APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Apr. 2004, p. 2464–2473 Vol. 70, No. 4
0099-2240/04/$08.00+0 DOI: 10.1128/AEM.70.4.2464–2473.2004

Phylogeny and Strain Typing of *Escherichia coli*, Inferred from Variation at Mononucleotide Repeat Loci

Eran Diamant,† Yniv Palti,†‡ Riva Gur-Arie, Helit Cohen, Eric M. Hallerman,§ and Yechezkel Kashi*

Department of Food Engineering and Biotechnology, Technion—Israel Institute of Technology, Haifa 32000, Israel

Received 10 September 2003/Accepted 13 January 2004

Multilocus sequencing of housekeeping genes has been used previously for bacterial strain typing and for inferring evolutionary relationships among strains of *Escherichia coli*. In this study, we used shorter intergenic sequences that contained simple sequence repeats (SSRs) of repeating mononucleotide motifs (mononucleotide repeats [MNRs]) to infer the phylogeny of pathogenic and commensal *E. coli* strains. Seven noncoding loci (four MNRs and three non-SSRs) were sequenced in 27 strains, including enterohemorrhagic (six isolates of O157:H7), enteropathogenic, enterotoxigenic, B, and K-12 strains. The four MNRs were also sequenced in 20 representative strains of the *E. coli* reference (ECOR) collection. Sequence polymorphism was significantly higher at the MNR loci, including the flanking sequences, indicating a higher mutation rate in the sequences flanking the MNR tracts. The four MNR loci were amplifiable by PCR in the standard ECOR A, B1, and D groups, but only one (*yaiN*) in the B2 group was amplified, which is consistent with previous studies that suggested that B2 is the most ancient group. High sequence compatibility was found between the four MNR loci, indicating that they are in the same clonal frame. The phylogenetic trees that were constructed from the sequence data were in good agreement with those of previous studies that used multilocus enzyme electrophoresis. The results demonstrate that MNR loci are useful for inferring phylogenetic relationships and provide much higher sequence variation than housekeeping genes. Therefore, the use of MNR loci for multilocus sequence typing should prove efficient for clinical diagnostics, epidemiology, and evolutionary study of bacteria.

Escherichia coli is a species of gram-negative bacterium that includes numerous strains and serotypes (1, 31). The species includes commensal strains and a variety of pathogenic groups, including enterohemorrhagic *E. coli* (EHEC), enteropathogenic *E. coli* (EPEC) and enterotoxigenic *E. coli* (ETEC) strains. EHEC strains, also known as Shiga-like toxin-producing *E. coli* strains, are associated with a variety of medical syndromes, including diarrhea, hemorrhagic colitis, and hemolytic uremic syndrome. EHEC serotype O157:H7 is one of the primary food-borne pathogenic threats in Europe and North America (40, 50). EPEC and ETEC strains are major causes of dehydrating infant diarrhea and infections correlated with adverse nutritional consequences in developing countries (22, 40).

Subdivision of bacterial strains is based on various O:H serotypes, discrete virulence and adherence properties, and distinct clinical phenotypes. Electrophoretic allozyme typing has been used to assess genetic relationships among *E. coli* strains in several studies (31, 32, 35, 41, 49, 50). In these studies, it was demonstrated that O:H serotyping is not sufficient for defining phylogenetic relatedness among *E. coli* strains. Analysis of the DNA sequences of housekeeping genes has also been used to study phylogenetic relationships among *E. coli* strains (21, 26, 35). Reid et al. (36) used sequence data for seven housekeeping genes to elucidate the evolution of pathogenic mechanisms by inferring phylogenetic relationships among 14 EHEC, EPEC, and K-12 strains.

Simple sequence repeats (SSRs, or microsatellites) are a class of DNA sequences consisting of simple motifs that are tandemly repeated at a locus (47). SSRs have long been known to be distributed throughout the genomes of eukaryotes, highly polymorphic (43, 48), and useful as tools for phylogenetic inference (5). The variability observed in SSRs is thought to be caused by slipped-strand mispairing. The abnormal tertiary structure of repetitive DNA allows mismatching of neighboring sequences, and repeats can be inserted or deleted during DNA duplication (reference 46 and references therein). Screening of prokaryotic genomes for SSRs has revealed large numbers of SSR tracts (7–9, 45, 46). Publication of the complete genome sequence for *E. coli* (2) provided the basis for characterizing SSR tracts in this organism, both genomewide and at particular loci (8, 9, 24). PCR-based assays have been developed in our laboratory for screening SSR polymorphism in different *E. coli* strains. Mononucleotide SSRs (mononucleotide repeats [MNRs]), consisting of at least five repeats, were found to be abundant and polymorphic in noncoding regions of the *E. coli* genome (9, 24).

Inference of evolutionary relationships within *E. coli* is not

* Corresponding author. Mailing address: Department of Food Engineering and Biotechnology, Technion — Israel Institute of Technology, Haifa 32000, Israel. Phone: 97248293074. Fax: 97248293399. E-mail: kashi@tx.technion.ac.il.

† E.D. and Y.P. contributed equally to this article.

‡ Present address: National Center for Cool and Cold Water Aquaculture, USDA-ARS, Kearneysville, WV 25430.

§ Permanent address: Department of Fisheries and Wildlife Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

TABLE 1. Details of *E. coli* strains analyzed in this study

E. coli group	Description and source or reference	Strains (including serotypes)
Wild types	Michigan State University (32)	ECOR collection strains 1–72
K-12, B	Technion Faculty of Food Engineering & Biotechnology collection	DH5α, W3110, W4100 (K-12) SR9b, SR9c (B)
EHEC	*E. coli* Reference Center, 90.032	O22:H8
	E. coli Reference Center, 88.0501	O42:H2
	E. coli Reference Center, 88.0015	O111:H$^-$
	E. coli Reference Center, 88.0632	O113:H2
	Centers for Disease Control, CDC 2239-69	O26:H11
	USDA-FSIS, MF7123A	O157:H$^-$
	USFDA, SEA13B88, Odwalla cider outbreak strain	O157:H7
	Ontario Public Health Laboratory, (1)	O157:H7 HER phage types 1057, 1058, 1261, 1265, 1266
ETEC	Haifa Public Health Dept., Rowe no. E10407	O78:H$^-$
	Central Laboratories, Jerusalem, Israel Ministry of Health	O8:H9, O9:H33, O86:H10, O86:H18, O153:H$^-$
EPEC	Haifa Public Health Dept., Rowe no. E639616	O111ac:H$^-$
	Central Laboratories, Jerusalem, Israel Ministry of Health	O26:H$^-$, O55:H7, O127:H21

trivial due to horizontal transfer of genomic sequences between strains and species, which leads to fragmentation of the "clonal frame" (25). The inferred evolutionary history of a particular lineage may differ among different parts of its genome. Thus, combining data from two loci may obscure the reconstruction of either history (21). Statistical methods for identifying regions of recombination and for assessing its impact on phylogenetic reconstruction rely on large numbers of polymorphic genetic markers distributed throughout the genome. Therefore, screening of multiple polymorphic MNR loci is appropriate for supporting inferences about evolutionary relationships within *E. coli* and as a model system for phylogenetic studies in bacteria.

