NT 12375  T  -  ------  --------  CCCCCC CC  --  C  -  ---
iA     NT 12456  G  A  AAAAAA  --------  ------ --  --  -  -  --A
#      NT 12457  A  C  CCCCCC  --------  ------ --  --  -  -  --C
#      NT 12463  C  -  ------  -AA---AA  ------ --  --  -  -  ---
#      NT 12467  A  -  ------  --------  ------ --  --  -  -  -T-
iC     NT 12468  T  -  ------  --------  ------ --  -C  -  -  ---
iA     NT 12469  G  -  ------  --------  ------ --  --  -  -  -A-
#      NT 12470  C  -  ------  --------  ------ --  --  -  -  -A-
iC     NT 12475  T  C  CCCCCC  --------  ------ --  --  -  -  ---
#      NT 12479  A  -  ------  ----T---  ------ --  --  -  -  ---
#      NT 12481  C  -  ------  ------G-  ------ --  --  -  -  ---
```

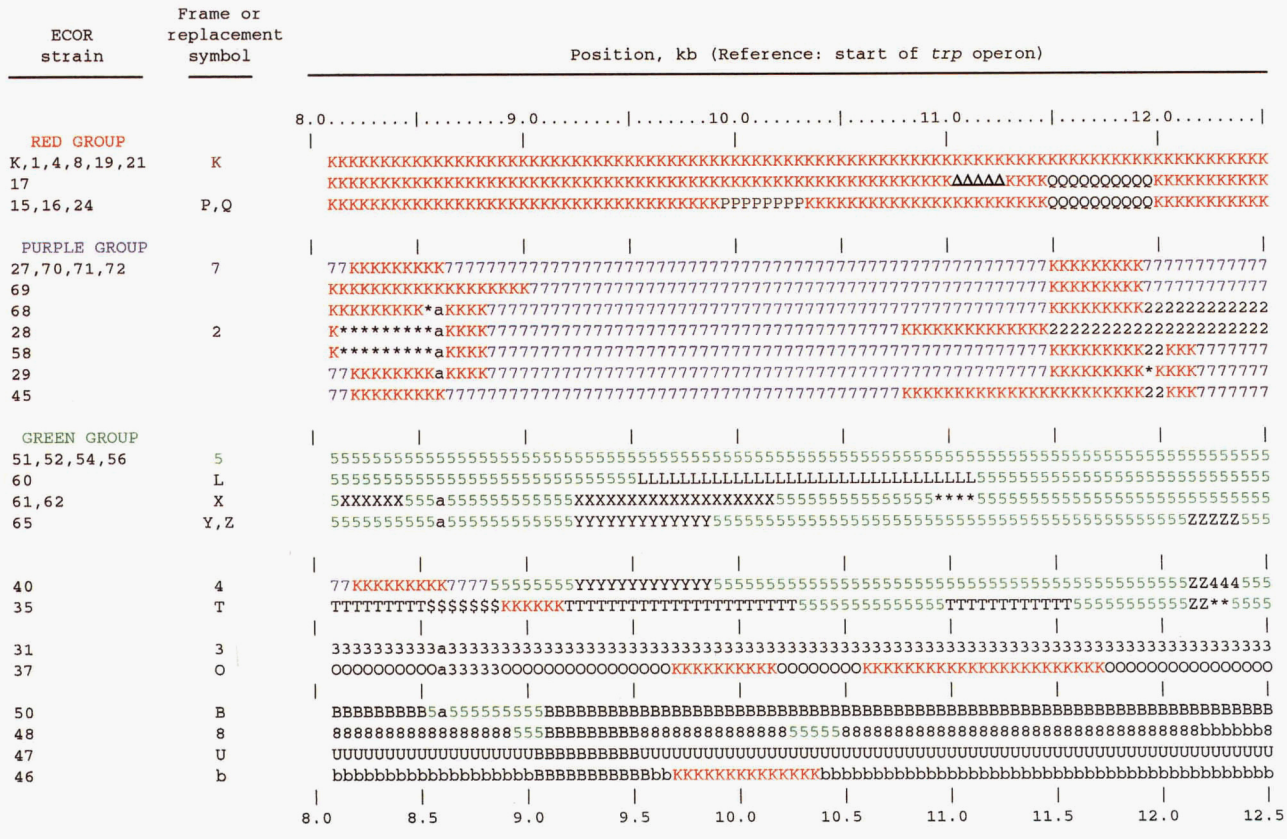

FIGURE 2.—Clonal segments as inferred from Figure 1. Each symbol represents 50 bp. Unlisted fragments {tentative major group} : 64 {51}; 66 {51}; 67 {71}. $ = IS1k (nonhomologous replacement); a = clonal segment of unknown affinity. * = interpretation uncertain. Δ is part of a deletion. K, 7 and 5 represent what are concluded to be clonal frames or their parts. Other symbols are used only to show similarity over homologous regions. Where **B** is written in corresponding parts of 50, 48, 47 and 46, it could just as well be **8**, **U** or **b**. There is no way of telling, from the present data set, which of the three sequence types is shared at that position.

sus base at the left. Then the first of six closely packed columns contains the sequence data for not one but two strains listed vertically, ECOR 4 and ECOR 19. These are identical throughout the region covered. The data are directly below the right hand digit of the strain number in all cases. Two strains, ECOR 1 and ECOR 8, head the next column in this close set; the position of "08" directly below "01" indicates that these two strains are similar enough to share a column. In this case, though, the strains do differ at positions 8,767 and 10,097: the details are signaled by "|." Reading across at any position demonstrates the great similarity of the first 10 strains listed (from K12 through ECOR 24). These strains are concluded, on the basis of these data and more extensive restriction analysis (Table 6 in MILKMAN and BRIDGES 1990, and subsequent unpublished data) to share a *Level II clonal frame*, which is referred to here and in Figure 2 as the "Red" or "K" clonal frame (without further qualification as to level). Each symbol in Figure 2 represents 50 bp. A look at the region between positions 9,929 and 10,278 reveals a break in the uniformity of these 10 sequences. The rightmost two columns in the set show that three strains, ECOR 15, ECOR 16 and ECOR 24, share a distinctly different sequence here–there are differences in 10 of some 350 positions. This is interpreted as a recombinational replacement having a more distant relationship to the other seven strains (K to ECOR 17) than these seven share in this region. It is represented in Figure 2 as a run of "P"s.

