A Specific Genetic Background Is Required for Acquisition and Expression of Virulence Factors in *Escherichia coli*

Patricia Escobar-Páramo, Olivier Clermont,* Anne-Béatrice Blanc-Potard,† Hung Bui,‡ Chantal Le Bouguénec,§ and Erick Denamur**

*INSERM E0339, Faculté de Médecine Xavier Bichat, Paris, France; †UMR CNRS-IRD 9926, Centre IRD, Montpellier, France; ‡Centre d'Etude du Polymorphisme Humain, Hôpital Saint Louis, Paris, France; and §Unité de Pathogénie Bactérienne des Muqueuses, Institut Pasteur, Paris, France

In bacteria, the evolution of pathogenicity seems to be the result of the constant arrival of virulence factors (VFs) into the bacterial genome. However, the integration, retention, and/or expression of these factors may be the result of the interaction between the new arriving genes and the bacterial genomic background. To test this hypothesis, a phylogenetic analysis was done on a collection of 98 *Escherichia coli/Shigella* strains representing the pathogenic and commensal diversity of the species. The distribution of 17 VFs associated to the different *E. coli* pathovars was superimposed on the phylogenetic tree. Three major types of VFs can be recognized: (1) VFs that arrive and are expressed in different genetic backgrounds (such as VFs associated with the pathovars of mild chronic diarrhea: enteroaggregative, enteropathogenic, and diffusely-adhering *E. coli*), (2) VFs that arrive in different genetic backgrounds but are preferentially found, associated with a specific pathology, in only one particular background (such as VFs associated with extraintestinal diseases), and (3) VFs that require a particular genetic background for the arrival and expression of their virulence potential (such as VFs associated with pathovars typical of severe acute diarrhea: enterohemorragic, enterotoxigenic, and enteroinvasive *E. coli* strains). The possibility of a single arrival of VFs by chance, followed by a vertical transmission, was ruled out by comparing the evolutionary histories of some of these VFs to the strain phylogeny. These evidences suggest that important changes in the genome of *E. coli* have occurred during the diversification of the species, allowing the virulence factors associated with severe acute diarrhea to arrive in the population. Thus, the *E. coli* genome seems to be formed by an "ancestral" and a "derived" background, each one responsible for the acquisition and expression of different virulence factors.

Introduction

Current models of bacterial evolution propose that pathogenic diversity is the result of the acquisition of pathogenic genes most likely through successive horizontal gene transfers (Ochman, Lawrence, and Groisman 2000). However, whether the acquisition of these pathogenic genes or virulence factors (VFs) is sufficient to give virulence or whether a specific genetic background is important for their integration, retention, and expression is a critical issue in understanding how virulence is spread in bacterial populations. *Escherichia coli* is a particularly suitable organism for addressing this question, as it is a normal inhabitant of the gut flora of vertebrates, including human, and the within-species genetic variability leads to the differential colonization of hosts (Gordon and Cowling 2003). It is also frequently isolated in a broad spectrum of intestinal and extraintestinal diseases (Donnenberg 2002). In *E. coli*, the existing diversity of pathogenic clones seems to be the result of the constant arrival, sometimes in parallel (Reid et al. 2000), of different VFs into the population (Ochman, Lawrence, and Groisman 2000). These VFs are generally carried on plasmids, pathogenicity islands (PAIs), or phages and are supposed to be highly interchangeable among bacterial strains through horizontal transfer (Hacker and Kaper 2000). Few attempts have been made to elucidate the evolutionary history of the ensemble of the diversity of pathogenic and nonpathogenic strains of this species. Most of the studies have focused on strains of the *E. coli* reference (ECOR) collection (Herzer et al. 1990; Escobar-Páramo et al. 2004) or concentrated on specific pathovars without taking into account the genetic structure of the species as a whole (Whittam et al. 1993; Brando et al. 1998; Czeczulin et al. 1999; Pupo, Lan, and Reeves 2000; Reid et al. 2000; Escobar-Páramo et al. 2003). The phylogenetic trees of Pupo et al. (1997), based on the neighbor-joining analysis of the data of 10 metabolic enzymes and the sequence of the *mdh* gene on a collection of strains combining representatives of the diversity of diarrheagenic, uropathogenic, and commensal strains have been, to the present, the only reference for the evolution of the pathogenic and nonpathogenic *E. coli* strains. However, it is difficult to make major conclusions from these data, as the phylogenetic trees derived from the analysis of both the 10 enzymes and the *mdh* sequence differ substantially among themselves, as does the topology of these two trees and the new trees obtained in recent studies on more robust data sets (Pupo, Lan, and Reeves 2000; Escobar-Páramo et al. 2003; 2004). In addition, enteroaggregative *E. coli* (EAEC) and diffusely adhering *E. coli* (DAEC) pathovars are not represented in their study.

In an effort to investigate the relationship between the genetic background and the virulence genes, we established the phylogenetic relationship of 98 *E. coli/Shigella* strains representing the pathogenic diversity of the species and determined the presence of different pathogenic determinants. In addition, the evolutionary histories of some of these pathogenic determinants were compared with the strain phylogeny to assess the single or multiple arrivals of such determinants.

Key words: *Escherichia coli*, bacterial evolution, virulence, phylogeny.

E-mail: denamur@bichat.inserm.fr.

Mol. Biol. Evol. 21(6):1085–1094. 2004
DOI:10.1093/molbev/msh118
Advance Access publication March 10, 2004

Materials and Methods

Strains

Besides the commensal, several pathovars of diarrheagenic *E. coli* have been differentiated on the basis of pathogenic features. Thus, enterohemorragic *E. coli* (EHEC), a subcategory of Shiga toxin–producing *E. coli* (STEC), enterotoxigenic *E. coli* (ETEC), and enteroinvasive *E. coli* (EIEC) are obligatory pathogens responsible for severe and acute diarrhea, because of the production of toxins and/or the invasion of the intestinal epithelium. *Shigella*, the bacillary agent of dysentery, constitutes an additional category of diarrhea-associated strains that must be considered as EIEC. Enteropathogenic *E. coli* (EPEC), EAEC, and DAEC are associated with chronic and mild diarrhea and are characterized by the adherence pattern on epithelial cells (for review, see Nataro and Kaper [1998]). Extraintestinal infections, including urinary tract infections (UTI), newborn meningitis (NBM), pneumonia, and bacteriemia are caused mainly by extraintestinal pathogenic *E. coli* (ExPEC) (Russo and Johnson 2000).

