Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains

Tony Le Gall,†[1] Olivier Clermont,*[1] Stéphanie Gouriou,‡ Bertrand Picard,* Xavier Nassif,§ Erick Denamur,* and Olivier Tenaillon**

*INSERM U722 and Université Paris 7 Denis Diderot, Faculté de Médecine, Paris, France; †INSERM U613 and Université de Bretagne Occidentale, CHU Brest, Brest, France; ‡Laboratoire de Microbiologie, CHU Brest, Brest, France; and §INSERM U570 and Université Paris 5 René Descartes, Faculté de Médecine, Paris, France

The selective pressures leading to the evolution and maintenance of virulence in the case of facultative pathogens are quite unclear. For example, *Escherichia coli*, a commensal of the gut of warm-blooded animals and humans, can cause severe extraintestinal diseases, such as septicemia and meningitis, which represent evolutionary dead ends for the pathogen as they are associated to rapid host death and poor interhost transmission. Such infectious process has been linked to the presence of so-called "virulence genes." To understand the evolutionary forces that select and maintain these genes, we focused our study on *E. coli* B2 phylogenetic group strains that encompass both commensal and pathogenic (extra- and intraintestinal) strains. Multilocus sequence typing (MLST), comparative genomic hybridization of the B2 flexible gene pool, and quantification of extraintestinal virulence using a mouse model of septicemia were performed on a panel of 60 B2 strains chosen for their genetic and ecologic diversity. The phylogenetic history of the strains reconstructed from the MLST data indicates the emergence of at least 9 subgroups of strains. A high polymorphism is observed in the B2 flexible gene pool among the strains with a good correlation between the MLST-inferred phylogenetic history of the strains and the presence/absence of specific genomic regions, indicating coevolution between the chromosomal background and the flexible gene pool. Virulence in the mouse model is a highly prevalent and widespread character present in all subgroups except one. Association studies reveal that extraintestinal virulence is a multigenic process with a common set of "virulence determinants" encompassing genes involved in transcriptional regulation, iron metabolism, adhesion, lipopolysaccharide (LPS) biosynthesis, and the recently reported peptide polyketide hybrid synthesis system. Interestingly, these determinants can also be viewed as intestinal colonization and survival factors linked to commensalism as they can increase the fitness of the strains within the normal gut environment. Altogether, these data argue for an ancestral emergence of the extraintestinal virulence character that is a coincidental by-product of commensalism. Furthermore, the phenotypic and genotypic markers identified in this work will allow further epidemiological studies devoted to test the niche specialization hypothesis for the B2 phylogenetic subgroups.

Introduction

Theoretical and empirical studies have created a convincing conceptual framework regarding the evolution of virulence for obligate pathogens. However, the evolution of virulence in facultative pathogens, such as *Escherichia coli*, is poorly understood. The selective advantages associated with colonization of blood in bacteriemia or of cerebrospinal fluid in meningitis are not clear, as such infections often lead to rapid host death and poor transmission to new hosts. Yet, the ability to initiate such infections requires that the cell possesses many elaborate traits usually termed "virulence factors." The apparent contradiction between the presence and maintenance of many virulence factors and the presumed poor selective advantage accruing to the cell from extraintestinal infection has lead to the idea that many virulence factors might be required for the colonization of niches, where disease does not result, or for transmission to new hosts. This view leads to the hypothesis that infections caused by facultative pathogens occur by accident. However, such a hypothesis is difficult to test in the laboratory as it is almost impossible to reconstruct the natural environment in which a bacteria species evolves. Nevertheless, a combination of genomic, phylogenetic, and epidemiological data provides a framework with which to study how virulence and commensalism evolves in facultative pathogens and give hints on the selective success of a given strategy.

To investigate the determinants of facultative pathogen evolution, we used *E. coli* as a model species. *Escherichia coli* is typically a commensal of the lower gastrointestinal tract of humans and other warm-blooded vertebrates, but it can cause a variety of devastating diseases, including diarrhea, cystitis, pyelonephritis, septicemia, and meningitis (Donnenberg 2002). *Escherichia coli* species can be considered as having mainly a clonal genetic structure, with a low level of recombination (Selander and Levin 1980; Desjardins et al. 1995). The majority of *E. coli* strains can be assigned to 1 of 4 main phylogenetic groups, A, B1, B2, and D (Herzer et al. 1990; Escobar-Páramo, Sabbagh, et al. 2004). Strains of the 4 phylogenetic groups differ in their phenotypic and genotypic characteristics and appear to have different ecological niches and life history characteristics (Gordon and Cowling 2003; Escobar-Páramo, Grenet, et al. 2004; Escobar-Páramo et al. 2006). In addition, a variety of putative virulence factors associated with extraintestinal infections are nonrandomly distributed among strains of the 4 phylogenetic groups, with strains from groups B2 and D harboring a greater frequency and diversity of virulence traits compared with strains of groups A and B1 (Bingen et al. 1998; Boyd and Hartl 1998; Johnson et al. 2001). Further, the majority of strains isolated from extraintestinal body sites are members of group B2 and to a lesser extent group D (Cherifi et al. 1991; Picard et al. 1999). By contrast, with the exception of some diffusely adherent *E. coli* and group 1 enteropathogenic *E. coli* (EPEC), diarrhea-causing strains

[1] Equally contribution to this work.

Key words: *Escherichia coli*, commensalism, extraintestinal virulence, coincidental by-product.

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

Mol. Biol. Evol. 24(11):2373–2384. 2007
doi:10.1093/molbev/msm172
Advance Access publication August 19, 2007

are not B2 strains (Reid et al. 2000; Escobar-Páramo, Clermont, et al. 2004; Wirth et al. 2006). Consequently, *E. coli* group B2 strains are an excellent model with which to investigate the evolution of virulence in a facultative pathogen. On the one hand, group B2 strains are responsible for many extraintestinal infections and are therefore a major public health concern (Donnenberg 2002), whereas on the other hand, group B2 strains can be the strains most frequently isolated from the feces of asymptomatic humans (Zhang et al. 2002; Escobar-Páramo, Grenet, et al. 2004; Escobar-Páramo et al. 2006). Moreover, it seems that in industrialized countries, the prevalence of B2 strains isolated in human feces has substantially increased over the last 2 decades (Escobar-Páramo, Grenet, et al. 2004; Nowrouzian et al. 2006).

Due to its contrasted profile of virulence and commensalism, we focused on B2 group and did an in-depth analysis of a panel of 60 B2 phylogenetic group strains chosen for their genetic and ecologic diversity, according to their host and geographic origin, commensality/pathogenicity, O type, and virulence factor content profiles. To quantify the intrinsic virulence of strains, rather than using the isolation conditions that can be strongly influenced by the host conditions, we used a standardized mouse model of extraintestinal virulence (Picard et al. 1999; Johnson, Clermont, et al. 2006). Bacterial evolution occurs mainly by clonal divergence through the modification of existing genetic information and by acquisition of new sequences through horizontal gene transfer (HGT), followed by periodic selection (Ochman et al. 2000). This has led to the concept that bacterial genome are composed of a conserved "core" genome, which contains the genetic information that is required for essential cellular functions, and of a "flexible" gene pool, which encodes additional traits that can be beneficial under certain circumstances (Dobrindt et al. 2004). In *E. coli*, phylogenetic (Escobar-Páramo, Clermont, et al. 2004) and molecular (Dobrindt et al. 2002) data support the idea that the fitness of the organism is optimum when there is a "fine-tuning" between the chromosomal backbone and the genes newly acquired by horizontal transfer. We therefore decided to study both aspects of genome evolution: the phylogenetic relationships through the sequencing of 7 genes from the core genome and the analysis of presence/absence of genes of the B2 flexible gene pool using macroarrays.

