J Mol Evol (2003) 57:140–148
DOI: 10.1007/s00239-003-2460-3

JOURNAL OF MOLECULAR EVOLUTION

The Evolutionary History of *Shigella* and Enteroinvasive *Escherichia coli* Revised

Patricia Escobar-Páramo,[1] Catherine Giudicelli,[2] Claude Parsot,[3] Erick Denamur[1]

[1] INSERM U458, Hôpital Robert Debré, 48 Boulevard Sérurier, 75019, Paris, France
[2] Centre d'Etude du Polymorphisme Humain, Hôpital Saint Louis, Paris, France
[3] Unité de Pathogénie Microbienne Moléculaire, Institut Pasteur, Paris, France

Received: 21 October 2002 / Accepted: 14 January 2003

Abstract. In *Shigella* and enteroinvasive *Escherichia coli* (EIEC), the etiologic agents of shigellosis in humans, the determinants responsible for entry of bacteria into and dissemination within epithelial cells are encoded by a virulence plasmid. To understand the evolution of the association between the virulence plasmid and the chromosome, we performed a phylogenetic analysis using the sequences of four chromosomal genes (*trpA*, *trpB*, *pabB*, and *putP*) and three virulence plasmid genes (*ipaB*, *ipaD*, and *icsA*) of a collection of 51 *Shigella* and EIEC strains. The phylogenetic tree derived from chromosomal genes showed a typical "star" phylogeny, indicating a fast diversification of *Shigella* and EIEC groups. Phylogenetic groups obtained from the chromosomal and plasmidic genes were similar, suggesting that the virulence plasmid and the chromosome share similar evolutionary histories. The few incongruences between the trees could be attributed to exchanges of fragments of different plasmids and not to the transfer of an entire plasmid. This indicates that the virulence plasmid was not transferred between the different *Shigella* and EIEC groups. These data support a model of evolution in which the acquisition of the virulence plasmid in an ancestral *E. coli* strain preceded the diversification by radiation of all *Shigella* and EIEC groups, which led to highly diversified but highly specialized pathogenic groups.

Correspondence to: Patricia Escobar-Páramo; *email:* pacopa@idf.inserm.fr; or Erick Denamur; *email:* denamur@bichat.inserm.fr

Key words: *Shigella* — *Escherichia coli* — Virulence plasmid — Phylogeny — Radiation

Introduction

In *Escherichia coli*, the diversity of pathogenic strains seems to be the result of the constant arrival of different virulence factors encoded by pathogenecity islands, phages, or plasmids (Lecointre et al. 1998; Ochman et al. 2000; Reid et al. 2000). In enteroinvasive *E. coli* (EIEC) and members of the genus *Shigella*, the etiologic agents of shigellosis in humans, the determinants responsible for entry of these bacteria into and dissemination within epithelial cells are encoded by a virulence plasmid (VP) (Parsot and Sansonetti 1996; Sansonetti et al. 1982; Sasakawa et al. 1992).

For epidemiological purposes, the genus *Shigella* is divided into four nomen-species, *S. boydii*, *S. flexneri*, *S. dysenteriae* (each of which is subdivided into different serotypes), and *S. soneii* (that contains only one serotype). Multilocus enzyme electrophoresis and ribotyping analyses indicated that *Shigella* genotypes are interspersed within *E. coli* genotypes, irrespective of their nomen-species (Goullet and Picard 1987; Ochman et al. 1983; Pupo et al. 1997; Rolland et al. 1998). Phylogenetic analysis of the sequences of several chromosomal genes showed that *Shigella* strains belong to at least seven different groups (Pupo et al. 2000). Thus, *Shigella* and EIEC strains represent diverse, but highly specialized, pathogenic *E. coli*