FIGURE 1.—Polymorphic sites in the ALO region. General symbols: Heading: Ch = Change; Ph = phase (remainder of N/3); C*P* = Codon Position; *n* = position of nucleotide with reference to the start of the *trp* operon; CO*n* = consensus (most frequent) base; K = strain K12. Body: # = transversion; iC = C-T transition; iA = A-G transition; M = multiple; Δ = deletion; > = insertion; NT = not translated; R = transcribed in reverse direction; - = consensus base; small letters = amino acid symbols. Special symbols: $ = insertion sequence IS1k substituted; @ = atlas insertion; = polymorphism within column; ? = uncertain base; J = jam, order uncertain, often (as in 11,101–11,104) with a change whose location is unclear. Incomplete sequences: [<75 bp missing] ECOR 8: 8,849–8,888, 9,700–9,708; ECOR 24: 8,816–8,888; ECOR 35: 11,814–11,849; ECOR 37: 8,853–8,888. [>75 bp missing] ECOR 48: 8,849–8,899, 10,646–10,750; ECOR 64 start-8,088; 11,066-end; ECOR 17 10,981–11,254 (includes apparent 235-bp deletion); ECOR 66: start-9,708 (includes large, incompletely defined deletion). ECOR 49: 8,889-end (completion not planned). Positions 9,672–9,708 were not sequenced in most strains–primers overlap.

Further on, ECOR 17 has a deletion beginning about 11,000 (Δs in Figure 2). Finally, strains 15, 16 and 24, and also 17 share a sequence that differs noticeably from the other six in the group between about 11,500 and 12,000 ("Q"s in Figure 2). These are the only evident segmental differences within the K group in the ALO region: two of them appear to be replacements shared by three and by four strains, respectively, and thus attributable to events in respective common ancestors; deletions of the sort seen in ECOR 17 have been noted elsewhere, including a 319-bp deletion in CAF shared by six ECOR strains (MILKMAN and BRIDGES 1990).

The next group of eight columns covers 10 strains sharing the "Purple" clonal frame. Four of the strains, 27, 70, 71 and 72, evince a specific sequence type over the whole 4,400-bp range, except for stretches of K12-type DNA from about 8,200 to 8,600, and from about 11,500 to 11,900. There are more shared replacements. Some are coextensive; others vary in length (again, see Figure 2).

The next nine columns describe 11 strains, for which only nine sequences are complete; ECOR 64 (column 7) and ECOR 66 (column 9) are not. These 11 strains share the "Green" sequence type over most of the region; the restriction data, which again are broadly distributed over the chromosome, suggest that this sequence type is their clonal frame, with the possible exception of ECOR 40.

The next two columns refer to strains that share some similar sequences; ECOR 37 also is similar to K12 over two stretches. Finally, six more strains are listed, of which ECOR 49 is fragmentary. These do not belong to large groups, but restriction analysis (MILKMAN and BRIDGES 1990; and unpublished results) indicates close chromosome-wide similarities between ECOR 35 and 36; among 38, 39, 40 and 41; and between 49 and 50.

Between the groups discussed individually so far, local similarities are occasionally noted. Sequence types seen in the form of clonal frames also appear to have been exported as clonal segments. In Figure 2, the symbols K, 7 and 5 are each seen outside their groups. Also, small regions of similarity such as one between 8,578 and 8,589 (shown as "a" in Figure 2) occasionally appear in a broad array of strains. All of the comparisons referred to so far clearly imply homology, but there is one *replacement* (symbolized by $ beginning at 8,536) that is not homologous: it is the insertion sequence IS1k (STOLTZFUS, LESLIE and MILKMAN 1988). This replacement evidently arrived subsequent to what is now known to be the precise *insertion* of a set of lambdoid phages, collectively called Atlas, between positions 8,888 and 8,889 (MILKMAN and BRIDGES 1990; STOLTZFUS 1991; see also CAMPBELL, SCHNEIDER and SONG 1992). Only strains 35 and 36 share IS1k. In all, 23 of the 72 ECOR strains, all from primates and mostly from humans, have a variety of functional or partially deleted Atlas phages here. Their presence is indicated by "@". Note that sets of closely related sequences (70 and 71 *vs.* 72, as well as 51 *vs.* 52, 54 and 56) differ as to the presence of Atlas.

The *precise* extent of the discontinuities in degree of similarity is usually impossible to determine because of the small proportion of base differences involved, but ECOR 64, beginning at position 8,893, shows a striking replacement. About 24% of some 283 bases are different from those in ECOR 51 *et al.*, and from the consensus sequence. Most of these base differences are in the third position, but there are a good number at the other two. Numerous amino acid differences are seen; still, there is no doubt as to the qualitative homology relating this stretch to that in the other strains. The distal end of the replacement seems clearly marked within a few bases–if not indeed between 9,171 and 9,172. Proximally, Atlas intervenes: ECOR 64's Atlas sequence is not particularly different from most others (Figure 3), conforming to the previous evidence (STOLTZFUS 1991; MILKMAN and BRIDGES 1990) that Atlas has come and perhaps gone more recently than some homologous recombinational replacements. On the proximal side of Atlas, we have been unable so far to amplify DNA from ECOR 64 (using primers complementary either to Atlas DNA or to the flanking DNA), suggesting that the great sequence difference persists there as well. [The Atlas inserts themselves are quite uniform at the distal end, but extremely diverse, presumably due to assorted deletions, at the proximal end.]