A total of 98 *E. coli*/*Shigella* strains selected from previously reported collections of prototypes of the diversity of the different pathovars and commensal strains were considered in this study. Strains isolated in pathogenic conditions include 10 EAEC, 16 DAEC, 11 STEC (with nine EHEC), six EPEC, eight ETEC, nine ExPEC, five *Shigella*/EIEC, and two nonclassified diarrheagenic *E. coli*. Thirty-one commensal strains were also analyzed. The 30 strains of the ECOR collection (Ochman and Selander 1984) included in this study have been previously used in a phylogenetic study based on the sequence analysis of 11 genes (Escobar-Páramo et al., 2004). The 15 DEC strains are representatives of different intestinal pathogenic groups (ETEC, EPEC 1 and 2, EAEC, and EHEC 1 and 2) previously described by Whittam et al. (1993), Reid, Betting and Whittam (1999), and Reid et al. (2000). Strains EIEC85b, SB01-97, SS92a, SB11-56, and SD01-77 have been selected from a study on the evolution of *Shigella*/EIEC (Escobar-Páramo et al. 2003), and each one represents a monophyletic group showing the diversity of *Shigella*. Strains C1845 (isolated from a patient with diarrhea), IH11128 and EC7372 (isolated from urine of patients with UTI) are archetypes of DAEC strains producing Afa/Dr adhesins. The other DAEC strains, isolated from patients with diarrhea from Brazil (seven strains) and France (eight strains), represent the genetic diversity of this pathovar based on the hybridization with 10 probes derived from strain C1845 (Blanc-Potard et al. 2002). In addition, six DAEC strains isolated from asymptomatic patients in Brazil and France (DAECT2, T437, T179, T192, T14, and T19) were included. Strains EAEC 042, JM221, 17-2, and 55989 are reference strains that have been used to elucidate the pathogenicity of enteroaggregative *E. coli*. Additional strains were selected from different collections: EAEC strains 56390, 384P, 381A, 11097, and 11074; ETEC strains E2539-C1, EDL1493, TX-1, 469, 440, and H10407; EHEC/STEC strains EDL931, 255/1-1, 248/1-2, and H-19; and EPEC strain 2348/69. Strains EDL933, RIMD0509952, CFT073, and RS218 correspond to strains from which the complete genome sequence have been determined. EDL933 and RIMD0509952 are EHEC O157:H7 strains, whereas CFT073 and RS218 are ExPEC strains. A strain of *E. fergusonii* (ATCC35469T), which is the closest species to *E. coli* (Lawrence, Ochman and Hartl 1991), was used as the outgroup in the phylogenetic analysis. Specific information on each strain is available on table S1 in Supplementary Material online.

Gene Sequencing

Sequences of six essential chromosomal genes encoding the tryptophan synthase subunits A and B (*trpA* and *trpB*) of the tryptophan operon, the p-aminobenzoate synthase (*pabB*), the proline permease (*putP*), the isocitrate dehydrogenase (*icd*), and the polymerase PolII (*polB*) were obtained for phylogenetic reconstruction of the 99 strains described above. These genes have been shown to exhibit low levels of horizontal gene transfer in *E. coli* (Lecointre et al. 1998; Denamur et al. 2000) and, thus, are useful to assess strain phylogeny. ECOR sequences, as well as the sequences of the *Shigella*/EIEC strains and of *E. fergusonii*, were determined elsewhere (Escobar-Páramo et al. 2003; 2004). Sequences of EDL933, RIMD0509952, CFT073, and RS218 were from the *E. coli* genome sequence projects. Gene sequences of the remaining 58 strains were obtained by direct sequencing of PCR products as in Bjedov et al. (2003).

In addition, sequences of three VFs representing each category defined below (see *Results and Discussion*) were performed. *hlyCA* (505 bp) from *hly* operon, *afaD* (550 bp) from the *afa* operon, coding for an invasin and LT-I (614 bp) sequences were obtained by direct sequencing of PCR products. Sequences of all the used primers are available on table S2 in Supplementary Material online.

Virulence Factors

The presence of 17 VFs, characteristic of the different *E. coli* pathovars was determined in the entire collection of strains by standard hybridization protocols using digoxigenin (DIG) or radioactive ^{32}P-labeled probes. Determinants highly common in ExPEC strains include the outer membrane usher protein gene *papC* of the pyelonephritis-associated pilus system, *sfa*/*focDE* encoding S fimbrial adhesins, *hlyCA* genes encoding synthesis of active intracellular α-hemolysin, and the cytotoxin necrotizing factor–encoding gene *cnf1*. Two genes involved in iron uptake, *chuA* and *iucC*, were also analyzed. Probes *afaBC*/*daaC* and M030, specific to the subclass of DAEC strains encoding Afa/Dr adhesins, were also used. The presence of aggregative adhesion-encoding plasmid in EAEC was detected using the classical CVD432 probe (AA probe). The presence of the pathogenicity island containing the locus of enterocyte effacement (LEE) of EHEC and EPEC was evidenced with a probe specific for the *eae* gene coding for the outer membrane protein intimin. The gene coding for the pilin protein (*bfpA*) in the bundle-forming pilus (BFP) system of EPEC was used to determine the presence of the EPEC adherence factor (EAF) plasmid (pB171). We also looked at the presence of the heat stable toxin (STa) and the heat labile toxin (LT-I) defining typical ETEC pathovars

as well as the Shigalike toxins (*stx1* and *stx2*) of STEC/EHEC. In addition, we determined the presence of the 60-MDa plasmid defining typical EHEC by means of the gene *ehxA* coding for the enterohemolysin (of the RTX family). Finally, the *ipaB* gene, a gene of the type III secretion system of *Shigella* and EIEC encoded by the virulence plasmid, was determined. The sequences of the primers used for generating the PCR probes for hybridations were taken from the literature.

To assign the ExPEC VFs to specific PAIs, additional analyses were performed. The presence of *iroN* and *hra* genes coding for an enterobactin siderophore receptor protein and a heat-resistant agglutinin, respectively, was determined by PCR on strains exhibiting at least one ExPEC VF in the first step screening. *papG* alleles and S fimbrial adhesin types (*sfa* or *foc*) were determined by PCR in *papC* or *sfa*/*foc* positive strains, respectively. Sequences of the primers are available on table S2 in Supplementary Material online. The type of intimin allele was taken from the literature (Reid et al. 1999) or determined *in silico* for the RIMD0509952 strain.