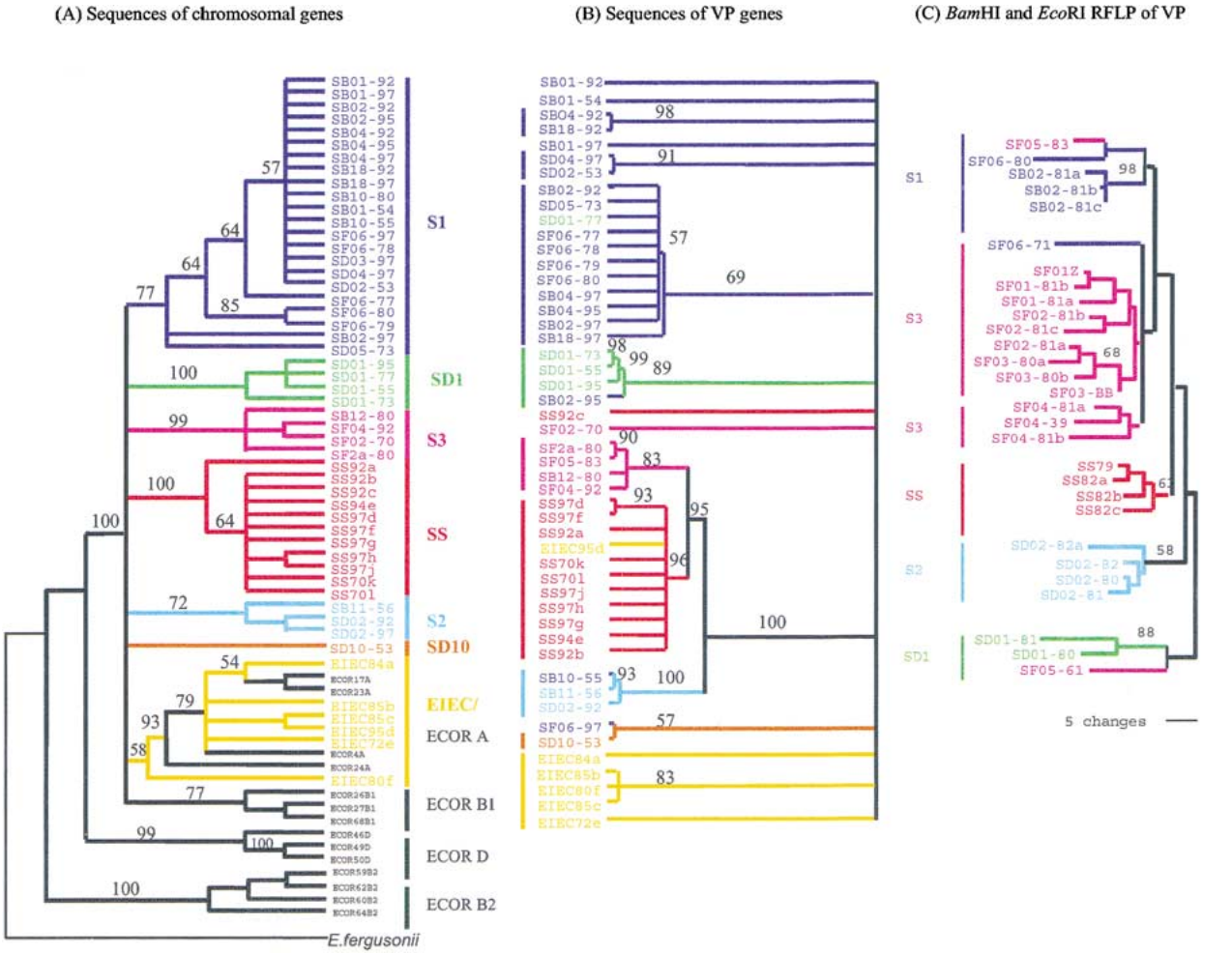

Fig. 1. Phylogenic trees of the chromosome and the virulence plasmid. **A** The 50% bootstrap majority-rule consensus tree based on the simultaneous analysis of the sequences of *trpA*, *trpB*, *pabB*, and *putP* genes of 65 *E. coli* and *Shigella* strains and rooted on *E. fergusonii*. **B** The unrooted 50% bootstrap majority-rule consensus tree based on the simultaneous analysis of the sequences of *ipaB*, *ipaD*, and *icsA* genes of 49 *Shigella* and EIEC strains. Strain SF05-83 corresponds to *S. flexneri* strain M90T (serotype 5), from which the plasmid pWR100 has been entirely sequenced (Buchrieser et al. 2000). **C** The neighbor-joining midpoint-rooted tree based on the *Bam*HI and *Eco*RI RFLP patterns of the VP present in a set of 29 strains that had been analyzed by Sansonetti et al. (1983). For the sake of consistency, the names of the strains used by Sansonetti et al. (1983) have been changed (see Materials and Methods). The color code used for strains in the plasmid phylogenic trees is the same as that used in the chromosomal phylogenic tree. For strain SF05-83 (M90T), the color code is according to serotype (Pupo et al. 2000). Bootstrap values greater than 50% are indicated above the branches. The nomenclature used for the strains is as follows: the letters indicate the nomen-species (SB, *S. boydi;* SD, *S. dysenteriae;* SF, *S. flexneri;* SS, *S. sonnei;* EIEC, enteroinvasive *E. coli*); the first two-digit number indicates the serotype; the second two-digit number indicates the year of isolation of the strains; for SS and EIEC strains, the two-digit number denotes the year of isolation and the last character is used to identify each strain.

groups in which the virulence genes are carried by a plasmid. Analysis of the complete sequence of the VP of one *S. flexneri* strain suggests that the VP is a mosaic of genes of various origins and contains traces of four different plasmids (Buchrieser et al. 2000). The purpose of this study is to determine the evolutionary history of the association of the bacteria and plasmid genomes.

Two hypotheses could account for the origin of the different *Shigella*/EIEC groups. These groups could derive from different groups of *E. coli* that acquired the VP plasmid independently, either from an unknown donor or from other *Shigella* strains that already had the VP. Phylogenetic analysis of chromosomal genes, which showed the diversity of genotypes within the *Shigella* genus, was interpreted as evidence for the independent acquisition of the VP by different *E. coli* groups (Pupo et al. 2000). Alternatively, the VP plasmid could have been transferred, or constructed, only once in an ancestral *E. coli* strain that subsequently gave rise to the different *Shigella* groups. Comparisons of the phylogenies of the chromosome and the plasmid can give clues to the history of the association between the plasmid and the chromosome. According to Souza and Eguiarte (1997), a clonal pattern, in which the phylogenies of