Comparison of the ALO sequences from 36 strains in Figure 2 yields the following information. First, three sets of 10, 10 and eight strains can be grouped by common clonal frames. Of these, six, none and four, respectively, share a given Level II clonal frame over the entire ALO region. Second, homologous replacements are frequently common to several strains. These are generally on the order of 250–1,000 bp in extent. Third, the other eight sequences vary in their affinities. All but ECOR 31 and ECOR 47 contain clonal segments similar to one or more of the three major clonal frames, and even these two strains share clonal segments with others. Moreover, in the adjacent CAF region (see MATERIALS AND METHODS), ECOR 31 and ECOR 47 each have several small Red clonal segments and three Purple ones (data not shown).

General observations: Three general observations can be drawn from Figures 1 and 2. First, chromosomes are frequently linear mosaics of small (ranging up to 1 kb or more) discrete segments of different sequence types, often embedded in evident clonal

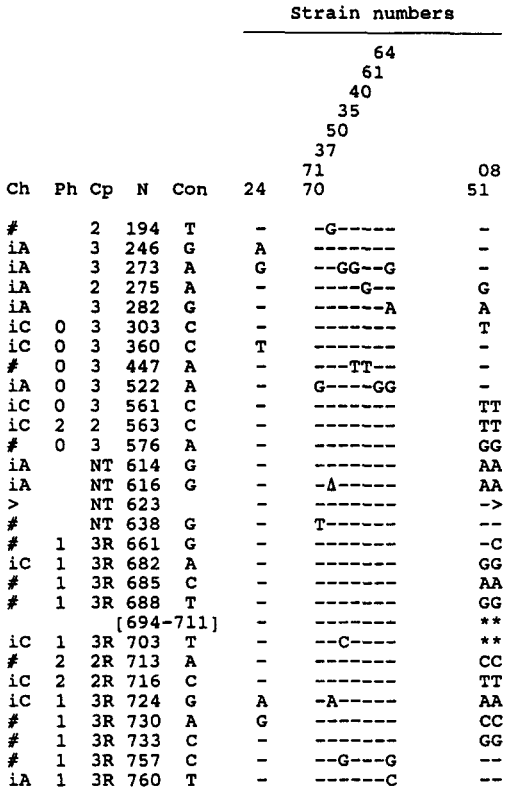

FIGURE 3.—Comparison of Atlas distal sequences (positions 67–761). General symbols as in Figure 1. Δ: in ECOR 37, the bases in positions 616–628 are deleted. >: in ECOR 8, AGAAGAGTTTTGC is inserted after 623 (ECOR 8 repeats *its* 611–623 once exactly). *: in ECOR 8 & 51, "ATT GTC GCT ATA ACT GCC TTT AGC CAG TTT ACG" replaces C^{694}GA GTC TTT TAC GGT AGT^{711}. This results in a replacement in ORF IV (which runs into Atlas from the right): t t v k d s → r k l a k g s y s d n. ORF IV ends at position 643 (end of stop codon). Integrase (see text) ends at 597 (end of first stop codon, which is followed by another) in all strains sequenced. Unreadable sections: 40:639–653; 51:659–674; 35:640–655; 37,50:638–654; 61:647–654; 24:650–653. ECOR 71 was sequenced by ARLIN STOLTZFUS and later by MELISSA BRIDGES.

frames. Second, among several strains (*e.g.*, the Red group) where a clonal frame is shared, sporadic variation characteristic of nucleotide substitution rather than recombinational replacement is often seen. And third, the sequence types are distributed broadly, but not randomly, among strains. For example, certain inserts may be shared by members of a given clonal frame group (P and Q in the Red group; K and "2" in the Purple group). On the other hand, Red and Green segments are not found close together frequently, though they are both present in ECOR 35 and in ECOR 40, for example.

The nonrandom distribution of sequence types is consistent with a historical nonrandom habitat distribution of clones and/or differential compatibility among clones. In this regard, it should be noted that recently collected wild strains (MILKMAN 1973 and unpublished results; MILKMAN and CRAWFORD 1983; CAUGANT, LEVIN and SELANDER 1981) frequently included genetically diverse strains from a single fecal sample. The ECOR strains show no general correlation between genotype (as expressed in MLEE or nucleotide sequence–Atlas is an exception) and host (see OCHMAN and SELANDER 1984); earlier times may have been different, however.

TABLE 1

Nucleotide differences (N) from clonal frame

Segment		Frame					
		Red		Purple		Green	
Symbol	Length	N	%	N	%	N	%
P	350 bp	9	3	6[a]	2	9	3
Q	450	5	1	4[a]	1	10	2
2	150	3[a]	2	4	2	4	2
L[b]	1,600	29	1.8	30	1.9	17[a]	1.1
X[c]	900	20	2.2	22	2.4	10[a]	1.1
Y[d]	600	14	2.3	15	2.5	6[a]	1.0
Z	200	9	4	8[a]	4	8[a]	4

[a] Lowest.
[b] L-X difference = 8/600 = 1.3%.
[c] X-Y difference = 4/600 = 0.7%.
[d] L-Y difference = 4/300 = 1.3%.

Among the strains, ECOR 40 is of interest because of the variety of sequence types contained in ALO. (This variety is seen also in CAF.) Presumably, it has had recombinational contact with a relatively great variety of genomes.

Some specific details from Figure 2 will now be considered in order. First, the top row: six strains (K12, 1, 4, 8, 19 and 21) share a sequence type that we have called a Level II clonal frame. Next, the three strains in the third row, ECOR 15, 16 and 24, share a stretch labeled "P"; further, with ECOR 17 they share "Q." Is P a different sequence type from Q, or are they of the same clonal origin? There is no way to tell, in the absence of a *reference sequence*, such as those shared by the respective six similar Red strains, or the four similar Purples, or the four Greens.