Materials and Methods

Bacterial Strains

Sixty strains belonging to the B2 phylogenetic group (Clermont et al. 2000) were studied. These strains originated from several collections (Ochman and Selander 1984; Picard et al. 1999; Duriez et al. 2001; Bonacorsi et al. 2003; Watt et al. 2003; Escobar-Páramo, Clermont, et al. 2004; Escobar-Páramo, Grenet, et al. 2004; Escobar-Páramo et al. 2006; Johnson, Clermont, et al. 2006) including the *E. coli* reference (ECOR) collection (Ochman and Selander 1984). Also included are those strains with fully (CFT073, EPEC 2348/69, 536) or partially (RS218, F11) sequenced genomes. These strains were isolated over a 25-year period from various continents, a variety of host species, and intestinal and extraintestinal body sites. Eight strains belonging to the other phylogenetic groups of *E. coli* (A, B1, D, and E) were also included. *Escherichia fergusonii*, the closest *E. coli* relative (Lawrence et al. 1991), was used as an outgroup. The origin and primary characteristics of the strains are given in supplementary table S1 (Supplementary Material online).

Multilocus Sequence Typing Analysis

The phylogenetic relationships among the strains were inferred using nucleotide sequence data from 7 essential genes (*trpA*, *trpB*, *pabB*, *putP*, *icd*, *polB*, and *dinB* [Gerdes et al. 2003]) that are thought to experience little recombination and to produce a strong phylogenetic signal (Escobar-Páramo, Sabbagh, et al. 2004). Sequence data for the 7 genes were concatenated (7,016 nucleotides) and analyzed using 3 methods of phylogenetic reconstruction (Neighbor-Joining, Maximum Parsimony, or Maximum Likelihood) as implemented in the molecular evolutionary genetic analysis (MEGA) program (Kumar et al. 2001). The presence of recombination events among the sequences was inferred with the Clonal Frame software (Didelot and Falush 2007).

Serotype and Virulence Factors

The O type was taken from the literature or determined using antisera (Gastroenteric Disease Center, Pennsylvania State University, University Park, PA) and/or an allele-specific polymerase chain reaction (PCR) method (Clermont et al. 2007). The presence of the K1 antigen was determined by a PCR of the *neuC* gene as in Diard et al. (2007). The presence of 9 virulence factors implicated in extraintestinal virulence [*sfa*/*foc*, *iroN*, *aer* (*iucC*), *papC*, *hly*, *cnf1*, *hra*, *fyuA*, and *irp2*], as well as the alleles of *papG* (I, II, or III), were determined by PCR as previously described (Bingen-Bidois et al. 2002). The presence of these genes can be used to infer if one or more of 4 *E. coli* pathogenicity islands (PAIs) (PAI I_{CFT073}, PAI II_{J96}, PAI III_{536}, high pathogenicity island [HPI]) and plasmids bearing *aer* and/or *iroN* are present in the genome of strains as reported in Bonacorsi et al. (2003) (supplementary table S1, Supplementary Material online).

DNA Macroarray Membrane

To perform comparative genomic hybridizations (CGH), a DNA macroarray containing part of the B2 flexible gene pool was developed based on in vitro and in silico comparative genome analyses between B2 and A group strains. DNA sequences specific for RS218 (neonatal meningitis, O18:K1) and CFT073 (pyelonephritis, O6:K2) strains were identified in silico by using sequences of clones (GenBank accession numbers AF222070–AF222307) obtained in vitro by subtractive hybridization between the C5 strain (a O18:K1 neonatal meningitis strain closely related to RS218) and nonpathogenic group A *E. coli* strains (Bonacorsi et al. 2000). This set of B2-specific DNA sequences was increased by in silico comparative analyses among the *E. coli* genomes that were available when this study was initiated (K12-MG1655, CFT073, and RS218).

These analyses gave rise to 1,172 specific genomic DNA fragments (open reading frames [ORFs] and large [>500 bp] intergenic regions [IRs]). The 1,172 DNA fragments were amplified by PCR using specific primers from either RS218 (300 from ORFs, 18 IRs) or CFT073 (764 from ORFs, 90 IRs) genomes. ORFs longer than 1 kbp were amplified in sections of about 500 bp. The amplicons were spotted in duplicate by a robot (Eurogentec, Seraing, Belgium) on nylon membranes and fixed by alkali treatment. Positive (16S rDNA and genomic DNA of RS218 and CFT073 strains) and negative (human alpha-globin gene DNA and no DNA) control spots were used for normalization.

To avoid misinterpretation due to the presence of repeated sequences, sequences were excluded from the analysis when they were found to be at least duplicated in one or more of the 4 B2 fully or partially sequenced genomes (CFT073, RS218, 536, and E2348/69). Duplications, complete or partial, were identified by blasting each array nucleotide sequence against the 4 genome sequences cited above. Considering all hit sequences found by those Blasts, a sequence was defined as duplicated if one or more hits satisfied all of the following minimal criteria: a distance of at least 1 kbp from the main sequence found, a Blast expect value (*E* value) $<10^{-3}$, an identity >85%, a length >100 nt, and a ratio length between the main sequence found and its putative duplicated counterpart >50%. These method and criteria lead us to discard 214 sequences. In addition, 142 other sequences were excluded from the analyses either because they were related to IS, phages, or transposons or because no or insufficient information was available for them. Finally, among the 1,172 spotted specific sequences, 816 (70%) were considered for subsequent analyses (supplementary table S2, Supplementary Material online). About half (49.7%) of these sequences are common to CFT073 and RS218.

Genomic regions were defined as a subset of sequences found consecutively in the reference strain without any disruption by any gene with shared homology and position in both K12 and the reference strain, that is, a conserved gene whose relative position in both genome is similar, revealing the backbone of the species. These fragments were displayed in 77 genomic regions (R), noted from R1 to R84 (supplementary table S3, Supplementary Material online). The distribution of these DNA fragments within each defined genomic region and along the CFT073 chromosome and the RS218 chromosome plus plasmid is shown in supplementary figure S1 (Supplementary Material online).

Genomic DNA Extraction, Labeling, Macroarray Hybridization, and Data Acquisition

CGH was performed on a set of 52 B2 group strains (supplementary table S1, Supplementary Material online). Bacteria were grown in 869 liquid media at 37 °C, under aerobic conditions with constant shaking. Genomic DNA was isolated from 1 ml of an overnight culture using the DNA Wizard Genomic DNA Purification Kit (Promega, Madison, WI) according to the manufacturer's protocol. Five hundred nanograms of genomic DNA was then used as the matrix in a polymerization reaction in presence of [α^{33}P] deoxycytidine triphosphate (PerkinElmer, Maanstraat, Germany), random hexamer probes, and the Klenow fragment of DNA polymerase I (Roche, Mannheim, Germany), as recommended by each manufacturers. Hybridizations were carried out for 15–18 h at 68 °C in ExpressHyb Hybridization solution (Clontech, Saint Quentin en Yvelines, France). Following hybridization, each membrane was rinsed with washing solution (0.5× saline sodium phosphate EDTA/0.2% sodium dodecyl sulfate [SDS]) at room temperature for 3 min, 3 times, followed by 3 washes in the same solution at 65 °C for 20 min each. The membranes were then exposed to a PhosphorImager screen (AGFA ADC plate MD30) for 48–72 h. The PhosphorImager screens were scanned on a Storm 860 PhosphorImager device (Molecular Dynamics, Sunnyvale, CA) at a pixel size of 100 μm. Before rehybridization, membranes were stripped in 500 ml of a boiling buffer (10 mM Tris–HCl pH 7.6/1 mM ethylenediaminetetraacetic acid pH 8.0/1% SDS) for 25 min at 100 °C and the efficacy of this stripping procedure was controlled following a reexposition of the filter to a PhosphorImager screen. In all, among the 52 processed strains (fig. 1*A* and supplementary table S1 [Supplementary Material online]), 34 were hybridized 1 time and 18 (including CFT073 and RS218 strains) were hybridized 2 or more times (up to 12 times for CFT073).

