Molecular Evolution of the *Escherichia coli* Chromosome. VI. Two Regions of High Effective Recombination

Roger Milkman,[1] Erich Jaeger[2] and Ryan D. McBride[3]

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

Manuscript received October 17, 2002
Accepted for publication October 29, 2002

ABSTRACT

Two 6- to 8-min regions, centered respectively near 45 min (O-antigen region) and 99 min (restriction-modification region) on the *Escherichia coli* chromosome, display unusually high variability among 11 otherwise very similar strains. This variation, revealed by restriction fragment length polymorphism (RFLP) and nucleotide sequence comparisons, appears to be due to a great local increase in the retention frequency of recombinant replacements. We infer a two-step mechanism. The first step is the acquisition of a small stretch of DNA from a phylogenetically distant source. The second is the successful retransmission of the imported DNA, together with flanking native DNA, to other strains of *E. coli*. Each cell containing the newly transferred DNA has a very high selective advantage until it reaches a high frequency and (in the O-antigen case) is recognized by the new host's immune system. A high selective advantage increases the probability of retention greatly; the effective recombination rate is the product of the basic recombination rate and the probability of retention. Nearby nucleotide sequences clockwise from the O-antigen (*rfb*) region are correlated with specific O antigens, confirming local hitchhiking. Comparable selection involving imported restriction endonuclease genes is proposed for the region near 99 min.

REEVES, Whitfield, and colleagues (Reeves 1993; Liu and Reeves 1994; Reeves 1995; Whitfield 1995; Lan and Reeves 1996; Amor *et al.* 2000; Li and Reeves 2000) have described in great detail the largely nonhomologous structural polymorphism of the O antigen of Salmonella and *Escherichia coli*. Illustrations of the surface lipopolysaccharide of gram-negative bacteria (Madigan *et al.* 2002, pp. 79–81) show the basal lipid (lipid A), the core polysaccharide, and the O-specific polysaccharide, which is the definitive O antigen. Each O antigen is characterized by specific sugars in specific linkages [Yao and Valvano 1994; Stevenson *et al.* 1994; *e.g.*, → 2)-β-D-galactofuranose-(1 → 6)-α-D-glucosamine-(1 → 3)-α-L-rhamnose-(1 → 3)-α-D-*N*-acetylglucosamine-(1 →]. Its specific structure is determined by a gene complex (the *rfb* region), whose functional polymorphism is nonhomologous as well, consisting of genes for particular sugar synthases and transferases. New O antigens in a species originate in the effective lateral transfer of new genes (rather than new alleles), mainly across considerable phylogenetic distances. This contrasts with the homologous genetic polymorphism of bacterial surface proteins, which is generally produced by mutation. The K-12 *rfb* region is entirely in genome section (G.S.) 184 (see materials and methods); it begins with *rfbB*, going counterclockwise, and ends just before *gnd*.

Because the entire flat surface of the cell is covered by a single type of O antigen among the hundreds known, each distinct type is extremely valuable when unrecognized, presumably because hosts tend to limit the growth of intestinal residents, even of beneficial ones. Evidently, as cells bearing a novel O antigen reach a high enough frequency, they become recognized by the immune system. Specific secretory immunoglobulin A molecules immobilize these cells in the mucin layer of the colon (Salyers and Whitt 2001), and they lose their growth advantage over previous, recognized arrivals but remain at high frequency. Intraspecific horizontal transfer, before or after egestion followed by ingestion, results in the repetition of this event in each new host, and the recent immigrant spreads in a geographic wave. Behind this wave lies a new opportunity for novel forms. Thus there is an insatiable demand for new nonhomologous variants, and these are more likely to arrive from distant evolutionary lines than to originate within the species.

Frequent retransmission among the *E. coli* strains of new gene complexes within the *rfb* region, together with hitchhiking flanking DNA, is believed to be responsible for the well-known hypervariability of *gnd*, a neighboring gene that makes 6-phosphogluconate dehydrogenase (Bisercic *et al.* 1991; Dykhuizen and Green 1991; Nelson and Selander 1994; Selander *et al.* 1996, p. 2695). After the finding (Milkman and McKane Bridges

[1]*Corresponding author:* 12 Fells Rd., Falmouth, MA 02540-1626. E-mail: rmilkman@mbl.edu

[2]*Present address:* The University of Chicago Urology Research Laboratory, 5812 South Ellis Ave., MC6038, Chicago, IL 60637.

[3]*Present address:* Temple University School of Podiatric Medicine, 801 Cherry St. #711, Philadelphia, PA 19107.

1990) that the adjacent histidine operon also seemed highly variable in the ECOR (Ochman and Selander 1984) strains, although not to the degree of *gnd*, we undertook the local restriction analysis of a tight group of 11 strains (see materials and methods) within which the patterns are ordinarily uniform or nearly so. We expected that an unusually high local effective rate of recombination among a broad range of *E. coli* strains would lead to the frequent appearance of detectable replacements in this normally uniform set. If this occurred, it would presumably decline with increasing distance from the *rfb* region, which is near 45 min.

