This is to certify that the
dissertation entitled

COMPARATIVE MOLECULAR EVOLUTIONARY ANALYSIS
OF VIRULENCE LOCI IN PATHOGENIC *ESCHERICHIA COLI*

presented by

David William Lacher

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Genetics

Thomas S. Whittam
Major Professor's Signature

April 16, 2007
Date

COMPARATIVE MOLECULAR EVOLUTIONARY ANALYSIS OF VIRULENCE LOCI IN PATHOGENIC *ESCHERICHIA COLI*

By

David William Lacher

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Graduate Program in Genetics

2007

ABSTRACT

COMPARATIVE MOLECULAR EVOLUTIONARY ANALYSIS OF VIRULENCE LOCI IN PATHOGENIC *ESCHERICHIA COLI*

By

David William Lacher

Escherichia coli is a diverse species of Gram-negative bacteria, some strains of which are pathogenic. The early stages of *E. coli* pathogenesis often involve bacterial attachment mediated by the expression of surface proteins. It has been hypothesized that pathogens alter their surface proteins in order to evade detection by their host's immune system. Therefore, it is likely that natural selection is acting to generate new allelic variants. It is the goal of this research to examine the allelic diversity of genes that encode a variety of surface structures in different classes of pathogenic *E. coli* (pathotypes). The specific aims are to: 1) develop a method to quickly and accurately subtype a highly polymorphic locus responsible for the hallmark phenotype of the attaching and effacing *E. coli*, 2) characterize enteropathogenic *E. coli* (EPEC) through multilocus sequence typing (MLST) and restriction fragment length polymorphism (RFLP) analyses, 3) assess the level of genetic polymorphism in a region of the operon encoding the type 1 fimbriae of *E. coli*, and 4) examine various genes encoding surface structures for the actions of positive selection and recombination.

To address specific aim 1, a new method to quickly and accurately type the *eae* locus was developed. This new technique addresses the limitations of existing typing schemes and was applied to a set of *E. coli* capable of the attaching and effacing phenotype. For specific aim 2, a system was designed to detect and identify the alleles of three EPEC virulence genes. The distribution of these virulence gene alleles was assessed in a collection of strains representing a variety of EPEC serotypes and then compared to a phylogenetic framework generated from MLST analysis of conserved housekeeping loci. To address specific aim 3, a segment of DNA encompassing a regulatory region of the type 1 fimbrial operon was sequenced in various types of pathogenic *E. coli*. This region contains an invertible genetic element responsible for the phase variability of the type 1 fimbriae and has been shown to be inactive in some strains. For specific aim 4, a collection of allelic sequences was assembled for genes encoding five different surface proteins from several *E. coli* pathotypes. These genes were analyzed for evidence of positive selection and homologous recombination. The results of this work will give us a better understanding of how different types of pathogenic *E. coli* have evolved and may have important public health implications.

ACKNOWLEDGEMENTS

I wish to thank my mentor, Thomas S. Whittam for his guidance and support. I thank my committee members, Michael Bagdasarian, Richard Lenski, and Vincent Young for their suggestions and guidance.

I would also like to thank my fellow lab members by providing suggestions and support. I thank Galeb Abu-Ali, Teresa Bergholz, Alyssa Bumbaugh, Susan Francisco, Sara Kienzle, Shannon Manning, Adam Nelson, Suzanne Nelson, Lindsey Ouellette, Weihong Qi, James, Riordan, Hans Steinsland, Cheryl Tarr, Sivapriya Kailasan Vanajaa, Seth Walk, and Lukas Wick.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

LITERATURE REVIEW

Escherichia coli is a diverse species of Gram-negative bacteria. Most strains are nonpathogenic and do not harm their host, but some may cause a variety of harmful intestinal and extra-intestinal infections. A common theme in the pathogenesis of the different types of *E. coli* is bacterial attachment mediated by the expression of surface proteins. A wide range of surface proteins are expressed by these strains, some of which are ubiquitous while others are specific to certain pathogenic types (pathotypes). It has been hypothesized that these surface proteins must alter their three-dimensional structure or surface epitopes in order to confer an evolutionary advantage and evade detection by antibodies generated from previous antigen exposure. Therefore, it is likely that natural selection drives repeated amino acid replacements in surface proteins by acting upon the gene or genes that encode them.

Natural selection and recombination in gene evolution. Natural selection is the evolutionary mechanism by which the relative genotype frequencies in a population change according to their relative fitness. There are two types of natural selection, positive and negative. These types of selection can be detected at the DNA sequence level by comparing the rates of synonymous (silent) and nonsynonymous (amino acid changing) nucleotide substitutions. According to the neutral theory of molecular evolution (68), synonymous and nonsynonymous substitutions should accumulate at approximately the same rate in the absence of selection.

Most coding genes show an excess of synonymous substitutions (19). This is indicative of negative selection acting to maintain the existing function and structure of the encoded protein. Under this type of selective pressure, mutants have a lower fitness than their parental genotype(s) and their frequency decreases in the following generations. Positive selection, in contrast, favors amino acid changes and is characterized by a greater rate of nonsynonymous than synonymous substitution. Under positive selection, a newly produced mutant has a higher fitness than the average in the population, and its frequency increases in the following generations.

Methods have been developed for estimating the numbers of synonymous and nonsynonymous substitutions (22, 44, 57, 86, 87, 99, 103). These methods calculate the substitution rate for an entire gene by computing the average number of substitutions over a particular length of codons. A limitation of these methods is that they could potentially fail to identify genes under positive selection if high nonsynonymous substitution rates occur at only a few codons. More recently, methods have been developed to detect positive selection at single amino acid sites. These approaches have been classified into three categories: counting methods, fixed effects models, and random effects models (73).

Counting methods reconstruct the ancestral sequences and then estimate the number of nonsynonymous and synonymous changes that have occurred at each codon throughout the evolutionary history of the data set (106, 107, 142, 143). Fixed effects models estimate the ratio of nonsynonymous to synonymous substitutions on a site-by-site basis (142, 159). Like counting methods, fixed effects models make no assumption regarding the distribution of substitution rates across sites. Random effects models fit a distribution of substitution rates across sites and then infer the rate at which individual sites evolve (51, 108, 158).

