INFECTION AND IMMUNITY, Sept. 2004, p. 5080–5088 Vol. 72, No. 9
0019-9567/04/$08.00+0 DOI: 10.1128/IAI.72.9.5080–5088.2004

Molecular Evolutionary Relationships of Enteroinvasive *Escherichia coli* and *Shigella* spp.

Ruiting Lan,[1]* M. Chehani Alles,[2] Kathy Donohoe,[2] Marina B. Martinez,[3] and Peter R. Reeves[2]*

School of Biotechnology and Biomolecular Sciences, University of New South Wales,[1] and School of Molecular and Microbial Biosciences, University of Sydney,[2] Sydney, Australia, and Faculdade de Ciencias Farmaceuticas, Universidade de Sao Paulo-SP, Sao Paulo, Brazil[3]

Received 9 December 2003/Returned for modification 10 May 2004/Accepted 7 June 2004

Enteroinvasive *Escherichia coli* (EIEC), a distinctive pathogenic form of *E. coli* causing dysentery, is similar in many properties to bacteria placed in the four species of *Shigella*. *Shigella* has been separated as a genus but in fact comprises several clones of *E. coli*. The evolutionary relationships of 32 EIEC strains of 12 serotypes have been determined by sequencing of four housekeeping genes and two plasmid genes which were used previously to determine the relationships of *Shigella* strains. The EIEC strains were grouped in four clusters with one outlier strain, indicating independent derivation of EIEC several times. Three of the four clusters contain more than one O antigen type. One EIEC strain (an O112ac:H− strain) was found in *Shigella* cluster 3 but is not identical to the *Shigella* cluster 3 D2 and B15 strains with the same O antigen. Two forms of the virulence plasmid pINV have been identified in *Shigella* strains by using the sequences of *ipgD* and *mxiA* genes, and all but two of our EIEC strains have pINV A. The EIEC strains were grouped in two subclusters with a very low level of variation, generally not intermingled with *Shigella* pINV A strains. The EIEC clusters based on housekeeping genes were reflected in the plasmid gene sequences, with some exceptions. Two strains were found in the pINV B form by using the *ipgD* sequence, with one strain having an *mxiA* sequence similar to the divergent sequence of D1. Clearly, EIEC and *Shigella* spp. form a pathovar of *E. coli*.

Enteroinvasive *Escherichia coli* (EIEC) is a pathogenic form of *E. coli* that can cause dysentery (25). Making a distinction between EIEC and *Shigella* spp. has been known for a long time to be difficult and depends on a very limited number of characteristics. *Shigella* spp. have been shown unequivocally to be clones of *E. coli* by sequencing of housekeeping genes (references 32 and 33 and references therein).

Historically, EIEC was first described in 1944, when it was called paracolon bacillus, but it was later identified as *E. coli* O124. In the 1950s, another group of *E. coli* strains was found to cause experimental keratoconjunctivitis in guinea pigs by the Serény test—a trait common with *Shigella*. These strains were initially classified under *Shigella* as *Shigella manolovi*, *S. sofia*, *Shigella* strain 13, and *S. metadysenteriae* and were later placed in the *E. coli* subgroup EIEC as *E. coli* O164 (2, 23, 35). EIEC and *Shigella* spp. bear remarkable phenotypic likeness, with a reduction in the number of substrates utilized relative to commensal *E. coli* strains. These similar phenotypes may be attributed to the fact that these organisms spend much of their lifetime within eukaryotic cells and have a different nutrient supply from most *E. coli* strains (19). Most EIEC strains are Lac$^-$, nonmotile and lysine decarboxylase negative (11, 38). The few biochemical properties differentiating *E. coli* and *Shigella* spp. include mucate and acetate. EIEC may be positive for one or both of the properties, while, with rare exceptions, *Shigella* strains are negative for both and more than 90% of typical *E. coli* strains are positive for both (4, 7).

A limited set of O antigens are found in EIEC strains, including O28ac, O29, O112ac, O121, O124, O135, O136, O143, O144, O152, O159, O164, O167, and O173 (5, 10, 14, 15, 21, 25,30). Three of these EIEC-associated O antigens are identical to O antigens present in *Shigella* spp. (5), namely, O112ac, O124, and O152, with *Shigella* O antigens of *S. boydii* serotype 15/*S. dysenteriae* serotype 2, *S. boydii* serotype 3, and *S. dysenteriae* serotype 12, respectively. EIEC strains with those shared O antigens show a higher metabolic activity than the *Shigella* strains (10). Nonetheless, differentiation between EIEC and *Shigella* organisms with the same serotype often proves very difficult due to their almost identical biochemical and physiological traits. EIEC strains can be distinguished from other *E. coli* strains by testing their invasion capacity by the Serény test or by identification of bacterial invasion-associated proteins or genes via specific tests. However, these methods are not generally used for routine diagnosis, and EIEC strains are only provisionally identified by O serotyping with commercially available antisera in routine diagnostic laboratories (3).

EIEC and *Shigella* strains harbor a common 220-kb plasmid, collectively termed pINV, although specific names were given for some pINV plasmids, for example, pWR100 of *S. flexneri* 5, pMYSH6000 of *S. flexneri* 2a, and pSS120 of *S. sonnei*. In our prevous analysis of *Shigella* strains (18), two major forms of the virulence plasmid with consistent sequence differences were revealed and referred to as pINV A and pINV B. The distribution of the two forms correlated well with the variation in chromosomal genes found in our earlier study, in which the

* Corresponding author. Mailing address: School of Molecular and Microbial Biosciences, University of Sydney, Sydney, NSW 2006, Australia. Phone: 61-2 9385 2095. Fax: 61-2 9385 1591. E-mail: r.lan@unsw.edu.au or reeves@angis.usyd.edu.au.