In fact we have now observed such variation, declining with distance, both near *rfb* and near the immigration control region (ICR; Raleigh 1992), a similar center of nonhomologous polymorphism where restriction-modification gene complexes reside near 99 min (Barcus and Murray 1995; Barcus *et al.* 1995). Novel restriction endonucleases have a likely advantage in the arms race against bacteriophages, and again novel imports appear to be the key. In a survey of the rest of the chromosome, variation among these strains is infrequent, but not absent. It does include occasional cases of clear-cut replacements (Table 1 and see results). Evidently, when one of the two centers of hypervariability includes a rare new import, any normal intraspecific recombinant that contains the import has a vastly greater chance of persisting. This is due to the power of strong selection to overcome random genetic drift, which is an overwhelming eliminator of new mutant alleles or recombinational replacements that range from deleterious to weakly favorable (Crow and Kimura 1970; Milkman 1999). Sequence variation observed up to some 150 kb clockwise from *rfbB* is evidently due to the previously mentioned hitchhiking of DNA flanking the *rfb* region import.

MATERIALS AND METHODS

Strains: From the ECOR collection (Ochman and Selander 1984), the strains ECOR 1, 2, 3, 5, 8, 9, 10, 11, 12, and 25 were chosen for analysis. These strains are closely grouped in a well-known multilocus enzyme electrophoresis phenogram (Herzer *et al.* 1990), where they compose a subset of the "A" group, which corresponds to the "K12" *meroclone* (McKane and Milkman 1995; Milkman 1996, 1997). To this set of 10 wild strains, the very similar laboratory strain K12 W3110 was added, making a group called the Big Ten (by analogy with the 11-member Big Ten universities).

Restriction fragment length polymorphism analysis: From genomic DNA or a cell culture of each strain, PCR fragments ordinarily ~1500 bp in length were amplified in a large number of chromosomal regions (shown in Tables 1 and 2). The PCR fragments were digested typically with each of six commercial restriction endonucleases (New England Biolabs, Beverly, MA), mainly those having 4-bp recognition sites. Methods are similar to those described in Milkman and McKane Bridges (1990) and Milkman and McKane (1995). In the nonhypervariable region, when a restriction fragment from a strain contained more than one restriction site difference, nearby fragments from the differing strains were analyzed to determine whether the difference extended further, implying lateral transfer. Nucleotide sequencing was employed for broader and more detailed analysis.

Nucleotide sequencing: PCR-primer extension sequencing employed the addition of a ^{33}P-labeled specific dideoxynucleotide to each of four reaction mixes, which were loaded individually on sequencing gels and electrophoresed according to directions for the ThermoSequenase-radiolabeled terminator cycle sequencing kit from Amersham (Piscataway, NJ).

Chromosomal locations: Positions are expressed in minutes for use with genetic maps (Berlyn 1998). For use with physical maps (Rudd 1998), the sequence of *E. coli* K12 MG1655 (Blattner *et al.* 1997) is divided into 400 genome sections, whose regularly updated detailed descriptions are available from the National Center for Biotechnology Information at http://www.ncbi.nlm.nih.gov/; see also Figure 1 legend.

Number of pairwise comparisons: The number of combinations of n distinct things taken m at a time, $C(n, m)$, is calculated as $n!/m!(n - m)!$ (Beyer 1966). Thus the number of all pairwise comparisons of 11 sequences is $11!/(2! \times 9!) = 55$.

RESULTS

The restriction fragment length polymorphism data are summarized in two tables: Table 1 covers the two nonhypervariable regions, and Table 2 focuses on the two hypervariable regions, which flank, respectively, the O-antigen region (*rfb*, near 45 min) and the ICR (near 99 min). In each table are listed the map positions in minutes, the *E. coli* K12 MG1655 genome section numbers (Blattner *et al.* 1997), and the gene affiliations of the PCR fragments. The two tables illustrate the sharp difference, between the nonhypervariable and hypervariable regions, in the degree of variation contained.

In both tables, each illustrated region shares a common uniform boundary fragment, labeled E, with each adjoining region. Because the chromosome is circular, Table 1 runs from 2.1 to 40.7 min and from 49.0 to 95.1 min. Table 2 runs from 95.1 to 2.1 min and from 40.7 to 49.0 min. Variation is sporadic in the two nonhypervariable regions. Occasional strains (underlined) do show infrequent local differences from the others, as expected. The illustrated cluster of differences from 66.9 to 68.2 min resulted from an attempt to determine the extent of two evident replacements at 67.7 min. Subsequent nucleotide sequencing has shown that, for the fragments in this table, each of the strains with more than one restriction site difference contains a recombinational replacement, and those with only one restriction site difference were found to differ in only a single nucleotide, most likely due to mutation. Between 66 and 68 min, ECOR 8 and ECOR 11 have distinct overlapping replacements, ~91 and 137 kb long, respectively (R. McBride and R. Milkman, unpublished results). Near 12 min, K12 and ECOR 11 share a lengthy replacement, and subsequent extensive studies throughout the nonhypervariable region have shown sporadic differences (not illustrated; R. Milk-

TABLE 1

Variant strains in nonhypervariable regions. Results of six restriction digests on the Big Ten strains