Another mechanism that affects gene evolution is homologous recombination. In this process, genetic information is exchanged between related segments of DNA.

Numerous methods have been developed to test for the occurrence and boundaries of recombinational events. Drouin et al. (33) have classified these techniques into four general categories: similarity, phylogenetic, compatibility, and nucleotide substitution distribution methods.

Similarity methods infer recombination when synonymous substitutions at variable genes (or regions of genes) exceed those at conserved genes (98, 110). Phylogenetic methods infer recombination when the phylogenies from different parts of a genome or gene result in conflicting topologies (50, 91, 154). Recombination can also be detected under these methods when orthologous genes do not reflect the evolutionary relationships of their species. Compatibility methods test for phylogenetic incongruence in a site-by-site basis using only informative sites but do not require a phylogeny for the sequences being analyzed to be known in advance (58, 59). A site is said to be parsimoniously informative when there are at least two different nucleotides, each represented at least twice, at that position in the set of sequences under investigation. These informative sites are defined as compatible when their evolutionary histories are congruent with a single tree topology. Recombination is detected when a region of incompatible sites is found. In the final category, nucleotide substitution distribution, sequences are analyzed for a random distribution of substitutions along the sequences (130, 139). Recombination is inferred if a significant clustering of substitutions is detected.

Attaching and effacing *E. coli*. Strains of attaching and effacing *E. coli* (AEEC) are capable of intimately attaching to intestinal epithelial cells (38, 62). Once the bacterium attaches to the host cell, it manipulates the host cytoskeleton to form pedestal structures beneath the bacterium and, in the process, effaces the microvilli of the intestinal mucosa (21, 49). This phenotype, termed attaching/effacement, is due to the locus of enterocyte effacement (LEE), a ~35-kb pathogenicity island that is inserted into the chromosome of AEEC (36, 93, 120). The LEE island encodes approximately 40 genes, including the components of a type III secretion system (TTSS) (36, 120). The TTSS acts as a molecular syringe to inject bacterial effector proteins, many of which are also LEE-encoded, into the host cell.

A main component of the LEE-encoded TTSS is EspA, which enables initial attachment of the bacterium to the host cell (21) and acts as a conduit for bacterial effector proteins to enter the enterocyte (25, 67, 71, 134). Since these filaments are exposed to the host immune system, they may experience selection pressures that could lead to an increase in *espA* polymorphism. Another important locus on the LEE island is *eae*, a highly polymorphic gene that encodes that intimin adhesin. Intimin protein plays a crucial role in the attaching/effacing phenotype and studies have indicated that different groups of pathogenic *E. coli* possess different *eae* alleles. This has been used to help classify strains into pathogenic types, so a reliable intimin typing scheme is necessary.

Enteropathogenic *E. coli*. Enteropathogenic *E. coli* (EPEC) infections are a leading cause of infantile diarrhea in developing nations (63, 82, 83). EPEC are capable of attaching/effacement and are therefore members of the AEEC pathotype. Typical EPEC strains are differentiated from other types of pathogenic *E. coli* by their ability to form microcolonies on the surface of epithelial cells (6, 100, 141). This phenotype, termed localized adherence, is due to the presence of a large (~70 kb) virulence plasmid called the EPEC adherence factor (EAF) plasmid (29). The EAF plasmid encodes the bundle-forming pilus (BFP), a member of the type IV family of fimbriae. An operon of 14 genes is necessary for expression of the BFP (31, 140), and the major structural subunit, termed bundlin, is encoded by *bfpA*. Another important locus on the EAF plasmid is the plasmid-encoded regulator (Per) consisting of three genes (*perA*, *perB*, and *perC*). PerA shows homology to members of the AraC family of transcriptional activators (45), whereas PerB and PerC show no significant homology to any known prokaryotic proteins (45) and their exact role in EPEC pathogenesis is still under investigation. Most typical EPEC strains fall into one of two phylogenetically distinct groups, designated EPEC 1 and EPEC 2 (156). Little is known about the allelic distributions of *bfpA*, *perA*, and *eae* among EPEC 1, EPEC 2, and other clonal lineages of EPEC.

Shiga toxin-producing *E. coli*. Strains of Shiga toxin-producing *E. coli* (STEC) are defined by their ability to produce one or more variants of Shiga toxin. These potent cytotoxins inhibit protein synthesis, resulting in cell death, and are usually encoded by bacteriophages. The main clinical manifestations of STEC infection are hemorrhagic colitis (HC) and hemolytic-uremic syndrome (HUS). HC is a distinctive gastrointestinal illness characterized by severe abdominal pain with cramps, watery diarrhea followed by grossly bloody diarrhea, and little or no fever (125). In some cases, infection with STEC can progress to HUS, which is characterized by acute renal failure, decreased platelet count, and hemolytic anemia (75).

Enterohemorrhagic *E. coli* (EHEC) belong to both the AEEC and STEC pathotypes. These strains are capable of attaching/effacement, express one or more Shiga toxins, cause HC and HUS, and possess a ~60-MDa plasmid (82, 83). Like EPEC, most

EHEC strains fall into one of two phylogenetically distinct groups, designated EHEC 1 and EHEC 2 (124). In many parts of the world, the EHEC serotype most frequently associated with severe disease is O157:H7 (64).

O157:H7 and type 1 fimbriae. Type 1 fimbriae are filamentous structures composed primarily of the structural subunit FimA (69)and are expressed on the surface of most clinical *E. coli* isolates (5). These fimbriae bind to mannose-containing receptors on epithelial cells (109), have been shown to be required for colonization of the urinary tract (23, 53, 65, 132), and may play a role in the colonization of the intestinal tract (76, 77). *E. coli* O157:H7 is unusual in that does not express type 1 fimbriae. Genetic analysis has revealed that these strains possess a 16-bp deletion within the invertible genetic element (known as the *fim* switch) that controls fimbrial expression (85). A study by Roe et al. (127) found that the deletion was responsible for the lack of type 1 fimbrial expression in O157:H7. Their results demonstrated that the *fim* switch was permanently locked in the "off" orientation so that transcription of *fimA* did not occur. It is unknown if this deletion is unique to O157:H7 or if it is also present in other closely related strains. Another area of underdeveloped research concerns polymorphism in *fimA*. Extensive *fimA* sequence variation has been described for strains isolated from a wide range of animal hosts (119), but a similar analysis in a diverse set of human isolates has yet to be performed. Recombination also appears to play a role in *fimA* allelic diversification. Evidence for horizontal transfer of multiple *fimA* alleles has been reported in a set of closely related extraintestinal strains (155), but it remains unclear if recombination is responsible for *fimA* sequence variation in other classes of strains.