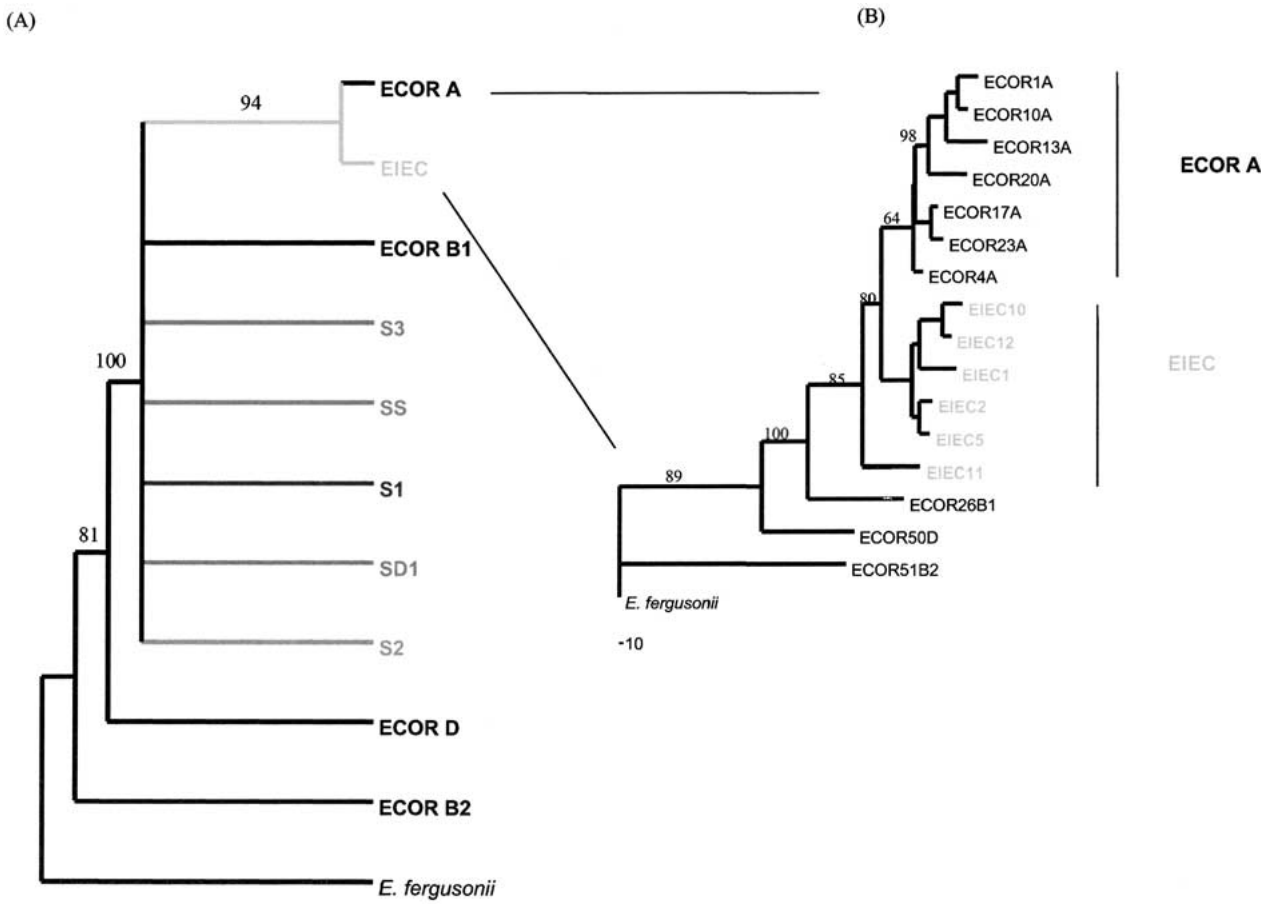

Fig. 2. Phylogenetic trees based on the simultaneous analysis of the sequences of 11 genes (7 metabolic—*trpA*, *trpB*, *pabB*, *putP*, *icd*, *purM*, and *thrB*—and 4 polymerases—*polB*, *dinB*, *dnaE*, and *umuC*). **A** The 50% bootstrap majority-rule consensus tree of 10 strains representing different *Shigella*, EIEC/ECOR A, and other ECOR phylogenetic groups rooted on *E. fergusonii*. Only bootstrap values higher than 50% are shown. **B** The semistrict consensus tree showing the phylogenetic relationship among strains of the EIEC/ECOR A clade. Only bootstraps above 50% are indicated.

the chromosome and the plasmid are similar, indicates that there is no or limited transfer of the plasmid between different phylogenetic groups. Such a pattern would be consistent with the hypothesis that the VP was acquired only once and that all groups descend from the same ancestor. In contrast, a panmictic pattern, in which groups defined by the analysis of chromosomal and plasmid markers are unrelated, indicates that different plasmids are dispersed among different chromosomal backgrounds, and an epidemic pattern, in which the same plasmid is present in strains belonging to different chromosomal groups, indicates that a particular plasmid can disperse in a wide range of populations. These patterns would be consistent with the hypotheses that the *Shigella* and EIEC groups derive from different *E. coli* groups that acquired the VP independently or that the VP can disperse among these groups.

To understand the evolution of the association between the VP and the chromosome, we performed a phylogenetic analysis using sequences of genes carried by either the chromosome or the VP of a large set of *Shigella* and EIEC strains. The results show that the phylogenic trees derived from the chromosomal and plasmidic genes are very similar, indicating similar evolutionary histories of the VP and the chromosome. The results also suggest that the single arrival of the VP in an ancestral *E. coli* strain gave rise to a radiation from which all *Shigella* groups and EIEC descend.

Materials and Methods

Strains

The collection of 51 strains, which were isolated in Europe, Asia, and Africa, consisted of (i) 15 strains of *S. boydii*, including 3 strains of serotypes 1, 2, and 4, 2 strains of serotype 10 and 18, and 1 strain of serotype 11 and 12; (ii) 11 strains of *S. dysenteriae*, including 1 strain of serotypes 3, 4, 5, and 10, 4 strains of serotype 1, and 3 strains of serotype 2; (iii) 11 strains of *S. sonnei;* (iv) 8 strains of *S. flexneri*, including 1 strain of serotypes 2, 2a, and 4 and 5 strains of serotype 6; and 6 EIEC strains. All the strains had the VP except strains SB10-80, SD03-97, and SD02-97, which had lost the VP plasmid after isolation. However, chromosomal sequences

