

Comparison of Two Major Forms of the *Shigella* Virulence Plasmid pINV: Positive Selection Is a Major Force Driving the Divergence

Ruiting Lan,[†] Gordon Stevenson, and Peter R. Reeves*

School of Molecular and Microbial Biosciences, University of Sydney, Sydney, Australia

Received 20 December 2002/Returned for modification 31 March 2003/Accepted 5 August 2003

All *Shigella* and enteroinvasive *Escherichia coli* (EIEC) strains carry a 230-kb virulence plasmid (pINV) which is essential for their invasiveness. There are two sequence forms, pINV A and pINV B, of the plasmid (R. Lan, B. Lumb, D. Ryan, and P. R. Reeves, *Infect. Immun.* 69:6303-6309, 2001), and the recently sequenced pINV plasmid from *Shigella flexneri* serotype 5 is a pINV B form. In this study we sequenced the majority of the coding region of the pINV A form from *S. flexneri* serotype 6 other than insertion sequence or related sequences and compared it with the pINV B form. More than half of the genes sequenced appear to be under positive selection based on their low ratio of synonymous to nonsynonymous substitutions. This high proportion of selected differences indicates that the two pINV forms have functional differences, and comparative studies of pathogenicity in different *Shigella*-EIEC strains could be informative. There are also genes absent in the *S. flexneri* serotype 6 plasmid, including the *sepA* gene encoding serine protease, the major secreted protein of *S. flexneri* serotype 2a, and the *stbAB* genes, which encode one of the two partition systems found in *S. flexneri* serotype 5. The incompatibility of the two pINV forms appears to be due to either small differences in the *mvpAT* postsegregational killing system or the presence of an unknown system in pINVA.

Shigella strains are the causative agent of shigellosis or bacillary dysentery, a diarrheal disease of humans (1). Molecular evidence derived from studies involving DNA hybridization, multilocus enzyme electrophoresis, and sequencing of housekeeping genes indicates that *Escherichia coli* and all members of the genus *Shigella* belong to the same species (8, 30, 35). Enteroinvasive *E. coli* (EIEC) and *Shigella* have long been known to be similar (28) and can in fact be treated as comprising a single pathovar of *E. coli* (21). Most *Shigella* strains fall into three clusters (clusters 1, 2, and 3) based on the sequence of housekeeping genes (36), but *Shigella dysenteriae* serotypes 1, 8, and 10 and *Shigella sonnei* are not in any of the clusters while *Shigella boydii* serotype 13 is outside *E. coli*.

The pathogenesis of EIEC and *Shigella* involves invasion of mucosal epithelial cells of the large intestine (1, 34). Virulence in *Shigella* strains is dependent on the presence of a large 210- to 230-kb plasmid. The virulence plasmids pWR100 in *S. flexneri* serotype 5, pMYSH6000 in *S. flexneri* serotype 2a, and pSS120 in *S. sonnei*, together with those of other *Shigella* bacteria, have been shown to carry determinants for invasiveness and the ability to cause disease. These large plasmids are collectively termed pINV plasmids (13), which are also present in EIEC strains. The cell invasion capacity of *Shigella*-EIEC is determined by a cluster of 38 genes within a 32-kb segment of the pINV plasmid, often referred to as the entry or invasion region, which includes genes for invasins, molecular chaper-

ones, motility, regulation, and a specialized type III secretion apparatus (32).

In a previous study three pINV genes, *ipgD*, *mxiC*, and *mxiA*, were sequenced from strains representing a range of *Shigella* isolates and two EIEC isolates and showed that the plasmid exists in two related but clearly distinct sequence forms (pINV A and pINV B) (20). The phylogenetic relationships of the plasmid and chromosomal genes of *Shigella* strains are largely consistent. The cluster 1 and cluster 3 strains tested have pINV A and pINV B plasmids, respectively. However, of the three cluster 2 organisms, *S. boydii* serotypes 9 and 15 have pINV A while *S. boydii* serotype 11 has a pINV B plasmid. *S. dysenteriae* serotypes 8 and 10 and *S. sonnei* and also EIEC strains, none of which fall into the three clusters, were all found to have either pINV A or pINV B, except for *S. dysenteriae* serotype 1, which has a distinct pINV form. The outlier organisms other than *S. dysenteriae* serotype 1 must have obtained the plasmid relatively recently, after divergence of the two forms.

The invasion plasmid from *S. flexneri* serotype 5 strain M90T has been sequenced by two groups (9, 50). The two sequences are essentially identical but use different numbering and often different gene names. We use the numbering and gene names of Buchrieser et al. (9). The plasmid contains 93 segments, totaling 58 kb, which are homologous to known or putative insertion sequences (ISs), suggesting a remarkable history of IS-mediated acquisition of DNA. Analysis of the GC content, position, and function of non-IS-related genes indicates that the plasmid contains blocks of genes of various origins. There are three partition systems, two functional and one remnant, and the GC content of the replication system genes is different from that of the partition genes, indicating yet another source for this region (9). It appears that pINV was assembled from several different plasmids (9).

In this study we sequenced the non-IS-related coding re-

* Corresponding author. Mailing address: School of Molecular and Microbial Biosciences, Bldg. G08, University of Sydney, Sydney, New South Wales 2006, Australia. Phone: 61 2 9351 6045. Fax: 61 2 9351 4571. E-mail: reeves@angis.usyd.edu.au.

[†] Present address: School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia.