Macroarray Data Processing

The resulting images were analyzed in order to determine the presence of the DNA sequences in each strain. A program was developed based on a statistical procedure previously described for "presence/absence decisions" (Kim et al. 2002). The program first grids the image and records the signal intensity associated with each spot. The background threshold is calculated as the mean intensity of the negative spots (empty spots and human DNA). Hybridizations of reference strains were performed twice for RS218 and 12 times for CFT073 to monitor membrane quality. Second, the program calculates for each spot a value called expected probability of presence (EPP) (Kim et al. 2002) for the corresponding sequence in the evaluated strain genome. EPP is calculated as follows: the intensity of a spot signal is divided by the spot signal obtained when hybridizing a reference strain carrying the gene spotted at that position. The distribution of the logarithm of all such ratios is then plotted. In an ideal situation, a bimodal distribution should be obtained, one where the right-hand side corresponds to the sequences present in the strain under study and where the left-hand side corresponds to the absent sequence in such strain. Unfortunately, computed data reveal a Gaussian-like peak with a long tail on the left corresponding to an excess of poorly hybridizing spots with the strain under study. However, using the right parts of the distribution, one can estimate the parameter of the unskewed normal distribution that corresponds to spots present in the strain under study. We used a fit (Press 1992) taking into account the whole right-hand side distribution and not just height and width of the peak as initially proposed by Kim et al. (2002). EPP of a spot is then the chance that its signal ratio is due to the presence of a sequence as predicted from the normal distribution (see fig. 4 in Kim et al. 2002).

$$\%\mathrm{EPP} = 100 \times (\text{expected normal value}/\text{observed value}).$$

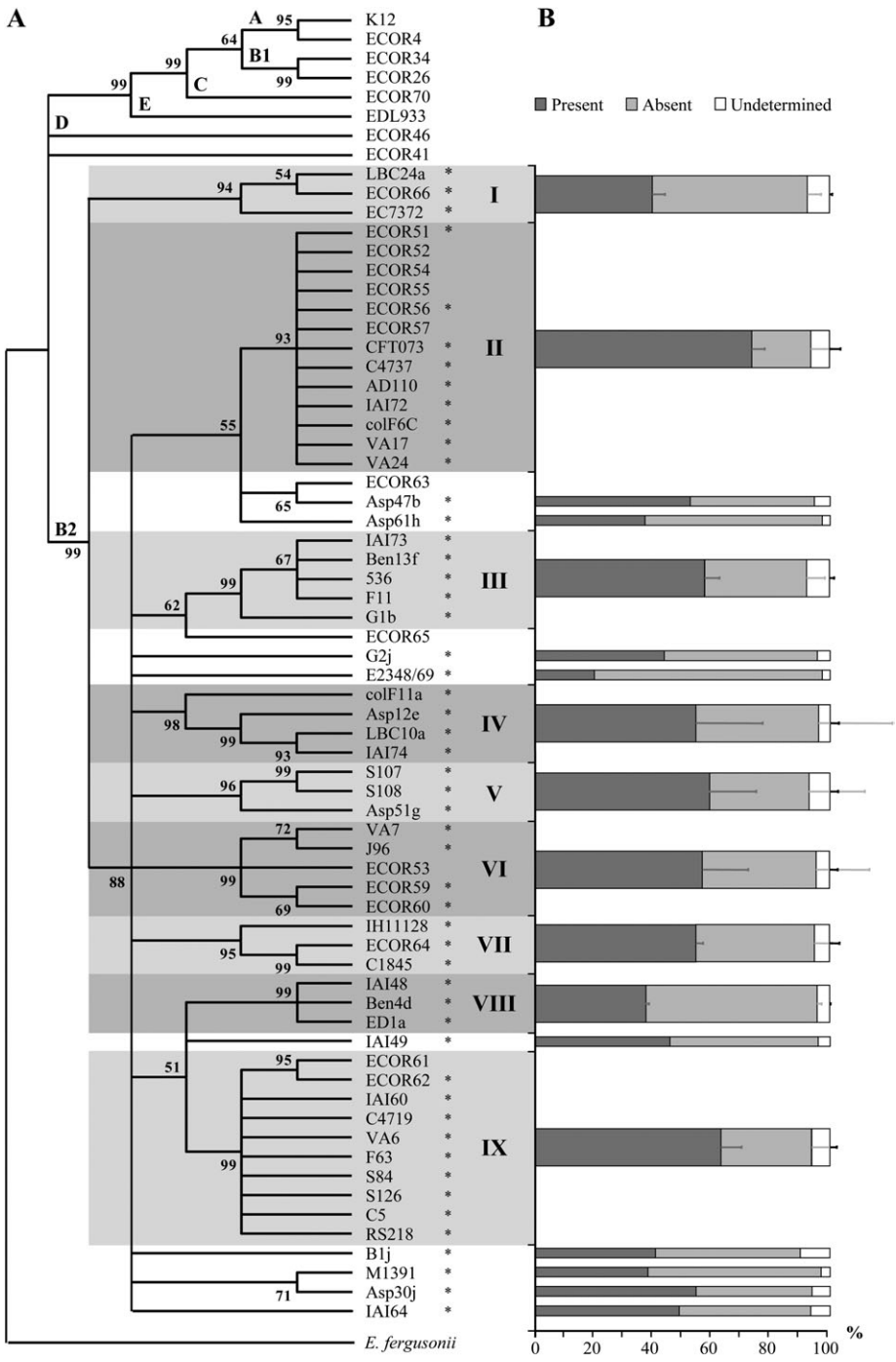

FIG. 1.—(*A*) Phylogenetic tree of the 68 *Escherichia coli* strains used in this study, *Escherichia fergusonii* as an outgroup, showing the delineation of 9 B2 subgroups (I–IX). This bootstrap consensus tree is based on the simultaneous analysis of 7 essential chromosomal genes (*trpA*, *trpB*, *pabB*, *putP*, *icd*, *polB*, and *dinB*) using parsimony procedure. Total characters: 7,016; informative sites: 591; total number of trees: 638; tree length: 2,242; consistency index: 0.49; retention index: 0.74. Bootstrap values, calculated on 500 replicated trees, are shown if higher than 50%. The studied strains belong either to B2 phylogenetic group (60 strains) or to A, B1, C, D, and E phylogenetic groups (8 strains). The stars indicate the strains that have been studied by macroarrays. (*B*) Sequence contents (in percentages) of the strains determined by the macroarrays of the 9 B2 subgroups (I–IX) and of the ungrouped strains. For the strains belonging to one of the 9 subgroups, the means are indicated with SDs. A gradient in the gene content is observed from subgroup VIII containing the less to subgroup II containing the more.

For instance, if 40 spots have a signal ratio of x and the expected number of positive sequences for the x value is 10, then the 40 spots have an EPP of 25%. For a given strain, the EPP of each sequence spotted on the membrane was averaged over comparisons to all reference strains. Such a method is interesting in that it does not require the definition of cutoff values for all sequences. The macroarray sequence data were then coded as follows: "0" for undetectable (EPP $<20\%$), "1" for detectable (EPP $>80\%$), and "?" for unknown (EPP between 20% and

80%). However, thresholds of 50% or 5–95% gave similar results (data not shown). Genomic regions were considered present within a strain when a least 50% of the sequences within that region were found to be present in this strain. Ratios of 30% or 70% gave similar results (data not shown). More extreme values will make our analysis more sensitive to false positive, to cross hybridizations, or to partial deletions occurring after large segment acquisition.

The output data sets obtained were then exported to a Microsoft Excel spreadsheet for subsequent manipulations. They were used 1) to reconstruct a phylogenetic Neighbor-Joining tree and 2) to assess global statistical analyses by factorial analysis of correspondence (FAC) (see below).