Metzgar et al. (24) found poor consistency between phylogenetic trees constructed by amplification fragment size analysis of SSRs and the standard *E. coli* reference (ECOR) tree of Herzer et al. (12). They concluded that individual SSRs mutate too frequently to retain meaningful phylogenetic information at the evolutionary scale represented by the standard ECOR tree. However, we hypothesize that by combining sequence data from as many loci as possible, MNRs may be used to construct phylogenetic trees which are consistent with those reported in previous studies that used other genetic markers. Additionally, we hypothesize that sequence variation at the flanking regions of the MNRs contains important information that can aid in the reconstruction of evolutionary relationships. Hence, in this study, we addressed two questions: Are randomly selected noncoding loci that contain MNRs more polymorphic at the sequence level than noncoding loci that do not contain SSRs in *E. coli*? How useful are those MNRs and their flanking sequences for inferring evolutionary relationships in *E. coli* when analyzed for sequence polymorphism?

Seven noncoding loci (four MNRs and three non-SSRs) were sequenced in 27 EHEC, EPEC, ETEC, B, and K-12 strains to address the first question. The MNR loci were also examined in the 72 strains of the ECOR collection (31) to enable comparison of the inferred evolutionary relationships with those of previous studies (12, 24, 35, 36, 50). The underlying goal of this study was to test the utility of MNRs for phylogenetic studies and strain identification in *E. coli* as a model system for prokaryotes.

MATERIALS AND METHODS

Bacterial strains. The study included 99 pathogenic and nonpathogenic strains of *E. coli* (Table 1). To verify that the ECOR strains are identical to the clones used in previous studies, we sequenced a PCR-amplified fragment of the *trpC* gene (21) in nine of them, ECOR strains 4, 8, 26, 27, 30, 52, 54, 60, and 70. The primers used to amplify a 500-bp *trpC* fragment were 5′-GGATTAACACCTA TGGTCAG-3′ (forward) and 5′-CCGCCAGTAACACATTATC-3′ (reverse). All sequences were identical to the GenBank sequences of these strains.

Locus selection and primer construction. The complete genomic sequence of *E. coli* K-12 (2) was obtained from http://mol.genes.nig.ac.jp/ecoli/. PCR primers for fragment amplification were designed from one open reading frame to the adjacent downstream open reading frame as previously described (9). All primer pairs were based on the K-12 sequence (Table 2). Loci were named after the downstream open reading frame. The loci were selected at random from various sites of the *E. coli* genome. The locations of the loci on the circular chromosome are shown in Fig. 1. Four of the loci (*b2345*, *ykgE*, *ycgW*, and *yaiN*) contain MNRs, and three (*pepD*, *galS*, and *osmB*) do not contain SSRs.

TABLE 2. PCR primers used for amplification of given loci, based on genomic sequence in *E. coli* K-12

Locus	MNR type	Size (bp)	MNR length (bp)	Forward primer, 5′-to-3′	Reverse primer, 5′-to-3′	T_m (°C)	GenBank accession no.
yaiN	Poly(G)	250	10	AATTTATCCGGTGAATGTGGT	GGACGCCAGAAACACGCTAC	64	AY208980
ycgW	Poly(C)	205	8	GTCAATTACTATCCACTTCA	TTAATTACAGGATGTTCAGTC	57	AY208979
b2345	Poly(C)	206	8	GTTCTGCTCGTCCTTCTC	GTTGATTGAAATAGATGGTAGC	58	AY208974
ykgE	Poly(G)	203	7	ATTGCATTTGACGTTTGG	AGGGCGTCACCAATACA	52	AY208975
osmB	No SSR	140		TCACGGAAGTAAGCTCT	CTAAGATGATTCCTGGTTG	51	AY208976
galS	No SSR	120		CGCTACATCACGAATGGTG	AAACATCAAGATAACGATCTGG	59	AY208977
pepD	No SSR	251		GGAGATAATTGAGACAGTTCAG	ATGTCCCAGGTGACGATG	56	AY208978

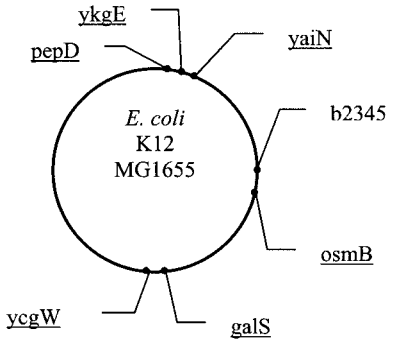


FIG. 1. Genomic locations of the noncoding loci sequenced. The locations are based on the sequenced *E. coli* K-12 genome (2) as follows: *pepD*, kb 254; *ykgE*, kb 321; *yaiN*, kb 379; *ycgW*, kb 1211; *osmB*, kb 1341; *gals*, kb 2239; and *b2345*, kb 2461.

DNA extraction and PCR amplification. Extraction of genomic DNA and the PCR amplification protocol were performed as previously described (9).

DNA sequencing. PCR products were purified (QIAquick; Qiagen) and sequenced on both strands on an ABI 310 automated DNA sequencer with the BigDye terminator cycle sequencing kit (Applied Biosystems, Inc.), following established procedures (3). Only sequences with complete agreement between the two strands were used for further analysis. Multiple alignment of the sequences was performed with the Sequence Navigator program (version 1.0.1; Applied Biosystems, Inc.).

Sequence polymorphism: comparison between MNR-containing loci and non-SSR loci. Two methods were used to assess the level of sequence variation at each locus. In the first, the percentage of polymorphic positions in a locus (single nucleotide polymorphisms [SNPs]) divided by the number of strains that were amplified in that locus was determined: polymorphism = [(SNP/SEQ) × 100]/no. of strains, where polymorphism is the fraction of polymorphic positions, SNP is the number of SNPs, SEQ is the length of the sequence in base pairs, and the number of strains is the number of strains that were amplified by PCR. The second method was the same except that the core repeat motifs within the MNRs were not included in the calculation of polymorphism. This was done to quantify the level of polymorphism in the sequences flanking MNR loci.