The ECOR 37 sequence, in contrast, is concluded to contain two replacements of the same sequence type (K), as well as three more in CAF. We know they are the same sequence type because they are like the K reference sequence, and we infer that they are replacements from the following line of evidence. It appears that these K segments have been introduced locally on an "O" (ECOR 37) background, rather than vice versa, because the broader comparisons of restriction analyses and MLEE show ECOR 37 to be very different from the Red group–indeed, from all major groups (MILKMAN and BRIDGES 1990).

In the Red, Purple and Green groups, various apparent replacements are seen. Table 1 shows that some (P, Q, 2) are not more similar to the frame in which they are embedded than to other frames. L, X and Y are clearly close to the Green frame and also to one another in their respective common regions.

TABLE 2

Summary of nucleotide substitutions

	Transitions			Transversions					All changes	Length (bp)
	C↔T	A↔G	Total	A↔C	G↔T	A↔T	G↔C	Total		
Translated (NTS)	108	89	197	22	27	24	24	97	294	3684
Nontranslated	26	24	50	10	4	4	5	23	73	686
Total	134	113	247	32	31	28	29	120	367	4370

Replacements of similar length (Figure 2) include Y (strains 65 and 40). In the Purple set, some of the K and "2" replacements vary in length. Some other examples of variation in replacement length are seen in Z (strains 65, 40 and 35) and in 5 (strains 50 and 48). It has not yet been possible to estimate the *maximum replacement size*, since the 4,400-bp ALO region described here is a small window. Large replacements are therefore not yet easy to define by sequencing, and restriction analysis does not provide the necessary resolution. *But clearly there are small discontinuities.*

Variation in nucleotides, amino acids and length: Table 2 summarizes the nucleotide substitutions observed. There is a statistically insignificant excess of C-T over A-G transitions in both translated (NTS) and nontranslated (the strand continuous with *trp* NTS was chosen arbitrarily) DNA. Also total changes are significantly more frequent in nontranslated DNA than in translated DNA, but significantly *less* frequent than would be predicted by the existence of neutral alternatives for all nontranslated bases. There are local variations in the frequency and nature of substitutions as well. Figure 1 suggests that structural polymorphism in TonB [the protein] is adaptive: changes involving proline are frequent, and two small phase-constant indels (one 3 bp and the other 6 bp in extent) are seen in *tonB* at 12,020 and 12,057, respectively. Aside from the ends of the ECOR 17 deletion, no other indels have been seen in translated regions of ALO. Table 3 summarizes amino acid differences by sequence type and by ORF. Compared to the five replacements involving proline in *tonB*, the rest of the translated sequence (about 4 times as long) contains only three–two at the same position–as well as one nonsense replacement. The individual amino acid differences are listed in the rightmost column of Figure 1.

Atlas. Figure 3 compares sequences of a 695-bp stretch of an arbitrarily numbered region at the distal end of Atlas. An integrase quite similar to that of Coliphage 21 (SCHNEIDER 1992; Genbank/EMBL Accession Number M61865) ends at position 597 in this region. A brief nontranslated stretch is followed by the terminal portion of ORF IV coming in from the right flank [a corresponding terminal portion of ORF IV is seen both in non-Atlas- and Atlas-containing strains proceeding left from position 8,888 (STOLTZFUS 1991; STOLTZFUS, LESLIE and MILKMAN 1988). While the sequences in Figure 3 are moderately similar in general, the incidence of indels of various sorts is considerable.

DISCUSSION

Agreement among MLEE, restriction analysis and sequencing: The deduced clonal frames correlate well with expectations developed from our restriction analyses, and they show a striking correspondence to the major ECOR strain classification derived from multiple locus enzyme electrophoresis (MLEE). In a phenogram "based on polymorphisms at 38 enzyme-encoding loci" (HERZER *et al.* 1990; see also SELANDER, CAUGANT and WHITTAM 1987), the 25-strain Group A contains all nine ECOR strains that we classify by sequencing as having the "Red" clonal frame. Group B_1's 16 strains include all 11 strains with "Purple" clonal frames, and the 15 strains in group B_2 include all 10 that we have assigned the "Green" clonal frame. We are not sure whether ECOR 35/36 and ECOR 38/39/40/41 should be grouped in an extended green set. We agree that the other six strains we have sequenced seem to fall outside of these major groupings.

The closeness of the agreement between the results of sequencing in a 4,400-bp region and the aggregates of broadly distributed MLEE and RFLP analyses is greater than might be expected. More precisely, it would be expected that sequencing in other regions would not conform to this pattern, just as the RFLP affinities vary from region to region (Table 6 in MILKMAN and BRIDGES 1990; additional unpublished results). While the Green group overlaps rarely with the Purple and never with the Red, the Purple and Red groups are often close. And indeed in the CAF region (data not shown), all members of the Purple group have both Red and Purple segments, with Purple predominating only in ECOR 70, 71 and 72 and Red in the others. The other strains appear much as they do in ALO.

There is a discrepancy in the order of branching between the MLEE major groups and ours: we find that our Red ["A"] and Purple ["B_1"] clonal frames differ by about 1% of their nucleotides, and that each

TABLE 3

Amino acid differences from consensus

ORF	Length (bp)	Start	Clonal Frame			Other Segments			
			Red	Purple	Green	Single	Close-dual	Multiple[a]	ECOR 64
II(part)	75	(7658)	1	0	0	1	0	0	—
III	504	8204	1	0	1	4	1	3	—
IV(←)	636	9408	0	0	1	1	0	1	17
V	741	9765	0	1	2	5	0	2[b]	—
VI	537	10538	0	0	2	6	0	3[b]	—
P14	396	11182	0	0	0	1	0	0	—
tonB(←)	732	12354	2	2	1	11	0	3[c]	—
Total	3621		4	3	7	28	1	12[d]	17

[a] Or two sequences not closely related (cf. "Close-dual").