Curli fimbriae. *E. coli* express thin, coiled surface structures, designated curli, that mediate bacterial binding to a variety of extracellular matrix and serum proteins (113, 114, 136). Research also suggests that curli may play a role in the development of biofilms on inert surfaces (121, 152). Curli fibers are encoded by two divergently transcribed chromosomal operons, *csgBAC* and *csgDEFG*. The assembly of the fibers is unique and involves extracellular self-assembly of the curlin subunit (CsgA), dependent on a specific nucleator protein (CsgB) (47, 48). The *csgD* gene encodes the transcriptional regulator of curli production (47) and Uhlich et al (148, 149) identified mutations within the promoter region of *csgD* responsible for enhanced curli expression. However, an extensive study of the genetic polymorphism in *csgA* has not been performed. Natural selection could be acting upon *csgA* to generate allelic variants that form more effective biofilms, thereby enhancing bacterial survival in the external environment.

Purpose. The primary objective of this research is to examine the allelic diversity of genes that encode a variety of surface proteins in different classes of pathogenic *E. coli*. This diversity will be uncovered primarily through direct sequencing of the genes under investigation. Once an allelic database has been assembled, a more rapid and cost-effective typing method will be assessed for some of the loci by examining additional isolates to determine the level of variation that can be resolved by digestion with one or more restriction enzymes.

The observed allelic variation in the surface protein-encoding genes will be compared to a phylogenetic framework generated from polymorphisms within conserved housekeeping genes. The questions to be addressed are: How does the level of sequence polymorphism observed within the surface protein-encoding genes compare to that seen within the conserved housekeeping genes? Do the sequence data indicate that these two categories of loci are experiencing similar or different selective pressures? Is there evidence of recombination both within the surface protein-encoding genes and between these loci and the chromosomal backbone represented by the housekeeping loci? It is expected that the answers to these questions will provide insights into the evolution of pathogenic *E. coli* at the level of individual genes as well as at the strain level.

CHAPTER 2

ALLELIC SUBTYPING OF THE INTIMIN LOCUS (*EAE*) OF PATHOGENIC *ESCHERICHIA COLI* BY FLUORESCENT RFLP

SUMMARY

Intimin is a highly polymorphic protein encoded by the *eae* gene and plays a crucial role in the attaching-effacing phenotype of diarrheagenic *Escherichia coli* and related pathogens. A method to quickly and accurately uncover allelic variation at the *eae* locus was developed through the use of fluorescent RFLP (fRFLP). Application of fRFLP to 151 *eae*-positive strains (including the newly described *Escherichia albertii*) revealed 26 different fRFLP types that correspond to 20 of the 28 previously described *eae* alleles. Two sequence variants of the γ, ι, κ, and ζ alleles and three variants of ε were also observed. In addition to being reliable and accurate, the method can be easily adapted to accommodate new *eae* allelic sequences, as they become known.

INTRODUCTION

Strains of attaching and effacing *E. coli* (AEEC) are capable of intimately attaching to intestinal epithelial cells (38, 62). Once the bacteria attach to the host cells, they manipulate the host cytoskeleton to form pedestal structures beneath the bacterial cells and, in the process, efface the microvilli of the intestinal mucosa (21, 49). This process creates a characteristic intestinal histopathology, termed attaching effacing (AE) lesions. The ability of pathogenic *E. coli* to form AE lesions is encoded in a pathogenicity island referred to as the locus of enterocyte effacement or LEE (93, 94). The LEE island is ~35-kb in length and can be inserted into one of several chromosomal locations in AEEC clonal groups (36, 120, 137, 157). The LEE island comprises ~40 genes including those that encode the structural components of a type III secretion system (TTSS) (36, 120). The TTSS acts as a molecular delivery system to translocate bacterial effector proteins, many of which are also LEE-encoded, into the host cell (26).

Intimin plays a crucial role in the AE phenotype and is encoded by *eae*, one of the most highly polymorphic genes of the LEE island (20). Intimin is composed of six domains: a periplasmic domain, a transmembrane domain, three extracellular immunoglobulin-like domains, and an extracellular lectin-like domain (89). Much of the transmembrane domain is homologous to the invasins of pathogenic *Yersinia* and this part of the molecule has been termed the central conserved domain (95). The extracellular domains of intimin interact with cellular receptors (135) including Tir, the translocated intimin receptor that is encoded in LEE and moves to the eukaryotic cell via the LEE encoded TTSS (67). An analysis of 27 intimin alleles from GenBank indicates that the four C-terminal extracellular domains contain more than 75% of the nucleotide variation present within those sequences (unpublished data). The AE strains represent a variety of enteric pathotypes of both humans and animals, including both typical and atypical enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), *Escherichia albertii*, and *Citrobacter rodentium* (28, 153). Among the AEEC, there is a striking association of different *eae* alleles with specific clones or clonal groups of pathogenic *E. coli* (2, 144, 157). This association of *eae* alleles with pathogenic lineages has been used to help classify strains into pathotypes, so a rapid and reliable intimin typing scheme is a valuable tool.

Here a method to identify allelic variants of the *eae* locus through the use of fluorescent RFLP (fRFLP) (80) is described. In this method, the entire highly variable 3' half of *eae* is amplified in a standard PCR reaction using primers that are located in the conserved central domain of *eae* and in the conserved downstream gene *escD*. The exact size of the amplicon depends on the specific *eae* allele, but is typically about 2 kb in length. The PCR amplicon is then digested with one or more restriction enzymes that leave a 5' overhang that acts as a template for the incorporation of a fluorescent dye-terminator nucleotide from a standard cycle sequencing kit. The multiple labeled restriction fragments are then separated on a capillary-based sequencer and their sizes estimated to within a few base pairs in length. This fRFLP system provides a rapid method for uncovering allelic variation in *eae* and for classifying *eae* subtypes based on complex restriction digests.