The number of microhaplotypes (set of specific mutations within limited chromosomal regions) observed at each locus was normalized as a fraction of the sequence length and number of strains that were amplified as in the first method. Two sequences were considered different genotypes if they contained at least one position at which they differed.

Student's *t* test for unequal variances (27) was conducted to compare the level of polymorphism between MNR and non-SSR loci as calculated by each of the two methods.

The comparison of sequence polymorphism between MNR and non-SSR loci was conducted in 27 strains (not including ECOR strains [Table 1]).

Phylogenetic analysis. The program Reticulate was used to identify putative recombination between loci through the construction of a compatibility matrix (13). The program START, version 1.0.4, written by Keith Jolley, University of Oxford, 2000 (Index of Association implementation modified from code provided by John Maynard-Smith, United Kingdom), was used to calculate the index of association (I_A) value for the four MNR loci. An I_A value that is not significantly different from 0 indicates that the loci may be incompatible (42). Loci that show incompatibility should not be combined for phylogenetic analysis, since their incompatibility may be the result of horizontal gene transfer (21).

Phylogenetic trees were inferred for each of the four MNR loci and for the combined sequence of all four loci. Trees were constructed by the unweighted pair-group method with arithmetic means (UPGMA) and neighbor-joining (NJ) algorithms (MEGA software) (29) and by the method of maximum parsimony (MP) (6). Bootstrap confidence values for the UPGMA and NJ algorithms were based on 1,000 simulated trees.

RESULTS

PCR amplification and sequence analysis. Twenty of the 72 ECOR strains were sequenced at the four MNR loci for phylogenetic analysis. The strains selected represent the different phylogenetic groups of Herzer et al. (12) and also supported maximum PCR amplification at the four MNR loci (Table 3); 27 non-ECOR strains were sequenced at each of seven noncoding loci (four MNRs and three non-SSRs) that produced PCR amplification.

The presence or absence of PCR amplification varied among loci; while *yaiN* was the only locus which was amplified in the B2 group, *b2345* and *ycgW* were amplifiable in almost all A, B1, D, and E strains but not in B2 strains (Table 3).

In order to simplify the sequence analysis, all the polymorphic sites were joined to create an artificial core sequence alignment. For example, the core sequence of *yaiN* is 13-bp point mutations different from the consensus 198-bp sequence. The alignment results for four MNR loci in 41 strains is presented in Fig. 2.

Sequence polymorphism. MNR tracts showed high levels of polymorphism with up to five alleles despite their short length. The polymorphism of the MNR loci was significantly higher than the polymorphism of the noncoding loci that did not contain SSRs, as evidenced by comparing both the number of polymorphic sites and the number of microhaplotypes at each locus (Table 4). The sequences flanking the MNRs were found to be more polymorphic than those flanking non-SSR loci at a nearly significant level ($P = 0.053$). These findings suggest that the presence of MNR tracts at these loci correlates with their increased variable nature.

Phylogenetic analysis. A compatibility matrix (Fig. 3a) was derived from comparison of all possible pairs of 68 sites. Two sites were defined to be compatible if all nucleotide changes at both sites can be inferred to have occurred only once, indicating a common evolutionary history of the sequences (13).

The compatibility can be separated into two components, a within-locus component based on the comparison of pairs of sites at the same locus, and a between-loci component based on the comparison of sites between loci. Overall, 92.7% of the pairs within a locus were compatible, and 79.6% were compatible between loci. The lowest level of compatibility for both components was presented by *yaiN* and the highest by *ykgE* (Fig. 3b). The mean I_A value for the four MNR loci was 0.46, with a standard error of 0.069, indicating that the loci were compatible.

A phylogenetic tree was constructed based on each of the loci separately (supplemental material may be found at http://www.technion.ac.il/technion/food/) and on the multilocus sequence of the four MNR loci (Fig. 4).

DISCUSSION

Previous studies demonstrated that MNRs in noncoding regions of the *E. coli* and other prokaryotic genomes are abundant and polymorphic (8, 9, 15, 24, 46). These characteristics suggest that SSRs may be markers of choice for evolutionary studies, as was previously shown for eukaryotes (5). However, an attempt to use fragment size analysis of mononucleotide SSR amplicons produced phylogenetic trees that were inconsistent with the accepted tree (24).

Our goal in this phylogenetic analysis was to utilize polymorphic DNA sequences by combining four MNR loci for multilocus sequence typing (MLST) analysis. We found that se-

TABLE 3. PCR amplification data for four MNR loci in the 72 ECOR strains

Group[a]	Strain	Amplification[b]					Group[a]	Strain	Amplification[b]				
		YaiN	*ycgW*	*b2345*	*ykgE*	Seq			*YaiN*	*ycgW*	*b2345*	*ykgE*	Seq
A	EC01	+	+	+	+	Y		EC68	+	+	+	+	Y
	EC02	+	+	+	+			EC69	+	−	+	+	
	EC03	+	+	+	+			EC70	+	+	+	+	Y
	EC04	+	+	+	+	Y		EC71	+	+	+	+	
	EC05	+	+	+	+			EC72	+	+	+	+	
	EC06	+	−	+	−								
	EC07	+	+	+	+		B2	EC51	+	−	−	−	
	EC08	+	+	+	+	Y		EC52	+	−	−	−	
	EC09	+	+	+	+			EC53	+	−	−	−	
	EC10	+	+	+	+	Y		EC54	+	−	−	−	
	EC11	+	+	+	+			EC55	+	−	−	−	
	EC12	+	+	+	+			EC56	−	−	−	−	
	EC13	+	+	+	+			EC57	+	−	−	−	
	EC14	+	+	+	+	Y		EC59	+	−	−	−	
	EC15	+	+	+	+	Y		EC60	+	−	−	−	
	EC16	+	+	+	+			EC61	+	−	−	−	
	EC17	+	+	+	+			EC62	+	−	−	−	
	EC18	+	+	+	+			EC63	+	−	−	−	
	EC19	+	+	+	+			EC64	+	−	−	−	
	EC20	+	−	+	+			EC65	+	−	−	+	
	EC21	+	−	+	+								
	EC22	+	+	+	+		D	EC35	+	+	+	−	
	EC23	+	+	+	+			EC36	+	+	+	−	
	EC24	+	+	+	+			EC38	+	−	+	−	
	EC25	+	+	+	+			EC39	+	+	+	−	
								EC40	+	+	+	−	Y
B1	EC26	+	+	+	+			EC41	+	+	+	−	
	EC27	+	+	+	+	Y		EC44	+	+	+	+	Y
	EC28	+	+	+	+	Y		EC46	+	+	+	+	Y
	EC29	+	+	+	+			EC47	+	+	+	−	
	EC30	+	+	+	+			EC48	+	+	+	−	
	EC32	+	+	+	+	Y		EC49	+	+	+	+	Y
	EC33	+	+	+	+			EC50	+	+	+	+	Y
	EC34	+	+	+	+								
	EC45	+	+	+	+		E	EC31	−	+	+	+	
	EC58	−	−	+	+			EC37	−	+	+	+	Y
	EC66	−	−	−	−			EC42	−	+	+	+	Y
	EC67	+	+	+	+			EC43	−	−	+	+	