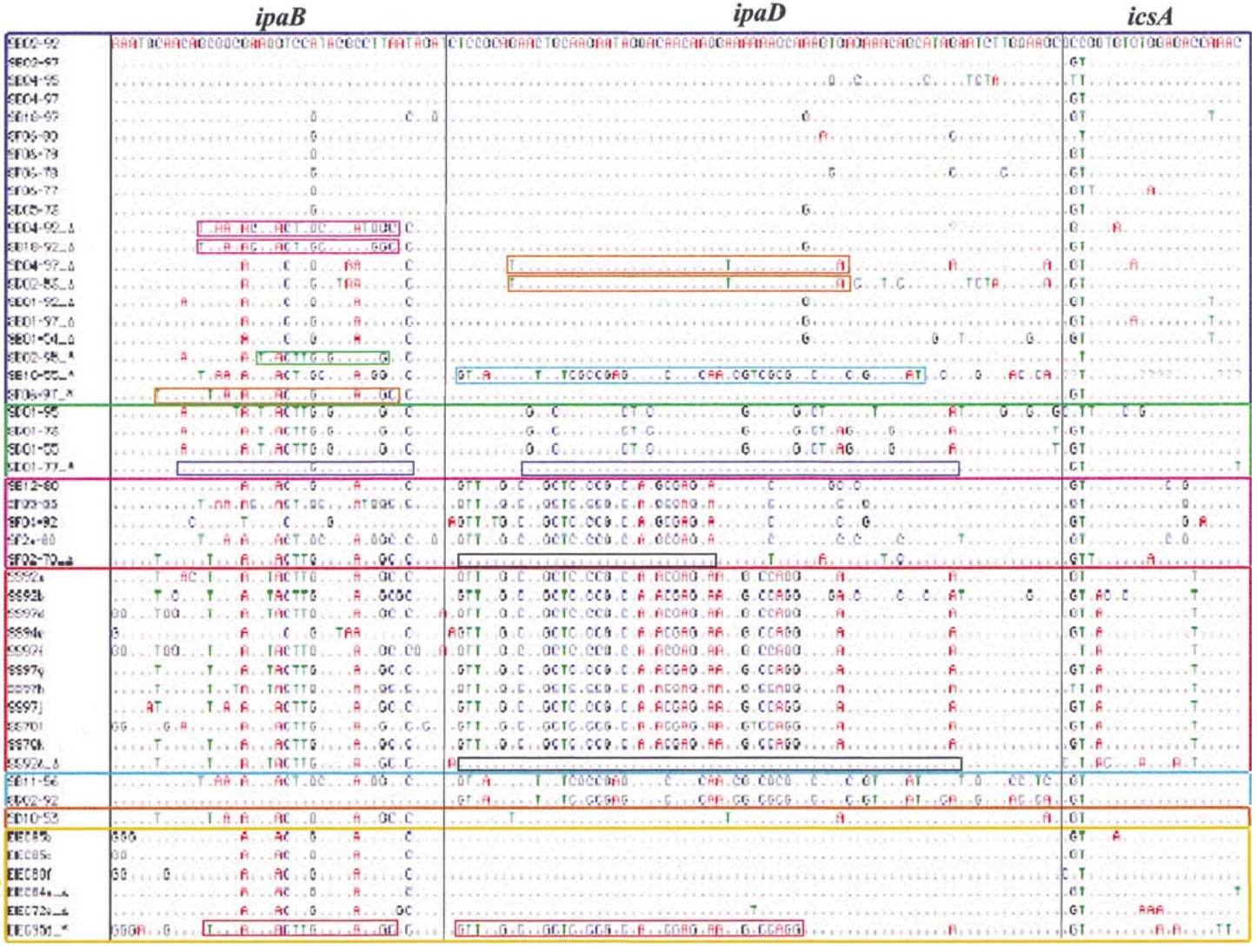

Fig. 3. Alignment of the phylogenetically informative sites in the *ipaB*, *ipaD*, and *icsA* genes. Strains are grouped by colors according to the groups defined by the phylogeny of the chromosome. Segments of genes that appear to have been exchanged between different plasmids are *boxed* and the color of the box refers to the putative donor group. A *black box* indicates that the donor could not be identified unambiguously. Symbols * and Δ indicate strains that correspond to incongruent taxa in Fig. 1: (*) strains that were not placed in the "right" plasmidic group; (Δ) strains that subtended new branches in the tree based on plasmidic genes.

from these strains were obtained, as they have underrepresented serotypes in the collection. In addition, several strains from each group of the *E. coli* reference (ECOR) collection (four strains from group A, three strains from group B1, four strains from group B2, and three strains from group D) (Herzer et al. 1990) were analyzed to assess the phylogenetic position of *Shigella* and EIEC strains within the *E. coli* species. A strain of *E. fergusonii* (ATCC35469^T), which is the closest species to *E. coli* (Lawrence et al. 1991), was used as the outgroup in the analysis.

Gene Amplification and Sequencing

Sequences of two genes of the tryptophan operon, *trpA* and *trpB*, encoding for tryptophan synthase subunits A and B, respectively, the *p*-aminobenzoate synthase gene (*pabB*) and the proline permease gene (*putP*) were obtained for phylogenetic reconstruction of the 66 strains described above. These genes have been shown to exhibit low levels of horizontal gene transfer in *E. coli* (Denamur et al. 2000; Lecointre et al. 1998), allowing us to perform simultaneous gene analysis (Escobar-Páramo et al. 2003).

In an attempt to increase the resolution of the internal branches of the chromosomal gene tree, we selected one strain representing each of the different *Shigella* phylogenetic groups (as defined by the phylogenetic analysis of the four genes) and of the ECOR groups B1, D, and B2 as well as six EIEC and seven ECOR A strains. Sequences of 11 genes including 7 housekeeping genes (homoserine kinase, *thrB*; phosphoribosylaminoimidazol synthase, *purM*; and isocitrate dehydrogenase, *icd*; and the 4 genes described above) and 4 polymerases (*polB*, *umuC*, *dinB*, and *dnaE*) were obtained from these strains. These genes have been shown to be very useful for phylogenetic reconstruction of the *E. coli* species because they are polymorphic and exhibit low levels of recombination (Escobar-Páramo et al. 2003). Primers used for PCR amplification of these genes are described elsewhere (metabolic genes [Escobar-Páramo et al. 2003], polymerases [Bjedov et al. 2003]).