[b] Including two clonal frames (already listed).
[c] Including three clonal frames (already listed).
[d] Including seven clonal frames (already listed).
(←): transcribed in direction opposite to *trp*.

differs from Green ["B_2"] by 2%, while the MLEE phenogram places B_1 and B_2 closer to one another than to A. Note also that the MLEE phenogram is not a phylogenetic tree, since it is now clear that the loci compared do not share a common genealogy. Thus, the branch lengths do not represent single genome-wide evolutionary distances.

Clonal levels: The data reported here are largely consistent with the previous assignments to clonal levels (MILKMAN and BRIDGES 1990). Thus, K12 and ECOR 1, 4, 16, 19 and 21 are still thought to share a Level II clonal frame; strains K12, ECOR 15, 68 and 71 are still thought to share a Level III clonal frame.

Note that Level III clonal frames were defined to include differences at the 1% level, so that a Level III clonal frame can include different sequence types, but a Level II clonal frame cannot. The Red, Purple and Green groups in Figures 1 and 2 are based on common sequence types in the ALO region, as opposed to a larger sample of the genome, so they may not correspond precisely to level II clonal frames. The specific Level II affinities of such strains as ECOR 15 and 68 should become clear with more extensive sequencing. It is clear that the Red and Purple groups taken together still evince a single Level III clonal frame. By these definitions, the red and purple sequence types are different; there are distinct red and purple Level II clonal frames, which may or may not turn out to be the prominent sequence type for ECOR 15, 24 and 68 (for example); and the K12 Level III clonal frame includes the red and purple sequence types, but not the others.

Interpretation of the replacement patterns: The DNA fragments introduced into *E. coli* cells by general transduction are on the order of 50–100 kb in length; conjugation can introduce single strands of considerably greater length. The discontinuities in sequence type reported here, however, are often at least two orders of magnitude shorter. This contrast in lengths reminds us that entry into the cell is followed by a separate process, incorporation into the chromosome. With this in mind, three possible explanations (not mutually exclusive) are considered.

1. Small entrant DNA fragments have been incorporated directly to form the observed replacements.

2. Large entrant fragments have been incorporated *in toto*; successive overlapping incorporations result in short mosaic elements.

3. Large entrant molecules have been incorporated in series of small discontinuous segments, which are the replacements observed. (*Cascade* hypothesis.)

The first alternative is the simplest conceptually, but it has no known mechanism in *E. coli*. Nevertheless, it may be unwise to rule out the possibility of natural transformation in *E. coli* in the absence of a sufficiently broad investigation of natural strains and conditions. Natural transformation is widespread in bacteria (STEWART 1989), and there is evidence that this process can lead to the chromosomal incorporation of short stretches of homologous DNA. Specifically, STEWART, SINIGALLIANO and GARKO (1991 and G. STEWART, personal communication) have demonstrated natural transformation of rifampicin resistance via linear genomic fragments averaging less than 1,000 bp in *Pseudomonas stutzeri* at a rate of 3×10^{-8} per potential recipient cell, above background. Also, Frederick Cohan (personal communication) reports routine transformation of rifampicin resistance in *Bacillus subtilis* using PCR-generated 3.4-kb DNA fragments with the same efficiency (5×10^{-3} to 1×10^{-2}) as genomic DNA. This mechanism thus appears conceivable for *E. coli*. In addition, a plasmid intermediary is not out of the question. For an interesting experimental case, see SCHNEIDER *et al.* (1981).

The second alternative, the production of short discontinuities in degree of homology by overlapping

replacements, is more plausible at first glance. Transduction (ROBESON *et al.* 1980) and conjugation (UMEDA and OHTSUBO 1989) are likely mechanisms of recombination in *E. coli*. Presumably, progressively shorter discontinuous sequence types would accumulate as one already-mosaic fragment replaces part of an already-mosaic chromosome. Computer simulations of this process, still in their initial stages, have now routinely produced banding patterns qualitatively consistent with the observations reported here.

Third, the cascade hypothesis suggests that a single large entrant molecule gives rise to a large number of separate incorporations, perhaps because of endonuclease activity in wild strains, or possibly local sequence dissimilarities. This would account for rare local runs of discontinuous segments of a particular sequence type; it would not exclude superimposed replacements of the same type. There is evidence that large entrant molecules may be incorporated discontinuously after conjugation in *E. coli*. For example, PAUL and RILEY (1974) used a combination of radioactive and density isotopes to show the incorporation of Hfr strand fragments ranging from 4,000 to 17,000 nt in length and separated by gaps of no more than 450 nt, which would presumably be repaired subsequently. Previously, PITTARD (1964) and HARRIS and CHRISTENSEN (1966) (see also ARBER and MORSE 1965) had observed that "restricted" (in the classic sense of the term, but now understandable in terms of digestion by restriction endonucleases) DNA showed far less linkage between selected and unselected markers following conjugation than did unrestricted DNA. The procedure was to select for one marker and screen the colonies for other markers. The interpretation is that the entrant DNA was broken up before being incorporated, and that some fragments were destroyed. This would presumably increase the level of discontinuity above that observed in conjugation between K12 strains (which presumably did not differ in restriction-modification systems) by PAUL and RILEY. Recently, COHAN, ROBERTS and KING (1991) showed that transformation was reduced (but not eliminated) between groups of *B. subtilis* strains differing in their restriction enzymes. These enzymes digested the DNA of strains outside their respective groups but not within them.