[a] Group classification follows Herzer et al. (12).
[b] +, successful PCR amplification; −, no PCR amplification. SeqY, PCR fragments were further sequenced for the strain.

quence analysis of MNR loci produced phylogenetic trees that are in good agreement with those constructed by use of other genetic markers.

A panel of the ECOR strains that were amplifiable at most of the MNR loci and that represented the five phylogenetic groups of Herzer et al. (12) was chosen for the phylogenetic study. Based on the rooted phylogenetic analysis of Lecointre et al. (21), B2 is the most ancient *E. coli* group, followed by D and then the sister groups A and B1. Since *ycgW*, *b2345*, and *ykgE* were not amplifiable in the B2 group (Table 3), they likely were transferred into the genome of the common ancestor of the D group (*ycgW* and *b2345*) or during the early evolution of D group before the segregation of the A and B1 clades (*ykgE*). These PCR amplification results (Table 3) can support the hypothesis that *yaiN* is the oldest locus of the four, followed by *ycgW* and *b2345* and then *ykgE*.

The divergence of bacterial strains in nature is accelerated by the high rate of recombination, which may indicate horizontal gene transfer, resulting in fragmentation of the clonal frame of each strain (10, 25). Loci identified as recombination hot spots should be removed from the multilocus analysis because they may represent different evolutionary histories (21). Several approaches for inferring recombination and horizontal gene transfer in bacteria from DNA sequence data have been described (4, 13, 19, 20, 33, 42). Due to the availability of appropriate software and their wide use in other studies, we used the compatibility matrix approach (13) and calculated an association index for the four loci (42). Both approaches aim to examine the sequence compatibility between loci. A low level of compatibility indicates that the loci have experienced several changes due to recombination or repeated mutations at specific sites. The analysis of our sequence data by the two methods suggested that the four loci are compatible. The order of sequence compatibility plotted in Fig. 3b is in agreement with the hypothesis that we drew from the PCR amplification analysis: the longest-evolving locus, *yaiN*, had the lowest within-locus and between-loci compatibility levels, followed by *b2345* and *ycgW* and then *ykgE*, which had the highest between-loci compatibility level. However, it is important to note that the source of *yaiN*'s low between-loci compatibility is its hy-

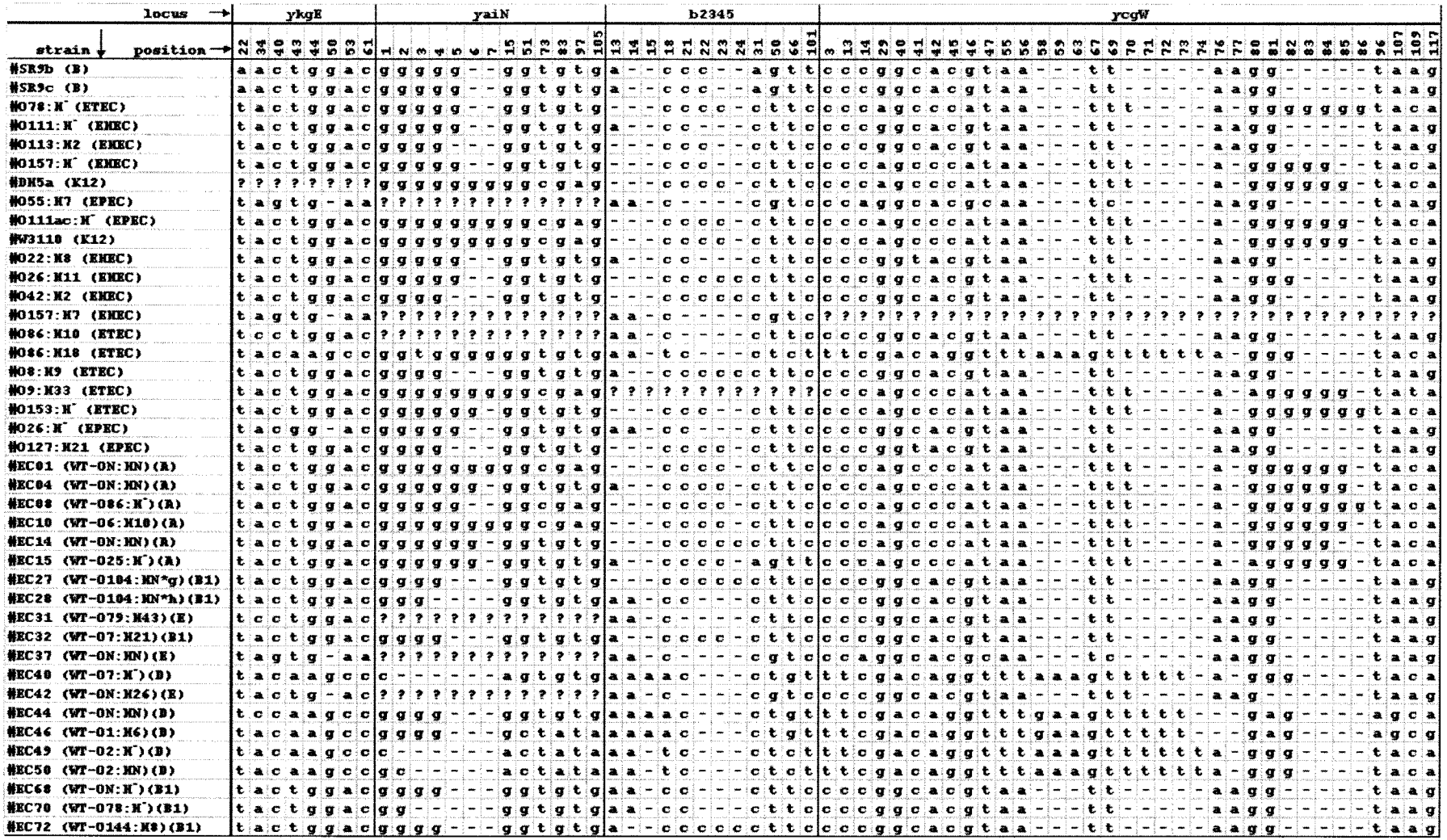


FIG. 2. Core artificial sequence alignment. The polymorphic positions (SNPs, insertions, and deletions) at four MNR loci were joined to create a reduced sequence alignment (68 polymorphic points along 534 bp) for 41 *E. coli* strains. Numbers above each column indicate the position at the specific locus.?, entire locus was not amplified in this strain; −, the specific position was absent (i.e., deleted).