TABLE 1. EIEC strains used in this study

Strain name	Serotype	Place of isolation	Yr of isolation	Source[a]	Original laboratory no.
M519	O124:H−	Argentina	Unknown	SHL	C906-91
M520	O28ac:H−	United States	Unknown	IEKC	C499-89
M2326	O124:H30	Unknown	Unknown	BML	R411/55
M2327	O136:H9	Unknown	Unknown	BML	111/55
M2328	O143:H−	Unknown	Unknown	BML	223
M2329	O144:H−	Unknown	Unknown	BML	1642-56
M2330	O152:H−	Unknown	Unknown	BML	1184-68
M2331	O28ac:H−	Unknown	Unknown	BML	152/82
M2332	O112ac:H−	Unknown	Unknown	BML	329/81
M2333	O124:H30	Unknown	Unknown	BML	I314
M2334	O144[b]	Unknown	Unknown	BML	6/84
M2335	O167[b]	Unknown	Unknown	BML	223/83
M2338	O136:H−	Brazil	1964	USP	FBC13601
M2339	O124:H7	United States	Unknown	USP	FBC12406
M2340	O143:H−	Brazil	1979	USP	FBC14302
M2341	O143:H−	Brazil	1985	USP	FBC14310
M2342	O144:H−	Brazil	1966	USP	FBC14402
M2343	O136:H−	Brazil	1977	USP	FBC13616
M2344	O124:H−	Brazil	1985	USP	FBC12436
M2345	O28ac:H−	Brazil	1977	USP	FBC2801
M2346	O28ac:H−	Unknown	Unknown	USP	FBC28
M2347	O164:H−	Brazil	1985	USP	FBC16406
M2348	O136:H−	Brazil	1977	USP	FBC13619
M2349	O29:H−	Brazil	1976	USP	FBC2901
M2350	O143:H−	Brazil	1986	USP	FBC14314
M2351	O167:H−	Brazil	1979	USP	FBC16702
M2352	O152:H−	Brazil	1964	USP	FBC15202
M2353	O135:H−	Brazil	1980	USP	FBC13501
M2354	O152:H−	Brazil	1965	USP	FBC15206
M2355	O124:H−	Brazil	1977	USP	FBC12405
M2356	O124:H30	Brazil	1977	USP	FBC12430
M2357	O164:H−	Brazil	1985	USP	FBC16405

[a] SHL, State Health Laboratory, Perth, Australia; IEKC, International *Escherichia* and *Klebsiella* Centre, Copenhagen, Denmark; BRL, Biomedical Reference Laboratory, Fairfield Hospital, Melbourne, Australia; UPS, Universidade de Sao Paolo, Brazil.
[b] H antigen form is not known.

majority of the *Shigella* strains were grouped into three clusters with a few outliers (33). All of the cluster 1 strains, consisting of the majority of *S. dysenteriae* and *S. boydii* serotypes plus *S. flexneri* serotype 6, contain the pINV A form, and cluster 3 strains, consisting of *S. flexneri* serotypes 1 to 5 and *S. boydii* serotype 12, have pINV B. Of the cluster 2 strains (seven *S. boydii* serotypes and *S. dysenteriae* serotype 2), pINV sequences could be obtained from only three strains, of which two contain the pINV A form and one contains the pINV B form. Of the outliers not in the three clusters, all contain either the pINV A or the pINV B form, with the exception of *S. dysenteriae* serotype 1, which contains a plasmid with a divergent sequence not corresponding to these two forms, and the *S. boydii* serotype 13 strain, which does not contain a plasmid.

The two EIEC strains included in the above study contain pINV A. In this study, we examine 32 EIEC strains to determine their relationships to *Shigella* strains by using both chromosomal and plasmid genes.

MATERIALS AND METHODS

Strains and genes. A total of 32 EIEC strains were used, and details are given in Table 1. Primers for PCR and sequencing are listed in Table 2. Primer sets for amplification of *thrB* and *purN* were designed by Pupo et al. (33), and those for amplification of *ipgD* and *mxiA* were designed by Lan et al. (18). Primer sets for the *mdh-argR* region and *trpB* were designed specifically for this study. The *mdh-argR* gene region has low-level variation, and we selected the most variable segment, including the intergenic region between *mdh* and *argR*, for amplification and sequencing.

PCR and sequencing. The sequence was obtained directly from the PCR product. Double-stranded PCR product was purified with the Wizard PCR purification system (Promega, Madison, Wis.) to remove excess PCR primer and eluted in 30 μl of sterile distilled water, and the sequence was determined by the dye-terminator technique using a thermal cycler (Perkin-Elmer Cetus, Norwalk, Conn.) and automated 377 DNA sequencer (Applied Biosystems, Burwood, Victoria, Australia), as specified by the manufacturers.

Sequence analysis. DNA sequences were assembled and edited using PHRED, PHRAP, and CONSED software (13). Further analysis was undertaken using software available from the Australian National Genomic Information Service at The University of Sydney. Sequence comparisons were analyzed using the MULTICOMP package (34). MULTICOMP calculates nucleotide diversity (π) by the method described by Nei and Miller (28) and average pairwise percentage difference. Molecular evolutionary relationships among each of the genes studied were examined by the neighbor-joining method of tree construction (27, 36), based on distances estimated using the two-parameter method of Kimura (17). Phylogenetic trees and bootstrap analysis to determine the statistical stability of each node were done using PHYLIP (version 3.5; Joseph Felsenstein, Department of Genetics, University of Washington, Seattle, Wash.).

Biochemical tests. Utilization of acetate and mucate fermentation was tested using the methods described by Ewing (10).

Nucleotide sequence accession numbers. The GenBank accession numbers for the nucleotide sequences determined in this study are AY627050 to AY627276.

RESULTS

Sequence variation in housekeeping genes. A total of 3,390 bp of sequence data was analyzed for each strain, comprising

TABLE 2. Primers used in this study

Gene amplified	Primer (laboratory no.)	Primer sequence (5′–3′)	5′ start position[a]	Amplicon size (bp)
mdh-argR[b]	4351	TTCATCCTAAACTCCTTATT	+363	1,049
	4352	TAATCGCAGGGGAAAGCAGG	+1412	
thrB	1408	ATGAGCGTCGGGTTTGATGT	+36	1,088
	1409	GGATTTCCTGTGGGATTTCA	+1124	
trpB	4389	ACAGAGATTCGTAGCGTT	+1181	1,174
	4390	GAAAGGAACAATGACAAC	+7	
purN	1410	AATGCCAACGGTGAAAACGC	−77	1,074
	1411	GGAGCCTTGTTCTTCGCTAA	+997	
ipgD	1660	TATCAGGCTCGGAGTGTTAT	−355	960
	1662	TCGGCGTCAGAAGAGAAGTC	+605	
	1661	GACCAGAGTTATTATCACAG	+547	927
	1663	TATTAGCACATCATCATCAA	+1474	
mxiA	1817	GAGATGAAAGTGATCCAG	−3	1,002
	1818	AAATGTACCAGTATAGCC	+999	
	1819	ATTATAAAAAGGTCGTAG	+946	1,058
	1820	AGCATACGATATAACGAC	+2004	

[a] Designation of primer locations relative to the ATG start codon, where A is position 1 in the *mdh*, *thrB*, *trpB*, *purM*, *ipgD*, and *mxiA* genes. For the *purN* gene, the primer location is relative to the GTG start codon, where G is position 1.
[b] The region amplified is part of *mdh* and argR genes and intergenic variable region.