Phylogenetic Analyses

Sequences were aligned using the Clustal program (Higgins, Bleasby, and Fuchs 1992) from the Sequence Navigator™ package. Neighbor-joining analyses were performed using the BioNJ method of PAUP* version 4.0 (Swofford 2002). The semistrict consensus trees, as well as the bootstrap trees, were obtained using maximum parsimony as the optimality criteria, with the heuristic search of PAUP* 4.0 with 1,000 iterations. The starting tree for the analyses was constructed via stepwise addition with the TBR branch-swapping algorithm. Maximum-likelihood and Bayesian analyses were performed using the PHYML (Guindon and Gascuel 2003) and MrBayes version 2.01 (Huelsenbeck and Ronquist 2001) programs, respectively.

Results

The Strain Phylogeny

The general topology of the semistrict consensus tree of the 98 *E. coli*/*Shigella* strains analyzed (with *E. fergusonii* as the outgroup) based on simultaneous analysis of the sequence data of the six essential genes (*trpA*, *trpB*, *pabB*, *putP*, *icd*, and *polB*) is shown in figure 1. Six major groups of *E. coli* (A, B1, C, E, D, and B2), in addition to the different *Shigella* monophyletic groups form the core of the *E. coli* species. Groups A, B1, D, and B2 have been reported previously as the major groups of the species (Herzer et al. 1990; Lecointre et al. 1998; Escobar-Páramo et al. 2004), with B2 being the most ancestral group, followed by group D. Groups C and E (Herzer et al. 1990) represent two additional monophyletic groups. The group C, put on evidence through this analysis, is a sister group of A and B1 groups and of *Shigella* groups, all emerging during the radiation (Escobar-Páramo et al. 2003). Group E emerged after group D but before the radiation. Despite the low bootstrap values for groups A, B1, and D (less than 50%), we are confident in the accuracy of the phylogeny as the general topology of the phylogenetic tree shown in figure 1 is in agreement with that obtained by neighbor-joining, maximum-likelihood, and Bayesian analyses (data not shown) and with previously reported phylogenies of individual collections (Czeczulin et al. 1999; Reid et al. 2000; Escobar-Páramo et al. 2003; in press).

The Distribution of Pathovars and Virulence Factors

The presence of 17 VFs associated to the seven *E. coli* pathovars was determined in the ensemble of strains (fig. 2). Pathovars were assigned to the strains based on the syndrome of isolation and the presence of specific VFs. Strains DEC6a and DEC14a, although isolated from patients with diarrhea do not have any of the VFs described above; therefore, no pathovar has been indicated.

Typical EPEC strains are distributed in two major clusters, which correspond to previously reported EPEC 2 and 1 complexes belonging to groups B1 and B2, respectively (Reid et al. 2000). These two groups are differentiated by the type of *eae* gene of the LEE PAI. The first group possesses the *eae* β type, whereas the second has the *eae* α type. Strain DEC5d is an atypical EPEC belonging to group E and having an *eae* γ type. EHEC strains are distributed among groups A and B1, but the major concentration of these pathovars is found in strains of serotype O157:H7 of group E. EHEC strains of group A belong to the EHEC 2 complex, whereas O157:H7 strains represent the EHEC 1 complex (Reid et al. 2000). ETEC strains are found in groups A, B1, and C. EAEC and DAEC strains are distributed all over the phylogenetic tree, except in group E. *Shigella* and EIEC are highly localized and represent highly specialized monophyletic groups. ExPEC strains are clustered mostly in group B2, but some strains are found in group D.

The distribution pattern of the pathovars does not always coincide with the distribution of the VFs (fig. 2). For example, in contrast to the localized distribution of ExPEC strains in groups B2 and D, the ExPEC VFs, although more frequent in B2 and D, can be found all over the phylogeny, in both commensal and diarrheic strains. This is the same case for Afa/Dr and M030, the VFs typical of DAEC strains, which are found in other pathovars (STEC, EHEC, and ExPEC) and in commensal strains. However, the presence of toxins (ST, LT, *stx1*, and *stx2*) and other VFs (*ehxA* and *ipaB*) of strains causing severe diarrhea, as well as the AA gene of EAEC strains, is exclusive to the particular pathovars that harbor these determinants (except strain H10407, an ETEC carrying the AA gene).

The two iron-uptake genes, *iucC* and *chuA*, are not associated to a particular pathovar. Nevertheless, whereas *iucC* is found in all genetic backgrounds, the distribution of *chuA* is restricted to phylogenetic groups B2, D, E, and SD01, regardless of the pathogenic nature of the strains, which support its value as a phylogenetic marker (Clermont, Bonacorsi, and Bingen 2000).

The Virulence Factor Evolutionary Histories

To test whether the correlation between genetic background and the presence of VFs is due to a single