Mouse Lethality Assay

A mouse model of systemic infection was used to assess the intrinsic virulence of the 60 B2 strains (Picard et al. 1999). For each strain, 10 outbred female Swiss OF1 mice (3–4 weeks old, 14–16 gm) were challenged subcutaneously in the abdomen with a standardized bacterial inoculum (10^9 cfu/ml of log-phase bacteria in 0.2 ml Ringer solution). Mortality was assessed over 7 days postchallenge. Previously published lethality assay results were available for 17 of the strains (Picard et al. 1999; Johnson, Clermont, et al. 2006). Assays were performed for the balance of the strains using the urosepsis strains CFT073 as a positive control and the fecal-derived strain K12-MG1655 as a negative control. In this model system, lethality is a rather clear-cut parameter and strains were usually classified as nonkiller (strains killing <2 mice out of 10) or killer (strains killing >8 mice) (Johnson, Clermont, et al. 2006). The results for 4 of the strains did not fall into either category, and so these strains were scored as intermediate killers.

Global Statistical Analyses

FAC was used to describe associations among different data sets. FAC uses a covariance matrix based on Khi2 distances (Greenacre 1992). This computation method determines a plane defined by 2 principal axes of the analysis; the first axis (F1) accounts for most of the variance and the second axis (F2), orthogonal to F1, accounts for the largest part of the variance that is not accounted by F1. FAC was conducted with SPAD.N software (Cisia, Saint Mandé, France) from 2 two-way tables. The first table had 1) 52 rows, one for each B2 *E. coli* strain for which multilocus sequence typing (MLST) and CGH data are available, and 2) 846 columns corresponding to 846 variables: extraintestinal pathogen/intestinal pathogen/commensal origin; 10 phylogenetic subgroups among the B2 phylogenetic group as defined by MLST (B2 I to B2 IX and B2 UG); 8 O types (O2, O4, O6, O16, O18, O25, O75, O81, other O types [distinct O types were considered as a variable when they were present in at least 2 strains, otherwise all the other O types were assigned to an unique variable = "other O types"]), K1 antigen, 4 PAIs (PAI I_{CFT073}, PAI II_{J96}, PAI III_{536}, and HPI) and the plasmid bearing *aer* and/or *iroN* genes, all deduced from the presence/absence of 10 virulence genes (supplementary table S1, Supplementary Material online); mouse killer, mouse intermediate killer, mouse nonkiller; and 816 spotted sequences on the macroarray (supplementary table S2, Supplementary Material online). The second table had 1) 52 rows, one for each B2 *E. coli* strain, and 2) 107 columns corresponding to 107 variables: the first 30 variables were identical to those of the first table, 77 genomic regions as defined from the macroarray data (supplementary table S3, Supplementary Material online). In these tables, data were binary coded: "1" for present and "0" for absent.

Results

Phylogenetic History of the B2 Strains

To assess the phylogenetic relationships among the strains, a MLST approach based on validated (Escobar-Páramo, Sabbagh, et al. 2004) genes was performed using *E. fergusonii* as an outgroup. All B2 strains grouped together and, as previously reported (Lecointre et al. 1998; Escobar-Páramo, Clermont, et al. 2004; Escobar-Páramo, Sabbagh, et al. 2004), the B2 group appears basal, with A and B1 groups being the more recently diverged (fig. 1*A*). Based on the presence of at least 3 strains and a bootstrap value $>75\%$, 9 B2 subgroups (I–IX) were clearly delineated (the lower bootstrap value being 93%). Forty-nine strains were clustered in these subgroups (from 13 to 3 strains per subgroup), whereas 11 stand outside. Each subgroup encompasses strains isolated over a 25-year period on 2 continents (supplementary table S1, Supplementary Material online). The 2 most represented subgroups are the subgroups II and IX with 13 and 10 strains, respectively. The EPEC strain E2348/69 stands alone, as a representative of the EPEC 1 group (Reid et al. 2000). ECOR65 strain, although not in a subgroup according to the bootstrap value criteria, is closely related to the subgroup III strains (fig. 1*A*). These subgroups were retrieved with the same high confidence whatever the method used to reconstruct the tree (NJ, MP, or ML) (data not shown). By contrast, the interrelationships between these groups are not robust and the subgroups emerge as in a star phylogeny in the bootstrap tree (fig. 1*A*). However, the subgroup I always appears as basal using the 3 reconstruction methods with a bootstrap value $\geq 50\%$ (fig. 1*A* and data not shown).

As bacterial genomes are subject to recombination, phylogenetic approaches are to be taken with care. The vast majority of recombination events involve the acquisition of short tracks of DNA in a process similar to gene conversion. Such a process of recombination does not affect the global genealogic relationships between strains, so that a phylogenetic approach is still valid as long as a long enough sequence is studied to reveal the global topology and not its local distortion due to the short gene conversions. A recent software, Clonal Frame (Didelot and Falush 2007), was developed to take into account the bacterial patterns of recombination. The use of Clonal Frame on our MLST scheme (with standard parameters) revealed the presence of recombination of on average 126-bp long (95% confidence interval [CI] 103–160). Even if the ratio of recombination over mutation was close to unity (ratio of 0.75, 95% CI:

0.56–0.98), a 95% consensus tree performed with such an approach recovered all the subgroups we have defined. The basal position of subgroup I was not recovered with this method. As the ancestral sequences are predicted with many recombination events, we question the reliability of this very new method for deep branches of the tree. The use of Clonal Frame confirms the idea that recombination is occurring in our data set, that it is mainly due to very short gene transfers, and that such recombination is not strong enough to destroy the observed pattern: the existence of subgroups within the B2 group. It also suggests that the absence of robust topological information on the intergroup relationships is due to low levels of information for deep branches of the tree as well as the existence of recombination that creates more inconsistencies in deeper branches.

In sum, MLST data show that the evolutionary history of B2 group is dominated by the emergence of at least 10 clonal complexes, one being represented by the EPEC 1 group.

B2 Flexible Gene Pool Content

A macroarray was developed with a set of sequences present in the B2 extraintestinal pathogenic strains RS218 (neonatal meningitis) and CFT073 (pyelonephritis) but not in the commensal derived K12 strain. These 2 pathogenic strains belong to B2 subgroups IX and II, respectively (fig. 1*A*). In a first step validation of the array used, in silico analyses based on a Blast procedure of the spotted DNA sequences on the whole-genome sequences of 4 B2 group strains (E2348/69, 536, CFT073, and RS218) were performed in order to correlate the predicted data (presence/absence) to the experimental hybridization data sets obtained with these 4 strains. The correlation was excellent because, according to the strain, 91–95% of sequences were in full agreement between the in silico and the experimental data sets. The second validation step involved extracting DNA from several cultures of CFT073 and RS218 strains and 2 repeats using the same DNA for 15 other strains. As pairwise correlations always grouped together all the hybridizations performed for a same strain, values from repeated experiments were averaged and then used for analyses. Of the 816 sequences retained for analysis (supplementary table S2, Supplementary Material online), 69 (8%) were present in all the 52 B2 tested strains, whereas 56% were present in at least 25 of these strains. The 69 B2 common sequences are mainly involved in the transport of sugars (phospho-transferase systems, sorbose or sorbitol specific, etc.) or iron (TonB receptor, heme/hemoglobin transport, etc.) with some also coding for transcriptional regulators.

EPEC-E2348/69 (subgroup UG) and LBC24a (subgroup I) strains contain the fewest B2 flexible gene pool sequences, with 20% and 35% of the sequences present in their genomes, respectively. By contrast, the genomes of CFT073 and C4737 strains are the most enriched, containing 80% and 82% of the sequences, respectively. These 2 strains belong to the B2 subgroup II, a group of 9 strains of which 8 are among the most enriched strains studied ($74 \pm 5\%$). B2 subgroup VIII includes 3 strains that harbor very few sequences ($38 \pm 1\%$) (fig. 1*B*). By considering the complete genome sequences available for CFT073 and RS218 strains, the genomic location of the 816 sequences was then used in order to identify those sequences that are physically linked (see Material and Methods). Thus, we were able to define 77 genomic regions (supplementary fig. S1 and table S3, Supplementary Material online): 65 encompassing from 2 up to 81 sequences and 12 regions encompassing only one sequence.