1,032 bp of *thrB*, 898 bp of *purN*, 526 bp of *trpB*, and 934 bp of *mdh-argR*. The four regions are within regions sequenced previously for *Shigella* strains, comprising 2,032 bp of the *thrB-thrC* region, 2,101 bp of the *purM-purN* region, 1,486 bp of the *trpB-trpC* region, and 1,541 bp of the *mdh-argR* region, respectively (Table 2). The sequence alignment for informative sites (sites that affect the tree topology because they have at least two different bases present in at least two strains each) is shown in Fig. 1 (see supplementary Fig. 1 for informative sites of *Shigella* and EIEC strains combined at http://www.mmb.usyd.edu.au/archives/). It is very clear that most EIEC strains fall into one of four groups and that the same strains are in each group for all four regions of the chromosome, although the *purN* and *mdh-argR* regions have few informative sites for separation of the groups. One strain, M2339, appears to be a recombinant, and another, M2332, fits within cluster 2 *Shigella* strains.

Phylogenetic relationships based on housekeeping genes. A tree for the combined data is shown in Fig. 2. Published data for the 46 *Shigella* and 8 *E. coli* reference (ECOR) *E. coli* strains (33) were included in the analysis. We also sequenced an additional 12 ECOR set strains to increase their representation. The three *Shigella* clusters (clusters 1, 2, and 3) identified by the previous study were retained. All but one (M2339) of the EIEC strains are in clusters with at least three strains, named clusters 4, 5, 6, and 7. Cluster 7 consists of strains M2329, M2334, and M2342, with the same O antigen, while the other clusters contain two or more O-antigen types. Bootstrap analysis using 1,000 replicates gave strong support to each cluster.

Sequence variation within pINV plasmid genes. The pINV genes *ipgD* and *mxiA* were sequenced for 32 strains. Both genes have been used for the study of *Shigella* pINV (18). The *ipgD* gene is 1,617 bp long, is located at the beginning of the *mxiA* operon, and encodes a product that is involved in entry of the host cell on bacterial invasion (29), while the 2,006-bp *mxiA* gene encodes an essential component of the type III secretory machinery (31). Only a partial sequence of *ipgD* for M2357 was obtained, and *ipgD* of M2334 failed to be amplified by PCR. Two strains (M519 and M520) sequenced previously (18) were included.

A 1,630-bp region containing *ipgD* (1,617 bp) was analyzed. A comparison of the EIEC sequences obtained revealed an average pairwise difference of 0.51% and indicated the presence of 44 polymorphic sites, of which 43 are informative and are summarized in Fig. 3. M2339 and M2355 are very similar, with a highly divergent sequence which is similar to the pINV B form.

An 1,866-bp stretch of the *mxiA* sequence was analyzed in each strain. Comparison of the sequences obtained indicated an average pairwise difference of 0.1%. Analysis of the *mxiA* alleles within all EIEC strains showed that the region contained 26 polymorphic bases but only 2 informative bases (Fig. 3). Of the singularly polymorphic bases, 22 are in M2339, which has the same highly divergent sequence as D1. However, M2339 has an *ipgD* sequence similar to those of the pINV B form. M2355 is identical to M2339 in its *ipgD* gene, while the *mxiA* gene is similar to those of the pINV A form.

Phylogenetic analysis of virulence plasmid genes. The data obtained in this study were combined with previous results (18) to generate a phylogenetic tree for all *Shigella* and EIEC strains. The combined tree for plasmid genes *ipgD* and *mxiA* is shown in Fig. 4. Only strains for which sequence data for at least *ipgD* were obtained are included. The EIEC strain with

		*thr*B	*pur*N	*trp*B	*mdh-arg*R
		11			
		11222223444678900	113577	1111111122222233344445 5	457
		1256017893179849622	223158	8116788991166771390346 22	964
		4874958795738339867	073104	8250247061725149413683 39	918
Cluster	Consensus	TTATATTGTAAGATGTCAC	TAGGGG	TGTGCACGCCTGGGGTCTTCGGCT	CCA
2	M2332 (O112ac H-)	C..GGCC..T..CC....G	C.....	.....CTTA.....A..	T.G
4	M2327 (O136 H9)	.....G.....	.....	.....T...AA..	.A.
4	M2331 (O28ac H-)	.....G.....	.....	.....T...AA..	.A.
4	M2344 (O124 H-)	.....G.....	.....	.....T...AA..	.A.
4	M2347 (O164 H-)	.....G.....	.....	.....T...AA..	.A.
4	M2348 (O136 H-)	.....G.....	.....	.....T...AA..	.A.
4	M2349 (O29 H-)	.....G.....	.....	.....T...AA..	.A.
4	M520 (O28ac H-)	.....G.....	.....	.....T...AA..	.A.
4	M2338 (O136 H-)	.....G.....G	.....	.....T...AA..	.A.
4	M2343 (O136 H-)	.....G.....G	.....	.....T...AA..	.A.
4	M2345 (O28ac H-)	.....G.....G	.....	.....T...AA..	.A.
4	M2346 (O28ac H-)	.....G.....G	.....	.....T...AA..	.A.
5	M2353 (O135 H-)	.C.....G	.G.A..	.....TC	...
5	M519 (O124 H-)	.C.....G	.G.A..	.....TC	...
5	M2326 (O124 H30)	.....	.G.A..	.....TC	...
5	M2330 (O152 H-)	.....	.G.AC.	.....TC	...
5	M2333 (O124 H30)	.....	.G.A..	.....TC	...
5	M2354 (O152 H-)	.....	.G.AC.	.....TC	...
5	M2356 (O124 H30)	.....	.G.A..	.....TC	...
5	M2357 (O164 H-)	.....	.G.A..	.....TC	...
5	M2355 (O124 H-)	.....	.G.A..	.....TC	...
5	M2352 (O152 H-)	.....GA	.G.AC.	.....TC	...
6	M2335 (O167)	C.TGGCCACT.ACCACT..	.....	C...TGTC.GCCTTAG.G.T....	T.G
6	M2341 (O143 H-)	C.TGGCCACT.ACCACT..	..A...	C...TGTC.GCCTTAG.G.T....	T.G
6	M2350 (O143 H-)	C.TGGCCACT.ACCACT..	..A...	C...TGTC.GCCTTAG.G.T....	T.G
6	M2328 (O143 H-)	C.TGGCCACT.ACCACT..	..A...	C...TGTC.GCCTTAG.G.T....	T.G
6	M2340 (O143 H-)	C.TGGCCACT.ACCACT.G	..A...	C...TGTC.GCCTTAG.G.T....	T.G
6	M2351 (O167 H-)	CCTGGCCACT.ACCACTGA	.....	C...TGTC.GCCTTAG.G.T....	T.G
7	M2329 (O144 H-)	.....C.....	.....	CACA.GT.AGCCTTA...A...TC	.A.
7	M2334 (O144)	.....C.....	.....A	CACA.GT.AGCCTTA...A...TC	.A.
7	M2342 (O144 H-)	.....C....G	.....A	CACA.GT.AGCCTTA...A...TC	.A.
Outlier	M2339 (O124 H7)	.....CACT..	C.....	.....CTTA...A...TC	.A.