TABLE 4. Levels of polymorphism at MNR loci compared to non-SSR type loci among 27 *E. coli* strains[a]

Polymorphic site	Locus	Length of sequence (bp)	No. of strains sequenced	MNR + flanking sequence[b]		Flanking sequence only[c]		Microhaplotypes	
				No. of polymorphic sites	Polymorphism level	No. of polymorphic sites	Polymorphism level	No. of microhaplotypes	Polymorphism level
MNR	*yaiN*	198	21	5	0.12	2	0.05	5	0.12
	ycqW	118	21	17	0.69	10	0.40	10	0.40
	b2345	128	26	11	0.33	6	0.18	11	0.33
	ykqE	90	25	6	0.27	5	0.22	6	0.27
	Mean ± SD				0.351 ± 0.015		0.214 ± 0.005		0.280 ± 0.004
Non-SSR	*osmB*	140	26	4	0.11	4	0.11	7	0.19
	qalS	120	27	1	0.03	1	0.03	3	0.09
	pepD	251	27	0	0.00	0	0.00	1	0.01
	Mean ± SD				0.047 ± 0.001		0.047 ± 0.001		0.100 ± 0.003
	P[d]				0.046		0.053		0.036

[a] Non-SSR loci were sequenced in a 27-strain panel; hence, the comparison to MNR loci was done for the same 27 strains.
[b] Mean number of polymorphic sites when the MNRs are considered as multiple sites.
[c] Mean number of polymorphic sites in sequences flanking the MNR.
[d] One-tailed *P* value (Student's *t* test) with unequal variances.

pervariable poly(G) MNR. Both *ycgW* and *b2345* had higher within-locus compatibility values. This high level of within-locus conservation may be related to their functions, which may be subject to selection pressures.

A multilocus tree was constructed from the combined sequence of the four loci (Fig. 4), and for each locus separately (http://www.technion.ac.il/technion/food/). We found very good agreement among the three methods of phylogenetic analysis that we used (NJ, UPGMA, and MP). Groups A and D clustered as expected, and group B1 strains branched separately from A and D but did not cluster with each other. Group B2 is not present because it did not amplify at three of the four loci.

The six O157:H7 outbreak serotypes that we examined had completely identical sequences at the loci examined. O157:H7 clustered tightly with O55:H7 and the ungrouped EC37, which is consistent with findings in previous studies (35, 50). O55:H7 and O157:H7 evolved recently from a common ancestor and are more likely to be distinguished from each other by the presence or absence of the PCR amplicon due to the high rate of insertion and deletion events in the O157 genome (17, 30). EC42 was closely related to this cluster in the MP and NJ trees but not in the UPGMA tree (Fig. 4). It was found by Pupo et al. (35) to cluster with O157:H7 and EC37.

The K-12 strains O111ac:H$^-$ (EPEC) and O78:H$^-$ (ETEC) clustered with group A strains in the multilocus analysis and in each of the trees constructed for the individual loci. The clustering of these strains with group A is also consistent with the results of previous studies (11, 35). The clustering of O86:H18 (ETEC) with group D was a consensus in the three multilocus trees and was also evident in all the single-locus trees except that for *yaiN*.

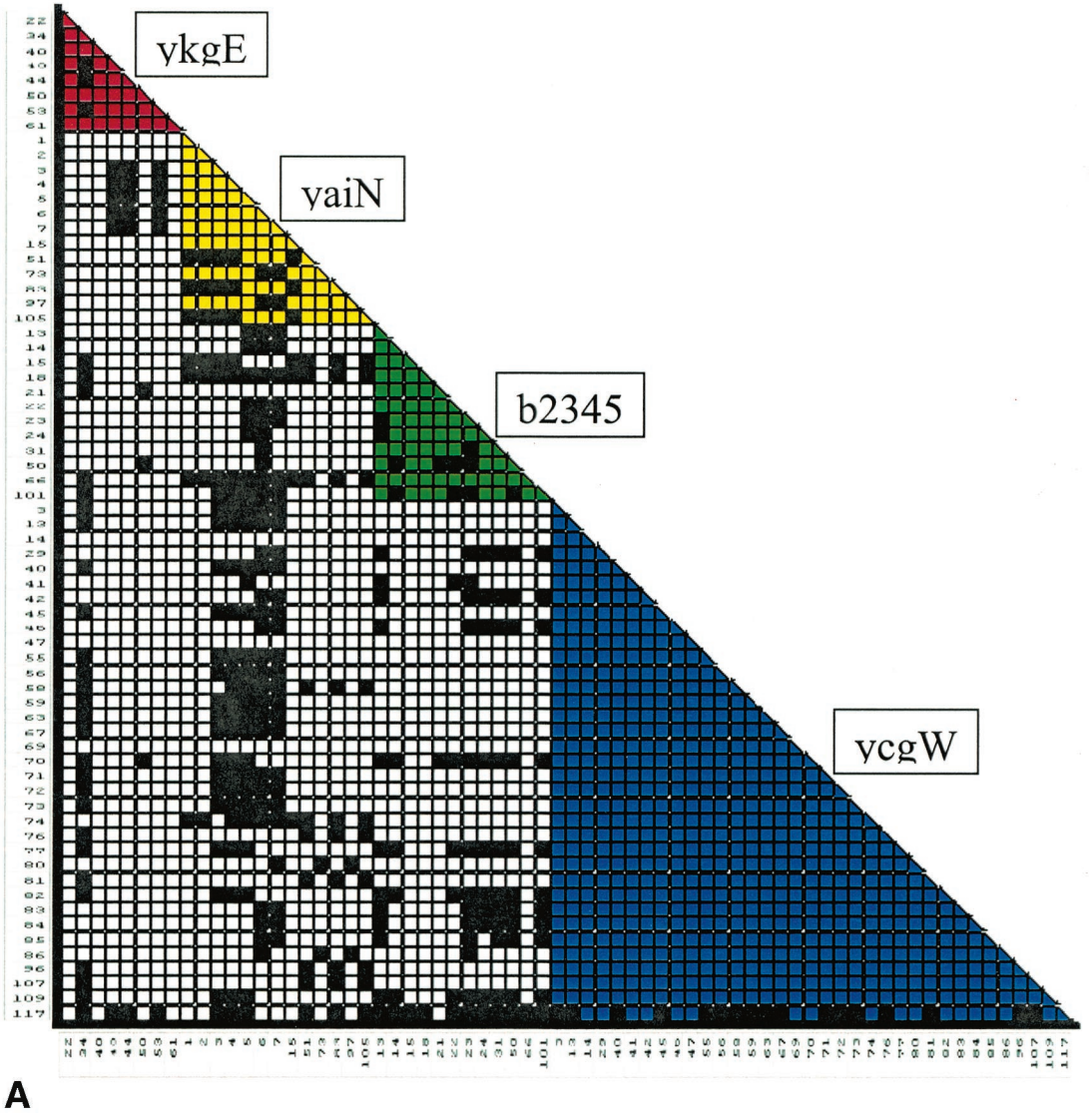
An illustration of the increased resolving power of the multilocus analysis was found in the clustering of EC08 with O78:H$^-$ (ETEC) and O157:H$^-$ (EHEC) with O153:H$^-$(ETEC). The two clusters were not distinct in any of the single-locus trees but were distinct in each of the three multilocus trees. Further support for the advantage of multilocus over single-locus phylogenetic analysis was recently reported by Rokas et al. (38)