In the past, the required use of genetic markers limited the resolution of multiple recombination events, both in number and in distance. Even so, multiple exchanges have been observed very frequently (often, earlier, described with reference to "negative interference"). Now the availability of extensive sequence differences permits the high-resolution observation of discontinuous incorporation in the absence of selection for more than a single marker, thus presumably under relatively natural circumstances.

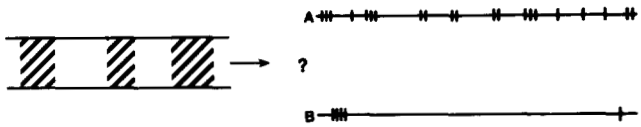

FIGURE 4.—Chromosomal distribution of replacements predicted by recombination mechanisms 1 and 2 (A: random) and mechanism 3 (B: clustered).

The availability of extensive comparative sequence data now makes it possible to test the cascade hypothesis experimentally, and to examine other aspects of recombination in *E. coli* at a finer level than before. Bacteriophage P1 *trp* transductants are now being produced, locally amplified by PCR, and sequenced to determine whether donor DNA has been incorporated continuously or discontinuously. Corresponding conjugation experiments will be undertaken shortly.

Chromosomal distribution of replacements from a given source: While the first two mechanisms would predict a random chromosomal distribution of replacements from a given source (Figure 4A), the cascade mechanism could produce a demonstrable spatial clustering (Figure 4B). In practical terms, this might be seen in a large series of K12 replacements in ECOR 37 in the regions currently being sequenced and none in one or more other regions. Thus, further sequencing in distant regions can provide evidence as to which mechanism is most likely, since the cascade mechanism implies *local* discontinuous series of a particular replacement type.

The three mechanisms imply different recombination rates: Estimation of the recombination rate depends in part on the number of events required to produce an observed replacement. For example, the direct chromosomal incorporation of a small DNA fragment would mean one event per observed replacement. In contrast, the requirement of many overlapping incorporations to pare large original replacements down to pieces of 1 kb or less may mean several events per observed replacement, and thus a higher recombination rate. But the cascade hypothesis leads in the opposite direction: the discontinuous incorporation of many small pieces of a single entrant molecule would imply one initial event for many replacements.

Intraclonal replacements: A high recombination rate, especially if intraclonal replacements were favored, would predict frequent replacements within a sequence type, where they would often be too subtle to detect. These would invalidate the use of Level II clonal frames as a means of measuring divergence among members of a Level II clone, since each frame could be homogenized into a cryptic hodgepodge of segments from related genomes. This possibility could not previously be evaluated using RFLP data alone.

Now, in principle, differences among different versions of a Level II clonal frame can be compared with differences among their respective inclusions. Four variables, all potentially dependent on divergence time, are of interest: Clearly, the frequency of individual nucleotide differences in clonal frames should correlate with the frequency of individual nucleotide differences in shared replacements, with length variation in shared replacements, and with the frequency of unshared replacements.

The length variation would presumably reflect otherwise undetectable intraclonal replacements within the frame that had abridged the extraclonal replacements. It would be of interest to find such length variation in a clonal frame showing little sequence variation, especially few individual nucleotide variants. This would argue for the homogenization of the clonal frame. The available data, however, are not sufficient to support analysis.

A steady-state pattern: The observed pattern of sequence discontinuities must be taken, it would seem, as representative of a complex steady state. Repeated extraclonal recombination would compromise a newly formed genome-wide clone with replacements whose initial size is not yet certain. On a longer time scale, it is also clear that this mechanism would lead to the eventual extreme reduction of *all* observed replacement lengths, were it not for the occasional redefinition of a sequence type, or "color," as the result of the formation of a new large clone by periodic selection, as illustrated in Figure 2 of MILKMAN and BRIDGES (1990).

We now return to the very small replacement, labeled "a" in Figure 2, whose distribution is surprisingly broad. It consists of a central set of three nucleotide substitutions between positions 8,578 and 8,589 (resulting in an isoleucine replacing a threonine) in four purple and three green sequences, as well as in ECOR 31, 37, 49 and 50. In three of the purple sequences it extends one position earlier (and a valine replaces alanine); in the three green sequences and in 49 and 50 it extends to 8,590 (the greens go further to 8,611, where glutamic acid replaces aspartic). Is this (or its complement) a coincidentally uniform remnant of a large replacement that has been completely effaced in other strains? Is there some selective advantage, present or past, to the central amino acid replacement? Or is this polymorphism the result of intragenomic conversion rather than recombinational replacement? As to the last possibility, a Genbank search of *E. coli* DNA sequenced to date has revealed no sequences that could be responsible, but a potentially more decisive approach (genomic Southerns with a probe containing "a") has not been undertaken.

The occurrence of small retained replacements evidently due to recombination is not limited to intra-

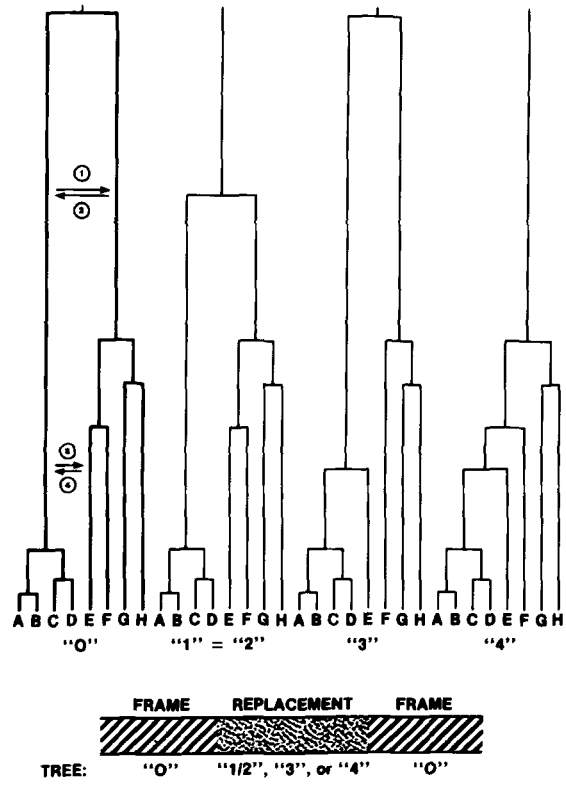

FIGURE 5.—Local phylogenies. Phylogeny (left) of clonal frames and phylogenies that would result from particular replacements (numbering of trees corresponds to numbering of events).