Map position, min	G.S.[a]	R.F.[b]	Variant strains[c]
E[d] 2.1	8	ftsW	—
8.6	35	phoA	—
11.9	48	cysS	K = 11(3)
15.2	61	nagE	10(1)
16.4	65	sucAB	3(1),12(1)
22.3	89	hyaCDEF	—
28.1	113	yciU	—
28.2	113	tonB	—
28.4	114	trpED	3(1)
28.7	115	topA	—
30.1	121	fnr	9(1)
33.2	133	narZ	—
36.3	146	fumA	—
E[d] 40.7	165	fad	—
E[d] 49.0	197	yejF	—
49.5	199	nap	—
50.3	202	gyrA	—
53.4	215	emrY	—
58.9	235	clp	—
62.7	252	relA	—
66.9	269	speC	8(2),11(5)
67.3	270	glcB	8(6),11(4)
67.7	272	hybA	8(9),11(13),25(1)
68.2	274	parC	11(4)
71.6	288	hflB	3(1)
75.3	302	nirB	—
79.9	321	dppA	—
84.3	340	glmS	—
88.7	357	glpK	—
93.1	372	phnD	—
93.6	375	fumB	—
E[d] 95.1	380	aidB	—

Variant strains are underlined, and their respective number of restriction site differences from the remaining uniform set are shown in parentheses. Big Ten strains are: K12, ECOR 1, 2, 3, 5, 8, 9, 10, 11, 12, and 25. See text.

[a] G.S., genome section, one of 400 divisions of the *E. coli* K-12 MG1655 genome sequence (Blattner *et al.* 1997).

[b] R.F., restriction fragment, named for source gene(s).

[c] Dashes indicate that all strains are identical.

[d] E, location is next to edge of a hypervariable region.

man, J. Harrington and M. Thompson, unpublished results).

In Table 2, the variation in each hypervariable region decreases somewhat irregularly with increasing distance from its center. A rough index of variation, described in the footnote, is compiled for each fragment from the restriction fragment length polymorphism analyses of the Big Ten strains. The index value is 0 where all these strains are identical (*e.g.*, *fad* at 40.7 min and *yejF* at 49.0 min, which border the "45 min" hypervariable region), and additional cases of uniformity are seen farther out in each direction (Table 1, dashes).

Subsequent intermittent comparative nucleotide sequencing of the Big Ten strains on each side of the O-antigen region (which is located between ~45.25 and 45.51 min) showed dramatic variation. Figure 1 displays the variation on the clockwise side. Over a total of 15,680 nucleotide sites, the variation is again seen to decline with distance from *rfb*. To measure the nucleotide variation in a typical portion of the hypervariable region, an arbitrary breakpoint was chosen to exclude the outer regions in which the nucleotide variation has clearly declined. The chosen sequences consist of 12,213 nucleotide sites, of which 11,599 are monomorphic. The remaining 614 are polymorphic, each with one or occasionally two different substitutions, each present in up to 5 of the 11 strains. The excluded terminal portion of Figure 1 begins with G.S. 192: DLD 7765–9071.

Quantitative comparison of nucleotide variation in the hypervariable and nonhypervariable regions: To compare the amount of nucleotide variation typical of the hypervariable region with that of the nonhypervariable region, it was useful to quantify the variation in each region and to determine the ratio of the two amounts. Three *regional* comparisons (between hypervariable and nonhypervariable regions) were made and led to ratios of the order of 50.

In the first regional comparison, all 55 possible pairwise comparisons of the 11 sequences (see materials and methods) were performed. In the chosen hypervariable 12,213 sites, the 55 pairs contained a total of 13,521 differences, for a frequency of 1.107 differences per site. Similar pairwise comparisons for intermittent sets of 11 Big Ten sequences totaling ~126,800 bp assembled from throughout the nonhypervariable regions (R. Milkman, J. Harrington and M. Thompson, unpublished results) revealed 2699 differences, or 0.0213 per site. The ratio $1.107/0.0213 \cong 52$.

A second, similar comparison was made, using a different and much smaller sample running counterclockwise from *trpC* to *cls* in the nonhypervariable *trp* region (Milkman 1996; Rudd 1998). Here, the comparison of sequences from four Big Ten strains (K12, ECOR 1, ECOR 8, and ECOR 12) yielded 21 pairwise differences over 12,686 sites. Note that the four strains form only 6 possible pairs $[4!/(2 \times 2)]$ as opposed to the 55 pairs possible with 11 strains. For comparison, 21 was multiplied by 55/6, yielding 192.5 or 0.0152 differences per site. The ratio $1.107/0.0152 \cong 73$.