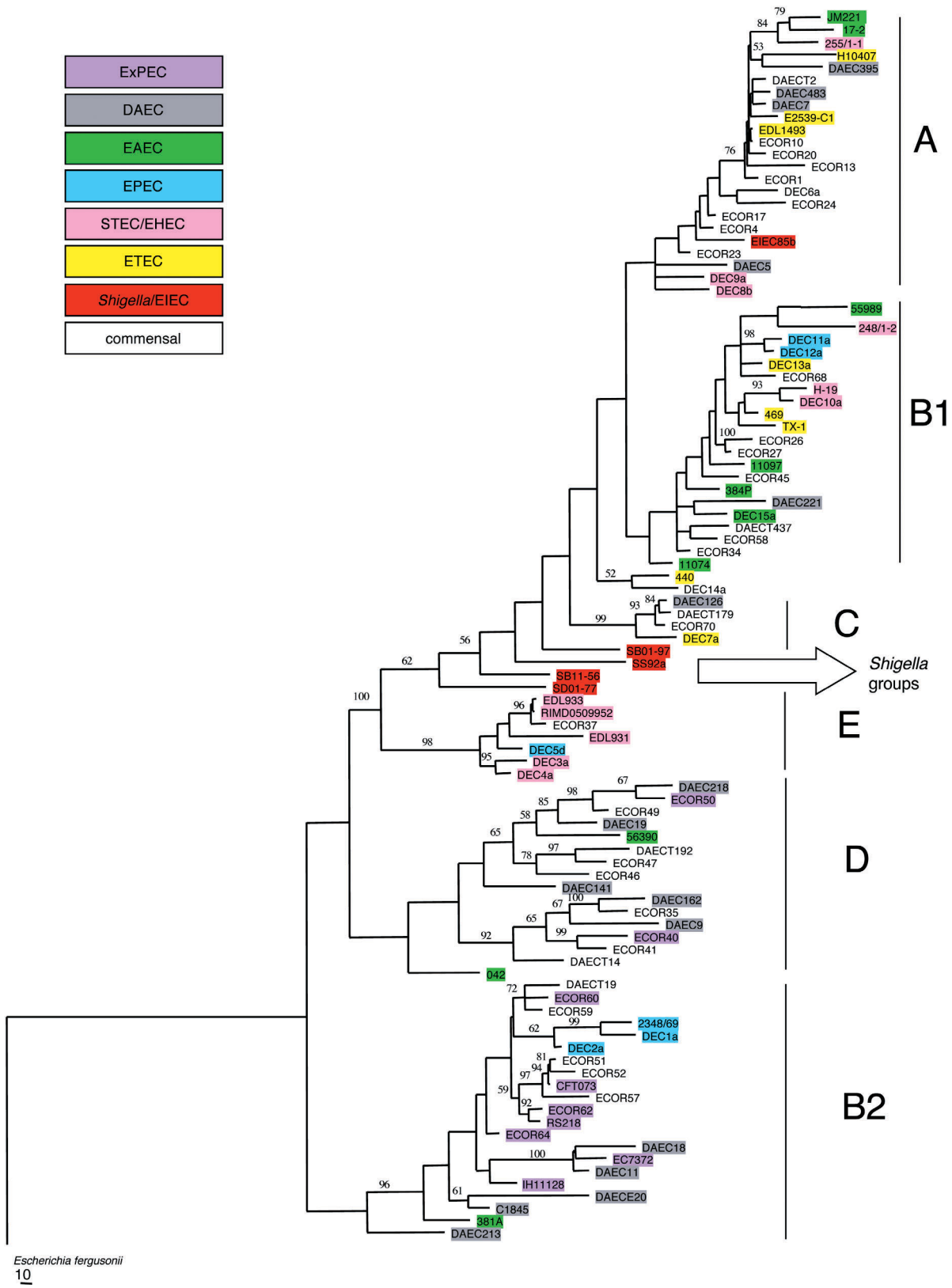


FIG. 1.—Semistrict consensus tree based on the simultaneous analysis of six essential chromosomal genes (*trpA*, *trpB*, *pabB*, *putP*, *icd*, and *polB*) using parsimony on a collection of 98 *E.coli*/*Shigella* strains, rooted on *E. fergusonii*. Total characters: 5,901; informative sites: 764; length: 4,620; number of most-parsimonious trees: 31; consistency index (CI): 0.37; retention index (RI): 0.7. Bootstrap values higher than 50% are indicated above the nodes. The vertical bars delineate the major phylogenetic groups A, B1, C, E, D, and B2.

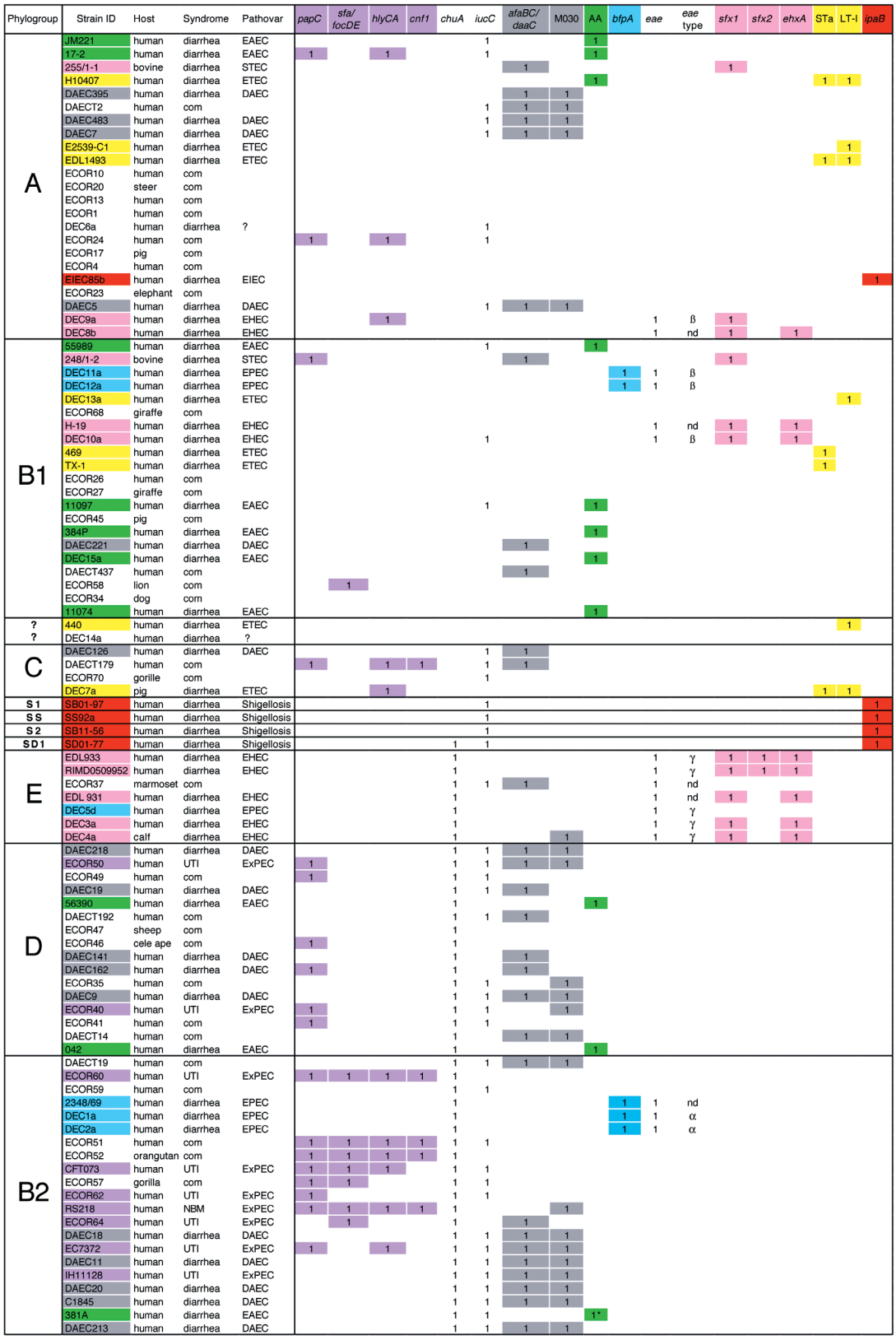
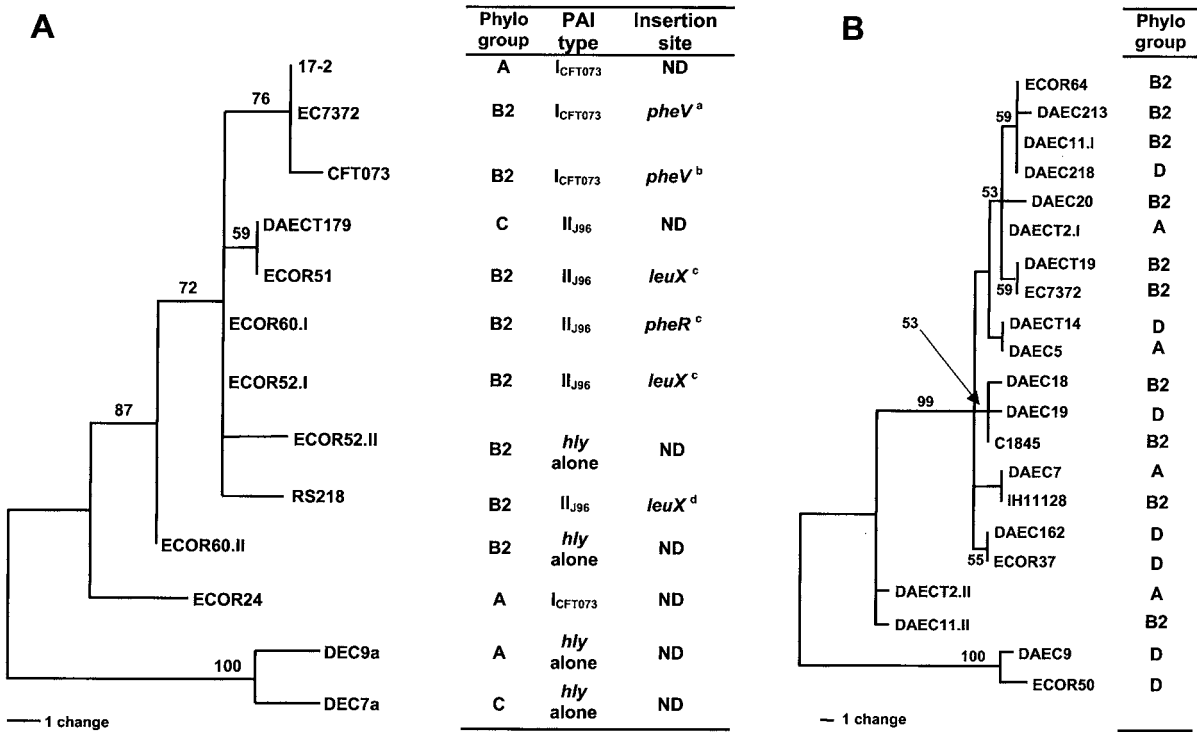