As expected, the strains and B2 subgroups mentioned above span the minimum and maximum numbers of the 77 genomic regions (supplementary fig. S2, Supplementary Material online). The fact that strains of subgroups B2 II and IX encompass the higher numbers of sequences/genomic regions is expected as the macroarrays were developed based on comparisons between RS218 and CFT073 and K12. Of note, 11 genomic regions are common to all the B2 strains. These common regions encompass genes, as stated above for the B2 common sequences, mainly involved in sugar and iron metabolisms (supplementary fig. S3, Supplementary Material online).

Clearly, there is a high degree of polymorphism in the flexible gene pool content among B2 strains at the level 1) of individual sequences, 2) of sequences within a genomic region, and 3) of genomic regions.

MLST versus Comparative Genomic Hybridization Data, PAIs, and O Types

The search for associations between the phylogenetic history of the strains based on the MLST data and the presence/absence of sequences from the B2 pool was undertaken in 2 ways.

First, unrooted NJ trees were reconstructed from the macroarray data, either using sequences or genomic regions, and compared with the MLST tree. NJ, a clustering method without model was preferred as other population genetics model of evolution apply to nucleotide data and not to gene acquisition and loss (a complex asymmetrical model linking sites would be required). The resulting macroarray trees are influenced by different patterns of evolution. On the one hand, genomic region analysis predominately reflects the occurrence of HGT throughout the history of the sample. Whether a single HGT resulted in the acquisition of one sequence or a 100, the same weight is given to the presence of the genomic region as both reflect a single genomic event. The bootstrap consensus tree obtained with genomic regions (fig. 2) revealed a high congruence with the one obtained with MLST, although with lower bootstrap values, due to smaller number of sites (77 regions), horizontal transfer among strains of the species or from other species and the greater error associated with the macroarray data compared with the MLST data. Among the 9 B2 subgroups identified with MLST, 6 were completely recovered. Of the other 3 subgroups, 1 was completely disrupted (subgroup V) and other 2 (subgroups IV and IX) were separated in 2 clusters, but in each case one of the new cluster was homogeneous in its serotype (O2 vs. O83 in subgroup IV and O18 vs. O1/O2 in subgroup IX) (fig. 2), an observation revealing a biological meaning to the clustering pattern observed. As expected, the trees made on presence/absence of sequences were poorly

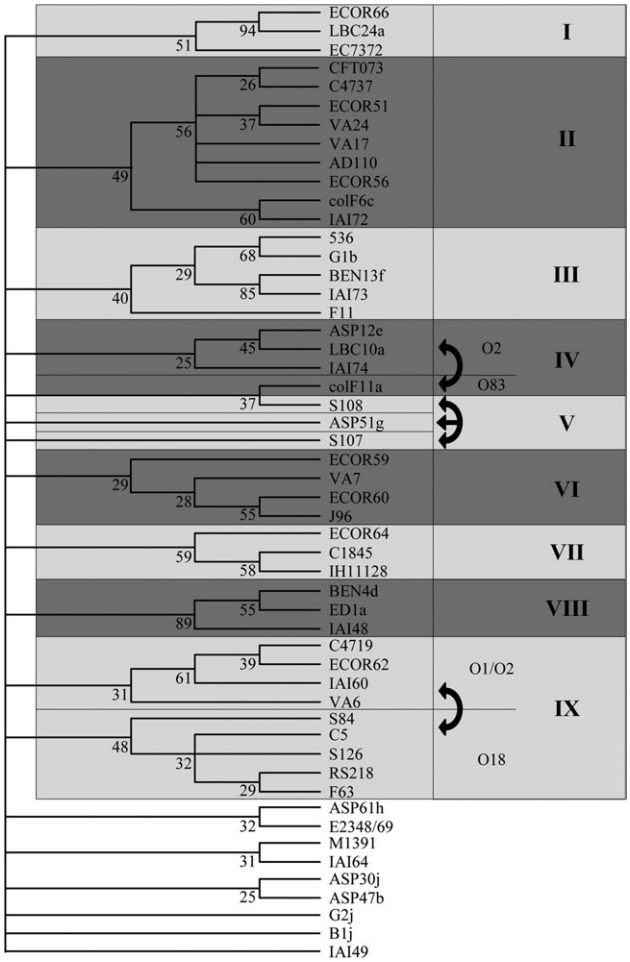

FIG. 2.—Phylogenetic tree of the 52 B2 *Escherichia coli* strains used in this study, based on presence and absence of genomic regions. This Neighbor-Joining bootstrap consensus tree is based on the simultaneous analysis of 77 genomic regions, whose presence was assigned if at least 50% of the genes composing the contig were found present according to macroarray analysis. Bootstrap values, calculated on 1,000 replicated trees, are shown if higher than 25%. The 9 subgroups identified in figure 1*A* are presented. Subgroups IV, V, and IX are not monophyletic as pointed out by arrows on the figure; however, subgroups IV and IX are split according to serotypes, O2 versus O83 for subgroup IV and O1/O2 versus O18 in subgroup IX.

correlated to the phylogeny obtained through the analysis of genomic regions or MLST as only 4 B2 subgroups were recovered (I, II, III, and VIII) (data not shown). This outcome occurs because the resulting phylogeny is largely influenced by the acquisition of few HGT events of large size. Moreover within a large region, the analysis revealed a strong dynamic of deletions (data not shown).

Second, FACs considering either the 816 sequences or the 77 genomic regions as active variables and the 9 B2 phylogenetic subgroups as illustrative variables were conducted. The projections of the variables on the 8 planes (F1/F2 to F1/F9) of the FAC conducted with the genomic region data, allowed a clear distinction of the 9 B2 subgroups (I–IX). Moreover, the projections of the strains on these planes showed that they were correctly grouped according to their B2 subgroups. The distinction of the 9 B2 subgroups obtained by the projections of the variables and the strains on the planes F1/F2 to F1/F5 with the sequences was less clear. The most clearly distinguished subgroups remained B2 II and B2 VIII on the plane F1/F2 (data not shown).

A good correlation is also observed between the phylogenetic history of the strains and the serotype as 4 B2 subgroups (III, VI, VII, and VIII) are characterized by unique O types (O6, O4, O75, and O81, respectively) (supplementary table S1, Supplementary Material online). Conversely, the PAI content (determined from the virulence gene pattern and/or the macroarray data) is highly variable within each subgroup (except subgroup III) and the same PAI content can be found in numerous subgroups (supplementary table S1, Supplementary Material online), in accordance with the already known PAIs mobility (Dobrindt et al. 2004; Escobar-Páramo, Clermont, et al. 2004).

Virulence Phenotype/Genotype Association Analysis

Overall, 75% of the strains are highly virulent (killer) in the mouse model of extraintestinal virulence, 6.6% exhibit a moderate virulence (intermediate killer), and 18.4% are avirulent (nonkiller). No difference is observed in the virulence pattern between commensal and extraintestinal pathogenic strains (75% and 80.7% of killer, 15.6% and 15.3% of intermediate, and 9.4% and 4% of nonkiller strains, respectively). The 2 diarrheic strains are not virulent. When virulence is analyzed in regard to the B2 subgroups, it appears that in the majority of the cases, both virulent and avirulent strains belong to the same subgroup (subgroups II, V, VI, VII, and IX). In 3 subgroups (I, III, and IV), all the strains are virulent, whereas all the VIII subgroup strains are not virulent.