Another pragmatic measure of nucleotide variation is $n_e - 1$ (Milkman 1996), where n_e is the effective number of nucleotides (a measure analogous to the effective number of alleles, also symbolized as n_e; Crow and Kimura 1970). In this context, n_e is equal to $1/\Sigma p_i^2$, standing here for the inverse of the sum of the squared frequencies of the respective nucleotides at each given site of the 11 strains compared. Since the effective number of nucleotides in a monomorphic site is 1, the excess of n_e above 1 represents nucleotide variation. This quan-

TABLE 2

Levels of restriction site variation in hypervariable regions

Map Location				
Min.	G.S.	R.F.	I.V.	DNA sources (K12, ECOR strains)
95.1	380	aidB	0	K=1=2=3=5=8=9=10=11=12=25
96.1	385	tre	1.2	K=11;1=2=3=5=8=9=10=25*12
96.5	386	valS	4.2	K*1*3=5=8=10=25;2*9=12*11
97.2	389	fecA	0.4	2=8;K=3=5=9=10=11=12=25;1
97.8	392	fimBE	2.7	K=11=12#9*3;1=2=5=8*10=25
97.9	392	fimCD	3.4	K=11=12*10=25*5;1=2=8;9*3
98.3	394	yjiLM	2.2	3#K=2=8=9=11=12;5#1*10=25
	395-6	ICR		(extent 98.57–98.85 min)
98.9	396	tsr	4.5	K=9*1=3*2=8*5#11*10=12=25
99.3	398	prfC	6.2	K=9*1*2*3*5;11*8*10=12=25
0.7	3	carB	4.9	K=9#1=2=3;10=12=25*5;11*[No 8]
1.9	8	ilvI	0.5	K=1=2=3=5=9=10=11=12=25#8
E 2.1	8	ftsW	0	K=1=2=3=5=8=9=10=11=12=25
E 40.7	165	fad	0	K=1=2=3=5=8=9=10=11=12=25
40.8	165	sdaA	0.2	K=1=2=3=5=8=9=10=11=12;25
42.8	173	araG	1.5	K=1=2=3=12*5=8=25#9=10=11
43.2	175	fliD	2.6	2=3=12;K;8=25;10*5*1=9=11
43.8	177	dcm	4.9	K;5*3#1=8=11;9=10*2*12*25
45.0	183	hgd	6.6	K;11*1#9*3;(8;)25#5*2*10*12
45.1	183	hai	10.5	K*1*2*3#8#25*5#10*11*12*[No 9]
45.2	184	gnd	C	K*1*2*5*8*9*11*12 [No 3,10,25]
	184	*rfb*	D	(extent 45.25–45.51 min)
45.7	185	cpsB	9.0	K*1*2*3*5=12*8*9*10*11*25
45.8	185	fclgmd	3.1	1;K,2#8;11#10#5=12;3=9*25
45.8	185	wcaDE	4.9	K;1=5=12;8=11*2#9*3*10*25
45.9	185/6	wcaA	8.4	K*1*2;8*3=5=12*9*11;25*[No 10]
45.9	186	wzCBA	4.4	5=12#1;11;8*K=9=10#3*25*2
46.1	186	yeg1	1.7	3#K=1=2=5=8=12;9=10=11*25
46.2	186	yeg2	5.9	9;K=10#2*1=12;5*3*11*8*25
46.2	187	ak$^+$	2.4	11;K=9=10;25*1=3=5=8=12*2
46.7	188	gatc	2.2	5=12#1=2=3#K=10=11;8=9*25
47.0	189	189A$^+$	1.6	3;11;(9;)K;1=8=10*2=5=12=25
47.2	190	metG	4.0	11*1*K=2*3=5=12*8=9=10=25
48.2	194	mgl	1.2	25#(11#)K=2=3=5=8=9=10=12;1
48.3	194	cirA	0.7	9;K=1=2=3=5=8=10=12=25#11
48.4	195	cadR	0.2	K=1=2=3=5=8=9=10=12=25;11
48.6	196	fru	1.0	K=1=2=3=5=8=9=10=12=25*11
48.8	196	yeiQ	0.2	K=1=2=5=8=9=10=11=12=25;3
48.9	197	yejA	0.2	K=1=3=5=8=9=10=11=12=25;2
E 49.0	197	yejF	0	K=1=2=3=5=8=9=10=11=12=25

Restriction digests and terms as in Table 1. Both tables show the boundary fragments aidB, ftsW, fad, and yejF. $^+$See figure caption. IV, index of variation. ";" is one site difference (0.2 index point); #, two differences (0.5 index point); *, three or more differences (1.0 index point). Null fragment counts 2 index points. Parentheses contain branched connections: *e.g.*, at 48.2 in MGL, ECOR 25 and ECOR 11 each differ by one site from K12. Hypervariable region centers: ICR, immigration control region; *rfb*, O-antigen gene complex. C, includes variation at interspecific level; D, includes nonhomologous variation. Tables 1 and 2 both show the boundary fragments aidB, ftsW, fad, and yejF.

tity can be averaged over all observed sites, both polymorphic and monomorphic.

In this case, the 614 polymorphic sites have a total $n_e - 1$ of 393.63, while at each of the 11,599 monomorphic sites $n_e - 1$ is of course 0. The mean value of $n_e - 1$ for all sites is 393.63/12,213 = 0.032. This can be compared with a similar estimate for the nonhypervariable region, using the 126,800-bp sample referred to previously. Here, for all polymorphic sites containing nucleotide substitutions due to replacements or mutations, the values of $n_e - 1$ totaled 137.8. The mean value for all 126,800 sites, 137.8/126,800, came to 0.0011. The

ratio 0.032/0.0011 ≅ 29. The three ratios are thus 52, 73, and 29, with a mean ~51.

These calculations provide higher resolution and greater rigor than would comparisons of restriction analyses, given the rather arbitrary index of variation (IV) values and the occasional nonrandom placements of fragments chosen for analysis (see the second paragraph of RESULTS).