TABLE 1. Gene coding regions of pINV F6

Region	First gene	Start ^a	End ^a	Length (bp)	Sequenced in F6	Status in F6 ^b	Accession no.
1	<i>icsP</i>	991	1938	948	948	Complete	AY206427
2	<i>ospB</i>	3485	5420	1,936	1,936	Complete	AY206428
3	<i>ospD2</i> (part a)	9501	13971	4,471	2,860	9501–12361	AY206429
	<i>ospD2</i> (part b)				581	13393–13971	AY206430
4	<i>ospD1</i>	20964	21641	678	678	Complete	AY206431
5	<i>orf22</i>	22153	23315	1,163	1,163	Complete	AY206431
6	<i>parA</i>	29020	31199	2,180	2,180	Complete	AY206432
7	<i>virF</i>	38511	39299	789	789	Complete	AY206433
8	<i>ospE2</i>	40311	40577	267	267	Complete	AY206450
9	<i>ipaH2.5</i>	43257	44948	1,692		Southern positive	
10	<i>orf47</i>	46738	47513	776	776	Complete	AY206434
11	<i>ospC2</i>	49695	51149	1,455	1,455	Complete	AY206435
12	<i>sepA</i>	54594	58688	4,095		Absent	
13	<i>ipaH7.8</i>	64062	65759	1,698	1,698	Complete	AY206436
14	<i>ipaH4.5</i>	66187	67911	1,725		Poor sequence	
15	<i>ospD3</i>	74911	78350	3,440	3,440	Complete	AY206437
16	<i>orf81</i>	80644	82400	1,757		No amplification	
17	<i>orf85</i>	85586	86343	758		Absent	
18	<i>ospC3</i>	90851	92305	1,455	1,455	Complete	AY206438
19	<i>orf94</i>	94791	95567	777		Absent	
20	<i>ipaJ</i>	99403	131402	32,000	32,000	Complete	AY206439
21	<i>orf136</i>	136624	137028	405	405	Complete	AY206440
22	<i>orf137</i>	137580	138356	777	777	Complete	AY206440
23	<i>virA</i> (part a)	144615	149654	5,040	1,203	144615–145817	AY206441
	<i>virA</i> (part b)				3,309	146346–149654	AY294290
24	<i>ushA</i>	151839	153491	1,653	1,653	Complete	AY206442
25	<i>phoN1</i>	155789	156538	750		Poor sequence	
26	<i>orf157</i>	157598	163394	5,797		Absent	
27	<i>orf169a</i>	169678	170550	873	873	Complete	AY206443
28	<i>orf171</i>	171912	173950	2,039	526	Not attempted	AY206444
29	<i>ipaH9.8</i>	174343	177299	2,957	1,638	174343–175980	AY206445
30	<i>orf182</i>	181974	184886	2,913		No amplification	
31	<i>orf185</i>	185369	189266	3,898	3,815	185452–189266	AY206446
32	<i>trbH</i>	191768	199128	7,375	7,375	193484–199128	AY206447
33	<i>orf201</i>	201725	203963	2,239	3,836	201725–204398	AY206448
34	<i>ipaH1.4</i>	206811	209194	2,384	1,728	206811–208538	AY206449
35	<i>orf212</i>	212330	212896	567		No amplification	
Total				103,727	79,364		

^a Start and end of pINV F5 gene coding sequences are as in the work of Buchrieser et al. (9). The first gene of the region is indicated for easy identification.

^b If a region was only partially sequenced, the start and end positions of the sequenced segment are given. For regions where no sequence was obtained, the reasons are given: Southern positive means that PCR failed but that Southern hybridization indicated that the region is present. No amplification means that PCR failed but that Southern hybridization was not done. Absent means that one or more genes of the region were shown to be absent by Southern hybridization.

gions of an A form of pINV from *S. flexneri* serotype 6 (designated the F6 plasmid) to compare with the sequenced *S. flexneri* serotype 5 pINVB form (designated the F5 plasmid). The divergence ranged from 0 to 5.56%, with most genes differing by 0 to 1.0%, consistent with divergence levels in housekeeping genes. However, one group of genes and a few isolated genes showed much higher levels of divergence. The comparison indicated that divergence in these genes is driven by selection pressure, which aids our understanding of virulence variation and mechanisms of pathogenesis.

MATERIALS AND METHODS

Bacterial strains. An *S. flexneri* serotype 6 isolate (M1382) was used for the study, with *S. flexneri* serotype 5 isolate M1356 included as a control for PCR amplification. Both isolates were used in previous studies (20, 36).

DNA sequencing strategy. pINV plasmid DNA was prepared as described previously (42). Primers were designed to amplify overlapping 1-kb segments for each non-IS region based on *S. flexneri* serotype 5 pINV (9).

PCR and DNA sequencing. The following cycling profile was used: initial denaturation at 94°C for 2 min and then 35 cycles of 94°C for 15 s, 55 to 60°C for 30 s, and 72°C for 1 min with a final extension at 72°C for 5 min. The annealing

temperature was between 55 and 60°C depending on primer pair. PCR products were purified with the Wizard purification system (Promega). Samples were sequenced using dye terminator technology (Perkin-Elmer) through the Sydney University and Prince Alfred Hospital Macromolecule Analysis Center and with an automated 377 DNA sequencer (Applied Biosystems).

DNA sequence editing and analysis. Sequences were edited using the PHRED, PHRAP, and CONSED programs (<http://bozeman.mbt.washington.edu>) (12). Sequence comparisons were done using MULTICOMP (39). Open reading frame (ORF) analysis and database searches were done using ORF finder and BLAST search facilities at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>). The synonymous and nonsynonymous substitution rates were calculated using a program provided by Li (23).

RESULTS AND DISCUSSION

Comparison of pINV B and pINV A coding regions. The pINV plasmid genome has a high proportion of IS DNA, and the non-IS DNA was treated as 35 regions for which we give the start position of the first gene and the end position of the last gene in Table 1. We have used the gene annotation of Buchrieser et al. (9), and their genetic map will be of assistance in reading this paper. We attempted to sequence from *S. flex-*

neri serotype 6 all non-IS coding regions defined by Buchrieser et al. (9), a total of 103,727 bp. Both PCR and sequencing depended on the use of primers based on *S. flexneri* serotype 5 pINV, and for a number of regions we failed to obtain a PCR product, even after attempts with alternative primers. We obtained a sequence of 79,364 bp as 25 segments (Table 1), and six genes (see below) were shown by Southern blotting to be absent, accounting for a further 6,986 bp. We considered this to be more than sufficient for a comparison of the two plasmids and did not attempt to determine the nature of the changes that led to the failure to amplify the remaining 17 kb. We are naming the plasmid pINV_F6_M1382 in the GenBank entries to include both serotype and strain name, as we anticipate further sequencing of pINV genes, but in this paper refer to the two sequences as pINV F5 and pINV F6.

There were 511 polymorphic nucleotides, 311 (60.86%) of the substitutions involving amino acid substitution. The average percentage difference is 0.74%, ranging from 0 (11 genes-ORFs) to 5.56% for individual genes (Table 2; Fig. 1). There are 65 genes with differences in the range of 0 to 0.99% with an average of 0.35%, resembling *mxiA* and *mxiC*, sequenced previously from multiple *Shigella* strains (20) with 0.15 and 0.56% differences, respectively, and 15 genes with differences in the range of 1.0 to 5.56% and an average of 2.43%, including *ipgD* of our previous study, with a 3.4% difference.

There are some systematic patterns in the levels of difference (Fig. 1). The 19 genes with differences above the average of 0.71% are clustered, the majority being in the 32-kb entry region. Ten of these genes are in two groups of four (*ipaD*, *ipaC*, *ipaB*, and *ipgC*) and six (*icsB*, *ipgD*, *ipgE*, *ipgF*, *mxiG*, and *mxiH*) separated by only two genes. This pattern could have arisen by recombination. However, the *ipgD*, *mxiA*, and *mxiC* genes, which we studied previously for multiple *Shigella* strains, cover most of the range in level of sequence differences, but the difference between *ipgD* and the others is seen at all levels, from divergence of pINV A and B forms to divergence between closely related strains (20). This is consistent with these two regions, which include *ipgD*, having a mutation rate three to four times higher than the average for the plasmid genome. If the clustering of divergent genes were due to recombination, one would expect a long branch length at the point where the recombination occurred, but otherwise similar levels of variation as for other genes. As described below these genes are thought to be under selective pressure which drives the faster divergence. Perhaps it is more efficient in genome organization for these genes to be clustered for both functional and mutational purposes.