MATERIALS AND METHODS

Bacterial strains and DNA isolation. The strains in this study included 144 AEEC representing 38 O-serogroups, many of which include EPEC serotypes, and 7 strains from the *Escherichia albertii* and *Shigella boydii* 13 clonal lineage (55, 56). All 151 AE strains were grown overnight at 37°C in 10 ml of LB broth with moderate shaking. Genomic DNA was isolated using the Puregene DNA isolation kit (Gentra Systems Inc., Minneapolis, MN). DNA concentrations were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Rockland, DE). Working template concentrations of 50 ng/μl were used for PCR.

PCR and *eae-escD* primer design. PCR primers were designed in the central conserved domain of *eae* (eae-F1: 5'-ACT CCG ATT CCT CTG GTG AC-3') and the conserved downstream gene *escD* (escD-R1: 5'-GTA TCA ACA TCT CCC GCC CA-3') based on available sequences for the α, β, γ, ε, ζ, and θ intimin alleles. The eae-F1 and escD-R1 primers are located at positions 25986-26005 and 27918-27937, respectively, of the complete LEE sequence from strain E2348/69 (GenBank accession number AF022236). Each 25-μl reaction contained 2.5 μl 10X buffer II (Applied Biosystems, Foster City, CA), 2.5 μl 2 mM dNTP, 2.0 μl 25 mM $MgCl_2$, 0.5 μl 10 μM eae-F1 primer, 0.5 μl 10 μM escD-R1 primer, 1.5 units AmpliTaq Gold (Applied Biosystems), 1 μl 50 ng/μl genomic DNA template, and 15.7 μl ddH_2O. Amplification utilized an initial denaturing step at 94°C for 10 min., followed by 35 cycles of 92°C for 1 min., 55°C for 1 min., and 72°C for 2 min. A final step of 72°C for 5 min. was used for final completion of any partially extended product. PCR products (5 μl) were visualized on ethidium bromide-stained 0.8% agarose gels by illumination with UV light.

fRFLP development. *In silico* digestions were performed to find suitable restriction endonucleases for fRFLP. To be considered for fRFLP, the restriction enzymes had to 1) leave a 5' overhang, 2) mainly produce fragments in the range of the labeled size standard (60 to 640 bp), and 3) had to produce fragments that could be labeled with fluorescent tags that were different than that of the size standard. Out of 532 restriction enzymes tested *in silico*, *Mse*I was chosen for fRFLP because it meets all of the above criteria and was found to have the greatest power to discriminate known *eae* alleles (Figures 1 and 2, Table 1).

To confirm the *eae* allele assignments based on the *Mse*I findings, a second restriction digestion was designed. Multiple enzymes were needed to give a similar amount of pattern variation as seen with *Mse*I. For simplicity, restriction enzymes were selected that use the same reaction conditions (reaction buffer, incubation temperature, and labeled ddNTP). Three sets of triple-enzyme digests were found. A combination of *Ase*I, *Dde*I, and *Sal*I was chosen over the others based on lower enzyme cost and the ddNTP used is the same as that for the *Mse*I digest (Table 2). The two digests (*Mse*I and *Ase*I / *Dde*I / *Sal*I) were used to subtype the intimin alleles from a diverse set of 151 AE strains.

PCR clean-up and restriction enzyme digestion. Samples that were positive for the *eae-escD* PCR were treated with ExoSAP-IT (USB Corporation, Cleveland, OH) to remove unincorporated dNTPs and PCR primers (5 µl of PCR product and 2 µl of ExoSAP-IT). All restriction endonucleases, reaction buffers, and bovine serum albumin (BSA) solutions were obtained from New England Biolabs Inc. (Ipswich, MA). *Mse*I digests were set up so that each 30-µl reaction contained 7.0 µl ExoSAP-treated PCR

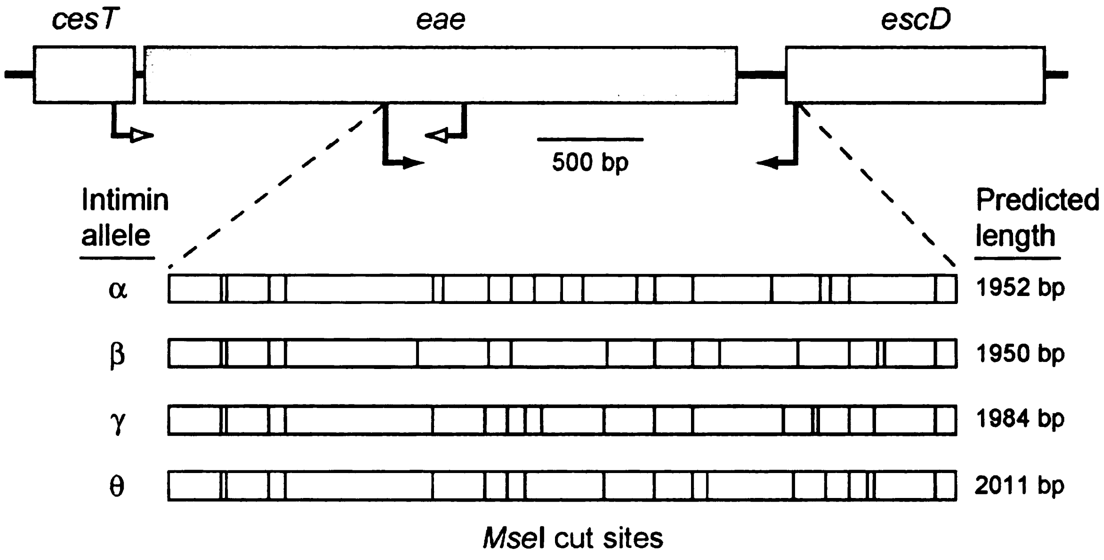

Figure 1. Location of PCR primers and predicted *Mse*I restriction sites for 4 major intimin alleles associated with human disease. The forward primer (eae-F1) is located in the central conserved domain of *eae* and the reverse primer (escD-R1) is located in the conserved downstream gene *escD*. The resulting amplicon is ~1800-2100 bp depending on the allele.