We also sequenced three genes, *ipaB*, *ipaD*, and *icsA*, which are carried by two different regions of the VP, are involved in different steps of pathogenesis, and exhibit different GC contents. The *ipaB* and *ipaD* genes are encoded by the same operon, are necessary for entry of bacteria into epithelial cells, and have a GC content of 34% (Menard et al. 1993; Sasakawa et al. 1989). The *icsA* gene is not linked to the *ipaB* and *ipaD* genes, encodes a protein essential for the intracellular actin-based movement, and has a GC content of 41% (Lett et al. 1989; Makino et al. 1986). The sequences of the *ipaB*, *ipaD*, and *icsA* genes were used to determine the phylogenetic relationship of the VP present in 48 strains of the collection described above. Plasmid sequences for strain SF05-83 (M90T) were taken from GenBank (Buchrieser et al. 2000) and added to the data set.

Primers used for PCR amplification and sequencing were as follows:

pabB-F, 5′TTTTACACTCCGGCTATGCCGATCA;
pabB-R, 5′GCTGCGGTTCCAGTTCGTCGATAAT (Guttman and Dykhuizen 1994);
putP-F, 5′GCGACGATCCTTTACACCTTTATTG;
putP-R, 5′CGCATCGGCCTCGGCAAAGCG;
trpA-F, 5′ATGGAACGCTACGAATCTCTGTTTGCCC;
trpA-R, 5′TCGCCGCTTTCATCGGTTGTACAAA;
trpB-F, 5′ACAATGACAACATTACTTAACCCCT;
trpB-R, 5′TTTCCCCTCGTGCTTTCAAAATATC;
icsA-F, 5′TTATTAACACCTGTTGATGGACTTT;
iscA-R, 5′GCCAAATGACATATTTCCAGAAGC;
ipaB-F, 5′GGACGCAATTCAGGATATCAAGGAG;
ipaB-R, 5′GCCCCAAGTATTTTCCCAACACCAC;
ipaD-F, 5′CCCTTACTATGCTCAACGACACCCT; and
ipaD-R, 5′TTTTTCATTGTTTCATCTTCGGCAG.

PCR amplifications were performed in a 50-µl reaction mixture containing template DNA (5 µl of boiled lysate), 1.5 m*M* $MgCl_2$, a 250 µ*M* concentration of each of four dNTPs, a 0.2 µ*M* concentration of each primer, and 1 U of *Taq* polymerase in 1× PCR buffer. Reactions were heated at 94°C for 4 min, followed by 30 cycles consisting of 30 s of denaturation (94°C), 30 s of annealing (56°C for *putP* primers, 55°C for *ipaB* and *ipaD* primers, and 50°C for the other primers), and 30 s of extension (68°C). Sequences were obtained by direct sequencing of PCR products and have been deposited in the GenBank database with the following accession numbers: *trpA*, AF330263–AF330320; *trpB*, AF330378–AF330435; *pabB*, AF330321–AF330377; *putP*, AF330543–AF330594; *ipaB*, AF330490–AF330542, *ipaD*, AF330436–AF330489; and *icsA*, AF336749–AF336792.

Phylogenetic Analysis

Sequences were aligned using the CLUSTAL program (Higgins et al. 1992) from the Sequence Navigator package. Bootstrap trees were constructed using maximum parsimony as the optimality criterion, with the heuristic search of PAUP*4.0 (Swofford 2002) with 1000 iterations. The starting tree for the bootstrap analyses was constructed via stepwise addition with the TBR branch-swapping algorithm. Neighbor-joining analyses were also applied to the combined data set of the chromosomal and plasmid sequences of the entire collection as well as to the combined data of the 11 chromosomal genes of 11 strains using the BioNJ method of PAUP*4.0. Maximum likelihood analysis was also performed on the latter data set. Distance and likelihood parameters were estimated with the PAUP*4.0 program.

For phylogenetic analysis of the plasmid as a whole, we used the results of a study performed by Sansonetti et al. (1983) on the restriction fragment length polymorphism (RFLP) pattern of the VP digested by *Eco*RI and *Bam*HI. For the sake of clarity, the nomenclature of the strains used by Sansonetti et al. (1983) has been changed as follows: 138-80, SF06-80; 31-81, SB02-81a; 33-81, SB02-81b; 33-81, SB02-81c, 2924-71, SF06-71; 01-82, SD02-82a; 01-82, SD02-82b; 13-80, SD02-80; 36-81, SD02-81; M25-8, SF01Z; 177-81, SF01-81a; 6154-61, SF05-61; M90T, SF05-83, 291-81, SF04-81a; M76-39, SF04-39; 351-81, SF04-81b; 43-81, SF02-81a; 08-80, SF03-80a; 109-80, SF03-80b; J17-B, SF03-BB; 02-81, SF02-81b; 07-81, SF02-81c; 330-81, SF01-81b; 31-81, SD01-81; 47-80, SD01-80; 482-79, SS79; 01-82, SS82a; 07-82, SS82b; and 11-82, SS82c. Phylogenetic analysis was performed using only the neighbor-joining method of PAUP*4.0 because the number of informative sites was too low to be analyzed using the parsimony method.