Most of the EPEC and EHEC strains did not cluster with either group A or D. With the addition of loci to the analysis and increase in resolution, they may cluster with B1 strains or with strains from the rapidly evolving ungrouped (E) category. However, since they were amplified at all loci, it is unlikely that they are closely related to the B2 group.

MLST of a number of fragments from housekeeping genes is widely used for evolutionary studies and has been put forward as a powerful tool for "global" epidemiology (23, 30, 36). DNA sequencing provides far more variation per locus than any other method currently used for bacterial strain typing, and it provides a uniform platform for comparison between laboratories and for database storage. Noncoding loci that contain mononucleotide SSRs were significantly more polymorphic at the sequence level than loci that did not contain SSRs (Table 4). Combining several polymorphic SSR loci enables the use of these sequences for SSR-based MLST. In this study, we demonstrated that MLST of MNRs from randomly chosen noncoding regions is as consistent and reliable as MLST of housekeeping genes. The advantage of MNRs is that they provide much higher variation per base.

The level of resolution that requires sequencing thousands of base pairs from housekeeping genes can be achieved by sequencing hundreds of base pairs from MNR loci. This should be most cost effective in clonal species such as *E. coli* and should make MLST a more rapid and affordable tool for epidemiology and clinical diagnostics.

Recent studies demonstrating that O157:H7 is rapidly evolving by unknown mechanisms of insertion and deletion of genomic fragments (17, 34) suggested a method for typing of O157 strains based on the absence or presence of specific PCR amplifications. In this study, we found that the amplification of randomly chosen sites of the *E. coli* genome can be useful for "local" typing of closely related strains (e.g., O55:H7 and O157:H7 with *ycgW*) as well as "global" typing for distinguishing between subspecies (of the four MNRs, only *yaiN* was



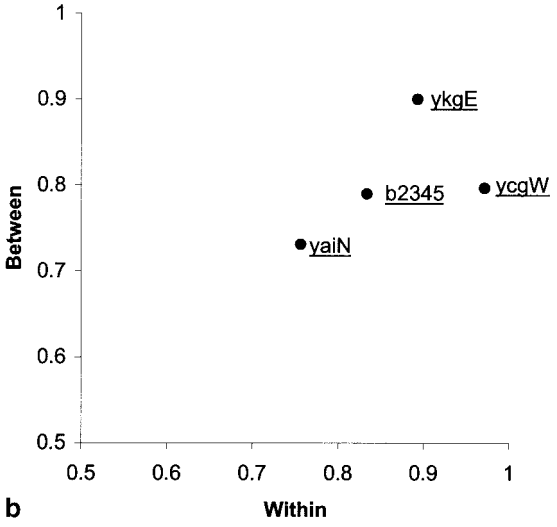


FIG. 3. (a) Compatibility matrix of polymorphic positions. The matrix was constructed based on 68 polymorphic positions at four MNR loci of 41 *E. coli* strains with the Reticulate software. Each square describes pairwise compatibility relationships between the 68 sites. Black squares, incompatibility; white squares, compatibility between two polymorphic sites at two different loci (between loci); colored squares, compatibility between two polymorphic sites at the same locus (within locus). (b) Compatibility at four MNR loci. Levels of between-loci compatibility are plotted against with-in locus compatibility.

amplified in B2 strains). This amplicon presence or absence approach was successful for strain typing in *E. coli* and *Listeria* spp. in our laboratory (L. Somer, E. Diamant, R. Gur-Arie, Y. Palti, Y. Danin-Poleg, and Y. Kashi, submitted for publication), and in combination with MNR, MLST can provide an efficient tool for epidemiology and for the development of rapid diagnostic kits for bacterial pathogens.

There is accumulating evidence that SSRs serve a functional role, affecting gene expression, and that polymorphism of SSR tracts may be important in the evolution of gene regulation