Several FACs were conducted in order to assess global relationships between the B2 flexible gene pool (sequences or genomic regions), the PAIs, the *iroN*/*aer* plasmid, and the virulence phenotype as determined either according to the origin of the strain (commensal, intraintestinal pathogenic, and extraintestinal pathogenic) or by its intrinsic extraintestinal virulence in the mouse model (nonkiller, intermediate killer, and killer). A first FAC was conducted using the sequence (presence/absence) data set. On the plane F1/F2, which accounted for 30.43% of the total variance, the variables "presence of sequence" were grouped in an area mainly distinguished by the negative and low positive values of F1 and F2, whereas the variables "absence of sequence" were grouped in an area mainly distinguished by the positive and low negative values of F1 and F2 (fig. 3*A*). The variables "mice killer," "intermediate mouse killer," "extraintestinal pathogen origin," "presence of PAIs I_{CFT073}, II_{J96}, III_{536}, and HPI" were associated with the variables presence of sequence, whereas the variables "commensal origin," "intestinal pathogen," "nonmouse killer," and "presence of the plasmid" were associated with the variables absence of sequence. A second FAC was conducted using genomic region (presence/absence) data set. On the F1/F2 plane, which accounted for 26.77% of the total variance, the variables were clearly distinguished by the F1 values (fig. 3*B*). The variables mouse killer; intermediate mouse killer; extraintestinal pathogen origin; presence

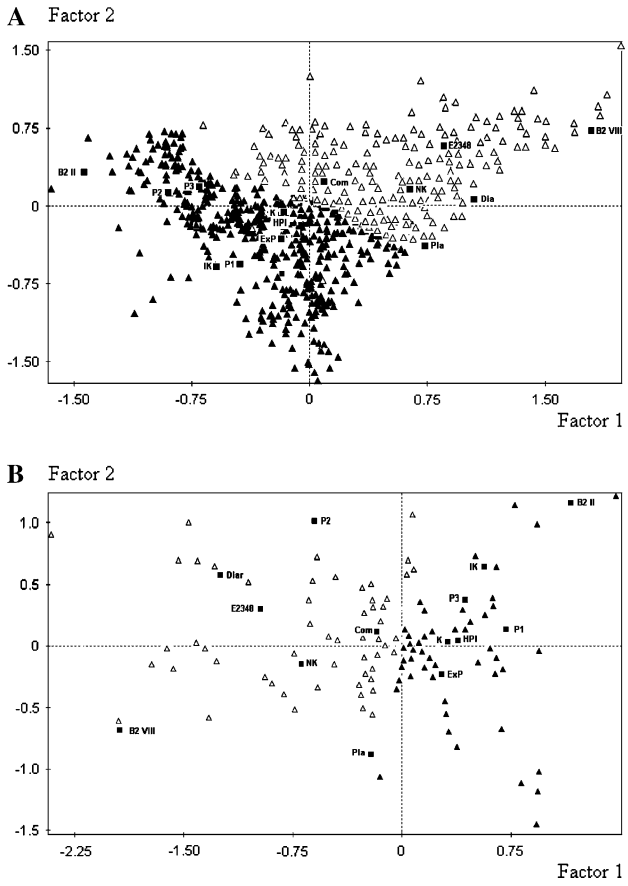

FIG. 3.—Factorial analyses of correspondence of the genetic and virulence phenotypic data of the 52 B2 *Escherichia coli* strains used in this study. (*A*) Projections of the 816 sequence variables and 14 selected other bacterial variables characterized in the 52 *E. coli* strains on the plane F1/F2 computed from the FAC. The abbreviations were as follows: K, mouse killer; IK, intermediate mouse killer; NK, mouse nonkiller; ExP, extraintestinal pathogen origin; Com, commensal origin; Dia, intestinal pathogen origin; P1, PAI I_{CFT073}; P2, PAI II_{J96}; P3, PAI III_{536}; HPI, HPI; *iroN* plasmid, Pla; B2 II, B2 subgroup II; B2 VIII, B2 subgroup VIII; E2348, EPEC E2348/69 diarrheic strain. Closed triangle, presence of sequence; Open triangle, absence of sequence. When several sequences are projected at the same point, only one is indicated. (*B*) Projections of the 77 genomic region variables and the 14 other bacterial variables characterized in the 52 *E. coli* strains on the plane F1/F2 computed from the FAC. The abbreviations were as in (*A*). Closed triangle, presence of genomic region; Open triangle, absence of genomic region. When several genomic regions are projected at the same point, only one is indicated.

of PAIs I_{CFT073}, III_{536}, and HPI; and presence of genomic region were distinguished by the F1 positive values. The variables nonmouse killer, commensal origin, intestinal pathogen origin, presence of the plasmid, and "absence of genomic region" were distinguished by the F1 negative values. Moreover, a detailed analysis of the projections of the variables mouse killer and presence of sequence or genomic region and of nonmouse killer and absence of sequence or genomic region showed some specific associations (supplementary table S4, Supplementary Material online). It emerged that genes involved in at least 5 classes of functions are always associated with virulence in the mouse model: transcriptional regulation, iron metabolism (HPI, *tonB*), adhesion (type I fimbriae), LPS biosynthesis, and the recently reported peptide polyketide hybrid synthesis system (*pks* island) (Nougayrede et al. 2006). Beside these identified functions, some genes coding for proteins of unknown functions were correlated with the mouse killer phenotype (supplementary table S4, Supplementary Material online).

Of note, we have identified a B2 subgroup (subgroup VIII) that encompasses only avirulent strains in our mouse model. This subgroup is clearly projected on the positive (fig. 3*A*) and negative (fig. 3*B*) values of the first axis with the absence of sequence and absence of genomic region characters, respectively. It is associated with the absence of genes involved in iron metabolism, adhesion, and LPS biosynthesis (supplementary fig. S3 and table S5, Supplementary Material online).

These data indicate that extraintestinal virulence is a multigenic process with different gene combinations leading to virulence. Among these genes, a common set encompassing genes involved in transcriptional regulation, iron metabolism, adhesion, LPS biosynthesis, and the recently reported peptide polyketide hybrid synthesis system can be identified.

Discussion

The selective pressures leading to the evolution and maintenance of virulence in the case of facultative pathogens are quite unclear. The mixture of phylogenetic, epidemiologic, genetic, and phenotypic data provides a framework to understand how virulence and commensalism evolved and propagated within a species. In this work, we have performed for the first time an in-depth study of the B2 phylogenetic group of *E. coli*, a group of strains involved in commensalism and in intestinal and extraintestinal pathogenesis.

The B2 Phylogenetic Group Is Well Structured

Our analysis, based on a panel of B2 phylogenetic group strains exhibiting a large genetic and ecologic diversity, reveals the existence of some genetic structure. Robust phylogenetic subgroups were clearly identified by our MLST approach (fig. 1*A*) even when recombination was taken into account using Clonal Frame (Didelot and Falush 2007). They are also retrieved by a MLST schema based on another set of genes (Johnson, Owens, et al. 2006; Wirth et al. 2006) as, for example, subgroups II and IX correspond to sequence type (ST) 73 and ST95 complexes, respectively, whereas ECOR 63 belongs to ST62 complex (Wirth et al. 2006). The interrelationships between these subgroups are unclear due to the lack of phylogenetic signal and to the effect of recombination; however, it seems that subgroup I is basal. These subgroups are also comforted by specific associations of genes from the B2 flexible gene pool, despite an evident mobility of these genes within and between subgroups. This is evidenced by the fact that the phylogenetic subgroups are better retrieved when contigs of genes rather that the genes themselves are taken into account. These data indicate a certain level of coevolution between the chromosomal background (assessed by the MLST) and the B2 flexible gene pool, a reminiscence of the link

between the chromosomal background and virulence genes that has been observed at the species level (Escobar-Páramo, Clermont, et al. 2004).