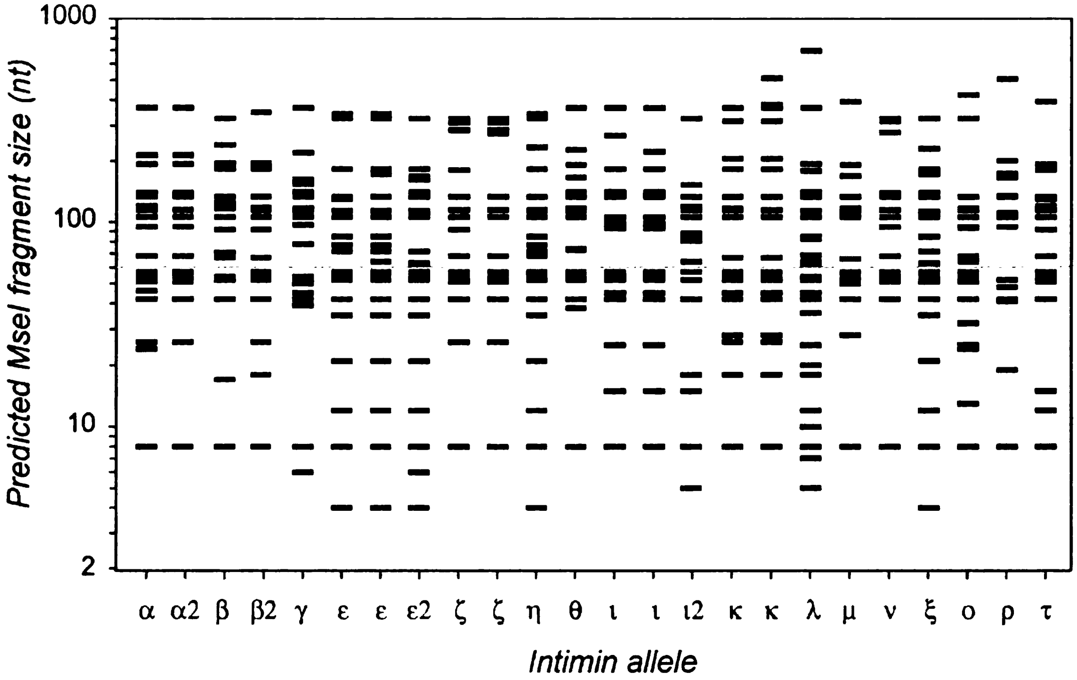

Figure 2. Fragment sizes of DNA resulting from *Mse*I digestion of the *eae-escD* amplicon. The number of restriction fragments ranges from 16 to 22 with an average of 19.1 fragments per amplicon. Two additional patterns were resolved for lane 5 (γ1.1 and γ1.2) and lane 7 (ε1.2 and ε1.3) by *Ase*I / *Dde*I / *Sal*I restriction digestions. The gray line denotes the 60 nt length that is the lower limit of fragment size determined by fRFLP.

Table 1. *Mse*I digestion of *eae-escD* PCR product.

fRFLP pattern	Amplicon size (bp)	Predicted fragment sizes (nt)[a]
α 1.1	1952	68, 95, 106, 115, 120, 134, 139, 193, 214, 366
α 2.1	1952	68, 95, 106, 115, 134, 135, 139, 193, 214, 366
β 1.1	1950	67, 71, 92, 106, 117, 123, 125, 134, 183, 195, 240, 324
β 2.1	1940	67, 92, 106, 115, 117, 118, 134, 183, 190, 195, 348
γ 1.1[b]	1984	78, 97, 106, 114, 117, 134, 141, 154, 162, 219, 366
γ 1.2[b]	1984	78, 97, 106, 114, 117, 134, 141, 154, 162, 219, 366
ε 1.1	2036	72, 77, 85, 106, 114, 130, 134, 183, 183, 324, 343
ε 1.2[b]	2142	64, 72, 77, 85, 106, 114, 134, 172, 183, 183, 324, 343
ε 1.3[b]	2142	64, 72, 77, 85, 106, 114, 134, 172, 183, 183, 324, 343
ε 2.1	2137	61, 63, 72, 106, 113, 114, 134, 141, 162, 163, 168, 183, 324
ζ 1.1	1946	68, 92, 106, 115, 134, 181, 284, 310, 324
ζ 1.2	1946	68, 106, 115, 134, 273, 284, 310, 324
η 1.1	2140	68, 72, 77, 85, 106, 114, 115, 134, 183, 234, 324, 343
θ 1.1	2011	73, 74, 106, 114, 117, 134, 141, 166, 192, 227, 366
ι 1.1	1824	93, 95, 99, 106, 134, 141, 183, 267, 366
ι 1.2	1824	93, 95, 99, 106, 134, 141, 183, 222, 366
ι 2.1	1831	64, 81, 84, 88, 106, 115, 119, 134, 153, 324, 324
κ 1.1	1939	67, 106, 115, 117, 134, 183, 206, 315, 366
κ 1.2	2716	67, 106, 115, 134, 183, 206, 315, 366, 382, 512
λ 1.1	2665	62, 64, 69, 83, 85, 106, 114, 134, 141, 179, 194, 366
μ 1.1	2006	66, 106, 106, 111, 114, 115, 117, 134, 169, 192, 392
ν 1.1	1944	68, 68, 95, 106, 115, 134, 139, 276, 313, 324
ξ 1.1	2141	63, 72, 85, 106, 113, 134, 140, 172, 183, 183, 230, 324
ο 1.1	2020	64, 68, 68, 94, 95, 106, 115, 117, 134, 324, 423
ρ 1.1	2013	95, 106, 111, 134, 135, 166, 174, 174, 201, 507
τ 1.1	1942	68, 92, 106, 115, 120, 120, 130, 134, 181, 193, 392

[a] Fragments outside the range of the size standard (<60 nt or >640 nt) are not shown.

[b] Variants that are indistinguishable by fRFLP with the *Mse*I digest.