Values of K_s , K_a , and the ratio K_s/K_a are also given in Table 2. K_s is a measure of the rate of synonymous substitutions, being base substitutions without amino acid substitution, and K_a is a similar measure for substitutions which cause amino acid substitution. Synonymous substitutions are generally neutral or of very low adaptive value. Many amino acid substitutions are deleterious and some are neutral, but, critically, adaptive mutations are almost always in the latter category (18). The ratio K_s/K_a is often used as an indicator of neutral or adaptive variation. All genes are subjected to purifying selection to remove deleterious mutations, and it appears that this applies to most nonsynonymous substitutions, giving for an *E. coli-Salmonella enterica* comparison an average K_s/K_a ratio of

10 to 20 for most genes (45). Genes subject to selection for change have a lower value for K_s/K_a , as selection for adaptive amino acid substitutions increases K_a , which lowers the K_s/K_a ratio. In contrast, when strong conservation of amino acid sequences is required for biological function, K_s greatly exceeds K_a .

Of the 80 genes sequenced, 11 have only synonymous substitutions and 9 have only nonsynonymous substitutions, while for the others K_s/K_a ranges from 0.37 to 8.8. For nine genes K_s/K_a is under 1.0. Apart from the genes with only synonymous substitutions, there are only three genes, *spa24*, *ipgB2*, and *spa29*, with a ratio above 8, which is close to the range for housekeeping genes. Spa24 and Spa29 are putative inner membrane proteins (6). For the majority (35 genes) the ratio is between 1 and 8, and it is difficult to give a cutoff that indicates which are under selection pressure for change. For this discussion we use the value of 4.0 as the cutoff, which we consider to be conservative, given that most genes have K_s/K_a ratios between 10 and 20 in *E. coli-S. enterica*. We categorize the 30 genes with K_s/K_a ratios between 1 and 4 as being under positive selection pressure and the 5 genes with K_s/K_a ratios between 4 and 8 as being intermediate and not easily categorized. Therefore, of the 80 genes, a total of 48, including 9 with nonsynonymous changes only, 9 with K_s/K_a ratios less than 1, and 30 with K_s/K_a ratios between 1 and 4, appear to be under positive selection according to the K_s/K_a ratio. This is a very high proportion, indicating the adaptive nature of invasive virulence and its functional complexity. For many genes the total number of substitutions is very low, but the overall picture is of many genes under selection pressure for change.

Genes for secreted proteins. The patterns of variation in secreted proteins are of particular interest. Many of these proteins are effector molecules interacting with host cells, and such proteins have been reported to be more diverse in other type III secretion systems (19, 33). There are at least 25 genes encoding proteins secreted by the Mxi-Spa secretion apparatus (9), of which we were able to sequence 23. The average sequence difference is 0.87%, which is not much higher than average for the genes sequenced. However, five genes (*ospB*, *ospC2*, *ospD2*, *ospF*, and *mxiL*) have only nonsynonymous substitutions and another four (*ipaH9.8*, *ospC1*, *ipaA*, and *ipaD*) have a K_s/K_a ratio less than 1.0, an indication that these genes are under strong positive selection pressure for amino acid variation.

The *ipa* genes (*ipaA* to *ipaD*) encode proteins which are invasins (27, 34). They are humoral immunogens, the predominant antigens recognized by sera from human patients and animal models (15, 38, 49), and so it is not surprising to find that all are under positive selection pressure with K_s/K_a ratios for *ipaA* and *ipaD* less than 1.0 and for *ipaC* and *ipaB* around 2.5. IpaC and IpaB proteins are inserted into the host membranes to form a pore for translocation of other invasins into target cells (5). Variation in these genes may be due to adaptation for optimal interaction with the host cell.

Two of the secreted proteins are toxins. OspD2 is annotated as shET2-2, an enterotoxin, by Venkatesan et al. (50), and OspD3 (SenA) is a known enterotoxin (29). Both genes appear to be under positive selection. *ospD2* has no synonymous substitutions with a single amino acid substitution, and *ospD3* has a K_s/K_a ratio of 1.28. However, the enterotoxin gene *ospD3*

TABLE 2. Sequence variation of pINV between F6 and F5 and between F2A and F5 plasmids