specific comparisons. MAYNARD SMITH (1992) and SPRATT *et al.* (1992) have described interspecific replacements ranging mainly between 0.1 and 1 kb in length in Neisseria. LIU *et al.* (1991) have described similar replacements among Salmonella serovars [corresponding to what are considered separate species by others (LIU and SANDERSON 1992), but see LE MINOR and POPOFF (1987)]. Their presence and perhaps their size may have some strong adaptive basis, in relation to penicillin resistance (Neisseria) or surface antigen polymorphism (Salmonella). Possibly, too, they may provide a clue as to the mechanism of recombination in nature.

Local genomic phylogenies: As indicated already, knowing the *direction* of a replacement (knowing the source and the recipient) is important. Figure illustrates a simple hypothetical example of phylogenetically distinguishable and indistinguishable consequences of some replacement events. Here it is assumed that a clonal frame has been characterized by extensive (5–10 kb) and numerous (5–10 strains) comparisons, and that a distinct clonal segment is observed. The bold tree on the left is that of the frames. The three finer trees on the right apply to the replacement segment in the middle. All the trees would derive from sequence comparisons of the respective segments, and the numbered events would be deduced from the comparison of the trees. The replace-

ments numbered *1* and *2* unify the middle segment in all eight strains and cannot be distinguished, without further information. [A simple illustration of the way recombination changes a tree is given in Figure 2, DYKHUIZEN and GREEN 1991]. Incidentally, different clonal frames do not differ uniformly along their lengths. There may be numerous reasons for this, but one is certainly the independent occurrence of recombinational replacements in each *before* they gave rise to large clones.

Extent and degree of homology: Referring to the *extent* of homology in terms of chromosomal length breaks little new ground. Referring to the *degree* of homology may at first glance seem to undo an important concerted effort on the part of a number of evolution journal editors to discourage the pointless practice of saying "homology" when "similarity" is meant (REECK *et al.* 1987). Similarity is an observation, as Walter Fitch has said, and homology is a conclusion. But the conclusion *can* be put quantitatively. It is not enough to say that two sequences are more similar in one region than in another, if the reason is variation in the recency of common ancestry. REECK *et al.* (1987) have correctly pointed out that "homology" and "similarity" are not synonymous, in calling for more precise use of both terms in the comparison of sequences. But it would be wrong to conclude that because two "sequences are either homologous or they are not," no quantification is possible. Here is a contradictory example: a crossvein in the wing of *Drosophila melanogaster* is either defective or it is not. Yet several authors (WADDINGTON 1953; MILKMAN 1970 and references therein; MOHLER 1965) have all quantified the defects and used the analysis of the data to reach useful conclusions. Furthermore, the population genetics concept of "identity by descent" has little value without reference to a point in time. And while recognizing that the homology of bird wings and human forearms is important even without reference to time (in part because development provides evidence of common ancestry), the comparison of sequences calls for the quantification of homology. It is possible that all DNA is descended from the same ultimate template, or for that matter, one of a thousand initial templates, but this distant possible common ancestry does not cloud our "qualitative" concept of homology, simply because this concept has *an implicit statute of limitations*. That is, homology is useful only in reference to actual or predicted similarities (of which common ancestry is of course not the only possible cause). When the ancestral relationship between two structures is too distant to impose similarity, it is no longer relevant to the comparison. Finally, it is useful and simple to *think* of homology in quantitative terms, even though the divergence time (like the age of the universe) can only be estimated, and even though the estimates may often be crude or wrong.

In conclusion, it appears that the comparative sequencing of K12 and the ECOR strains has produced information consistent with MLEE and restriction analysis, leading to the interpretation that the observed linear discontinuities in similarity are due to recombination. Three specific mechanisms have been proposed, and it appears that the issue may be resolved by some fairly simple experiments.

We thank ALLAN CAMPBELL, J. ROGER CHRISTENSEN, MONICA RILEY, GREG STEWART, and ARLIN STOLTZFUS for valuable comments. TIM FLANIGAN and HAIYAN JIANG provided technical assistance, and SHELLEY PLATTNER helped R. M. generously with word-processing instruction. This work has been supported by National Institutes of Health Grant GM 33518, National Science Foundation Grant BSR 9020173, personal funds, and funds from The University of Iowa.

LITERATURE CITED

ARBER, W., and M. L. MORSE, 1965 Host specificity of DNA produced by *Escherichia coli*. VI. Effects on bacterial conjugation. Genetics **51:** 137–148.

BISERČIĆ, M., J. Y. FEUTRIER and P. R. REEVES, 1991 Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173:** 3894–3900.

CAMPBELL, A., S. J. SCHNEIDER and B. SONG, 1992 Lambdoid phages as elements of bacterial genomes. Genetica **86:** 259–267.

CAUGANT, D. A., B. R. LEVIN and R. K. SELANDER, 1981 Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics **98:** 467–490.

COHAN, F. M., M. S. ROBERTS and E. C. KING, 1991 The potential for genetic exchange by transformation within a natural population of *Bacillus subtilis*. Evolution **45:** 1383–1421.

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.

DYKHUIZEN, D. E., and L. GREEN, 1986 DNA sequence variation, DNA phylogeny, and recombination. Genetics **113:** s71.