Virulence Is an Ancestral Character in the B2 Phylogenetic Group

We then wanted to elucidate if the observed genetic structure would be associated with variations in virulence. To estimate intrinsic extraintestinal virulence of strains, we used a standardized mouse model of septicemia. Even though this model does not explore the full range of extraintestinal virulence pathophysiology as the first step of dissemination from the intestine reservoir is not reproduced, it is much more informative than isolation conditions that can be highly influenced by chance and host factors. Commensal strains can generate infection in immunodeficient patients, and virulent strains can be isolated in guts of healthy subjects. This model allows a clear-cut delineation of the strains as mouse killer or nonkiller (Johnson, Clermont, et al. 2006). The majority of the B2 strains studied (more than 80%) are virulent in the mouse model. Moreover subgroup I, the basal subgroup, is highly virulent. Virulence therefore seems to be ancestral within the B2 group and widely conserved within that group, which is the more basal of the species (Lecointre et al. 1998; Escobar-Páramo, Clermont, et al. 2004; Escobar-Páramo, Sabbagh, et al. 2004). Nevertheless, within several subgroups, some avirulent strains were isolated, and more importantly, subgroup VIII appeared to be fully avirulent in the mouse model. So, the transition from virulence to avirulence seems to have occurred several times within the B2 group and at least once with some evolutionary success as a whole subgroup, encompassing widespread strains, is avirulent.

To target the genes involved in the transition from virulence to avirulence, association between the B2 flexible gene pool and virulence was performed with multiple component analysis. Each subgroup of strains had a specific combination of the tested genes, and virulence appeared as a multigenic process resulting from numerous gene combinations and multiple redundancies. These observations in association with the phylogenetic structure of the population and the dispersed loss of virulence, suggest that different independent gene losses resulted in the loss of virulence throughout the group.

The most parsimonious scenario derived from our data is that extraintestinal virulence is an ancestral character within the B2 phylogenetic group and that the avirulence is a derived one that occurred through loss of some virulence genes. Alternatively, due to the apparent propensity for virulence-related genes to undergo HGT, it could be also that, after the ancestral emergence of virulence, virulence may occasionally be lost and later reemerge via HGT.

The Maintenance of "Virulence" Is a Coincidental By-Product of Commensalism

At the exception of subgroup VIII, which completely losses virulence, the maintenance of virulence within every other groups, even though loss of virulence can be easily achieved, suggests a strong selective pressure for the maintenance of the genes involved in virulence. The genes with known functions that we found associated to virulence in our multidimensional approach are mainly genes coding for adhesins, proteins involved in iron metabolism, LPS biosynthesis, peptide polyketide hybrid synthesis system, and transcriptional regulators (supplementary fig. S3 and table S4, Supplementary Material online). The 3 first classes of genes belong to the classical so-called virulence factors, based on epidemiological (comparison of prevalence of a given gene between commensal and pathogenic strains) and animal model (knock out of the gene) data. However, several studies have now pointed out that, within the *E. coli* species, genes coding for adhesins and iron capture systems are associated to the persistence of the strains carrying them in the commensal intestinal microbiota (Levin and Svanborg Eden 1990; Wold et al. 1992; Nowrouzian et al. 2006). It has also been demonstrated that O-antigen diversity in *Salmonella*, a bacterium of the intestine that can cause disease, is a key element for surviving in the intestine from protozoan predation (Wildschutte et al. 2004) and that LPS mutant of an *E. coli* K1 strain is unable to survive in *Acanthamoeba castellanii*, at the opposite of the wild-type strain (Alsam et al. 2006). In the same line, the recently reported *pks* island that causes DNA double-strand breaks leading to cell cycle arrest and eventually cell death (Nougayrede et al. 2006) can be viewed either as a commensalism island due to its effect on the intestinal stem cells or as a pathogenic island due to its effect on the infected tissue cells (Hayashi 2006). An interesting new feature emerging from our analysis is the association of transcriptional regulator genes and virulence. These genes can also been involved in commensalism as it has been shown that increase in growth rate, a critical element for intestinal colonization, seems to be achieved by a transcriptional adjustment (Herring et al. 2006). In addition, it has been reported that *E. coli* strains belonging to B2 phylogenetic group, the one which is the most involved in extraintestinal infections (Picard et al. 1999), have superior capacity to persist in the intestinal microflora of infants, independently of carriage of all investigated virulence genes (Nowrouzian et al. 2005).

These data and our observation on the maintenance and evolution of virulence within the B2 group suggest that genes associated to virulence are implicated in complex host commensal niche colonization and that virulence is a coincidental by-product (Levin 1996). Nevertheless, if B2 group strains achieve the ability to produce intestinal disease as in the case of group 1 EPEC strains (represented by the E2348/69 strain in our study) (fig. 1 and supplementary table S1 [Supplementary Material online]), loss of extraintestinal virulence genes (fig. 3) but arrival of mobile elements leading to the intestinal virulence phenotype (diarrhea with increased transmission) ensures the evolutionary success of the strains.

What Can Explain the Structure of the B2 Phylogenetic Group?

There are up to 2% divergence at the nucleotide level between the isolates of the B2 phylogenetic group, whereas

the maximum nucleotide divergence found within the *E. coli* species as a whole is estimated to be around 3% (Escobar-Páramo, Sabbagh, et al. 2004). This extensive diversity within the B2 phylogenetic group is consistent with its basal position within the species phylogeny (Escobar-Páramo, Clermont, et al. 2004; Escobar-Páramo, Sabbagh, et al. 2004). The existence of several subgroups within the B2 phylogenetic group could result from several scenarios. No geographic structure has ever been reported in the *E. coli* species; nevertheless, the existence of several ecological niches (different host species or within host compartmentalization) could be involved. It is also possible that such subgroups appeared by historical contingency due to the small amount of recombination within *E. coli* (Desjardins et al. 1995) and/or to the rapid spread of some epidemic clones (Maynard Smith et al. 1993). When diversity among groups was observed, apart from group I that is basal, the distance between groups appeared to be very homogeneous (mean 0.0065, standard deviation [SD] 0.0020) as it reflects the lack of resolution of the phylogeny among these groups. Such a pattern could result either from several scenarios, an expansion of B2 populations (as suggested by negative Tajima's *D*, data not shown), an adaptive radiation of the B2 strains in several niches, or simply be due to recombination that effect accumulates as we go deeper into the tree.

Even if we have limited amount of strains within each group, they all have the same time-space distribution (supplementary table S1, Supplementary Material online) and some simple analyses of their diversity can provide insights on the way subgroups have originated. The different subgroups present variable within group sequence diversity on the MLST scheme. Subgroup I is clearly the most diverse (1% divergence with Kimura 2-parameter distance), an observation consistent with the basal position of that subgroup. Four subgroups (II, III, VIII, and IX) showed very little within-group diversity (less than 0.1%). Although care was taken to isolate B2 strains as diverse as possible, 2 of these subgroups (II and IX) were more represented than others. Interestingly, these 2 subgroups are also recovered as dominant ones in the Max Planck Institute MLST database (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli): among 172 B2 group strains, 46.5% and 16.8% belong to the subgroups IX (ST95) and II (ST73), respectively. Such strong frequencies could reflect a recent epidemic success of these subgroups; however, extensive diversity in serotypes and PAIs present within those subgroups argues that the origin of such subgroups is not extremely recent. This overrepresentation could also reflect some bias in the sampling as pathogenic strains tend to be more isolated and studied than commensal ones. The 2 other subgroups (III and VIII) characterized by a low sequence diversity are homogenous in their serotype, a finding that could indicate some very recent emergence. Among the other subgroups showing some consistent level of diversity, it is interesting to notice that although some are very diverse in their serotype, others are completely homogeneous (subgroups VI and VII). As we have seen that serotypes could evolve very rapidly within a group (subgroups II and IX), the serotype homogeneity within a group suggests the action of natural selection to maintain a given serotype in a given niche. Indeed, epidemiologic studies have clearly demonstrated specific associations between serotypes and diseases in man and animals (Orskov F and Orskov I 1992).

Overall, our data suggest that the clustering of the B2 population into subgroups reflects the existence of several ecological niches, but further supports from epidemiological studies will be required to confirm these hypotheses. The present molecular characterization of the subgroups, at sequence, PAI, serotype and gene content levels will be useful to identify the genetic markers needed to develop those in-depth epidemiological studies.