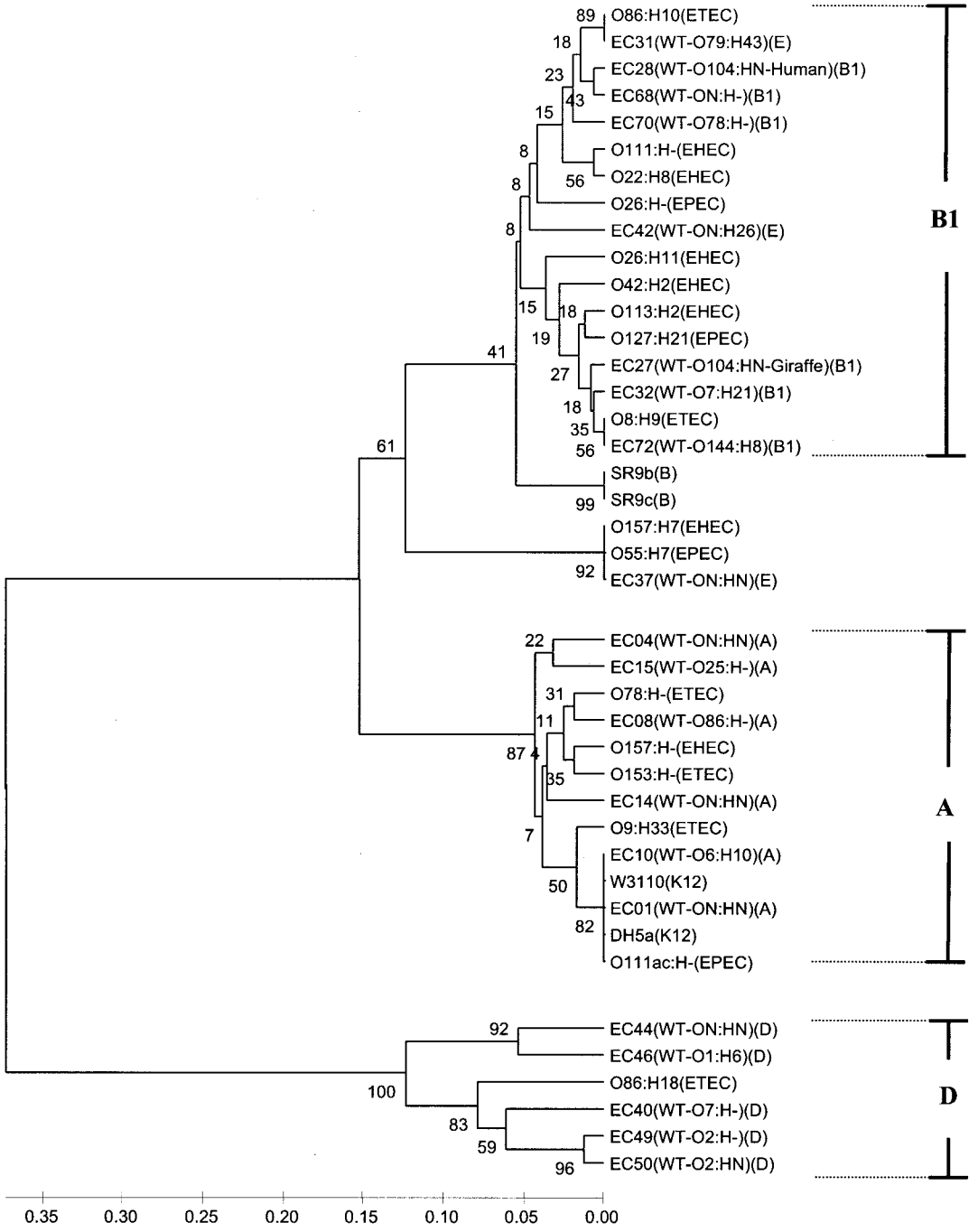


FIG. 4. Phylogenetic tree of 41 *E. coli* strains. The phylogeny was based on the polymorphic sites at the four MNR loci (multilocus analysis). The tree was constructed with the UPGMA method. The numbers at the nodes are bootstrap confidence values based on 1,000 replicates. The ECOR A, B1, and D groups are indicated.

(14, 16, 18, 28, 37, 39, 44–46). The markers used in phylogenetic studies should be as neutral as possible. Due to the potential involvement of SSRs in gene regulation, inference of SSR variation for evolutionary studies should be conducted with attention given to the ecological and epidemiological conditions. SSRs in genes that are known to contribute to ecological adaptation should be avoided in studies designed to infer evolutionary relationships. However, as demonstrated in this study, one way to overcome this problem is to conduct multilocus analysis, which dilutes the bias of individual loci.

In this study, we found that randomly selected noncoding loci that contain MNRs were significantly more polymorphic at the sequence level than noncoding loci that did not contain SSRs in *E. coli*. We also found that these polymorphic MNRs

were useful for inferring phylogenetic relationships and reconstructed trees that were consistent with the standard multilocus enzyme electrophoresis trees (12, 35, 50). The usefulness of SSRs for evolution studies and strain typing in less clonal species such as *Neisseria meningitidis* (23) should be tested in similar future studies.

ACKNOWLEDGMENTS

We thank Yael Danin Poleg, Technion, Israel, for helpful discussion and comments during preparation of the manuscript; Thomas and Beth Whittam, Michigan State University, for providing the samples of the original ECOR strains; and the Ministry of Health Central Laboratories, Jerusalem and Haifa Departments, Israel, for providing the bacterial strains.

This research was supported by the Grand Water Research Institute, Mitchel Soref Innovation Awards program, Technion, and by the Otto Meyerhof Center for Biotechnology, Technion, established by the Minerva Foundation, Germany. R. Gur-Arie was supported by the Food Control Administration in the Israel Ministry of Health. Eric Hallerman was supported by the Fulbright Senior Scholars Program and by the Virginia Polytechnic Institute and State University.