FIG. 2.—Characteristics of the studied strains indicating the presence of the 17 virulence factors as well as the *eae* type. The asterisk (*) for the AA probe in strain 381A indicates that it is positive for a AAF-II–like system. nd = not determined. Color code for the presence of VFs is according to the pathovar color (see figure 1) to which each VF is typical.



[a] Guignot et al. 2000 ; [b] Bingen-Bidois et al. 2002 ; [c] O. Clermont, unpublished data ; [d] Houdouin et al. 2002

FIG. 3.—Semistric consensus trees of the (*A*) *hlyCA* and (*B*) *afaD* gene sequences using parsimony with strain phylogenetic groups (*A* and *B*) and PAI characteristics (*A*). Trees are midpoint rooted. Characteristics of the trees are as follows. *afaD*: number of taxa: 21; total characters: 550; informative sites: 34; length: 45; number of most-parsimonious trees: 14; CI: 0.822; RI: 0.889. *hlyCA*: number of taxa: 13; total characters: 505; informative sites: 18; length: 39; number of most parsimonious trees: 5; CI: 0.897; RI: 0.893. The Roman numbers I or II after the name of the strains correspond to the two copies of the gene present in one strain (Labigne-Roussel and Falkow 1988; Swenson et al. 1996; Dobrindt et al. 2002*a*). Bootstrap values higher than 50% are indicated at the nodes. Identical topologies were obtained using neighbor-joining and maximum-likelihood (PHYML program of Guindon and Gascuel [2003]) analyses (data not shown).

arrival of the VF, we study the evolutionary history of some of the VFs.

The phylogenetic histories of the ExPEC VFs were assessed in two ways. First, we determined their localizations on known PAIs. It has been shown that the simultaneous presence of certain ExPEC VFs and *papG* alleles could be characteristic of PAI type (Bingen-Bidois et al. 2002; Bonacorsi et al. 2003). Thus, based on the specific combination of *pap* (including the *papG* allele), *sfa*/*foc*, *hly*, *cnf1*, *aer*, *iroN*, and *hra*, we were able to assign the ExPEC VFs to three known PAIs (I_{CFT073}, II_{J96}, and III_{536}) in the 22 strains exhibiting such genes (table S3 in Supplementary Material online). As multiple insertion sites have been described for the PAI II_{J96} (Bingen-Bidois et al. 2002), we also analyzed them in some representative strains exhibiting this PAI. Second, because the *hly* operon can be present within several PAIs or alone (or within an uncharacterized PAI), we chose this gene as a representative of ExPEC VFs to reconstruct its evolutionary history by sequencing *hlyCA*. Sequencing of *hlyCA* revealed that some strains exhibit two genes, as previously reported (Swenson et al. 1996; Dobrindt et al. 2002*a*). Indeed, the presence of two genes was suspected by the observation of two peaks in the same position on the electrophoregram. The individual full sequences were reconstructed (haplotypes) according to the known unique sequences. It is clear that the tree obtained with the *hly* sequences is clearly incongruent (i.e., in disagreement) with the strain phylogeny tree (fig. 3*A*). The arrival of the *hly* operon within the *E. coli* species corresponds to at least three different events, as *hlyCA* can be located on the PAI I_{CFT073} or II_{J96} or alone. The PAIs are not specific to phylogenetic groups: (1) the same PAIs harboring unique *hlyCA* sequences can be found in diverged strains (PAI II_{J96} in ECOR51 B2 and DAECT179 C strains, PAI I_{CFT073} in CFT073 and EC7372 B2 and 17-2 A strains) and (2) within the B2 group strains, *hly* gene can be found alone, in PAI I_{CFT073} or in PAI II_{J96} at two insertion sites, thus, corresponding at least to four arrival events (fig. 3*A*). All these data indicate multiple horizontal gene transfers for the acquisition of ExPEC VFs and demonstrate that the evolutionary histories of the VFs are distinct from the history of the strain.