Gene	Region	Position ^a	Length (bp)	No. of polymorphisms			K _s ^c	K _a ^c	Ratio	F2A-F5 comparison		
				Total poly ^b	Nonsyn ^b	%				Total poly ^b	Nonsyn ^b	%
<i>icsP</i>	1	991	948	5	4	0.527	0.003	0.006	0.51	0	0	0
<i>ospB</i>	2	3485	867	1	1	0.115	0	0.002	0	0	0	0
<i>phoN2</i>	2	4680	741	2	2	0.27	0	0.003	0	2	2	0.27
<i>ospD2</i>	3	9501	1,710	1	1	0.058	0	0.001	0	1	1	0.058
<i>ospF</i>	3	11642	720	1	1	0.139	0	0.002	0	2	2	0.278
<i>orf13</i>	3	13393	579	5	5	0.864	0	0.012	0	1	1	0.173
<i>ospD1</i>	4	20964	678	6	5	0.885	0.004	0.011	0.4	2	1	0.295
<i>orf22</i>	4	22153	315	3	1	0.952	0.019	0.005	3.78	1	0	0.317
<i>ipgB2</i>	4	22749	567	3	1	0.529	0.022	0.003	8.28	0	0	0
<i>parA</i>	6	29020	1,200	2	0	0.167	0.008	0	NA ^d	2	0	0.167
<i>parB</i>	6	30219	981	1	0	0.102	0.003	0	NA	2	0	0.204
<i>virF</i>	7	38511	789	1	0	0.127	0.004	0	NA	0	0	0
<i>ospE2</i>	8	40311	267	8	7	2.996	0.039	0.033	1.182	0	0	0
<i>orf47</i>	10	46738	483	3	2	0.621	0.006	0.006	0.9	0	0	0
<i>orf48</i>	10	47211	303	0	0	0	0	0	NA	0	0	0
<i>ospC2</i>	11	49695	1,455	3	3	0.206	0	0.003	0	3	3	0.206
<i>ipaH7.8</i>	13	64062	1,698	6	2	0.353	0.007	0.002	4	4	0	0.236
<i>ospD3</i>	15	74911	1,698	5	3	0.294	0.003	0.003	1.28	0	0	0
<i>ospC1</i>	15	76938	1,413	3	2	0.212	0.002	0.002	0.9	2	2	0.142
<i>ospC3</i>	15	90851	1,455	8	4	0.55	0.012	0.004	3.15	56	24	3.849
<i>ipaJ</i>	20	99403	780	0	0	0	0	0	NA	0	0	0
<i>virB</i>	20	101128	930	3	0	0.323	0.009	0	NA	2	0	0.215
<i>acp</i>	20	102274	237	0	0	0	0	0	NA	1	1	0.422
<i>ipaA</i>	20	102514	1,902	6	5	0.315	0.001	0.004	0.37	1	1	0.053
<i>ipaD</i>	20	104424	999	24	19	2.402	0.022	0.023	0.96	2	1	0.2
<i>ipaC</i>	20	105473	1,092	28	15	2.564	0.048	0.019	2.57	3	3	0.275
<i>ipaB</i>	20	106584	1,743	29	15	1.664	0.029	0.011	2.68	2	1	0.115
<i>ipgC</i>	20	108332	468	7	0	1.496	0.045	0	NA	1	0	0.214
<i>ipgB1</i>	20	108856	627	3	2	0.478	0.005	0.004	1.08	2	1	0.319
<i>ipgA</i>	20	109498	390	1	0	0.256	0.007	0	NA	0	0	0
<i>icsB</i>	20	109900	1,485	30	19	2.02	0.033	0.016	2.12	4	4	0.269
<i>ipgD</i>	20	111698	1,617	55	38	3.401	0.05	0.033	1.52	11	9	0.68
<i>ipgE</i>	20	113324	363	9	4	2.479	0.039	0.018	2.17	0	0	0
<i>ipgF</i>	20	113686	459	10	6	2.179	0.04	0.018	2.25	1	1	0.218
<i>mxiG</i>	20	114153	1,116	27	16	2.419	0.052	0.017	3.08	0	0	0
<i>mxiH</i>	20	115276	252	14	9	5.556	0.121	0.054	2.22	0	0	0
<i>mxiI</i>	20	115540	294	2	1	0.68	0.01	0.004	2.48	0	0	0
<i>mxiJ</i>	20	115839	726	3	3	0.413	0	0.005	0	0	0	0
<i>mxiK</i>	20	116561	528	3	1	0.568	0.011	0.002	4.99	0	0	0
<i>mxiN</i>	20	117030	696	2	1	0.287	0.004	0.002	2.47	1	0	0.147
<i>mxiL</i>	20	117718	408	1	1	0.245	0	0.004	0	1	1	0.245
<i>mxiM</i>	20	118109	429	0	0	0	0	0	NA	0	0	0
<i>mxiE</i>	20	118667	633	1	0	0.158	0.014	0	NA	0	0	0
<i>mxiD</i>	20	119286	1,701	9	5	0.529	0.007	0.004	1.7	3	1	0.176
<i>mxiC</i>	20	121005	1,068	6	4	0.562	0.011	0.006	1.97	1	1	0.094
<i>mxiA</i>	20	122085	2,061	3	1	0.146	0.003	0.001	3.55	1	0	0.049
<i>spa15</i>	20	124158	402	2	1	0.498	0.007	0.004	1.78	0	0	0
<i>spa47</i>	20	124563	1,293	9	5	0.696	0.012	0.006	2.01	2	0	0.155
<i>spa13</i>	20	125948	339	1	0	0.295	0.008	0	NA	0	0	0
<i>spa32</i>	20	126273	879	0	0	0	0	0	NA	0	0	0
<i>spa33</i>	20	127145	882	4	2	0.454	0.006	0.003	2.43	1	1	0.134
<i>spa24</i>	20	128031	651	11	3	1.69	0.057	0.007	8.77	0	0	0
<i>spa9</i>	20	128696	261	3	1	1.149	0.022	0.005	4.84	0	0	0
<i>spa29</i>	20	128956	771	4	1	0.519	0.017	0.002	8.1	1	1	0.13
<i>spa40</i>	20	129733	1,029	6	2	0.583	0.017	0.003	6.37	0	0	0
<i>orf131a</i>	20	130774	243	2	0	0.823	0.053	0	NA	0	0	0
<i>orf131b</i>	20	131133	270	0	0	0	0	0	NA	0	0	0
<i>orf136</i>	21	136624	405	0	0	0	0	0	NA	0	0	0
<i>orf137</i>	21	137580	777	4 (5) ^e		0.515				2	1	0.279
<i>virA</i>	23	144615	1,203	7	4	0.582	0.011	0.004	2.54	0	0	0
<i>icsA</i>	23	146346	3,309	9	8	0.272	0.001	0.003	0.33	2	1	0.06
<i>ushA</i>	24	151839	1,653	4	2	0.242	0.009	0.002	4.62	0	0	0
<i>orf169a</i>	27	169678	483	1	0	0.207	0	0.003	0	0	0	0
<i>orf169b</i>	27	169912	639	3	1	0.469	0.009	0.002	4.62	1	1	0.156
<i>ccdA</i>	28	173425	219	0	0	0	0	0		0	0	0
<i>ccdB</i>	28	173624	327	0	0	0	0	0		0	0	0

Continued on following page

movement within the host cell. There are nine substitutions in *icsA* with eight changing the amino acid. VirK is required for proper production or localization of IcsA, although its precise function is not known. *virK* shows high levels of divergence. IcsP is an outer membrane protease for cleavage of IcsA. Four of the five changes in *icsP* are nonsynonymous.

The *Shigella* type III secretion system is composed of a basal body and an external needle (6, 47). The needle interacts with the host cell to deliver effector proteins. MxiH and MxiI are the major and minor needle components, respectively. The *mxiH* gene, only 252 bp in size, has the highest variation (5.556%) with a K_s/K_a ratio of 2.22, while *mxiI* has two changes with one being nonsynonymous. Variation in these genes may confer a selective advantage because of direct interaction with host cells. It has been observed that, when MxiH was overexpressed, the type III secretion machinery protruded longer needles than did the wild type, and the bacteria invaded the host cells much more efficiently than the wild type did (47). However, genes encoding proteins for the basal body, *mxiD*, *mxiG*, and *mxiI*, also show low K_s/K_a ratios, for which we have no explanation.