REFERENCES

1. **Ahmed, R., C. Bopp, A. Borczyk, and S. Kasatiya.** 1987. Phage-typing scheme for Escherichia coli O157:H7. J. Infect. Dis. **155:**806–809.
2. **Blattner, F. R., G. Plunkett, III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of Escherichia coli K-12. Science **277:**1453–1474.
3. **Dietrich, W. F., J. L. Weber, D. A. Nickerson, and P.-Y. Kwok.** 1999. Identification and analysis of DNA polymorphisms, p. 135–186. *In* B. Birren, E. D Green, P. Hieter, S. Klapholz, R. M. Myers, H. Reithman, and J. Roskama (ed.), Genome analysis: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
4. **Farris, J. S., M. Kallersjo, A. G. Kluge, and C. Bult.** 1994. Testing significance of incongruence. Cladistics **10:**315–319.
5. **Feldman, M. W., J. Kumm, and J. Pritchard.** 1999. Mutation and migration in models of microsatellite evolution, p. 98–115. *In* D. B. Goldstein and C. Schlotterer (ed.), Microsatellites: evolution and applications. Oxford University Press, Inc., New York, N.Y.
6. **Felsenstein, J.** 1989. PHYLIP — phylogeny inference package (version 3.2). Cladistics **5:**164–166.
7. **Field, D., and C. Wills.** 1996. Long, polymorphic microsatellites in simple organisms. Proc. R. Soc. London B Biol. Sci. **263:**209–215.
8. **Field, D., and C. Wills.** 1998. Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. USA **95:**1647–1652.
9. **Gur-Arie, R., C. J. Cohen, Y. Eitan, L. Shelef, E. M. Hallerman, and Y. Kashi.** 2000. Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res. **10:**62–71.
10. **Guttman, D. S., and D. E. Dykhuizen.** 1994. Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science **266:**1380–1383.
11. **Guttman, D. S.** 1997. Recombination and clonality in natural populations of Escherichia coli. Trends Ecol. Evol. **12:**16–22.
12. **Herzer, P. J., S. Inouye, M. Inouye, and T. S. Whittam.** 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli.* J. Bacteriol. **172:**6175–6181.
13. **Jakobsen, I. B., and S. Easteal.** 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12:**291–295.
14. **Kashi, Y., D. King, and M. Soller.** 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. **13:**74–78.
15. **Keim, P., L. B. Price, A. M. Klevytska, K. L. Smith, J. M. Schupp, R. Okinaka, P. J. Jackson, and M. E. Hugh-Jones.** 2000. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J. Bacteriol. **182:**2928–2936.
16. **King, D. G., M. Soller, and Y. Kashi.** 1997. Evolutionary tuning knobs. Endeavour **21:**36–40.
17. **Kudva, I. T., P. S. Evans, N. T. Perna, T. J. Barrett, F. M. Ausubel, F. R. Blattner, and S. B. Calderwood.** 2002. Strains of *Escherichia coli* O157:H7 differ primarily by insertions or deletions, not single-nucleotide polymorphisms. J. Bacteriol. **184:**1873–1879.
18. **Kunzler, P., K. Matsuo, and W. Schaffner.** 1995. Pathological, physiological, and evolutionary aspects of short unstable DNA repeats in the human genome. Biol. Chem. Hoppe Seyler **376:**201–211.
19. **Lawrence, J. G., and D. L. Hartl.** 1992. Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics **131:**753–760.
20. **Lawrence, J. G., and H. Ochman.** 1998. Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. USA **95:**9413–9417.
21. **Lecointre, G., L. Rachdi, P. Darlu, and E. Denamur.** 1998. Escherichia coli molecular phylogeny using the incongruence length difference test. Mol. Biol. Evol. **15:**1685–1695.
22. **Levine, M. M.** 1987. Escherichia coli that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. J. Infect. Dis. **155:**377–389.
23. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95:**3140–3145.
24. **Metzgar, D., E. Thomas, C. Davis, D. Field, and C. Wills.** 2001. The microsatellites of Escherichia coli: rapidly evolving repetitive DNAs in a nonpathogenic prokaryote. Mol. Microbiol. **39:**183–190.
25. **Milkman, R.** 1997. Recombination and population structure in Escherichia coli. Genetics **146:**745–750.
26. **Milkman, R., and M. M. Bridges.** 1993. Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons. Genetics **133:**455–468.
27. **Montgomery D. C.** 1991. Design and analysis of experiments, p. 649. John Wiley and Sons, New York, N.Y.
28. **Moxon, E. R., and C. Wills.** 1999. DNA microsatellites: agents of evolution? Sci. Am. **280:**94–99.
29. **Nei, M., and S. Kumar.** 2000. Molecular evolution and phylogenetics, p. 352. Oxford University Press, Inc., New York, N.Y.
30. **Noller, A. C., M. C. McEllistrem, O. C. Stine, J. G. Morris, Jr., D. J. Boxrud, B. Dixon, and L. H. Harrison.** 2003. Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. J. Clin. Microbiol. **41:**675–679.
31. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli f*rom natural populations. J. Bacteriol. **157:**690–693.
32. **Ochman, H., T. S. Whittam, D. A. Caugant, and R. K. Selander.** 1983. Enzyme polymorphism and genetic population structure in Escherichia coli and Shigella. J. Gen. Microbiol. **129:**2715–2726.
33. **Ohta, T.** 1982. Linkage disequalibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79:**1940–1944.
34. **Perna, N. T., G. Plunkett, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner.** 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature **409:**529–533.
35. **Pupo, G. M., D. K. Karaolis, R. Lan, and P. R. Reeves.** 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. Infect. Immun. **65:**2685–2692.
36. **Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam.** 2000. Parallel evolution of virulence in pathogenic Escherichia coli. Nature **406:**64–67.
37. **Renders, N., L. Licciardello, C. IJsseldijk, M. Sijmons, L. van Alphen, H. Verbrugh, and A. van Belkum.** 1999. Variable numbers of tandem repeat loci in genetically homogeneous Haemophilus influenzae strains alter during persistent colonisation of cystic fibrosis patients. FEMS Microbiol. Lett. **173:**95–102.
38. **Rokas, A., B. L. Williams, N. King, and S. B. Carroll.** 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425:**798–804.
39. **Rosenberg, S. M., S. Longerich, P. Gee, and R. S. Harris.** 1994. Adaptive mutation by deletions in small mononucleotide repeats. Science **265:**405–407.
40. **Sears, C. L., and J. B. Kaper.** 1996. Enteric bacterial toxins: mechanisms of action and linkage to intestinal secretion. Microbiol. Rev. **60:**167–215.
41. **Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour, and T. S. Whittam.** 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. **51:**873–884.
42. **Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt.** 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:**4384–4388.
43. **Tautz, D.** 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17:**6463–6471.
44. **Tonjum, T., D. A. Caugant, S. A. Dunham, and M. Koomey.** 1998. Structure and function of repetitive sequence elements associated with a highly polymorphic domain of the Neisseria meningitidis PilQ protein. Mol. Microbiol. **29:**111–124.
45. **van Belkum, A.** 1999. Short sequence repeats in microbial pathogenesis and evolution. Cell. Mol. Life Sci. **56:**729–734.
46. **van Belkum, A., S. Scherer, L. van Alphen, and H. Verbrugh.** 1998. Short-

sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev. **62:**275–293.
47. **Vogt, P.** 1990. Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code." Hum. Genet. **84:**301–336.
48. **Weber, J. L.** 1990. Informativeness of human (dC-dA)n. (dG-dT)n polymorphisms. Genomics **7:**524–530.
49. **Whittam, T. S., H. Ochman, and R. K. Selander.** 1983. Multilocus genetic structure in natural populations of Escherichia coli. Proc. Natl. Acad. Sci. USA **80:**1751–1755.
50. **Whittam, T. S., M. L. Wolfe, I. K. Wachsmuth, F. Orskov, I. Orskov, and R. A. Wilson.** 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. Infect. Immun. **61:**1619–1629.