The evolutionary history of part of the *afa* operon, which is largely predominant all over the *E. coli* phylogenetic groups, was also determined by sequencing the *afaD* gene (known to be variable at the opposite of the *afaBC*/*daaC* gene [Garcia et al. 2000]). Here again, the phylogenetic tree of the gene is not congruent with the strain phylogeny (fig. 3*B*), suggesting multiple independent arrivals.

The evolutionary history of a gene highly restricted to specific phylogenetic groups was analyzed by sequencing the LT-I gene. The analyzed 614 bp of this gene are highly conserved, as only four mutations (three being nonsynonymous) were observed, one being informative for parsimony (data not shown). This indicates several recent arrival of this gene within the A, B1, and C phylogenetic group strains. The multiple independent arrivals of the LEE PAI harboring *eae* gene in EPEC have been previously established (Reid et al. 2000).

Discussion

Three important conclusions can be derived from the distribution of pathovars in the phylogenetic tree (fig. 1): (1) ExPEC strains belong preferentially to group B2 and, to a lesser extend, to group D, but they do not belong to any of other phylogenetic groups; (2) in contrast, obligatory pathogens responsible for acute and severe diarrhea and mostly associated with the production of toxins and/or invasiveness of eukaryotic cells (EHEC, ETEC, and *Shigella*/EIEC) are not found in groups D or B2, (3) pathovars linked to chronic and mild diarrhea, such as EPEC, EAEC and DAEC, are distributed all over the phylogeny.

The link between phylogeny and ExPEC strains has been evidenced in several occasions by different authors (Picard et al. 1993; Boyd and Hartl 1998*a*; Bingen et al. 1998; Johnson et al. 2001; Bingen-Bidois et al. 2002; Bonacorsi et al. 2003). These studies have shown that the majority of strains isolated from urine or newborn cerebrospinal fluid belong to phylogenetic groups B2 and D and that these strains harbor a greater number of ExPEC VFs as compared with strains from other phylogenetic groups isolated in UTI or NBM. Additionally, it has been shown that the number of ExPEC VFs on a strain is proportional to its pathogenic potential (Picard et al. 1999). Based on our results, there is no evidence that gain of one VF increases the likelihood of retention of another VF (data not shown), as it has been suggested for *Salmonella* (Ochman and Groisman 1996). Commensal strains from group B2 also seems to harbor more ExPEC VFs than their counterparts from other phylogenetic groups (Duriez et al. 2001; Zhang, Foxman, and Marrs 2002). Two hypotheses have been proposed to explain the concentration of ExPEC VFs within group B2: (1) the existence of preexisting features of the B2 genome that increase compatibility between the B2 genome and ExPEC VFs and, (2) chance and timing with the acquisition of ExPEC VFs by a B2 group ancestor with subsequent vertical inheritance or loss of the VFs (Johnson et al. 2000). Our data (fig. 3*A*), as well as the sequencing data of *papA* (Boyd and Hartl 1998*b*; Johnson et al. 2001) and *sfaA* (Boyd and Hartl 1998*b*) in the ECOR strains, demonstrate that the arrival of ExPEC VFs within the phylogenetic B2 group strains correspond in fact to numerous horizontal gene transfer events. Such repeatedly observed recombination events into the same genetic background argue most strongly for B2 genetic background being a critical player in the acquisition of ExPEC VFs. It has been shown, for example, that the efficient expression of the α-hemolysin determinant located on a PAI depends on a complex mechanism where several core chromosomal gene products as Hha, H-NS, RfaH, and $tRNA_5^{Leu}$ are required (Dobrindt et al. 2002*b*). One could easily imagine that polymorphisms in sequence or expression of these regulatory proteins will influence the expression of the VF.

The link between phylogeny and virulence is also observed among intestinal pathogens. As mentioned above, toxin-producing and/or enteroinvasive pathovars, such as EHEC, ETEC and *Shigella*/EIEC, as well as their specific VFs, are only found in groups A, B1, C, or E. The distribution of these VFs is not because of a lack of genetic exchange among these intestinal pathogens and strains from groups D and B2. In fact, analysis of the number of transferred fragments on the sequences of the six essential genes used in the phylogenetic analysis (Denamur et al. 2000) indicates that genetic exchange occurs between strains of different phylogenetic groups, commensal and pathogenic strains, and strains of different pathovars (O. Tenaillon, personal communication). In contrast, it is indicative of the lack of compatibility between these VFs and the D and B2 genetic background.

EHEC strains are mostly concentrated within group E, where strains of serotype O157:H7 are found (EHEC 1 complex [Reid et al. 2000]). There is evidence in support of a model in which O157:H7 evolved sequentially from an O55:H7 (DEC5d) ancestor (Whittam et al. 1993; Feng et al. 1998; Monday, Whittam, and Feng 2001). Our phylogeny does not contradict this scenario, as the position of the DEC5d strain within the E group is not supported by a significant bootstrap value (fig. 1), and individual gene phylogenies showed that DEC5d strain is basal according to the EHEC O157:H7 strains in three of five genes (*pabB*, *icd*, and *putP*; considering *trpA* and *B* linked) (data not shown). Interestingly, the ECOR37 strain, which has been isolated from a marmoset, seems to have secondarily lost *stx* genes. No particular phylogenetic clustering of ETEC strains is observed on the tree (fig. 1), but the distribution of the ST and LT genes (fig. 2) associated with these pathovars is limited to the groups A, B1, and C. This may indicate a predominantly horizontal mode of transmission of these VFs by the multiple independent arrivals of the plasmids carrying these genes from another species or may indicate within-species dissemination. The almost complete nucleotide conservation of the LT-I genes within diverged strains of A, B1, and C phylogenetic groups argues for this scenario of several recent arrivals. However, it seems that an A, B1, or C genetic background is necessary for the arrival and/or maintenance of these genes. The importance of the genetic background on the evolution of *Shigella*/EIEC has been suggested by the correlation between the lack of some phenotypic characters (such as motility, lysine decarboxylation, and lactose utilization) and the presence of the virulence plasmid responsible for the pathogenic nature of the bacteria (Maurelli et al. 1998; Pupo, Lan, and Reeves 2000; Escobar-Páramo et al. 2003). It has been demonstrated that cadaverine, the product of the activity of the lysine decarboxylase, blocks the action of *Shigella* enterotoxin (Maurelli et al. 1998) and prevents the escape of *S. flexneri* from the phagolysosome (Fernandez et al. 2001).