Virulence-associated regulatory genes under purifying selection. In contrast to the genes discussed above that are under selection pressure for change, several genes known to be associated with regulation of virulence are strongly conserved. *spa32* has no variation at all while *virF*, *virB*, *ipgC*, and *mxiE* all have only synonymous substitutions. Spa32 and IpgC have been discussed above. VirF is a transcription activator of the AraC family required for expression of genes of the entry region and *icsA* (31, 41). It positively controls synthesis of VirB, a DNA-binding protein which binds to promoter regions of virulence genes for activation of transcription (2). MxiE is a transcriptional regulator (17, 26) for at least six secreted pINV proteins (OspB, OspC1, OspE2, OspF, VirA, and IpaH9.8). This observation of regulatory genes being subject to purifying selection also gives indirect support to the conclusion above that the changes observed in other genes are a result of positive selection pressure, rather than accumulation of mildly deleterious mutations.

Genes absent in the F6 plasmid. A few genes were confirmed as absent in the F6 plasmid by hybridization (data not shown) after PCR failed to generate a product. These include *sepA*, *orf85*, *orf85b*, *orf94*, *stbA*, and *stbB*. The whole *stbA* and *stbB* region is likely absent. We discuss *stbA* and *stbB* in the replication section and *sepA* below, but the functions of the other three genes are unknown.

It is interesting that *sepA* is absent, as an *S. flexneri* serotype 5 *sepA* mutant exhibited attenuated virulence in the rabbit ligated ileal loop model (3). SepA is a serine protease of the immunoglobulin A1 protease family and the major protein secreted by *S. flexneri* serotype 5 (3). SepA is not required for entry into cultured epithelial cells, nor for intercellular dissemination (3), but might be involved in invasion and destruction of the host intestinal epithelium (4). We surveyed some other *Shigella* strains and found *sepA* to be variably present. Among pINV A plasmids it is present in those of *S. dysenteriae* serotype 10 and *S. boydii* serotype 9 but absent from those of *S. boydii* serotype 14 and *S. dysenteriae* serotype 3, while it is absent from the *S. sonnei* plasmid, the only other pINV B form tested (data not shown).

There are three genes, *impCAB*, that are known to be absent in *S. flexneri* serotype 5 (50) but present in some other strains (40). Southern hybridization showed them to be absent in *S. flexneri* serotype 6 (data not shown).

Comparison of replication, partition, and transfer regions. Several regions of the plasmid are involved in replication, partition, stable maintenance, or transfer (9, 50). In general genes in these regions have lower levels of difference than do those in other regions.

We sequenced the whole replication region which includes *oriR* (the origin of replication), *repA* (required for initiation of replication at *oriR*), *repB* (*copB*) (repressor of transcription of *repA*), *tapA* (upstream of *repA*, encoding leader peptide required for translation of *repA*), and *copA* (coding for antisense RNA that binds to leader region of *repA* for copy number control). The only sequence differences were four single base substitutions in the *oriR* region of 362 bases. We also sequenced downstream of *oriR* to the IS. Substantial differences were observed from the start of *repA4*, a pseudogene. The F5 plasmid has a segment of 220 bp after *repA4*, followed by IS 91.04 and part of IS 1294, whereas the F6 plasmid has a unique 577-bp segment after *repA4* followed by part of IS 1294, which precedes the IS 1294 in the F5 plasmid by 285 bp. Either this part of the sequence has undergone recombination, or one or both have undergone rapid change. However, there appear to be no functional genes in either the F6 or F5 sequence.

(i) Plasmid partition genes. Low-copy-number plasmids generally have more than one system for stable maintenance of the plasmid (14). These may affect partitioning or cause post-segregational killing (PSK), the latter being a "fail-safe" mechanism (14). pINV F5 has genes for two partition systems, *parAB* and *stbAB*, and two PSK systems, *ccdAB* and *mvpTA* (9, 50). Note a possible source of confusion through use of the name *stb* for different genes. The genes now known as *mvpA* and *mvpT* were previously called STBORF1 and STBORF2. The *stbA* and *stbB* genes referred to in this paper are the second of two partition systems and were first observed by Buchrieser et al. (9) when the plasmid genome was sequenced.

The *parA* and *parB* genes of *S. flexneri* serotypes 5 and 6 are identical, but the *stbAB* genes are absent in *S. flexneri* serotype 6. The *stbAB* genes are also absent in *S. sonnei* but present in *S. dysenteriae* serotypes 1 and 4 and *S. boydii* serotype 1, with presence or absence found in both pINV A and pINV B forms. For the two PSK systems in *S. flexneri* serotype 5, the *ccdA* and *ccdB* genes are identical in *S. flexneri* serotype 6, while for *mvpT* and *mvpA*, which are encoded on the opposite strand of *trbH*, *mvpT* has only one substitution (nonsynonymous), and *mvpA* has two substitutions (one nonsynonymous).

The only reported functional difference between the two pINV forms is incompatibility. An early study (25) showed that the *Shigella* pINV plasmids belong to two incompatibility groups, later found to correspond to pINV A and pINV B (20). The complete sequence comparison allows us to look for substitutions affecting incompatibility. There are several factors that determine plasmid incompatibility. pINV belongs to the IncFII family (46), in which the major incompatibility determinant is the *copA* (*inc*) RNA (14). However, *copA* is identical in *S. flexneri* serotypes 5 and 6 and so cannot be the determinant of their incompatibility difference.

We consider next the possibility that *mvpAT* is responsible

for the difference in compatibility. Compatibility was tested by introducing a plasmid (pMSYH6610) with kanamycin resistance and the compatibility region of the *S. flexneri* serotype 2a plasmid into *Shigella* strains with selection for kanamycin. The *mvpAT* system of the *S. flexneri* serotype 2a pINV B is present in plasmid MSYH6610 and in the compatibility test is retained in all surviving cells. When MSYH6610 is introduced into *Shigella* strains with kanamycin selection, the pINV plasmids of *S. flexneri* serotype 2a, 1b, or 2b (and presumably *S. flexneri* serotype 5) are lost, whereas pINV B of *S. flexneri* serotype 6 is stably retained. One would expect plasmids with the same copy number system to be lost under these conditions unless there was a PSK system operating on *S. flexneri* serotype 6 and a sufficient number of cells to retain two copies for these to maintain growth in the presence of kanamycin. MvpT is the toxin and MvpA is the antidote (44), so survival of pINV F6 indicates that it has an MvpT toxin that is not inactivated by the *S. flexneri* serotype 2a MvpA. It seems at first sight unlikely, as each differs from the *S. flexneri* serotype 2a form by only one amino acid residue. However, the F plasmid homologue is compatible with *S. flexneri* serotype 2a pINV (same form as that of *S. flexneri* serotype 5), and there are only four amino acid differences between MvpA of the F plasmid and that of *S. flexneri* serotype 5 and two differences in MvpT (37). One or more of these substitutions must determine the difference in incompatibility between F and pINV B. It is also possible that pINV of *S. flexneri* serotype 6 has a PSK system that is not present at all on pINV from *S. flexneri* serotype 2a, which would not have been detected by our use of only *S. flexneri* serotype 5-based PCR primers.