Results and Discussion

The Phylogeny of the Chromosome

The general topology of the 50% bootstrap majority-rule consensus tree based on simultaneous analysis of the sequence data of the four metabolic genes (*trpA*, *trpB*, *pabB*, and *putP*) (Fig. 1A) was robust and included 65 ingroup taxa, with the analysis of 3780 nucleotides and 350 informative sites (including the outgroup). The most parsimonious tree (MPT) had a length of 1395, a consistency index (CI) of 0.631, and a retention index (RI) of 0.755. The six *Shigella* groups defined by the simultaneous analysis were characterized by the serotypes rather than the nomen-species of strains belonging to each group. For example, *S. flexneri* strains of different serotypes

belonged to two groups (S1 and S3) and *S. dysenteriae* strains of different serotypes were scattered among four groups (S1, S2, SD1, and SD10), whereas all strains of a given serotype belonged to the same group. An exception was found to this pattern, strain SD02-53 belonging to group S1 instead of to S2. These results confirmed the classification of Pupo et al. (2000), which was based on the analysis of one strain of each serotype. In addition, the present analysis indicated that EIEC strains cluster together with group A of the ECOR collection. For *Shigella* strains, the observation that a given group contains almost all the strains of the same serotype suggests a vertical mode of transmission of serotypes. Thus, serotype variation among *Shigella* groups appears to have been concomitant with group diversification. In addition, after group diversification, four serotype-converting phages may have been responsible for the different serotypes observed among strains of the group S1 (Allison and Verma 2000). The neighbor-joining bootstrap tree (distance measure, general time reversal; *p*-var, 0.64; α, 0.44) obtained with the same data set gave a similar topology, differing only in the position of group ECOR B1, which appears to be derived from strain SD10-53 (data not shown).

The validity of the *Shigella*, EIEC/ECOR A, and ECOR B1 groups was supported by high bootstrap values. In contrast, the phylogenetic relationship among each of these groups was not resolved. A similar pattern was obtained with the independent analysis of each of the four metabolic genes (*trpA*, *trpB*, *pabB*, and *putP*) (data not shown). It is worth noting that the level of recombination in these genes is of the same order as that of commensal strains of the ECOR collection (Escobar-Páramo et al. 2003) despite their pathogenic nature. The low resolution of the tree obtained by simultaneous analysis of the four genes was not due to insufficient data because the chromosomal genes used in this analysis showed a high level of variation among *E. coli* groups (Lecointre et al. 1998). Indeed, combining the sequences of 11 genes (*trpA*, *trpB*, *pabB*, *putP*, *icd*, *thrB*, *purM*, *polB*, *dnaE*, *dinB*, and *umuC*) of the 10 strains representing each of the different phylogenetic groups (except SD10) obtained from the analysis of 4 genes also led to a phylogenic tree showing the same polytomy among the *Shigella*, EIEC/ECORA, and ECORB1 groups (10 ingroup taxa, 11,034 nucleotides, 403 informative sites; length of MPT, 1758; CI, 0.676; and RI, 0.276 [Fig. 2A]). The maximum likelihood bootstrap tree as well as the neighbor-joining bootstrap tree for this data set gave a similar topology (ti/ty, 2.53; α, 40.44; *p*-inv, 0.64) (data not shown). Such a chromosomal phylogenic tree, which exhibits highly resolved groups and low resolutions of the branches subtending the different groups, is typical of a "star" phylogeny (Hoelzer and Melnick 1994) and suggests that a phenomenon of fast diversification (radiation) has been responsible for the concomitant appearance of the *Shigella*, EIEC/ECOR A, and ECOR B1 groups.

The Phylogeny of the Virulence Plasmid

The sequences of the *ipaB*, *ipaD*, and *icsA* genes were used to determine the phylogenetic relationship of the VP present in 49 strains of the collection described above. The 50% bootstrap majority-rule consensus tree derived from this analysis (49 taxa, 2553 nucleotides, 121 informative sites; length of MPT, 539; CI, 605; and RI, 0.749 [Fig. 1B]) was largely consistent with that obtained for chromosomal genes, except for minor incongruences, i.e., disagreements (see below). The neighbor-joining bootstrap tree obtained with the same data set (distance HKY85; *p*-var, 0.88; α, 0.4) had a similar topology (data not shown). For phylogenetic analyses, we also used the RFLP data of the VP harbored by an additional set of 29 strains that had been analyzed by Sansonetti et al. (1983). The neighbor-joining tree derived from the *Bam*HI and *Eco*RI restriction patterns of these plasmids (Fig. 1C) defined five groups that corresponded to five of the six groups defined by the analysis of chromosomal genes, the sixth group being constituted by *S. dysenteriae* strains of serotype 10 for which no RFLP data were available. These results indicated that the phylogenies of both the chromosome and the plasmid defined similar groups, thereby suggesting that both genomes share similar evolutionary histories and that the VP had not been exchanged between groups.

Horizontal Transfer of Fragments of the VP

The two types of incongruences that were observed in the grouping pattern based on the plasmid phylogeny compared to that based on the chromosome phylogeny corresponded to either (i) strains that belonged to a plasmidic group that was different from the one that contained most other strains of the same chromosomal group (as shown by mixed colors in Fig. 1B) or (ii) strains that subtended branches in the plasmid-based tree but not in the chromosome-based tree. A detailed analysis of the sequences of plasmid genes (Fig. 3) suggested that these incongruences were due to exchanges of portions of different plasmids present in different groups. Transferred segments are defined by the presence of at least three mutations, contiguous or not, delineating a stretch of DNA with a sequence that is different from those of the strains belonging to the same phylogenetic group but is identical to those of another phylogenetic group (Denamur et al. 2000). Strains SB02-95 and SF06-97 (group S1) have *ipaD* and *icsA* genes that are