The nonhypervariable data were compared with Figure 1: From the nonhypervariable 126,800-bp sample's raw data it was calculated that a comparable 12,213-bp sample in the nonhypervariable regions would contain 52 polymorphic sites. Thus, a counterpart of Figure 1 for the nonhypervariable region would have 52 lines. Most lines would have a single variant dot, and a few would have two to four variant dots, reflecting a relatively low mean effective number of nucleotides, 65.4/52 = 1.26 per polymorphic site. These 52 lines contrast in number with the 614 lines in the compared hypervariable region (Figure 1), but the 12-fold difference in number of lines (614/52) is only part of the contrast. The rest is due to a greater-than-twofold difference in mean value of (n_e − 1) per polymorphic site between the hypervariable sample (394/614 = 0.64) and the nonhypervariable sample (13.4/52 = 0.26). The product (614/52) × (0.64/0.26) agrees with the ratio of 29 mentioned above.

A connection between specific O antigens and clockwise sequences was sought next: In six sets of two extremely closely related non-Big Ten ECOR strains, each pair was known to share a specific O antigen, and two of the pairs, which are not particularly related to one another, also share the same O antigen. This information is based on antigen identifications provided by T. S. Whittam (http://foodsafe.msu.edu/whittam/ecor) and independently confirmed more recently (AMOR *et al.* 2000). The supplementary appendix at http://www.genetics.org/supplemental/ shows sequence differences from K12 common to both members of each pair, as well as differences common over a limited range to the two distantly related pairs (ECOR 49/50 and ECOR 61/62) that share a specific O antigen. These sequences indicate that respective recent common ancestors of each of the six pairs received an O antigen and flanking DNA via intraspecific recombination. Furthermore, the common ancestor of ECOR 49 and 50 and the common ancestor of ECOR 61 and 62 received antigen O2 in independent but presumably contemporaneous replacements. Of these two independent replacements, at least one was quite short, since the sequence identity between the two pairs ends somewhere before genome section 187, in which at least one longer, previously acquired sequence is revealed. Note that an evident replacement closest to *rfb* is the most recent, and those more distant are remnants of progressively longer and older replacements. All of these replacements are anchored in the *rfb* region. Thus with increasing distance, the replacements that become newly evident are larger because of their common selected anchors, and they are less recent than the replacements that have ended. Shorter previous replacements would have been fully replaced.

One interesting feature of Figure 1 is the identity of ECOR 5 and ECOR 12 throughout: This detailed observation is supported in Table 2 from genome sections 185–190, except for one restriction-site difference in fragment YEG2 in G.S. 186. Elsewhere in Table 2, ECOR 5 and 12 show no special similarity counterclockwise to *rfb* or on either side of ICR. If ECOR 5 and 12 shared a given O antigen, the sequencing results would imply a shared ancestral gene transfer anchored in the *rfb* region, but the two strains have different O antigens. ECOR 5 has O79, and ECOR 12 has O7, according to T. S. Whittam (see above; not addressed by AMOR *et al.* 2000). A breakdown in identity was therefore sought in restriction digests closer to the heart of the region, and it was found in *galF*: a total of six differences were found in five of seven digests. This locus is immediately clockwise from *rfbB* (BERLYN 1998; RUDD 1998). However, in *cpsB*, which is 10 kb further clockwise, ECOR 5 and 12 show identical restriction patterns, as noted previously. This reflects an older, common replacement, which extends far into genome section 192. In G.S. 192 the majority of strains are alike, since most replacements have not extended that far.

Finally, the patterns of similarity in Figure 1 support the view that the observed variation is due largely to recombinational transfer, as opposed to recent mutation. Specifically, there are long tracts of deviant nucleotides in the last three genome sections illustrated (notably in ECOR 11), as well as easily identified patterns of similarity in sets of tracts throughout the figure. And the fact that K12 shows the same sorts of nucleotide distribution pattern here as do the other Big Ten strains confirms that its absence of an expressed O antigen "since at least the 1940s" (LIU and REEVES 1994), as well as its effective isolation during this period, corresponds to a mere instant in its individual evolution.

DISCUSSION

Repeated selection in the O-antigen region: Reeves and his group have developed a view of the evolution of the O antigen and its gene complex in Salmonella and *E. coli*, centering on the importation (spanning a great phylogenetic distance) of new nonhomologous genes involved in the synthesis and placement of sugars in the complex lipopolysaccharide and envisioning the continuing process of selection resulting in vast variability among naturally occurring O antigens (REEVES 1993). The recombinational retransmission of newly imported genes has also been inferred from the extreme and extensive variation in the DNA flanking the O-antigen


Compilation of Sequences in Region Clockwise from rfb
A C G T
Genome Section: Nucleotide Position Gene/Region
187: 109-883 AK
190: 7609-8861 METG (cont.)
190: 12994-14510 YEHI (cont.)
191: 8835-9324 YEHU
192: 4984-7264 BGLX (cont.)
192: 1900-3531; 3593-3725 YEHXY
192: 7265-7764 bridge-DLD
191: 5081-6666 YEHQ
190: 12994-14510 YEHI
192: 7765-9071 DLD
189: 2804-4265 189A
192: 4112-4695 YEHZ
194: 1269-1770; 2186-2677 MGLA
195: 469-778; 1058-1467 CADR
190: 7609-8861 METG
192: 4984-7264 BGLX
196: 3491-3936 FRU
K 1 2 3 5 8 9 10 11 12 25

region (Bisercic *et al.* 1991; Nelson and Selander 1994; Lan and Reeves 1996).