Table 2. *Ase*I / *Dde*I / *Sal*I digestion of *eae-escD* PCR product.

fRFLP pattern	Amplicon size (bp)	Predicted fragment sizes (nt)[a]
α 1.1	1952	106, 110, 121, 138, 213, 221, 278, 285, 428
α 2.1	1952	106, 110, 138, 221, 278, 285, 334, 428
β 1.1	1950	110, 114, 138, 213, 384, 408, 583
β 2.1	1940	106, 110, 114, 138, 164, 204, 213, 244, 260, 278
γ 1.1	1984	106, 110, 138, 213, 246, 278
γ 1.2	1984	106, 110, 138, 198, 246, 278
ε 1.1	2036	60, 106, 110, 138, 141, 141, 142, 213, 244, 267, 384
ε 1.2	2142	60, 110, 138, 141, 141, 142, 212, 213, 244, 267, 384
ε 1.3	2142	60, 110, 138, 141, 212, 213, 244, 267, 283, 384
ε 2.1	2137	106, 110, 123, 138, 144, 213, 244, 278, 334, 348
ζ 1.1[b]	1946	106, 110, 115, 138, 198, 237, 278
ζ 1.2[b]	1946	106, 110, 115, 138, 198, 237, 278
η 1.1	2140	110, 138, 141, 151, 183, 244, 262, 267, 597
θ 1.1	2011	106, 110, 120, 138, 222, 278, 357
ι 1.1[b]	1824	65, 106, 110, 138, 213, 238, 240, 278, 344
ι 1.2[b]	1824	65, 106, 110, 138, 213, 238, 240, 278, 344
ι 2.1	1831	73, 91, 110, 138, 239, 246, 285, 597
κ 1.1	1939	85, 106, 110, 138, 171, 198, 213, 278, 278, 310
κ 1.2	2716	85, 106, 110, 138, 171, 174, 198, 213, 278, 278, 330, 583
λ 1.1	2665	100, 106, 109, 110, 131, 138, 198, 278, 325, 481, 584
μ 1.1	2006	106, 110, 120, 138, 144, 206, 213, 278, 639
ν 1.1	1944	110, 121, 138, 582
ξ 1.1	2141	110, 138, 141, 151, 192, 213, 244, 263, 267, 384
ο 1.1	2020	85, 110, 136, 138, 158, 200, 278, 304, 611
ρ 1.1	2013	106, 110, 138, 165, 196, 218, 278
τ 1.1	1942	110, 138, 213, 384, 473, 624

[a] Fragments outside the range of the size standard (<60 nt or >640 nt) are not shown.

[b] Variants that are indistinguishable by fRFLP with the *Ase*I / *Dde*I / *Sal*I digest.

product, 0.3 μl 100X BSA, 0.5 μl 10 U/μl *Mse*I, 3.0 μl 10X NEB2 buffer, and 19.2 μl ddH$_2$O. *Ase*I / *Dde*I / *Sal*I digests were set up so that each 30-μl reaction contained 7.0 μl ExoSAP-treated PCR product, 0.3 μl 100X BSA, 0.5 μl 10 U/μl *Ase*I, 0.5 μl 10 U/μl *Dde*I, 0.25 μl 20 U/μl *Sal*I, 3.0 μl 10X NEB3 buffer, and 18.45 μl ddH$_2$O. Reactions were incubated overnight at 37°C.

fRFLP. fRFLP was performed using the CEQ DTCS standard kit (Beckman Coulter Inc, Fullerton, CA). Each reaction contained 2.0 μl of unpurified restriction enzyme digest, 1.5 μl 10X reaction buffer, 0.1 μl ddUTP, 0.1 μl *Taq* DNA polymerase, and 11.3 μl ddH$_2$O. Samples were incubated at 60°C for 1 hour, purified with Sephadex G-50 Fine columns (Amersham Pharmacia Biotech Inc., Piscataway, NJ), dried under vacuum centrifugation (Savant Instruments Inc., Holbrook, NY), and suspended in 10 μl of deionized formamide. Of this, 2 μl were mixed with 0.6 μl CEQ DNA Size Standard 600 (Beckman Coulter Inc., Fullerton, CA), and 39.4 μl deionized formamide, and run on a CEQ2000XL (Beckman Coulter Inc.) using a capillary temperature of 50°C, a denature step at 90°C for 2 min., injection at 2.0 kV for 30 sec., and separation at 4.8 kV for 65.0 min. (*Mse*I digest) or 90 min. (*Ase*I / *Dde*I / *Sal*I digest). Fragment sizes were determined with the CEQ2000XL software, version 4.3.9.

DNA sequencing. The complete *eae* gene was sequenced in at least one representative strain for each fRFLP pattern identified. The 5' half of *eae* was amplified using primers cesT-F9 (5'-TCA GGG AAT AAC ATT AGA AA-3') and eae-R3 (5'-TCT TGT GCG CTT TGG CTT-3') using the same PCR conditions described above. PCR products were purified using the QIAquick PCR purification kit (QIAGEN Inc., Valencia, CA) and quantified with a NanoDrop ND-1000 spectrophotometer. Cycle sequencing reactions contained 6.0 µl CEQ DTCS Quick Start premix (Beckman Coulter Inc.), 1.5 µl 20 µM primer, approximately 180 ng of *cesT/eae* product or 250 ng of *eae/escD* product, and ddH_2O to 15 µl. Amplification utilized an initial denaturing step at 94°C for 1 min., followed by 35 cycles of 96°C for 30 sec., 52°C for 30 sec., and 60°C for 2 min. Upon completion of cycle sequencing, samples were purified with Sephadex G-50 Fine columns, dried under vacuum centrifugation, suspended in 40 µl of deionized formamide, and sequenced using a Beckman CEQ2000XL DNA sequencer. Samples were analyzed using the CEQ2000XL software and then exported for further analysis with the SeqMan and MegAlign modules of the Lasergene software (DNASTAR Inc., Madison, WI). Internal sequencing primers were designed as new sequence data were generated.

RESULTS

The ability of the *eae* fRFLP method to resolve intimin alleles was first tested on 53 strains based on the expected patterns from the *in silico* digestion of available *eae* sequences from GenBank. This initial study included 34 AE strains, most of which had previously been known to have the α, β, or γ intimin alleles, most often found in strains associated with human infection (Table 3). Nineteen additional AE strains were selected and tested by the fRFLP method because their intimin allele had been determined by other researchers or by previous work in the Whittam laboratory based on DNA sequencing or conventional RFLP (Table 4). The intimin alleles of all 53 strains were confirmed by both the *Mse*I and *Ase*I / *Dde*I / *Sal*I digests.