FIG. 1. Informative bases for the *thrB*, *purN*, *trpB*, and *mdh-argR* gene regions for EIEC strains. The numbers at the top of the figure, reading vertically, are base positions. The phylogenetic grouping of each strain is indicated in the left column. O and H antigens are shown in parentheses after the strain name.

the *mxiA* sequence only, M2334, is excluded, since there is very little variation in the pINV A *mxiA* gene. It is evident that all but two EIEC strains have the pINV A form. The combined tree indicates the presence of two distinct groups within the A form: A1 and A2. However, the bootstrap value at the node of separation of these two subclusters within the A form is less than 50%. Nevertheless there are six sites in *ipgD* which support the division if only EIEC strains are considered (Fig. 3) and three sites for all strains (see supplementary Fig. 2 at http://www.mmb.usyd.edu.au/archives/). The majority of the cluster 4 and 5 strains are in the pINVA1 subcluster, with identical sequences which are slightly divergent from those of the *Shigella* strains. However, four and two strains of clusters 4 and 5, respectively, are in pINV subcluster A2. Four cluster 6 strains and two cluster 7 strains have identical pINV sequences. M2339 and M2355 *ipgD* sequences grouped together and appear to be an ancestral form of pINV B. However, M2339 grouped with D1 if only the *mxiA* sequence was used.

DISCUSSION

Multiple origins of EIEC. With the exception of one strain, which falls into *Shigella* cluster 2, the EIEC strains were found in four separate clusters of at least three strains each and one outlier strain, indicating that the EIEC phenotype has arisen several times. In our previous study, the *Shigella* phenotype was also shown to have originated separately several times during its evolution (33). The EIEC clusters had fewer O-antigen forms than the *Shigella* clusters did. While *Shigella* strains appear to have diverged more phenotypically from typical *E. coli* than EIEC strains have, none of the clusters defined contain both *Shigella* and EIEC, except for EIEC strain M2332, which could be a misclassification, as discussed below. EIEC clusters 4 and 5 could have a common origin since no ECOR set strains separate the two clusters. This may also be true for *Shigella* clusters 2 and 3. Since in each case the cluster branch lengths are long and the bootstrap value is high, the two clusters must

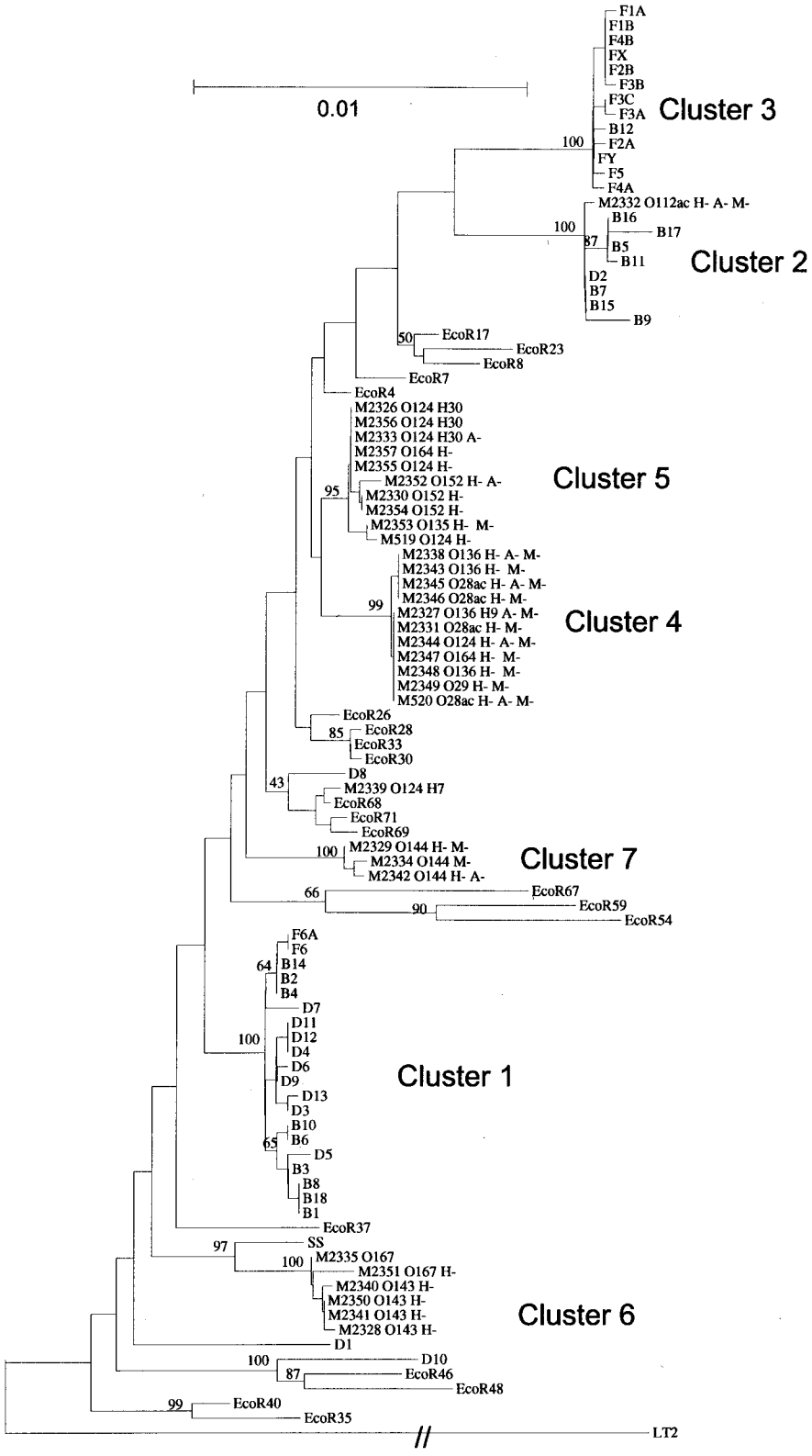

FIG. 2. Phylogenetic tree generated by the neighbor-joining method for the combined data for the *purN*, *thrB*, *trpB*, and *mdh-argR* regions sequenced to give a total of 3,390 bp. Strain names prefixed with M are EIEC and are followed by serotype and biochemical properties for mucate and acetate if negative as M- and A-respectively; for *Shigella* strains, F denotes *S. flexneri*, D denotes *S. dysenteriae*; B denotes *S. boydii*, and SS denotes *S. sonnei*, followed by the serotype number. Bootstrap values are percentages of 1,000 replications and, if 50% or greater, are indicated at the nodes. LT2 is used as the outgroup.