Circumstantial selection? The recent interpretation of the events following the arrival of a novel gene in the *rfb* region has generally included *frequency-dependent selection* (Reeves 1993; Selander *et al.* 1996; Milkman 1999). In fact, however, it has become apparent that there is no strict enduring rigorous relationship, direct or inverse, between the frequency of such a gene and its selective advantage. Rather, the course of events implies a series of step functions related to particular circumstances. When, for example, the *E. coli* cell bearing the novel gene arrives and produces an unrecognized O antigen in a bacterium in a host colon, it acquires a great selective advantage. This advantage is reduced to some extent with the rising frequency of the unrecognized cells, due to their mutual competition. The critical step, however, is the recognition of the new antigen by the host's immune system, causing a sudden drop to selective equality with the other, longer established *E. coli* cells in the colon. However, there is no indication that the new strain is now inferior or that its frequency declines. Instead, recombination among the various *E. coli* (and perhaps other) strains takes place, followed inevitably by egestion and occasional subsequent ingestion by other hosts. In a new, naive host, the recipient of the transferred novel gene regains its advantage—and often some flanking DNA differing in sequence from the DNA it has replaced. This sequence difference is likely to be homologous. And repetition of this series of events is likely to lead to increased nucleotide polymorphism, especially among strains that had not varied much before.

It makes sense that the imports come in small packages via plasmids, which are in general versatile agents of horizontal transfer, and it is evident that certain specific sequences regularly mediate the incorporation into the *rfb* region of the imported DNA (Hobbs and Reeves 1994). Recent (Vaisvila *et al.* 2001) and broader (Hall and Collis 1995) details that explicitly relate specific genetic systems (*integrons*) to horizontal gene transfer have been described.

Presumably the cost-benefit ratio of horizontal transfer across great phylogenetic distances increases sharply with the length of the incorporated segment. On the other hand, the inclusion of the imported gene in transferred *resident* flanking DNA incurs little cost. Thus the extent of high variation over 3–4 min (140–185 kb) on either side of the import site is not unexpected.

Presumably the longer stretches involve conjugation rather than transduction. Although bacteriophages with genomes in the 170-kb range that are potentially capable of transducing *E. coli* are known (Masters 1996, p. 2435), the abridgment (cutting and shortening) of incoming DNA fragments (McKane and Milkman 1995; Milkman *et al.* 1999) suggests that most transduced replacements are likely to be considerably smaller.

Rules relating to retention: Here is a simple extract (Crow and Kimura 1970, pp. 421–423; Milkman 1999) of some general parameters and rules that govern the entry and retention of a gene or allele in a large population of constant size (say 10 million or more individuals):

1. A new allele of a resident gene may arise by mutation or arrive by intraspecific recombination. A new gene (not a new form of a resident gene) may arrive directly or indirectly by recombination from a phylogenetically distant source.
2. A selection coefficient, s, is the differential rate of increase in numbers of a strain (*vs.* the rest of the population with which it competes) per unit time. In these circumstances, the time unit should be the mean generation time, to be comparable with the rate of random genetic drift. We are interested here only in a positive selection coefficient—a selective advantage for the cells containing the new allele or gene. If the population number is constant, and c is the average number of descendants that a novel gene (or cell) leaves in the next generation, then $s = \ln c$, and when s is small, $s = c - 1$. Particularly in bacteria, c is another way of expressing fitness, w.
3. The probability u_∞ of retention (survival for an indefinitely long time) of an individual mutant allele or arrival is approximately $2s$, if s is small and positive. Crow and Kimura (1970, p. 423) contains a table (*q.v.*) illustrating the probability that a novel gene will survive for a given number of generations as calculated iteratively from $u_t = 1 - e^{-cu_{t-1}}$ for various values of c. For the following higher values of c (in boldface type), the corresponding values of u_∞ and t^* (our notation for the generation when u_t first approximates u_∞) are **3**, 0.95048, 6; **4**, 0.98017, 4; **5**, 0.99302, 3; and **10**, 0.99995, 1. A "safe number" of

Figure 1.—Sequence variation in the Big Ten strains on the clockwise side of the O-antigen hypervariable region. Only polymorphic sites are displayed: each colored dot represents a specific nucleotide (red, A; yellow, C; lilac, G; and green, T). The 15 stretches total 15,680 nucleotide sites. Each heading includes genome section number, range of site numbers within the genome section, and an abbreviated reference to the sequenced region. For example, AK refers to *alkA* and 189A includes all of *yegU* (also called "b2099" in genome section 189) and about half of *yegV* ("b2100") in Rudd 1998. Symbols for K-12 and the 10 ECOR strains are on the line below. Fully annotated genome section files may be accessed from NCBI (http://www.ncbi.nlm.nih.gov/) by searching for "AE000" followed immediately by the genome section number + 110; thus genome section 189 corresponds to "AE000299" (see materials and methods). ▲, deleted nucleotide; ▼, TGG insert.

the alleles, N_s, giving a high probability of retention, is equal to about $1/u_\infty$ ($= 1/2s$).