Concluding Remarks

In this paper, we have shown that loss of facultative virulence occurred several times through different mechanisms and at least once was associated with an evolutionary success. The apparent evolutionary success of a derived avirulence has some interesting consequences on the evolution of virulence and commensalism. Such successful transition may suggest that the B2 chromosomal backbone is good at commensalism and not just virulence, which suggests that no irreversible specialization to virulence has occurred in that genome, in contrast to other virulent *E. coli* strains like *Shigella*. This observation comforts the idea that virulence within that group of strains might be coincidental to a commensal lifestyle. It also proves that if factors that contribute to intrinsic virulence may enhance colonization of the commensal niche, they are not necessary for the B2 strain to succeed as commensal. At a much shorter time scale, we here observed, within the B2 phylogenetic group, what seems to have previously occurred in the *E. coli* species, that is, the emergence of extraintestinal avirulent B1 and A phylogenetic groups from ancestral extraintestinal virulent B2 and D phylogenetic groups.

Supplementary Material

Supplementary tables S1–S5 and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Roland Quentin, Stéphane Bonacorsi, and Edouard Bingen for providing some of the strains used in this study and to David Gordon and Eric Oswald for critical reading of the manuscript. This work was partially funded by the "Fondation pour la Recherche Médicale."

Literature Cited

Alsam S, Jeong SR, Sissons J, Dudley R, Kim KS, Khan NA. 2006. *Escherichia coli* interactions with *Acanthamoeba*: a symbiosis with environmental and clinical implications. J Med Microbiol. 55:689–694.

Bingen E, Picard B, Brahimi N, Mathy S, Desjardins P, Elion J, Denamur E. 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, Clermont O, Bonacorsi S, Terki M, Brahimi N, Loukil C, Barraud D, Bingen E. 2002. Phylogenetic analysis and prevalence of urosepsis strains of *Escherichia coli* bearing pathogenicity island-like domains. Infect Immun. 70:3216–3226.

Bonacorsi S, Clermont O, Houdouin V, Cordevant C, Brahimi N, Marecat A, Tinsley C, Nassif X, Lange M, Bingen E. 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.

Bonacorsi SP, Clermont O, Tinsley C, Le Gall I, Beaudoin JC, Elion J, Nassif X, Bingen E. 2000. Identification of regions of the *Escherichia coli* chromosome specific for neonatal meningitis-associated strains. Infect Immun. 68:2096–2101.

Boyd EF, Hartl DL. 1998. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J Bacteriol. 180:1159–1165.

Cherifi A, Contrepois M, Picard B, Goullet P, Orskov I, Orskov F, De Rycke J. 1991. Clonal relationships among *Escherichia coli* serogroup 06 isolates from human and animal infections. FEMS Microbiol Lett. 64:225–230.

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

Clermont O, Johnson JR, Menard M, Denamur E. 2007. Determination of *Escherichia coli* O types by allele-specific polymerase chain reaction: application to the O types involved in human septicemia. Diagn Microbiol Infect Dis. 57:129–136.

Desjardins P, Picard B, Kaltenbock B, Elion J, Denamur E. 1995. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. J Mol Evol. 41:440–448.

Diard M, Baeriswyl S, Clermont C, Gouriou S, Picard B, Taddei F, Denamur E, Matic I. 2007. *Caenorhabditis elegans* as a simple model to study phenotypic and genetic virulence determinants of extraintestinal pathogenic *Escherichia coli*. Microbes Infect. 9:214–223.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics. 175:1251–1266.

Dobrindt U, Emody L, Gentschev I, Goebel W, Hacker J. 2002. Efficient expression of the alpha-haemolysin determinant in the uropathogenic *Escherichia coli* strain 536 requires the leuX-encoded tRNA(5)(Leu). Mol Genet Genomics. 267:370–379.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2:414–424.

Donnenberg. 2002. *Escherichia coli*. Virulence mechanisms of a versatile pathogen. Baltimore (MD): Academic press, Elsevier Science.

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

Escobar-Páramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur E. 2004. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. Mol Biol Evol. 21:1085–1094.

Escobar-Páramo P, Grenet K, Le Menac'h A, et al. (12 co-authors). 2004. Large-scale population structure of human commensal *Escherichia coli* isolates. Appl Environ Microbiol. 70:5698–5700.

Escobar-Páramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, Skurnik D, Denamur E. 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. Environ Microbiol. 8:1975–1984.

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

Gerdes SY, Scholle MD, Campbell JW, et al. (21 co-authors). 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J Bacteriol. 185:5673–5684.

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

Greenacre M. 1992. Correspondence analysis in medical research. Stat Methods Med Res. 1:97–117.

Hayashi T. 2006. Microbiology. Breaking the barrier between commensalism and pathogenicity. Science. 313:772–773.

Herring CD, Raghunathan A, Honisch C, et al. (11 co-authors). 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet. 38:1406–1412.

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

Johnson JR, Clermont O, Menard M, Kuskowski MA, Picard B, Denamur E. 2006. Experimental mouse lethality of *Escherichia coli* isolates, in relation to accessory traits, phylogenetic group, and ecological source. J Infect Dis. 194:1141–1150.

Johnson JR, Delavari P, Kuskowski M, Stell AL. 2001. Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*. J Infect Dis. 183:78–88.

Johnson JR, Owens KL, Clabots CR, Weissman SJ, Cannon SB. 2006. Phylogenetic relationships among clonal groups of extraintestinal pathogenic *Escherichia coli* as assessed by multi-locus sequence analysis. Microbes Infect. 8:1702–1713.

Kim CC, Joyce EA, Chan K, Falkow S. 2002. Improved analytical methods for microarray-based genome-composition analysis. Genome Biol. 3:RESEARCH0065.

Kumar S, Tamura K, Jakobsen IB, Nei M. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics. 17:1244–1245.

Lawrence JG, Ochman H, Hartl DL. 1991. Molecular and evolutionary relationships among enteric bacteria. J Gen Microbiol. 137(Pt 8):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.

Levin BR. 1996. The evolution and maintenance of virulence in microparasites. Emerg Infect Dis. 2:93–102.

Levin BR, Svanborg Eden C. 1990. Selection and evolution of virulence in bacteria: an ecumenical excursion and modest suggestion. Parasitology. 100(Suppl):S103–S115.

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? Proc Natl Acad Sci USA. 90:4384–4388.

Nougayrede JP, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, Buchrieser C, Hacker J, Dobrindt U, Oswald E. 2006. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science. 313: 848–851.

Nowrouzian FL, Adlerberth I, Wold AE. 2006. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. Microbes Infect. 8:834–840.

Nowrouzian FL, Wold AE, Adlerberth I. 2005. *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. J Infect Dis. 191:1078–1083.

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

Ochman H, Selander RK. 1984. Standard reference strains of *Escherichia coli* from natural populations. J Bacteriol. 157: 690–693.

Orskov F, Orskov I. 1992. *Escherichia coli* serotyping and disease in man and animals. Can J Microbiol. 38:699–704.

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.

Press WH. 1992. Numerical recipes in C: the art of scientific computing. Cambridge: Cambridge University Press.

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

Selander RK, Levin BR. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science. 210:545–547.

Watt S, Lanotte P, Mereghetti L, Moulin-Schouleur M, Picard B, Quentin R. 2003. *Escherichia coli* strains from pregnant women and neonates: intraspecies genetic distribution and prevalence of virulence factors. J Clin Microbiol. 41: 1929–1935.

Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG. 2004. Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonella*. Proc Natl Acad Sci USA. 101:10644–10649.

Wirth T, Falush D, Lan R, et al. (11 co-authors). 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 60:1136–1151.

Wold AE, Caugant DA, Lidin-Janson G, de Man P, Svanborg C. 1992. Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics. J Infect Dis. 165: 46–52.

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

William Martin, Associate Editor

Accepted August 9, 2007