characteristic of strains of group S1, whereas their *ipaB* genes are characteristic of strains of groups SD1 and SD10, respectively. This suggests that the *ipaB* genes present in strains SB02-95 and SF06-97 are the result of a horizontal transfer of the *ipaB* gene from strains of groups SD1 and SD10, respectively. Likewise, strains SB04-92 and SB18-92 (group S1) have *ipaD* and *icsA* genes that are characteristic of group S1, whereas a portion of their *ipaB* gene is identical to that of strain SF05-83 from group S3. Strains SD04-97 and SD02-53 (group S1) have an internal fragment of their *ipaD* gene that is characteristic of that from strain SD10-53 (group SD10), while strain SB10-55 from the same group has *ipaD* fragment similar to that of strains of group S2. Strain SS92c (group SS) and strain SF02-70 (group S3) have *ipaB* and *icsA* sequences that are characteristic of strains of groups SS and S3, respectively, whereas their *ipaD* gene is related to that of strains of groups S1 and EIEC. In strain EIEC95d (group EIEC), only a fragment of the *ipaB* and *ipaD* genes is related to the genes present in strains of group SS. The fact that strains SB01-92, SB01-97, and SB01-54 (group S1) and strains EIEC84a and EIEC72e (group EIEC) subtend new branches in the plasmid phylogeny is due to only a few point mutations that are not indicative of different lineages. In the case of strain SD01-77 (group SD1), the *ipaB* and *ipaD* genes correspond to genes from strains of groups S1, and the sequence of its *icsA* gene did not allow us to identify unambiguously the origin of the latter genes. This ambiguous case does not contradict the major conclusion that both genomes share similar evolutionary histories. Accordingly, most incongruences between the chromosomal and the plasmidic phylogenies appear to be due to the exchange of fragments between different VPs and not to the transfer of the entire VP between strains of different groups.

The level of transfer of genes carried by the VP does not appear to be different compared to that of most genes carried by the chromosome of *E. coli* (Fig. 3) (Denamur et al. 2000). Recently, Lan et al. (2001) analyzed the sequences of the *ipgD*, *mxiC*, and *mxiA* genes, which are carried by the region of the VP involved in entry of bacteria into epithelial cells, in 10 strains of group S1, 3 strains of group S2, 6 strains of group S3, 1 strain of each of groups SD1 and SD10, and 2 EIEC strains. They distinguished only two forms for the VP, which were designated pINV A and pINV B. pINV A was present in strains of groups S1, SD10, EIEC, and two of three strains of group S2; pINV B was present in strains of groups S3 and SS. The plasmid from the strain of group SD1 had an *ipgD* gene related to that of pINV A and *mxiA* and *mxiC* genes related to those of pINV B. Although the authors concluded that the clustering of *Shigella* strains based on plasmid genes was largely consistent with that based on chromosomal genes, they interpreted the grouping of one strain of group SD10 and one strain of group SS within groups S1 and S3, respectively, as support for the hypothesis that *Shigella* strains of different groups had acquired their VP by lateral transfer from other VP-carrying *Shigella* strains. Compared to that of Lan et al. (2001), the present study analyzed a larger number of strains, 49 vs. 23, and a larger number of phylogenetically informative sites, 121 vs. 79, which allowed us to identify a larger number of groups (7 vs. 2) among the VP present in *Shigella* and EIEC strains.

Since similar groups were defined by the analysis of the chromosomal and plasmid genes, the pattern of the association of the VP and the chromosome corresponds to neither the endemic nor the panmictic plasmid patterns (as proposed by Pupo et al. 2000), in which the same or different plasmids, respectively, are dispersed among different chromosomal backgrounds (Souza and Eguiarte 1997). Instead, the similar evolutionary histories of the chromosome and the VP correspond to a clonal pattern indicative of very limited, if any, transfer of the VP between different groups, which is consistent with its nonconjugative nature (Sansonetti et al. 1982; Buchrieser et al. 2000). This suggests that the VP was transferred into, or assembled within, an ancestral *E. coli* strain that gave rise to all the *Shigella* and EIEC groups. The relationship among groups S3, SS, and S2 is resolved in the phylogenic tree based on plasmid sequences, but not in that based on chromosomal sequences. In fact, only a portion of the *ipaD* gene is common to these three groups, while the rest of the sequence is either unique to each group or shared with other groups that do not belong to the S2–S3–SS clade. This suggests that two events of horizontal transfer of individual genes, followed by a process of selective sweep, might have occurred soon after group diversification.

The EIEC/ECORA Clade

Simultaneous analysis of the 11 genes described above were performed on six EIEC strains and six ECORA strains and one strain representing each of the B1, D, and B2 groups of the ECOR collection. The topology of the phylogenetic tree obtained with this analysis is well supported by high bootstrap values on the subtending branches as well as by high CI and RI scores (0.728 and 0.559, respectively) (16 ingroup taxa, 11,034 nucleotides, 387 informative sites; length of MPT, 1646 [Fig. 2B]). Since strains of *E. coli* group A are in a derived position within the clade EIEC, it seems that the loss of the VP in an ancestral EIEC strain gave rise to a group from which all *E. coli* strains of group A descend. The loss of the VP in this strain might have been followed by reactivation or reacquisition of some genes, such as the

lactose operon genes. This is suggested by the fact that the *lacY* genes of EIEC and ECOR A strains differ in only two positions, one of which is responsible for inactivation of the gene in EIEC, whereas the *lacY* gene of ECOR B1 strains differs from them in six synonymous sites (data not shown).

The Evolutionary Scenario

In conclusion, we propose the following scenario for the origin and diversification of *Shigella* and EIEC groups. The single arrival of the VP into an *E. coli* strain gave rise to a monophyletic group from which all *Shigella* groups, EIEC, and *E. coli* groups A and B1 descend. Persistence of the plasmid, due to the advantage it conferred to its new host, required adaptation of the chromosomal genome, which led to highly diversified but highly specialized pathogenic groups. Adaptation is also reflected by the convergent evolution toward the inactivation of several metabolic pathways of the different *Shigella* and EIEC groups including the ability to use lactose as carbon source (Ito et al. 1991; Escobar-Páramo and Denamur, unpublished data) and to decarboxylate lysine (Maurelli et al. 1998; McCormick et al. 1999). The absence of this process in ancestors of *E. coli* group B1 led to the loss of the VP. After the radiation, group A originated from within the EIEC clade, by losing the VP and reacquiring the metabolic functions. Accordingly, *E. coli* strains of groups A and B1, which correspond mainly to commensal strains (Picard et al. 1999), appear to be derivatives of an ancestor that had the VP and lost it. The possibility that commensal *E. coli* strains will become pathogenic *Shigella* strains by acquiring the VP seems unlikely, as the VP does not seem to be transmitted horizontally among *Shigella* and EIEC strains.