To further evaluate the new method, 98 *eae*-positive strains for which no allelic subtyping data existed were tested. These strains were originally recovered from two separate populations: a pediatric population in Seattle, Washington (27) and a cohort study of childhood diarrheal disease in Guinea-Bissau, West Africa (150). Among these 98 *eae*-positive strains, most strains (85%) exhibited known digestion patterns with both *Mse*I and the triple digest, and therefore could be easily subtyped to an allele. The 15 strains for which the *eae* allele could not be determined had one of seven previously unobserved digestion patterns. A representative of each pattern was sequenced and the alleles were either previously described or variants of previously described alleles (ε1.3, ε2.1, κ1.2, λ1.1, ν1.1, ρ1.1, and ξ1.1).

Among the 151 strains examined, there were a total of 24 different fRFLP patterns observed for the *Mse*I digests (Table 1, Figure 2). The *Ase*I / *Dde*I / *Sal*I digest also resolved 24 distinct fRFLP patterns (Table 2). For two alleles, ε1.2/ε1.3 and

Table 3. Allelic variation in *eae* based on fRFLP.

Accession number	Strain name	Serotype[a]	*eae* fRFLP pattern	Source and pathotype
TW00588	DEC 1a	O55:H6	α 1.1	Human EPEC
TW07884	E851/71	O142:H6	α 1.1	Human EPEC
TW07923	RN587/1	O157:[h45]	α 1.1	Human EPEC
TW04262	TB269C	O145:[h34]	α 2.1	Human atypical EPEC
TW01120	B170	O111:[h2]	β 1.1	Human EPEC
TW05355	13180-25	O111:H11	β 1.1	EHEC from food
TW00148	3448-87	O114:H2	β 1.1	Human EPEC
TW00389	29315	O119:H2	β 1.1	Human EPEC
TW07099	LT119-80	O119:H2	β 1.1	Human EPEC
TW01266	C342-62	O126:H2	β 1.1	Human EPEC
TW07896	E56/54	O128:H2	β 1.1	Human EPEC
TW01664	DEC 10i	O145:H16	β 1.1	Human EHEC
TW05149	BCL73	O145:[h-]	β 1.1	Bovine STEC
TW07860	314-S	O145:[h16]	β 1.1	Bovine STEC
TW08894	02-3422	O145:[h2]	β 1.1	Rabbit EPEC
TW09153	IH 16	O145:[h-]	β 1.1	Human STEC
TW07924	Z188-93	O110:H6	β 2.1	Avian EPEC
TW01225	1396/69	O119:H6	β 2.1	Human EPEC
TW03293	ECOR-37	O-:[h7]	γ 1.1	Marmoset atypical EPEC
TW00947	DEC 5d	O55:H7	γ 1.1	Human atypical EPEC
TW03064	B6820-C1	O145:[h28]	γ 1.1	Bovine STEC
TW07596	GS G5578620	O145:[h28]	γ 1.1	Human STEC
TW07865	IHIT0304	O145:H28	γ 1.1	Bovine STEC
TW08087	MT#66	O145:[h28]	γ 1.1	Human STEC
TW09356	4865/96	O145:[h28]	γ 1.1	Human STEC
TW08881	3556-77	Boydii 13	γ 1.2	Human atypical B13
TW08887	3557-77	Boydii 13	γ 1.2	Human atypical B13
TW08889	3053-94	Boydii 13	γ 1.2	Human atypical B13
TW07618	98ST607	O110:H28	ζ 1.1	Human STEC
TW00964	75-83	O145:[h25]	ζ 1.1	Human STEC
TW05307	LTO55-43	O55:H7	θ 1.1	Human atypical EPEC
TW07960	DA-34	O103:[h25]	θ 1.1	Human STEC
TW00970	DEC 8b	O111:H8	θ 1.1	Human EHEC
TW07888	010-311082	O76:H51	μ 1.1	Human EPEC

[a] Lower case H-types in brackets were inferred from *fliC* allele.

Table 4. Reference strains for defined intimin alleles and patterns determined by fRFLP and confirmed by DNA sequencing.

Pattern no.	Intimin allele	*eae* fRFLP pattern	Reference strain	Species and serotype[a]
1	α	α 1.1	TW06375 (E2348/69)	*E. coli* O127:H6
2	α2	α 2.1	TW01270 (C712-65)	*E. coli* O125:H6
3	β	β 1.1	TW07862 (413/89-1)	*E. coli* O26:[h11]
4	β2	β 2.1	TW07894 (0659-79)	*E. coli* O119:H6
5	γ	γ 1.1	TW08264 (Sakai)	*E. coli* O157:H7
6	γ	γ 1.2	TW08888 (3092-94)	*S. boydii* type 13
7	ε	ε 1.1	TW08101 (MT#80)	*E. coli* O103:H2
8	ε	ε 1.2	TW08023 (MT#2)	*E. coli* O121:H19
9[b]	ε	ε 1.3	TW10363 (83F4)	*E. coli* O-:[h8]
10[b]	ε2	ε 2.1	TW10371 (98B3)	*E. coli* O116:[h9]
11	ζ	ζ 1.1	TW07863 (537/89)	*E. coli* O84:[h2]
12	ζ	ζ 1.2	TW04892 (921)	*E. coli* O111:H9
13	η	η 1.1	TW07892 (012-050982)	*E. coli* O142:[h21]
14	θ	θ 1.1	TW01387 (CL-37)	*E. coli* O111:H8
15	ι	ι 1.1	TW01933 (1252-59)	*E. coli* O55:[h34]
16	ι	ι 1.2	TW04174 (TB227C)	*E. coli* O86:[h8]
17	ι2	ι 2.1	TW08839 (C-425)	*S. boydii* type 13
18	κ	κ 1.1	TW06584 (C295-53)	*E. coli* O86:H34
19[b]	κ	κ 1.2	TW10337 (64B4)	*E. coli* O49:[h10]
20[b]	λ	λ 1.1	TW10327 (57A1)	*E. coli* O33:[h34]
21	μ	μ 1.1	TW08260 (MA551/1)	*E. coli* O55:[h51]
22[b]	ν	ν 1.1	TW10376 (106A5)	*E. albertii*
23[b]	ξ	ξ 1.1	TW10334 (60A3)	*E. coli* O5:[h2]
24	ο	ο 1.1	TW07627 (Albert 19982)	*E. albertii*
25[b]	ρ	ρ 1.1	TW10366 (93I4)	*E. coli* O21:[h5]
26	τ	τ 1.1	TW08933 (K-1)	*S. boydii* type 7

[a] Lower case H-types in brackets were inferred from *fliC* allele.