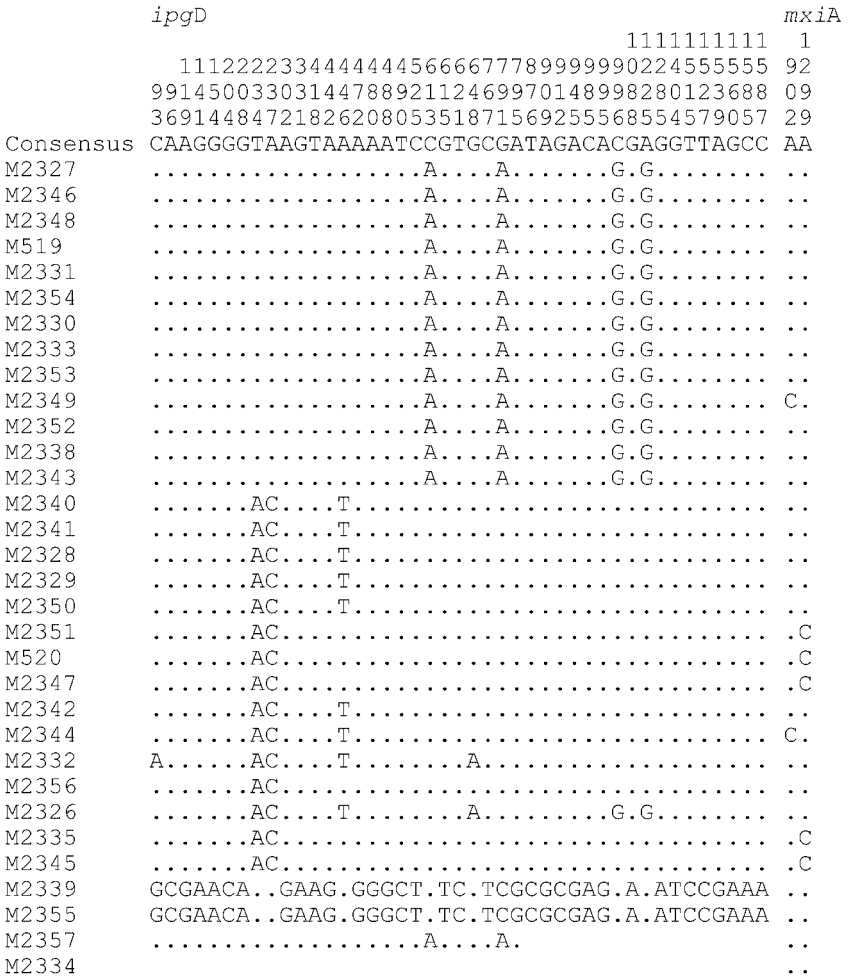

FIG. 3. Informative sites for the *ipgD* and *mxiA* gene for EIEC strains. Note that for some strains, where no sequence was obtained, the corresponding region appears blank. The numbers at the top of the figure, reading vertically, are base positions.

have been separated for a long time even if they were derived from a common pINV-carrying strain.

The diversification within *Shigella* clusters 1 and 2 is estimated to have originated 50,000 to 270,000 years ago, while cluster 3 apparently diversified later, at 35,000 to 170,000 years ago (33). On comparison of EIEC and *Shigella* clusters, it is clear that the EIEC clusters have diverged less than *Shigella* clusters. Assuming that the housekeeping genes are evolving at approximately equal rates within the two forms, we may hypothesize that the EIEC strains have had less time to diverge within their clusters and hence may have arisen at a later date than the *Shigella* clusters.

Two *Shigella* outlier strains (*Shigella* strain D8 and *S. sonnei*) were placed very close to EIEC strains but are possibly derived independently. M2339 is placed very close to *Shigella* strain D8, but they are separated by ECOR strains and M2339 is more similar to the ECOR strains than to D8 (Fig. 1 and supplementary Fig. 1 [http://www.mmb.usyd.edu.au/archives/]). Further, the plasmid gene sequences are very different. The *ipgD* gene of M2339 is similar to those of the B form, and *mxiA* is similar to that of D1. *S. sonnei* is close to cluster 6, with identical informative sites in *thrB*, and shares five characteristic informative sites of cluster 6 but is dissimilar in the other two gene regions. For the plasmid sequences, *S. sonnei* has a B-form plasmid while cluster 6 strains have the A form.

As evidenced in Fig. 2, the EIEC strains analyzed clustered separately from *Shigella*, with only one exception: M2332 grouped with *Shigella* cluster 2 strains in every gene studied. This, coupled with similar plasmid gene sequences and identical O antigen (O112ac) to the B15 and D2 O antigen, leads to the conclusion that M2332 is probably a *Shigella* strain that has been misclassified.

O-antigen diversity in EIEC strains. Our sample of strains contains 10 of the 14 O antigens reported to be associated with EIEC (5, 10, 14, 15, 21, 25, 30). The majority of the 10 O antigens sampled were grouped within a single cluster. Of the four EIEC clusters, three have more than one O antigen, with O28, O29, O124, O136, and O164 in cluster 4, O124, O135, O152, and O164 in cluster 5, and O143 and O167 in cluster 6. Cluster 7 has only O144. It is interesting that two O antigens are in two or more clusters, with O124 in clusters 4 and 5 and one outlier strain and O164 in clusters 4 and 5. Two of them are also found in *Shigella* strains (see below).