These rules lead to the following conclusions:

1. It doesn't take a very large selective advantage for an allele or gene to remain at a large absolute number in a population.
2. It does take a considerable selective advantage for a single allele or gene to remain in a population after arising by mutation or arriving by horizontal transfer. (The rates of horizontal transfer over great phylogenetic distances are not yet known, nor is the range of variation of the rates from case to case.)

With a sufficient selective advantage, which can be much greater than one in the case of drug resistance, the retention of a new gene is essentially certain. The new gene is likely to increase greatly in number, as will its relative frequency, p.

The amount of flanking DNA that increases in number as it hitchhikes with the new gene depends upon the nature and rate of the intraspecific recombination that retransmits the new gene. In the absence of recombination, the clonal sweep will be genome-wide, like that envisioned in the periodic selection model of Atwood *et al.* (1951). In the presence of recombination, selective sweeps of a chromosomally local nature are of course possible. Recall that the effective recombination rate depends upon selection, which in this case derives from presence of the new gene.

In addition, when the selective advantage diminishes with increasing gene or allele frequency, the systematic increase in numbers declines; it stops when the gene is neutral. In the case of the O-antigen region and the immigration control region, this accounts for the additional complexity of the repetitive acquisition (and retransmission) of new variants. Cases like these, combining great selective advantage with circumstantial selection, evidently occur rarely, but the large hypervariable regions in the *E. coli* genome are clear cumulative evidence of their existence and importance.

Finally, it seems clear that, while intraspecific recombination may operate at a relatively uniform rate throughout the genome, it is the combination of recombination and selection that results in effective, or retained, replacement. Thus in each of these two large hypervariable regions, the sequence variation reflects the *differentiation of the species genome* of *E. coli*.

A brief return to the meaning of the "effective number of nucleotides": The significance of n_e averaged over a large number of sites is altered by tracts of a given specific distribution, which imply a common single replacement event rather than a group of independent events. In the same way, the interpretation of the observation of a high effective number of alleles at many loci is altered when their distribution among strains is uniform over the loci. In this case, the estimation of the number of independent recombinational replacement events in the nonhypervariable regions is fairly easy, while comparable estimation in the hypervariable regions can be difficult due to complexities related to the number and nature of donors.

The observations reported here confirm in new detail and extent the insight of P. R. Reeves and colleagues, whose experiments established in the past decade a broad solution to the paradox raised by local hypervariability in the *E. coli* genome, which seemed to contradict the general prediction of genome-wide clonality by the periodic selection model (Atwood *et al.* 1951; see also Whittam 1996). The prediction was addressed in depth and with clarity by Levin (1981, 1986), but the solution had to await new molecular techniques, notably PCR and easy DNA sequencing, which resolved the paradox of the "bastions of polymorphism" (Milkman 1999) amid the uniform genomes of spreading clones.

Important relevant information on the ECOR strains is contained in T. S. Whittam's website (http://foodsafe.msu.edu/whittam/ecor/). We thank Richard Melvin and Glenda Trimble for technical contributions and Michael Feiss for references on T4 transduction. This work was supported by grants MCB 9420613 and MCB 9728230 from the National Science Foundation to R.M., under which E.J. and R.M. held REU stipends.

LITERATURE CITED

Amor, K., D. E. Heinrichs, E. Frirdich, K. Ziebell, R. Johnson *et al.*, 2000 Distribution of core oligosaccharide types in lipopolysaccharides from *Escherichia coli*. Infect. Immun. **68:** 1116–1124.

Atwood, K. C., L. K. Schneider and F. J. Ryan, 1951 Selective mechanisms in bacteria. Cold Spring Harbor Symp. Quant. Biol. **16:** 345–355.

Barcus, V. A., and N. E. Murray, 1995 Barriers to recombination: restriction, pp. 31–58 in *Population Genetics of Bacteria*, edited by S. Baumberg, J. P. W. Young, E. M H. Wellington and J. R. Sanders. Cambridge University Press, Cambridge, UK.

Barcus, V. A., J. B. Titheradge and N. E. Murray, 1995 The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli*. Genetics **140:** 1187–1197.

Berlyn, M. K., 1998 Linkage map of *Escherichia coli* K-12, edition 10: the traditional map. Microbiol. Mol. Biol. Rev. **62:** 814–984.

Beyer, W. H. (Editor), 1966 *Handbook of Tables for Probability and Statistics*, p. 339. The Chemical Rubber Co., Cleveland.

Bisercic, 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.

Blattner, F., G. Plunkett, III, C. Bloch, N. T. Perna, M. Riley *et al.*, 1997 The complete genome sequence of *Escherichia coli*. Science **177:** 1453–1474.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.

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

Hall, R. M., and C. M. Collis, 1995 Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol. Microbiol. **15:** 593–600.

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.

Hobbs, M., and P. R. Reeves, 1994 The JUMPstart sequence: a 39

bp element common to several polysaccharide gene clusters. Mol. Microbiol. **12:** 855–856.