The overall distribution of pathovars associated to mild and chronic diarrhea may be explained by the plasticity of these factors to adapt to different genetic backgrounds or by the interaction of these VF to an "ancestral background" of the bacterial genome. An ancient unique arrival of the VFs associated with these pathovars can be ruled out as the *afa* sequence data clearly suggest multiple horizontal gene transfer events (fig. 3*B*).

Conclusion

Although the evolution of pathogenicity in *E. coli* is mostly the result of the arrival of VFs in the population, the retention and expression of these factors is the result of the interaction between the bacterial genetic background and the new arriving genes. Virulence factors can be classified into three categories according to their interaction with the bacterial genomic background: (1) VFs that arrive and are expressed in different genetic backgrounds (such as those associated with mild chronic diarrheas), (2) VFs that arrive in different genetic backgrounds but that are preferentially found, associated with a specific pathology, in only one particular background (such as ExPEC VFs associated to extraintestinal diseases), and (3) VFs that require a particular genetic background for the arrival and expression of their virulence potential (such as those associated to EHEC, ETEC, and *Shigella*/EIEC strains).

This classification of the VFs genetic background interaction allows us to postulate the existence of two types of genomic backgrounds inside the *E. coli* genome. On one hand, there is the "ancestral background," which is present in all strains and allows the expression of VFs associated with mild and chronic diarrhea. On the other hand, a more "derived background" allows the expression of VFs associated with more severe pathologies. Interestingly, strains associated with severe and acute diarrheas all belong to groups originated after the differentiation of group D, suggesting that an important change in the *E. coli* genome took place at this point in the evolution of the species. Further modifications occurred after the split of group E, when the virulence plasmid of *Shigella*/EIEC arrived in the population, giving raise to a radiation from which groups A, B1, and C originated (Escobar-Páramo et al. 2003). Those modifications in the genomic background allow new VFs to arrive in the population, giving origin to the pathovars associated with severe acute diarrheas.

Comparative genomics among strains from different phylogenetic groups may help identify the changes that occurred in the genome after the split of group D (the "derived background"), as well as what may constitute the "ancestral background" of the *E. coli* genome. These comparisons are necessary for further understanding the implications of the genetic background in the evolution of pathogenicity in bacteria.

Supplementary Material

Data were deposited to GenBank under the following accession numbers: *icd*: AY245916 to AY245974; *trpB*: AY246054 to AY2246112; *trpA*: AY246113 to AY246170; *putP*: AY246275 to AY246333; *polB*: AY246334 to AY246388; *pabB*: AY255710 to AY255767; and *hlyCA* and *afaD*: AY525514 to AY525534.

Acknowledgments

We are grateful to Christiane Forestier and Alain L. Servin for providing DAEC strains, Thomas S. Whittam for providing the DEC strains, Laurence Du Merle for technical assistance, Olivier Tenaillon for the recombination analysis, Pierre Darlu for the maximum-likelihood and Bayesian analyses, Bertrand Picard and Claude Parsot for discussing the manuscript, and Jacques Elion for constant encouragements. This work was supported in part by the "Fondation pour la Recherche Médicale" and the "Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires."

Literature Cited

Bingen, E., B. Picard, N. Brahimi, S. Mathy, P. Desjardins, J. Elion, and E. Denamur. 1998. Phylogenetic analysis of *Escherichia coli* strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains. J. Infect. Dis. **177**:642–650.

Bingen-Bidois, M., O. Clermont, S. Bonacorsi, M. Terki, N. Brahimi, C. Loukil, D. Barraud, and E. Bingen. 2002. Phylogenetic analysis and prevalence of urosepsis strains of *Escherichia coli* strains bearing pathogenicity islands-like domains. Infect. Immun. **70**:3216–3226.

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

Blanc-Potard, A.-B., C. Tinsley, I. Scaletsky, C. Le Bouguenec, J. Guignot, A. L. Servin, X. Nassif, and M.-F. Bernet-Camard. 2002. Representational difference analysis between Afa/Dr diffusely adhering *Escherichia coli* and non-pathogenic *E.coli* K-12. Infect. Immun. **70**:5503–5511.

Bonacorsi, S., O. Clermont, V. Houdoin, C. Cordevant, N. Brahimi, A. Marecat, C. Tinsley, X. Nassif, M. Lange, and E. Bingen. 2003. Molecular analysis and experimental virulence of French and North American *Escherichia coli* neonatal meningitis isolates: identification of a new virulent clone. J. Infect. Dis. **187**:1895–1906.

Boyd, E. F., and D. L. Hartl. 1998*a*. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a clustered distribution. J. Bacteriol. **180**:1159–1165.

———. 1998*b*. Diversifying selection governs sequence polymorphism in the major adhesin proteins FimA, PapA, and SfaA of *Escherichia coli*. J. Mol. Evol. **47**:258–267.

Brando, 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.

Clermont, O., S. Bonacorsi, and E. Bingen. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic groups. Appl. Environ. Microbiol. **66**:4555–4558.

Czeczulin, J. R., T. S. Whittam, I. R. Henderson, F. Navarro-Garcia, and J. P. Nataro. 1999. Phylogeneic analysis of enteroaggregative and diffusely adherent *Escherichia coli*. Infect. Immun. **67**:2692–2699.

Denamur, E., G. Lecointre, P. Darlu et al. (12 co-authors). 2000. Evolutionary implications of the frequent horizontal transfer of mistmatch repair genes. Cell **103**:711–721.