(ii) The plasmid transfer (*tra*) region. It is known from tests on several *S. flexneri* strains that pINV is unable to initiate conjugation (43). However, there is a partial *tra* region in pINV F5, including the *trbH*, *traI*, *traX*, and *finO* genes and part of *traD*. This is only about 25% of the total *tra* operon as seen in the K-12 F factor, and as virtually the whole is needed for function, the remaining genes presumably have no function. We obtained sequences from *trbH*, *traI*, *traX*, and *finO* genes. There are four substitutions in *trbH*, of which three are nonsynonymous, giving a ratio of 0.613, but as *trbH* presumably has no role in the plasmid, divergence may result from selection on the *mvpAT* genes on the opposite strand. *traI* is nonfunctional in both F5 and F6. In comparison with the *traI* gene of the F plasmid, there are a deletion of seven bases and an insertion of seven bases in F5 and F6, respectively, in a region flanked by two palindromic sequences that disrupted the reading frame. *traI* also has 21 point mutation substitutions between F5 and F6. The *traI* gene in F6 suffered further damage with a point mutation of C to T at position 956, resulting in a stop codon. This indicates that the *traI* gene was inactivated independently after the divergence of the two plasmids. *traX* has a single synonymous substitution while *finO* has three substitutions, of which two are nonsynonymous.

ORFs of unknown function. Ten of the ORFs sequenced have no known function (*orf13*, *orf22*, *orf131a*, *orf131b*, *orf136*, *orf137*, *orf186*, *orf201*, *orf47*, and *orf48*), and two (*orf169a* and *orf169b*) are probably involved in plasmid maintenance (9). *orf48*, *orf131b*, and *orf136* have no substitutions while all or the majority of substitutions in *orf131a*, *orf169a*, and *orf169b* are synonymous. However, *orf13* and *orf47* are clearly under pos-

itive selection pressure. All five substitutions in *orf13* are nonsynonymous, and two of the three substitutions in *orf47* are also nonsynonymous with a K_s/K_a ratio less than 1. These ORFs may play an important role in virulence, making them candidates for functional analysis.

The only gene to have suffered a frameshift mutation is *orf137*, which is flanked by IS elements on both sides. It may be part of a gene that has been disrupted by an IS element, in which case the frameshift observed could be part of a continuing process of gene degradation.

Comparison of two pINV B plasmids. The pINV B plasmid of *S. flexneri* serotype 2a (plasmid F2A) has now been sequenced as part of a genome sequence (16). We compared it with the B-form pINV F5 plasmid to enhance our understanding of pINV evolution (Table 2). *S. flexneri* serotypes 2a and 5 belong to the same cluster based on chromosomal gene trees (36), and divergence is relatively recent. All 80 genes that we sequenced in F6 are all present in F2A. Of those, 37 are identical in F5 and F2A, 17 have a single base substitution, and 24 have from two to four substitutions. *traI* and *ospC3* have much bigger differences. *traI* has eight single nucleotide substitutions, a 10-bp deletion, and an inversion at a palindromic sequence. *ospC3* has 23 substitutions, with part of the sequence being similar to *ospC2*, most likely as a result of inter- or intrachromosomal recombination. If we consider *traI* and *ospC* to have undergone 10 events and 1 event, respectively, we have a total of 76 polymorphisms (65 if *traI* and *ospC* are excluded), to be compared with 494 polymorphic nucleotide sites in the F5-F6 comparison. The much greater divergence of F5 and F6 relative to F5 and F2A over 80 genes strongly supports the recognition of two forms of pINV, based originally on the study of only three genes (20).

We also looked at the number of synonymous and nonsynonymous substitutions between F2A and F5. The 11 genes that are identical in F5 and F6 have between them only one (nonsynonymous) substitution in F2A. The 11 genes with synonymous changes only in F5-F6 have 10 substitutions with three being nonsynonymous. The nine genes that have only nonsynonymous F5-F6 substitutions have a total of 10 F5-F2A substitutions, all nonsynonymous. For the genes with a K_s/K_a ratio of 4 or less, we excluded *ospC3* from the calculation, but for the 38 other genes in this category 250 of 402 F5-F6 substitutions and 36 of 59 F5-F2A substitutions are nonsynonymous. For the genes with a K_s/K_a ratio above 4 in the F5-F6 comparison, 12 of 37 F5-F6 substitutions are nonsynonymous, but two of two F5-F2A substitutions are nonsynonymous. The pattern of synonymous and nonsynonymous substitutions in the F5-F2A comparison of two pINV B plasmids mirrors the pattern observed in the F5-F6 comparison with the possible exception of the last category where the number of F5-F2A substitutions is too low to be meaningful. It appears that genes thought to be under positive selection in the F5-F6 comparison have similar characteristics in the F2A-F5 comparison.

Variation in the invasion gene homologues in *S. enterica*. The invasion ability of *S. enterica* is determined by a cluster of genes many of which are homologous to the invasion genes of pINV. Eight have been compared in multiple strains of *S. enterica* (7, 22). The K_s/K_a ratio of *invE* (pINV *mxiC*), *invA* (*mxiA*), *spaP* (*spa24*), and *spaQ* (*spa9*) is similar to that of housekeeping genes while *invH* (no pINV homologue), *spaM*

(*spa13*), *spaN* (*spa32*), and *spaO* (*spa33*) have a much lower K_s/K_a ratio. It was concluded (7, 22) that proteins that are membrane bound or membrane associated are relatively conserved whereas those that are exposed to the extracellular environment are hypervariable, reflecting the action of diversifying selection. Our data with a much larger number of genes also show that in general proteins exposed to the extracellular environment are more variable, but it is interesting that, for the seven homologues of the *S. enterica* genes used, the genes can differ in K_s/K_a ratio in *E. coli* and *S. enterica*. In fact only three genes, *spaP* (*spa24*), *spaQ* (*spa9*), and *spaO* (*spa33*), are consistent in the two species. *spaM* and *spaN* have a low K_s/K_a ratio in *S. enterica*, while there is only one (synonymous) change in *spa13* and no change in *spa32* between F5 and F6 pINV plasmids, and *mxiA* and *mxiC* have low K_s/K_a ratios of 3.55 and 1.97, respectively, while those of the *S. enterica* homologues are similar in this regard to housekeeping genes.

Spa32 and SpaN share only 15% amino acid identity, but the function of Spa32 can be complemented by SpaN (48), whereas there is known functional difference between these homologues (11). Spa32 is implicated in the release of the Ipa proteins but not their surface presentation. In contrast, SpaN is involved in both the secretion and the surface presentation of the Sip proteins (10). However, the earlier observation that Spa32, unlike SpaN, is not secreted to the culture supernatant (51) may not be entirely correct, as it was recently found that Spa32 is translocated through the type III secretion machinery into the medium (48). This may explain the difference in selection pressures between Spa32 and SpaN. It seems likely that there are functional differences for the other three homologues to account for the opposing trends.