Acknowledgments. The authors thank Bertrand Picard and Philippe Bouvet for provision of the strains used in this study, Hélène d'Hauteville for access to precious RFLP data, Guillaume Lecointre, Pierre Darlu, Guillermo Orti, Ehab Abouheif, Dan Dykhuizen, and Olivier Tenaillon for helpful discussions, and Jacques Elion for constant encouragement. This work was supported in part by the Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires and grants from the INSERM and the Fondation pour la Recherche Médicale.

References

Allison GE, Verma NK (2000) Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*. Trends Microbiol 8:17–23

Bjedov I, Lecointre G, Tenaillon O, Radman M, Taddei F, Denamur E, Matic I (2003) Polymorphism of gene encoding SOS polymerases in natural populations of *Escherichia coli*. DNA Repair 2:417–426

Buchrieser C, Glaser P, Rusniok C, Nedjari H, d'Hauteville H, Kunst F, Sansonetti P, Parsot C (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

Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I, Rothstein R, Elion J, Taddei F, Radman M, Matic I (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell 103:711–721

Escobar-Páramo P, Sabagh A, Darlu P, Pradillon O, Vaury C, Denamur E, Lecointre G (2003) Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. Mol Phylogenet Evol (in press)

Goullet P, Picard B (1987) Differentiation of *Shigella* by electrophoretic polymorphism. J Gen Microbiol 133:1005–1017

Guttman DS, Dykhuizen DE (1994) Detecting selective sweeps in naturally occurring *Escherichia coli*. Genetics 138:993–1003

Herzer PJ, Inouye S, Inouye M, Whittman TS (1990) Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J Bacteriol 172:6175–6181

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTALV: Improved software for multiple sequence alignment. CABIOS 8:189–191

Hoelzer GA, Melnick DJ (1994) Patterns of speciation and limits to phylogenetic resolution. TREE 9:104–107

Ito H, Kido N, Arakawa Y, Ohta M, Sugiyama T, Kato N (1991) Possible mechanisms underlying the slow lactose fermentation phenotype in *Shigella* spp. Appl Environ Microbiol 57:2912–2917

Lan R, Lumb B, Ryan D, Reeves PR (2001) Molecular evolution of large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*. Infect Immun 69:6303–6309

Lawrence JG, Ochman H, Hartl DL (1991) Molecular and evolutionary relationships among enteric bacteria. J Gen Microbiol 137:1911–1921

Lecointre G, Rachdi L, Darlu P, Denamur E (1998) *Escherichia coli* molecular phylogeny using the incongruence length difference test. Mol Biol Evol 15:1685–1695

Lett MC, Sasakawa C, Okada N, Sakai T, Makino S, Yamada M, Komatsu K, Yoshikawa M (1989) VirG, a plasmid-coded virulence gene of *Shigella flexneri*: Identification of the VirG protein and determination of the complete coding sequence. J Bacteriol 171:353–359

Makino S, Sasakawa C, Kamata K, Kurata T, Yoshikawa M (1986) A genetic determinant required for continuous reinfection of adjacent cells on a large plasmid in *S. flexneri* 2a. Cell 46:551–555

Maurelli AT, Fernandez RE, Bloch CA, Rode CK (1998) "Black holes" and bacterial pathogenecity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc Natl Acad Sci USA 95:3943–3948

McCormick BA, Fernandez MI, Siber AM, Maurelli AT (1999) Inhibition of *Shigella flexneri*-induced transepithelial migration of polynuclear leucocytes by cadaverine. Cell Microbiol 1:143–156

Menard R, Sansonetti PJ, Parsot C (1993) Nonpolar mutagenesis of the *ipa* genes defines IpaB, IpaC, and IpaD as effectors of *Shigella flexneri* entry into epithelial cells. J Bacteriol 175:5899–5906

Ochman H, Whittam TS, Caugant DA, Selander RK (1983) Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. J Gen Microbiol 129:2715–2726

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Parsot C, Sansonetti PJ (1996) Invasion and the pathogenesis of *Shigella* infections. Curr Top Microbiol Immunol 209:25–42

Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E (1999) The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. Infect Immun 67:546–553

Pupo GM, Karaolis DKR, Lan R, Reeves PR (1997) Evolutionary relationship among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* studies. Infect Immun 65:2985–2692

Pupo GM, Lan R, Reeves PR (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characters. Proc Natl Acad Sci USA 97:10567–10572

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406:64–67

Rolland K, Lambert-Zeschovsky N, Picard B, Denamur E (1998) *Shigella* and enteroinvasive *Escherichia coli* strains are derived from distinct ancestral strains of *E. coli*. Microbiology 144:2667–2672

Sansonetti PJ, Kopecko DJ, Formal SB (1982) Involvement of a plasmid in the invasive ability of *Shigella flexneri*. Infect Immun 35:852–860

Sansonetti PJ, d'Hauteville H, Ecobichon C, Pourcel C (1983) Molecular comparison of virulence plasmids in *Shigella* and enteroinvasive *Escherichia coli*. Ann Microbiol (Inst Pasteur) 13A:295–318

Sasakawa C, Adler B, Tobe T, Okada N, Nagai S, Komatsu K, Yoshikawa M (1989) Functional organization and nucleotide sequence of virulence Region-2 on the large virulence plasmid in *Shigella flexneri* 2a. Mol Microbiol 3:1191–1201

Sasakawa C, Buysse JM, Watanabe H (1992) The large virulence plasmid of *Shigella*. Curr Top Microbiol Immunol 180:21–44

Souza V, Eguiarte LE (1997) Bacteria gone native vs. bacteria gone awry?: Plasmidic transfer and bacterial evolution. Proc Natl Acad Sci USA 94:5501–5503

Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Suderland, MA