Dobrindt, U., G. Blum-Oehler, N. Gabor, G. Schneider, A. Johann, G. Gottschalk, and J. Hacker. 2002*a*. Genetic structure and distribution of four pathogenicity islands (PAI I_{536} to PAI IV_{536}) of uropathogenic *Escherichia coli* strain 536. Infect. Immun. **70**:6365–6372.

Dobrindt, U, L. Emody, I. Gentschev, W. Goebel, and J. Hacker. 2002*b*. Efficient expression of the alpha-haemolysin determinant in the uropathogenic *Escherichia coli* strain 536 requires the *leuX*-encoded $tRNA5^{Leu}$. Mol. Genet. Genomics **267**:370–379.

Donnenberg, M. S. 2002. *Escherichia coli*: virulence mechanisms of a versatile pathogen. Elsevier Science Edition, Academic Press, San Diego, Calif.

Duriez, P., O. Clermont, S. Bonacorsi, E. Bingen, A. Chaventré, J. Elion, B. Picard, and E. Denamur. 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographical distinct human populations. Microbiology **147**:1671–1676.

Escobar-Páramo, P., C. Giudicelli, C. Parsot, and E. Denamur. 2003. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. J. Mol. Evol. **57**:140–148.

Escobar-Páramo, P., A. Sabbagh, P. Darlu, O. Pradillon, C. Vaury, G. Lecointre, and E. Denamur. 2004. Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. Mol. Phylogenet. Evol. **30**:243–250.

Feng, P., K. A. Lampel, H. Karch, and T. S. Whittam. 1998. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. J. Infect. Dis. **177**:1750–1753.

Fernandez, I. M., M. Silva, R. Schuch, W. A. Welker, A. M. Siber, A. T. Maurelli, and B. A. McCormick. 2001. Cadaverine prevents the escape of *Shigella flexneri* from the phagolysosome: a connection between bacterial dissemination and neutrophil transepithelial signaling. J. Infect. Dis. **184**: 743–753.

Garcia, M. I., M. Jouve, J. P. Nataro, P. Gounon, and C. Le Bouguénec. 2000. Characterization of the AfaD-like family of invasins encoded by pathogenic *Escherichia coli* associated with intestinal and extra-intestinal infections. FEBS Lett. **479**:111–117.

Gordon D. M., and A. Cowling. 2003. The distribution and genetic structure of *Escherichi coli* in Australian vertebrates: host and geographic effects. Microbiology **149**:3575–3586.

Guignot, J., J. Breard, M. F. Bernet-Camard, I. Peiffer, B. J. Nowicki, A. L. Servin and A. B. Blanc-Potard. 2000. Pyelonephritogenic diffusely adhering *Escherichia coli* EC7372 harboring Dr-II adhesin carries classical uropathogenic virulence genes and promotes cell lysis and apoptosis in polarized epithelial caco-2/TC7 cells. Infect. Immun. **68**: 7018–7027.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

Hacker, J., and J. B. Kaper. 2000. Pathogenicity island and the evolution of microbes. Ann. Rev. Microbiol. **54**:641–679.

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

Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTALV: improved software for multiple sequence alignment. CABIOS **8**:189–191.

Houdouin, V., S. Bonacorsi, N. Brahimi, O. Clermont, X. Nassif, and E. Bingen. 2002. A uropathogenicity island contributes to the pathogenicity of *Escherichia coli* strains that cause neonatal meningitis. Infect. Immun. **70**:5865–5869.

Huelsenbeck, J. P., and F. Ronquist, 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics **17**:754–755.

Johnson, J. R., P. Delavari, M. Kuskowsi, and A. L. Stell. 2001. Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*. J. Infect. Dis. **183**:78–88.

Johnson, J. R., M. Kuskowski, E. Denamur, J. Elion, and B. Picard. 2000. Clonal origin, virulence factors, and virulence. Infect. Immun. **68**:424–425.

Labigne-Roussel, A., and S. Falkow. 1988. Distribution and degree of heterogeneity of the afimbrial-adhesin-encoding operon (*afa*) among uropathogenic *Escherichia coli* isolates. Infect Immun. **56**:640–648.

Lawrence, J. G., H. Ochman, and D. L. Hartl. 1991. Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. **137**:1911–1921.

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

Maurelli, A. T., R. E. Fernandez, C. A. Bloch, and C. K. Rode. 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.

Monday, S. R., T. S. Whittam, and P. C. H. Feng. 2001. Genetic and evolutionary analysis of mutations in the *gusA* gene that causes the absence of beta-glucoronidase activity in *Escherichia coli* O157:H7. J. Infect. Dis. **184**:918–921.

Nataro, J. P., and J. B. Kaper. 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. **11**:142–201.

Ochman, H., and E. A. Groisman. 1996. Distribution of pathogenicity islands in *Salmonella* spp. Infect. Immun. **64**:5410–5412.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

Ochman, H., and R. K. Selander. 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157**:690–692.

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

Picard, B., C. Journet-Mancy, N. Picard-Pasquier, and P. Goullet. 1993. Genetic structures of the B_2 and B_1 *Escherichia coli* strains responsible for extra-intestinal infections. J. Gen. Microbiol. **139**:3079–3088.

Pupo, G. M., D. K. R. Karaolis, R. Lan, and P. R. Reeves. 1997. Evolutionary relationship among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* studies. Infect. Immun. **65**:2985–2692.

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 characters. Proc. Natl. Acad. Sci. USA **97**:10567–10572.

Reid, S. D., D. J. Betting, and T. S. Whittam. 1999. Molecular detection and identification of intimin alleles in pathogenic *Escherichia coli* by multiplex PCR. J. Clin. Microbiol. **37**:2719–2722.

Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature **406**:64–67.

Russo, T. A., and J. R. Johnson. 2000. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli*: ExPEC. J. Infect. Dis. **181**:1753–1754.

Swenson, D. L., N. O. Bukanov, D. E. Berg, and R. A. Welch. 1996. Two pathogenicity islands in pathogenic *Escherichia coli* J96: cosmid cloning and sample sequencing. Infect. Immun. **64**:3736–3743.

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sinauer Associates, Sunderland, Mass.

Whittam, T. S., M. L. Wolfe, I. K. Wachsmuth, F. Orskov, I. Orskov, and R. A. Wilson. 1993. Clonal relationship among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. Infect. Immun. **61**:1619–1629.

Zhang, L., B. Foxman, and C. Marrs. 2002. Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group B2. J. Clin. Microbiol. **40**:3951–3955.

Brian Golding, Associate Editor

Accepted February 4, 2004