Concluding comments. The pINV plasmid of *Shigella*-EIEC occurs in two major forms as shown earlier by distribution of sequence forms of three genes sequenced from a range of strains. We have now sequenced 80 genes from the A-form plasmid, pINV F6, enabling us to make a detailed comparison of representative pINV A and pINV B plasmids. The three genes, *ipgD*, *mxiA*, and *mxiC*, studied previously (20) are shown to be representative of the range in level of difference between pINV A and pINV B plasmids. Generally genes for plasmid maintenance vary less than virulence-associated genes, those encoding secreted proteins that interact with host cells being particularly variable.

Many genes appear to be under selection pressure for change. In addition some of the F5 genes are absent in F6. Some of the plasmid-encoded proteins are immunogens, and for them selection may be in part at least for avoidance of the host immune system. More interesting is the possibility that differences may reflect ongoing adaptation to the *Shigella* niche, or selection for variants of that niche. *Shigella* strains infect only humans and as a group are a major human pathogen. Their mode of pathogenesis is thought to have become effective only 35,000 to 270,000 years ago, and its antecedents are not known. The high proportion of genes under selection pressure in the pINV plasmid that encodes many of the functions characteristic of this mode of pathogenesis may reflect continuing adaptation to achieve an optimal interaction with the human host. It is also possible that the many serotypes are adapted to different variants of the general niche occupied, and that the selection relates to competition between different *Shi-*

gella clones which differ in details of their adaptation to the niche. The observation that *senA* (29), a toxin gene apparently under selection pressure for change, and *sepA*, known to play a role in virulence in *S. flexneri* serotype 2a (3, 4), are absent in some *Shigella* strains supports the hypothesis that *Shigella* strains differ in the details of their adaptation to their mode of pathogenesis. A similar explanation may account for the observation that genes under selection pressure differ between *E. coli* and *S. enterica*.

As noted previously (20), not much has been reported on variation in virulence of *Shigella* clones, but only a few serotypes are responsible for many of the cases and there are also perceived differences between *Shigella* and EIEC clones. It could be interesting to relate variation in virulence within and between *Shigella* and EIEC strains to variation in the genes under selection pressure and genes present or absent in the pINV plasmids. This may help us to better understand the pathogenesis and epidemiology of *Shigella*-EIEC infections. Comparative functional studies of the genes that differ may point to the factors that give *S. flexneri*, *S. sonnei*, and *S. dysenteriae* serotype 1, for example, their distinctive disease characteristics and global distributions.

Incompatibility between the two plasmids was deduced to be due to very small differences: a single amino acid substitution in *mvpA* being the only difference likely to be involved. Its role in gene flow-recombination between the two forms and plasmid survival is not clear.

pINV plasmids are nonconjugative (25, 43). The substitution patterns in the incomplete *tra* region indicate that the remaining genes might be functional. This raises questions of transfer of pINV to other *E. coli* strains since many such events are presumed to have occurred to give rise to *Shigella*-EIEC strains of diverse genetic backgrounds (20). The remaining *tra* region genes may be involved in this process, complemented by genes on other plasmids in the recipient cells, leading to pINV transfer.

ACKNOWLEDGMENTS

This research is supported by a grant from the National Health and Medical Research Council of Australia.

We thank the anonymous referees for comments and suggestions.

REFERENCES

1. Acheson, D. W. K., and G. T. Keusch. 1995. *Shigella* and enteroinvasive *Escherichia coli*, p. 763–784. In M. J. Blaser, P. D. Smith, J. I. Ravdin, H. B. Greenberg, and R. L. Guerrant (ed.), *Infections of the gastrointestinal tract*. Raven Press, New York, N.Y.
2. Beloin, C., S. McKenna, and C. J. Dorman. 2002. Molecular dissection of VirB, a key regulator of the virulence cascade of *Shigella flexneri*. *J. Biol. Chem.* 277:15333–15344.
3. Benjelloun-Touimi, Z., P. J. Sansonetti, and C. Parsot. 1995. SepA, the major extracellular protein of *Shigella flexneri*: autonomous secretion and involvement in tissue invasion. *Mol. Microbiol.* 17:123–135.
4. Benjelloun-Touimi, Z., M. S. Tahar, C. Montecucco, P. J. Sansonetti, and C. Parsot. 1998. SepA, the 110-kDa protein secreted by *Shigella flexneri*: two-domain structure and proteolytic activity. *Microbiology* 144:1815–1822.
5. Blocker, A., P. Gounon, E. Larquet, K. Niebuhr, V. Cabiliaux, C. Parsot, and P. Sansonetti. 1999. The tripartite type III secretin of *Shigella flexneri* inserts IpaB and IpaC into host membranes. *J. Cell Biol.* 147:683–693.
6. Blocker, A., N. Jouihri, E. Larquet, P. Gounon, F. Ebel, C. Parsot, P. Sansonetti, and A. Allaoui. 2001. Structure and composition of the *Shigella flexneri* “needle complex,” a part of its type III secretin. *Mol. Microbiol.* 39:652–663.
7. Boyd, E. F., J. Li, H. Ochman, and R. K. Selander. 1997. Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica*. *J. Bacteriol.* 179:1985–1991.