Lan, R., and P. R. Reeves, 1996 Gene transfer is a major factor in bacterial evolution. Mol. Biol. Evol. **13:** 47–55.

Levin, B. R., 1981 Periodic selection, infectious gene exchange, and the genetic structure of *E. coli* populations. Genetics **99:** 1–23.

Levin, B. R., 1986 Restriction-modification immunity and the maintenance of genetic diversity in bacterial populations, pp. 669–688 in *Evolutionary Processes and Theory*, edited by S. Karlin and E. Nevo. Academic Press, New York.

Li, Q., and P. R. Reeves, 2000 Genetic variation of dTDP-l-rhamnose pathway genes in *Salmonella enterica*. Microbiology **146:** 2291–2307.

Liu, D., and P. R. Reeves, 1994 *Escherichia coli* K-12 regains its O antigen. Microbiology **140:** 49–57.

Madigan, M. T., J. M. Martinko and J. Parker, 2002 *Brock Biology of Microorganisms*, Ed. 10. Prentice-Hall, Upper Saddle River, NJ.

Masters, M., 1996 Generalized transduction, pp. 2421–2441 in *Escherichia coli and Salmonella Cellular and Molecular Biology*, edited by F. C. Neidhardt. American Society for Microbiology, Washington, DC.

McKane, M., and R. Milkman, 1995 Transduction, restriction and recombination patterns in *Escherichia coli*. Genetics **139:** 35–43.

Milkman, R., 1996 Recombinational exchange among clonal populations, pp. 2663–2684 in *Escherichia coli and Salmonella Cellular and Molecular Biology*, edited by F. C. Neidhardt. American Society for Microbiology, Washington, DC.

Milkman, R., 1997 Recombination and population structure in *Escherichia coli*. Genetics **146:** 745–750.

Milkman, R., 1999 Gene transfer in *Escherichia coli*, pp. 291–309 in *Organization of the Prokaryotic Genome*, edited by R. L. Charlebois. American Society for Microbiology, Washington, DC.

Milkman, R., and M. McKane, 1995 DNA sequence variation and recombination in *E. coli*, pp. 127–142 in *Population Genetics of Bacteria*, edited by S. Baumberg, J. P. W. Young, E. M. H. Wellington and J. R. Saunders. Cambridge University Press, Cambridge, UK.

Milkman, R., and M. McKane Bridges, 1990 Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics **126:** 505–517.

Milkman, R., E. A. Raleigh, M. McKane, D. Cryderman, P. Bilodeau *et al.*, 1999 Molecular evolution of the *Escherichia coli* chromosome. V. Recombination patterns among strains of diverse origin. Genetics **153:** 539–554.

Nelson, K., and R. K. Selander, 1994 Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. USA **91:** 10227–10231.

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

Raleigh, E. A., 1992 Organization and function of the *mcrBC* genes of *Escherichia coli* K-12. Mol. Microbiol. **6:** 1079–1086.

Reeves, P., 1993 Evolution of *Salmonella* O antigen variation by interspecific gene transfer on a large scale. Trends Genet. **9:** 17–22.

Reeves, P., 1995 Role of O-antigen variation in the immune response. Trends Microbiol. **3:** 381–386.

Rudd, K. E., 1998 Linkage map of *Escherichia coli* K-12, edition 10: the physical map. Microbiol. Mol. Biol. Rev. **62:** 985–1019.

Salyers, A. A., and D. D. Whitt, 2001 *Microbiology: Diversity, Disease and the Environment*, pp. 252–255. Fitzgerald Science Press, Bethesda, MD.

Selander, R. K., J. Li and K. Nelson, 1996 Evolutionary genetics of *Salmonella enterica*, pp. 2691–2707 in *Escherichia coli and Salmonella Cellular and Molecular Biology*, edited by F. C. Neidhardt. American Society for Microbiology, Washington, DC.

Stevenson, G., B. Neal, D. Liu, M. Hobbs, N. H. Packer *et al.*, 1994 Structure of the O antigen of *Escherichia coli* K-12 and the sequence of its *rfb* gene cluster. J. Bacteriol. **176:** 4144–4156.

Vaisvila, R., R. D. Morgan, J. Postal and E. A. Raleigh, 2001 Discovery and distribution of super-integrons among Pseudomonads. Mol. Microbiol. **42:** 587–601.

Whitfield, C., 1995 Biosynthesis of lipopolysaccharide O antigens. Trends Microbiol. **3:** 178–185.

Whittam, T. S., 1996 Genetic variation and evolutionary processes in natural populations of *Escherichia coli*, pp. 2708–2720 in *Escherichia coli and Salmonella Cellular and Molecular Biology*, edited by F. C. Neidhardt. American Society for Microbiology, Washington, DC.

Yao, Z., and M. A. Valvano, 1994 Genetic analysis of the O-specific lipopolysaccharide biosynthesis region (*rfb*) of *Escherichia coli* K-12 W3110: identification of genes that confer group 6 specificity to *Shigella flexneri* serotypes Y and 4a. J. Bacteriol. **176:** 4133–4143.

Communicating editor: S. Yokoyama