This study confirms the trend of rapid expansion of O anti-

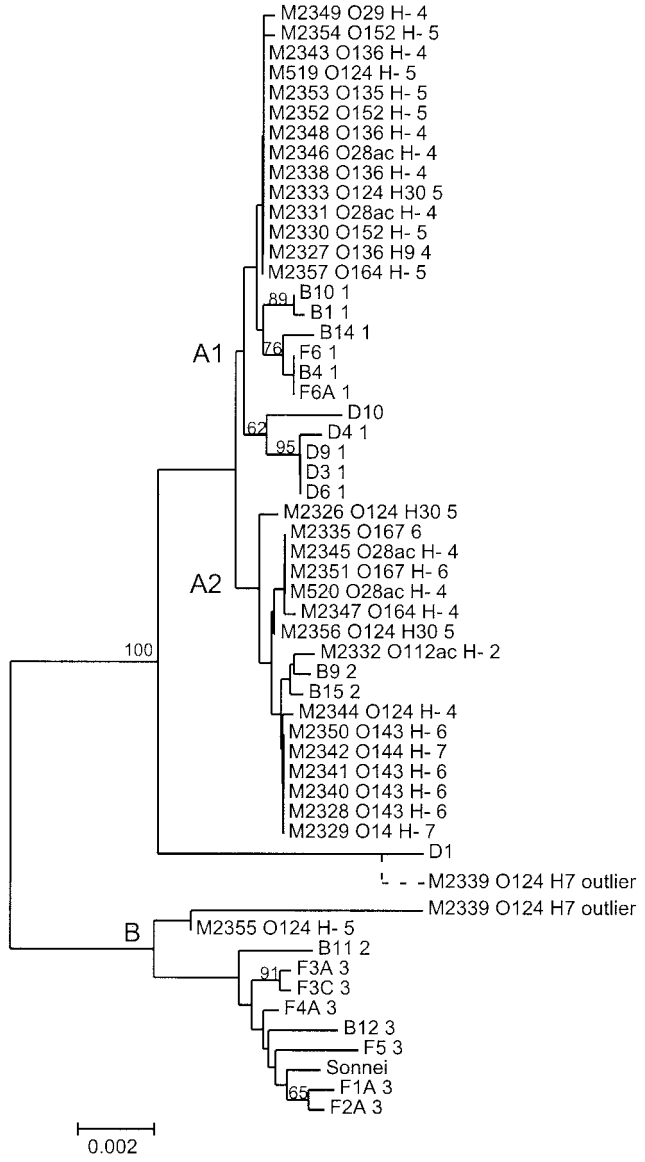

FIG. 4. Phylogenetic tree generated by the neighbor-joining method for the combined data for plasmid genes *ipgD* and *mxiA*. Strain names are prefixed with M for EIEC, F for *S. flexneri*, B for *S. boydii*, and D for *S. dysenteriae*, followed by serotype number and relevant cluster number. M2339 is shown twice on the tree. Its alternative position (marked by a dotted line) is based on the *mxiA* sequence. Bootstrap values are percentages of 1,000 replications and, if 50% or greater, are indicated at the nodes.

gens in the development of invasive clones of *E. coli* as observed in *Shigella* clones (19, 33). There are 46 O antigens described for *Shigella* strains. However, D2 and B15 are the same, and most *S. flexneri* forms are minor variants of one basic form; hence, there are 33 unique O-antigen forms (19), of which 12 are identical to known *E. coli* O antigens while 21 are unique to *Shigella* clones (33). The total number of novel O antigens in *Shigella* represents 11% of the known *E. coli*/*Shigella* O-antigen forms, which is disproportionatelly high considering that *Shigella* represents only a small fraction of the *E. coli* diversity. Of the 33 O antigens, 18 were identified in *Shigella* clusters 1 and 2. The EIEC clusters have far fewer O antigens. This may be because the EIEC clusters arose later than *Shigella* clusters, based on the level of sequence variation within clusters, and have had a shorter time for expansion of O antigens than *Shigella* has. However, it may be that EIEC strains have received less attention than *Shigella* strains and that less common forms have not been reported.

There are a number of *Shigella* O antigens common within EIEC, including O124 (D3), O112 (D2/B15), O143 (B8), O152 (D12), and O167 (B3). The only O112 strain included in this study seems to be a *Shigella* strain, as discussed above. This disproportionate representation of certain O-antigen forms in *Shigella* and EIEC strains supports the notion that O-antigen specificity may be important in pathogenicity (12, 24, 39).

Comparison between the clustering of plasmid genes and chromosomal genes. The clustering of *Shigella* clusters 1 and 3, as observed previously (18), and EIEC cluster 6 strains is consistent between chromosomal genes. However, there are several inconsistencies in the clustering of strains in other clusters, indicating lateral transfer of the pINV plasmid between clusters. Seven cluster 4 strains have pINV subcluster A1 plasmid, while four other cluster 4 strains have pINV subcluster A2 plasmid. Both groups are represented by three O-antigen forms. It is not known whether the primordial pINV form of cluster 4 is A1 or A2. Cluster 5 strains have either pINV subcluster A1 (seven strains) or A2 (two strains) or pINV B (one strain), suggesting that the last two cases are due to transfer of the virulence plasmid or recombination involving the two plasmid genes studied. The cluster 5 strain M2356 is likely to have a recombinant *ipgD* gene with two informative sites characteristic of pINV subcluster A1 (Fig. 3). Transfer of pINV is also seen in *Shigella* cluster 2, with two strains having the pINV A form and one having the pINV B form (18).

The plasmid forms found within *Shigella* tend to group away from those within EIEC, with a longer branch length in comparison to those of EIEC strains, which are tightly clustered, suggesting that *Shigella* has diverged to a greater extent within its clusters than EIEC has. This higher level of divergence is also reflected in the chromosomal genes, as discussed above.

Virulence variation of EIEC and *Shigella* spp. The long-standing demarcation of *Shigella* and EIEC is based in part on perceived differences in virulence. EIEC strains are generally said to be less virulent than *Shigella* strains, with a higher infectious dose of EIEC needed in volunteers (8, 9). However, the *Shigella* strains used in these studies were generally D1, *S. sonnei*, and *S. flexneri* strains, which are known to be the more virulent *Shigella* forms and more prevalent in epidemiological terms. We have now studied a large sample of EIEC strains and shown that the virulence plasmids of the EIEC strains are the pINV A form, which is very closely related to other pINV A plasmids of *Shigella* strains. There does not seem to be a case for separation of *Shigella* and EIEC as classes on the basis of the virulence plasmid. However, differences in pINV may play a role in virulence. We recently compared the pINV A plasmid from *S. flexneri* serotype 6 and the pINV B plasmid from *S. flexneri* serotype 2a and found that the majority of the virulence-associated genes are under strong selection pressure for change (20). There are also pINV genes that are variably present in *Shigella* and EIEC (20) strains, including *sepA*, encoding serine protease, the major secreted protein of *S. flexneri*

2a, and *ospD3* (*senA*), encoding an enterotoxin, which is found in only 75% of EIEC strains and 83% of *Shigella* strains (26). Variation in virulence within and between EIEC and *Shigella* strains should be further studied based on the phylogenetic relationships of the strains and the plasmids. This may help us to better understand the pathogenesis and epidemiology of *Shigella* and EIEC infections.

Phenotypic variation in EIEC and *Shigella* spp. EIEC strains are very similar to *Shigella* strains in biochemical properties but generally do not fit the full definition for the genus Shigella, since some are motile or lactose fermenting and are seen essentially as traditional *E. coli* strains. However, if one examines the profile of inactive *E. coli* (including EIEC) as described by Farmer et al. (11), EIEC is more similar to *Shigella* than to typical *E. coli* and some EIEC strains have essentially all the properties of *Shigella* strains. EIEC can be differentiated from *Shigella* only by a very limited number of tests including L-serine, D-xylose, and/or sodium acetate utilization and mucate fermentation (7). EIEC isolates may be positive for one or more of the tests, but *Shigella* strains are generally negative (7). With 12 independent derivations of EIEC and *Shigella* within *E. coli* (five for EIEC and seven for *Shigella*), it is remarkable that biochemical properties, although few, were found to distinguish *Shigella* strains from EIEC strains. Variation within EIEC strains and between EIEC and *Shigella* strains in phenotypic properties is now best reassessed based on their natural groupings. We tested all 32 EIEC strains for acetate utilization and mucate fermentation and found that 10 and 15 strains were negative for these two properties, respectively, and 4 strains were negative for both. The distribution of these properties is consistent with the phylogenetic clustering in two cases. All cluster 4 strains are mucate negative, while none of the cluster 6 strains is negative for either property (Fig. 2). The loss of acetate utilization seems to be sporadic, showing independent loss even within a cluster.