8. Brenner, D. J., G. R. Fanning, G. V. Miklos, and A. G. Steigerwalt. 1973. Polynucleotide sequence relatedness among *Shigella* species. *Int. J. Syst. Bacteriol.* **23**:1-7.
9. Buchrieser, C., P. Glaser, C. Rusniok, H. Nedjari, H. D'Hauteville, F. Kunst, P. Sansonetti, and C. Parsot. 2000. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. *Mol. Microbiol.* **38**:760-771.
10. Collazo, C. M., and J. E. Galan. 1996. Requirement for exported proteins in secretion through the invasion-associated type III system of *Salmonella typhimurium*. *Infect. Immun.* **64**:3524-3531.
11. Galan, J. E. 1996. Molecular genetic bases of *Salmonella* entry into host cells. *Mol. Microbiol.* **20**:263-271.
12. Gordon, D., C. Abajian, and P. Green. 1998. CONSED—a graphical tool for sequence finishing. *Genome Res.* **8**:195-202.
13. Hale, T. L. 1991. Genetic basis of virulence in *Shigella* species. *Microbiol. Rev.* **55**:206-224.
14. Helinski, D. R., A. E. Toukdarian, and R. P. Novick. 1996. Replication control and other stable maintenance mechanisms of plasmid, p. 2295-2324. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. American Society for Microbiology, Washington, D.C.
15. Hueck, C. J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol. Mol. Biol. Rev.* **62**:379-433.
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, and others. 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K-12 and O157. *Nucleic Acids Res.* **30**:4432-4441.
17. Kane, C. D., R. Schuch, W. A. Day, Jr., and A. T. Maurelli. 2002. MxiE regulates intracellular expression of factors secreted by the *Shigella flexneri* 2a type III secretion system. *J. Bacteriol.* **184**:4409-4419.
18. Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom.
19. Kresse, A. U., F. Beltrametti, A. Muller, F. Ebel, and C. A. Guzman. 2000. Characterization of SepL of enterohemorrhagic *Escherichia coli*. *J. Bacteriol.* **182**:6490-6498.
20. 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.
21. Lan, R., and P. R. Reeves. 2002. *Escherichia coli* in disguise: molecular origins of *Shigella*. *Microbes Infect.* **4**:1125-1132.
22. Li, J., H. Ochman, E. A. Groisman, E. F. Boyd, F. Solomon, K. Nelson, and R. K. Selander. 1995. Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica*. *Proc. Natl. Acad. Sci. USA* **92**:7252-7256.
23. Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and non-synonymous substitution. *J. Mol. Evol.* **36**:96-99.
24. Magdalena, J., A. Hachani, M. Chamekh, N. Jouihri, P. Gounon, A. Blocker, and A. Allaoui. 2002. Spa32 regulates a switch in substrate specificity of the type III secretion of *Shigella flexneri* from needle components to Ipa proteins. *J. Bacteriol.* **184**:3433-3441.
25. Makino, S., C. Sasakawa, and M. Yoshikawa. 1988. Genetic relatedness of the basic replicon of the virulence plasmid in shigellae and enteroinvasive *Escherichia coli*. *Microb. Pathog.* **5**:267-274.
26. Mavris, M., A. L. Page, R. Tournebize, B. Demers, P. Sansonetti, and C. Parsot. 2002. Regulation of transcription by the activity of the *Shigella flexneri* type III secretion apparatus. *Mol. Microbiol.* **43**:1543-1553.
27. Menard, R., C. Dehio, and P. J. Sansonetti. 1996. Bacterial entry into epithelial cells: the paradigm of *Shigella*. *Trends Microbiol.* **6**:220-226.
28. Nataro, J. P., and J. B. Kaper. 1998. Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* **11**:142-201.
29. 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.
30. Ochman, H., T. S. Whitlam, 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.
31. Parsot, C., and P. J. Sansonetti. 1999. The virulence plasmid of shigellae: an archipelago of pathogenicity islands?, p. 151-165. *In* J. B. Kaper and J. Hacker (ed.), *Pathogenicity islands and other mobile virulence elements*. American Society for Microbiology, Washington, D.C.
32. Parsot, C., and P. J. Sansonetti. 1996. Invasion and the pathogenesis of *Shigella* infections. *Curr. Top. Microbiol. Immunol.* **209**:25-42.
33. Perna, N. T., G. F. Mayhew, G. Pósfai, S. Elliott, M. S. Donnenberg, J. B. Kaper, and F. R. Blattner. 1998. Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. *Infect. Immun.* **66**:3810-3817.
34. Philpott, D. J., J. D. Edgeworth, and P. J. Sansonetti. 2000. The pathogenesis of *Shigella flexneri* infection: lessons from in vitro and in vivo studies. *Philos. Trans. R. Soc. Lond. B* **355**:575-586.
35. 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.
36. 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.
37. Radnedge, L., M. A. Davis, B. Youngren, and S. J. Austin. 1997. Plasmid maintenance functions of the large virulence plasmid of *Shigella flexneri*. *J. Bacteriol.* **179**:3670-3675.
38. Raqib, R., F. Qadri, P. Sarker, S. Mia, M., P. J. Sansonetti, M. J. Albert, and J. Andersson. 2002. Delayed and reduced adaptive humoral immune responses in children with shigellosis compared with in adults. *Scand. J. Immunol.* **55**:414-423.
39. Reeves, P. R., L. Farnell, and R. Lan. 1994. MULTICOMP: a program for preparing sequence data for phylogenetic analysis. *Comput. Appl. Biol. Sci.* **10**:281-284.
40. Runyen-Janecky, L. J., M. Hong, and S. M. Payne. 1999. The virulence plasmid-encoded *impCAB* operon enhances survival and induced mutagenesis in *Shigella flexneri* after exposure to UV radiation. *Infect. Immun.* **67**:1415-1423.
41. Sakai, T., C. Sasakawa, and M. Yoshikawa. 1988. Expression of four virulence antigens of *Shigella flexneri* is positively regulated at the transcriptional level by the 30-kilodalton virF protein. *Mol. Microbiol.* **2**:589-597.
42. Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: a laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
43. 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.
44. Sayeed, S., L. Reaves, L. Radnedge, and S. Austin. 2000. The stability region of the large virulence plasmid of *Shigella flexneri* encodes an efficient post-segregational killing system. *J. Bacteriol.* **182**:2416-2421.
45. Sharp, P. M. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *J. Mol. Evol.* **33**:23-33.
46. Silva, R. M., S. Saadi, and W. K. Maas. 1988. A basic replicon of virulence-associated plasmids of *Shigella* spp. and enteroinvasive *Escherichia coli* is homologous with a basic replicon in plasmids of IncF groups. *Infect. Immun.* **56**:836-842.
47. Tamano, K., S. Aizawa, E. Katayama, T. Nonaka, S. Imajoh-Ohmi, A. Kuwae, S. Nagai, and C. Sasakawa. 2000. Supramolecular structure of the *Shigella* type III secretion machinery: the needle part is changeable in length and essential for delivery of effectors. *EMBO J.* **19**:3876-3887.
48. Tamano, K., E. Katayama, T. Toyotome, and C. Sasakawa. 2002. *Shigella* Spa32 is an essential secretory protein for functional type III secretion machinery and uniformity of its needle length. *J. Bacteriol.* **184**:1244-1252.
49. van de Verg, L. L., C. P. Mallett, H. H. Collins, T. Larsen, C. Hammack, and T. L. Hale. 1995. Antibody and cytokine responses in a mouse pulmonary model of *Shigella flexneri* serotype 2a infection. *Infect. Immun.* **63**:1947-1954.
50. Venkatesan, M. M., M. B. Goldberg, D. J. Rose, E. J. Grotbeck, V. Burland, and F. R. Blattner. 2001. Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. *Infect. Immun.* **69**:3271-3285.
51. Watarai, M., T. Tobe, M. Yoshikawa, and C. Sasakawa. 1995. Contact of *Shigella* with host cells triggers release of Ipa invasins and is an essential function of invasiveness. *EMBO J.* **14**:2461-2470.