DYKHUIZEN, D. E., and L. GREEN, 1991 Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. **173:** 7257–7268.

HARRIS, D. J., and J. R. CHRISTENSEN, 1966 P1 lysogeny and bacterial conjugation. J. Bacteriol. **91:** 898.

HERZER, P. J., S. INOUYE, M. INOUYE and T. WHITTAM, 1990 Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. **172:** 6175–6181.

LE MINOR, L., and M. Y. POPOFF, 1987 Designation of *Salmonella enterica* sp. nov., nom. rev., as the type and only species of the genus *Salmonella*. Int. J. Syst. Bacteriol. **37:** 465–468.

LIU, D., N. K. VERMA, L. K. ROMANA and P. R. REEVES, 1991 Relationships among the *rfb* regions of *Salmonella* serovars A, B, and D. J. Bacteriol. **173:** 4814–4819.

LIU, S.-L., and K. E. SANDERSON, 1992 A physical map of the *Salmonella typhimurium* LT2 genome made by using *Xba*I analysis. J. Bacteriol. **174:** 1662–1672.

MAYNARD SMITH, J., 1992 Analyzing the mosaic structure of genes. J. Mol. Evol. **34:** 126–129

MILKMAN, R., 1970 The genetic basis of natural variation in *Drosophila melanogaster*. Adv. Genet. **15:** 55–114.

MILKMAN, R. 1973 Electrophoretic variation in *Escherichia coli* from natural sources. Science **182:** 1024–1026.

MILKMAN, R., and M. M. BRIDGES, 1990 Molecular evolution of the *E. coli* chromosome. III. Clonal frames. Genetics **126:** 505–517.

MILKMAN, R., and I. P. CRAWFORD, 1983 Clustered third-base substitution 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.

MOHLER, J. D., 1965 Preliminary genetic analysis of crossveinless-like strains of *Drosophila melanogaster*. Genetics **51:** 641–651.

NELSON, K., T. S. WHITTAM and R. K. SELANDER, 1992 Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. J. Bacteriol. **174:** 6886–6895.

NELSON, K., T. S. WHITTAM and R. K.SELANDER, 1991 Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc. Natl. Acad. Sci. USA **88:** 6667–6671.

OCHMAN, H., and R. K. SELANDER, 1984 Standard reference strains of *E. coli* from natural populations. J. Bacteriol. **157:** 690–693.

PAUL, A. V., and M. RILEY, 1974 Joint molecule formation following conjugation in wild type and mutant *Escherichia coli* recipients. J. Mol. Biol. **82:** 35–56.

PITTARD, J. 1964 Effect of phage-controlled restriction on genetic linkage in bacterial crosses. J. Bacteriol. **87:** 1256–1257.

REECK, G. R., C. DE HAËN, D. C. TELLER, R. F. DOOLITTLE, W. M. FITCH, R. E. DICKERSON, P. CHAMBON, A. D. MCLACHLAN, E. MARGOLIASH, T. H. JUKES and E. ZUCKERKANDL, 1987 "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell **50:** 667.

ROBESON, J. P., R. M. GOLDSCHMIDT and R. CURTISS III, 1980 Potential of *Escherichia coli* isolated from nature to propagate cloning vectors. Nature **283:** 104–106.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular Cloning, A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y.

SCHNEIDER, S. J., 1992 Site-specific recombination of lambdoid phage 21 into the *icd* gene of *Escherichia coli*. Ph. D. Thesis, Stanford University, Stanford, Calif.

SCHNEIDER, W. P., B. P. NICHOLS and C. YANOFSKY, 1981 Procedure for production of hybrid genes and proteins and its use in assessing significance of amino acid differences in homologous tryptophan synthetase α polypeptides. Proc. Natl. Acad. Sci. USA **78:** 2169–2173.

SELANDER, R. K., D. A. CAUGANT and T. S. WHITTAM, 1987 Genetic Structure and Variation in Natural Populations of *Escherichia coli*, pp. 1625–1648 in *Escherichia coli and Salmonella tymphimurium Cellular and Molecular Biology*, edited by F. C. Neidhardt. American Society for Microbiology, Washington, D. C.

SPRATT, B. G., L. D. BOWLER, Q.-Y. ZHANG, J. ZHOU and J. MAYNARD SMITH, 1992 Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. J. Mol. Evol. **34:** 115–125.

STEWART, G. J., 1989 The mechanism of natural transformation, pp. 139–164 in *Gene Transfer in the Environment*, edited by S. B. Levy and R. V. Millers. McGraw-Hill, New York.

STEWART, G. J., C. D. SINIGALLIANO and K. A. GARKO, 1991 Binding of exogenous DNA to marine sediments and the effect of DNA/sediment binding on natural transformation of *Pseudomonas stutzeri* strain ZoBell in sediment columns. FEMS Microbiol. Ecol. **85:** 1–8.

STOLTZFUS, A. B. 1991 A survey of natural variation in the *trp-tonB* region of the *E. coli* chromosome. Ph. D. Thesis, The University of Iowa, Iowa City.

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.

UMEDA, M., and E. OHTSUBO, 1989 Mapping of insertion elements IS*1*: IS*2* and IS*3* on the *Escherichia coli* K 12 chromosome. Role of the insertion elements in formation of Hfrs and F′ factors and in rearrangements of bacterial chromosomes. J. Mol. Biol. **208:** 601–614.

VANBOGELEN, R. A., P. SANKAR, R. L. CLARK, J. A. BOGAN and F. C. NEIDHARDT, 1992 The gene-protein database of *Escherichia coli:* Ed. 5. Electrophoresis 13 (in press).

WADDINGTON, C. H., 1953 Genetic assimilation of an acquired character. Evolution **7:** 118–126.

Communicating editor: W.-H. LI