Shigella and EIEC strains have many common characteristics, including the lack of several catabolic pathways widely present in *E. coli*, some of which are known to have resulted from independent loss of chromosomal properties. The loss of catabolic functions is presumably related to the life-style of *Shigella* and EIEC strains, because some properties are redundant since these strains spend much of their time within eukaryotic cells (19, 33), although at least one property, lysine decarboxylation, is also very deleterious for *Shigella* and EIEC, with strong selection against its presence (6, 22). As expected, sequencing of two *Shigella* genomes from *S. flexneri* serotype 2a strains 301 (16) and 2457T (40) has revealed many pseudogenes relative to other *E. coli* genomes, with 254 and 372 in strains 301 and 2457T, respectively, although some of the differences are reported to be due to annotation criteria (40). There are unique sets of pseudogenes in each strain, although they are the same serotype. Variation of phenotype properties in EIEC and *Shigella* strains is likely to be much larger than that seen from commonly used biochemical tests.

Concluding comments. EIEC and *Shigella* strains clearly form a distinctive *Shigella*-EIEC pathovar, with most EIEC and *Shigella* strains being found in distinct clusters. Acquisition of the pINV plasmid is a crucial step in the development of invasive forms of *E. coli*, followed by convergent evolution of properties such as the loss of specific catabolic pathways and motility and expansion of O-antigen diversity. There are many lineages of invasive *E. coli*. It appears that the EIEC lineages have been derived more recently than the *Shigella* strains. Resolution of the relationships provides a base for further studies of the rate of gene decay and virulence variation in EIEC and *Shigella*.

All but two EIEC strains have pINV A, which seems to be more frequently transferred to other *E. coli*. It is known from tests on several *S. flexneri* strains that pINV is unable to initiate conjugation (37). Our recent comparison of the F6 pINV A plasmid with the F5 pINV B form indicated that, from the substitution patterns in the incomplete *tra* region, the remaining *tra* genes must be functional, at least in some strains. No complete sequence for pINV A is available to determine whether it has retained more transfer-related genes to assist the process. It should also be noted that genes present in other plasmids or on the chromosome may complement the transfer functions present on the pINV plasmid.

Why has EIEC retained some *E. coli* properties that have been lost in multiple lineages of *Shigella?* It is likely that EIEC strains are in an intermediate stage and are a potential precursor of "full-blown" *Shigella* strains, as judged by the lower level of variation within EIEC clusters than within *Shigella* clusters. Motile strains were present within clusters 4 (O136:H9) and 5 (O124:H30), but the majority have lost the property or have it at retained reduced level (1). The selection pressure for loss of motility must be low in comparison to the gain of new O antigens, which has occurred several times in each cluster. It is also possible that EIEC is a distinctive form, perhaps differing from *Shigella* by being able to live in both commensal and epithelial mucosa niches. This will depend on the selection pressures involved. There might be a point of no return when EIEC strains lose sufficient commensal *E. coli* properties.

The elucidation of relationships of both housekeeping and plasmid genes of EIEC and *Shigella* warrants an end of the demarcation of the two forms, which should be regarded as a single pathovar of *E. coli*.

ACKNOWLEDGMENTS

This research is supported by grants from the National Health and Medical Research Council of Australia to P.R.R. and R.L. and by the National Council of Research of Brazil (CNPq) and the State of São Paulo Research Foundation (FAPESP) to M.B.M.

We thank Karl Bettelheim for providing strains and the anonymous referees for comments and suggestions.

REFERENCES

1. **Andrade, A., J. A. Giron, J. M. Amhaz, L. R. Trabulsi, and M. B. Martinez.** 2002. Expression and characterization of flagella in nonmotile enteroinvasive *Escherichia coli* isolated from diarrhea cases. Infect. Immun. **70:**5882–5886.
2. **Bando, S. Y., G. R. F. do Valle, M. B. Martinez, L. R. Trabulsi, and C. A. Moreira-Filho.** 1998. Characterization of enteroinvasive *Escherichia coli* and *Shigella* strains by RAPD analysis. FEMS Microbiol. Lett. **165:**159–165.
3. **Beutin, L., K. Gleier, I. Kontny, P. Echeverria, and F. Scheutz.** 1997. Origin and characteristics of enteroinvasive strains of *Escherichia coli* (EIEC) isolated in Germany. Epidemiol. Infect. **118:**199–205.
4. **Bopp, C. A., F. W. Brenner, P. I. Fields, J. G. Wells, and N. A. Strockbine.** 2003. *Escherichia*, *Shigella*, and *Salmonella*, p. 654–671. *In* P. R. Murray, E. J. Baron, J. H. Jorgensen, M. A. Pfaller, and R. H. Yolken (ed.), Manual of clinical microbiology, 8th ed., vol. 1. ASM Press, Washington, D.C.
5. **Cheasty, T., and B. Rowe.** 1983. Antigenic relationships between the enteroinvasive *Escherichia coli* O antigens O28ac, O112ac, O124, O136, O143, O144, O152, and O164 and Shigella O antigens. J. Clin. Microbiol. **17:**681–684.

6. **Day, W. A., R. E. Fernandez, and A. T. Maurelli.** 2001. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. Infect. Immun. **69:**7471–7480.
7. **Doyle, M. P., and V. V. Padhye.** 1989. *Escherichia coli*, p. 235–281. *In* M. P. Doyle (ed.), Foodborne bacterial pathogens. Marcel Dekker, Inc., New York, N.Y.
8. **DuPont, H. L., S. B. Formal, R. B. Hornick, M. J. Snyder, J. P. Libonati, D. J. Sheahan, E. H. LaBrec, and J. P. Kala.** 1971. Pathogenesis of *Escherichia coli* diarrhea. N. Engl. J. Med. **285:**1–9.
9. **DuPont, H. L., M. M. Levine, R. B. Hornick, and S. B. Formal.** 1989. Inoculum size in shigellosis and implications for expected mode of transmission. J. Infect. Dis. **159:**1126–1128.
10. **Ewing, W. H.** 1986. Edwards and Ewing's identification of the *Enterobacteriaceae*, 4th ed. Elsevier Science Publishers, Amsterdam, The Netherlands.
11. **Farmer, J. J., B. R. Davis, F. W. Hickman-Brenner, A. McWhorter, G. P. Huntley-Carter, M. A. Asbury, C. Riddle, H. G. Wathen-Grady, C. Elias, G. R. Fanning, A. G. Steigerwalt, C. M. OHara, G. K. Morris, P. B. Smith, and D. J. Brenner.** 1985. Biochemical identification of new species and biogroups of *Enterobacteriaceae* isolated from clinical specimens. J. Clin. Microbiol. **21:**46–76.
12. **Gemski, P. J., D. G. Sheahan, O. Washington, and S. B. Formal.** 1972. Virulence of *Shigella flexneri* hybrids expressing *Escherichia coli* somatic antigens. Infect. Immun. **6:**104–111.
13. **Gordon, D., C. Abajian, and P. Green.** 1998. CONSED—a graphical tool for sequence finishing. Genome Res. **8:**195–202.
14. **Gross, R. J., L. V. Thomas, T. Cheasty, N. P. Day, B. Rowe, M. R. Toledo, and L. R. Trabulsi.** 1983. Enterotoxigenic and enteroinvasive *Escherichia coli* strains belonging to a new O group, O167. J. Clin. Microbiol. **17:**521–523.
15. **Guth, B. E., R. M. Silva, M. R. Toledo, T. M. Lima, and L. R. Trabulsi.** 1989. Virulence factors and biochemical characteristics of serotypes of *Escherichia coli* serogroup O29. J. Clin. Microbiol. **27:**2161–2164.
16. **Jin, Q., Z. Yuan, J. Xu, Y. Wang, Y. Shen, W. Lu, J. Wang, H. Liu, J. Yang, F. Yang, X. Zhang, J. Zhang, G. Yang, H. Wu, D. Qu, J. Dong, et al.** 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res. **30:**4432–4441.
17. **Kimura, M.** 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **116:**111–120.
18. **Lan, R., B. Lumb, D. Ryan, and P. R. Reeves.** 2001. Molecular evolution of the large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*. Infect. Immun. **69:**6303–6309.
19. **Lan, R., and P. R. Reeves.** 2002. *Escherichia coli* in disguise: molecular origins of *Shigella*. Microbes Infect. **4:**1125–1132.
20. **Lan, R., G. Stevenson, and P. R. Reeves.** 2003. Comparison of two major forms of the *Shigella* virulence plasmid pINV: positive selection is a major force driving the divergence. Infect. Immun. **71:**6298–6306.
21. **Matsushita, S., S. Yamada, A. Kai, and Y. Kudoh.** 1993. Invasive strains of *Escherichia coli* belonging to serotype O121:NM. J. Clin. Microbiol. **31:**3034–3035.
22. **Maurelli, A. T., R. E. Fernandez, C. A. Bloch, C. K. Rode, and A. Fasano.** 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc. Natl. Acad. Sci. USA **95:**3943–3948.
23. **Monolov, D. G.** 1959. A new type of the genus *Shigella*—'*Shigella* 13'. J. Hyg. Epidemiol. Microbiol. Immunol. **3:**184–190.
24. **Morona, R., C. Daniels, and L. Van Den Bosch.** 2003. Genetic modulation of *Shigella flexneri* 2a lipopolysaccharide O antigen modal chain length reveals that it has been optimized for virulence. Microbiology **149:**925–939.
25. **Nataro, J. P., and J. B. Kaper.** 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. **11:**142–201.
26. **Nataro, J. P., J. Serawatana, A. Fasano, D. R. Maneval, L. D. Guers, F. Noriega, F. Dubrovski, M. M. Levine, and J. G. Morris.** 1995. Identification and cloning of a novel plasmid-encoded enterotoxin of enteroinvasive *Escherichia coli* and *Shigella* strains. Infect. Immun. **63:**4721–4728.
27. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York, N.Y.
28. **Nei, M., and J. C. Miller.** 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics **125:**873–879.
29. **Niebuhr, K., N. Jouihri, A. Allaoui, P. Gounon, P. J. Sansonetti, and C. Parsot.** 2000. IpgD, a protein secreted by the type III secretion machinery of *Shigella flexneri*, is chaperoned by IpgE and implicated in entry focus formation. Mol. Microbiol. **38:**8–19.
30. **Orskov, I., I. K. Wachsmuth, D. N. Taylor, P. Echeverria, B. Rowe, R. Sakazaki, and F. Orskov.** 1991. Two new *Escherichia coli* O groups: O172 from "Shiga-like" toxin II-producing strains (EHEC) and O173 from enteroinvasive *E. coli* (EIEC). APMIS **99:**30–32.
31. **Parsot, C., and P. J. Sansonetti.** 1996. Invasion and the pathogenesis of *Shigella* infections. Curr. Top. Microbiol. Immunol. **209:**25–42.
32. **Pupo, G. M., D. K. R. 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.
33. **Pupo, G. M., R. Lan, and P. R. Reeves.** 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. USA **97:**10567–10572.
34. **Reeves, P. R., L. Farnell, and R. Lan.** 1994. MULTICOMP: a program for preparing sequence data for phylogenetic analysis. CABIOS **10:**281–284.
35. **Rowe, B., R. J. Gross, and D. P. Woodroof.** 1977. Proposal to recognise serovar 145/146 (synonyms: 147, *Shigella* 13, *Shigella sofia*, and *Shigella manolovii*) as a new *Escherichia coli* O group, O164. Int. J. Syst. Bacteriol. **27:**15–18.
36. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogentic trees. Mol. Biol. Evol. **4:**406–425.
37. **Sansonetti, P. J., D. J. Kopecko, and S. B. Formal.** 1982. Involvement of a plasmid in the invasive ability of *Shigella flexneri*. Infect. Immun. **35:**852–860.
38. **Silva, R. M., M. R. F. Toledo, and L. R. Trabulsi.** 1980. Biochemical and cultural characteristics of invasive *Escherichia coli*. J. Clin. Microbiol. **11:**441–444.
39. **Wang, L., W. Qu, and P. R. Reeves.** 2001. Sequence analysis of four *Shigella boydii* O-antigen loci: implication for *Escherichia coli* and *Shigella* relationships. Infect. Immun. **69:**6923–6930.
40. **Wei, J., M. B. Goldberg, V. Burland, M. M. Venkatesan, W. Deng, G. Fournier, G. F. Mayhew, G. Plunkett III**, D. J. Rose, A. Darling, B. Mau, N. T. Perna, S. M. Payne, L. J. Runyen-Janecky, S. Zhou, D. C. Schwartz, and F. R. Blattner. 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect. Immun. **71:**2775–2786.

Editor: V. J. DiRita