[b] fRFLP pattern discovered in AEEC strains from Guinea-Bissau.

γ1.2/γ1.3, variants are resolved by the *Ase*I / *Dde*I / *Sal*I digests and are indistinguishable with *Mse*I digestion. In contrast, two variants that are resolved by *Mse*I (ι1.1/ι1.2 and ζ1.1/ζ1.2) are indistinguishable with the triple digest. Combined there are 26 distinct fRFLP patterns (Table 4). In all, the new fRFLP method is able to identify 20 of the 28 previously described alleles, and differentiated two new variants of the γ, ι, κ, and ζ alleles, and three new variants of ε.

To gauge the accuracy of the fragment size estimation, the expected fragment sizes based on *in silico* digestions were compared to those observed from the capillary sequencer. Overall the fragment scoring was accurate with most fragments over 100 nt in length less than 2% different in size from their expected values (Figure 3). Examination of the plot reveals that for fragments less than ~100 nt and greater than ~550 nt, the estimated fragments size deviates from the expected (Figure 3). For the 83 fragments observed in the *Mse*I digest, the average deviation from expected size is 1.76%. The triple digest was more accurate with an average deviation of 1.01% across the 70 distinct fragments.

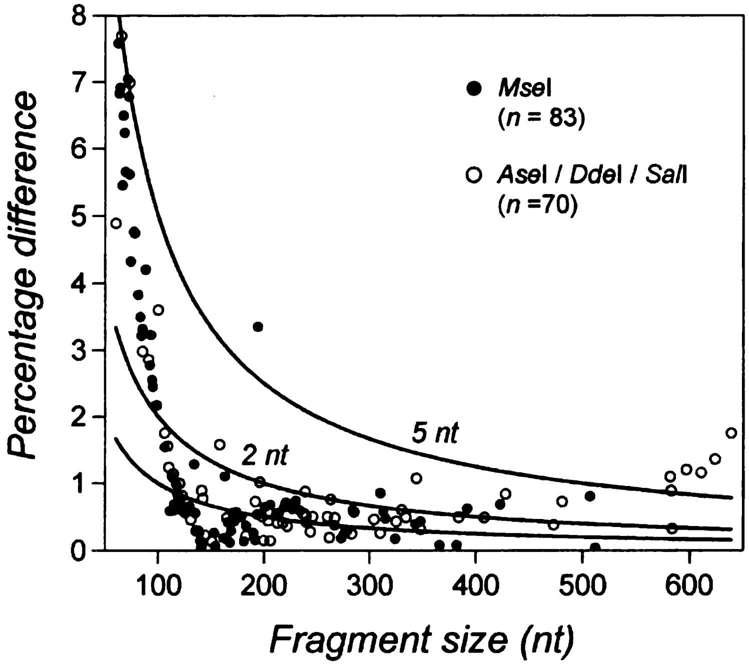

Figure 3. Accuracy of fragments sizes estimated by fRFLP. The percentage observed difference from the expected fragment size is plotted against the expected size. The lines mark the percentage deviation for 1, 2, and 5 nt respectively.

DISCUSSION

Current *eae* typing schemes either focus on allele-specific PCR amplification (2, 9, 61, 115, 118, 123, 160) or conventional RFLP analysis (60, 118, 122, 133). Sequence analysis of intimin alleles has revealed that many *eae* alleles are mosaics with segments having different evolutionary histories (95, 144). Therefore, allele-specific PCR amplification can lead to erroneous typing results. For example, the μ allele would be erroneously typed as γ by the Reid method (123) because the γ allele primer is located in a region that is shared between these two alleles (data not shown). In addition, as new alleles are discovered, new primers are necessary to amplify these alleles and the subtyping scheme becomes more complex.

Conventional RFLP analysis also has its limitations. The 5' half of *eae* is relatively conserved among the different alleles, so there may not be sufficient variation in the amplicons to accurately and reliably differentiate the alleles. For example, EPEC strain 1396/69 (O119:H6) was typed by Jenkins et al. (60) as possessing the γ allele of *eae* instead of the β2 allele. *In silico* analysis revealed that the β2, γ, and λ alleles of *eae* are virtually indistinguishable under the Jenkins system. Another limitation of conventional RFLP is that a system based on the highly variable 3' half of the gene may be difficult to score since small differences in the banding patterns of different alleles may not be easily discernible under standard electrophoretic conditions.

The fRFLP method described here addresses many of the limitations of the existing *eae* typing methods. A drawback of the devised method, however, is that it requires the location of *escD* relative to *eae* to be conserved. If this location changes such that *escD* is either upstream or far downstream of *eae*, PCR amplification will not occur and the strain will be nontypeable by this method. However, most of the known alleles for *eae* have been amplified and observed, so this situation does not appear to be common. The new subtyping method has been tested against a diverse panel of 87 *eae*-positive isolates originally recovered from children in West Africa and also uncovered seven additional variants. The remaining alleles (β3, ε3, ε4, η2, π, σ) still need to be tested. Another limitation is that some of the *eae* alleles in GenBank do not contain their associated downstream *escD* sequence, so the expected amplicon size and fRFLP profile cannot be fully determined in all cases. When new fRFLP patterns are observed, *eae* needs to be completely sequenced to verify the allele, but then the new pattern can be added to the fRFLP database for future reference.

ACKNOWLEDGEMENTS

I would like to thank Lindsey Ouellette, Dr. Weihong Qi, and Dr. Hans Steinsland for technical assistance. This work appears in the August 2006 issue of FEMS Microbiology Letters (261:80-87). This project has been funded in part with Federal funds from the NIAID, NIH, DHHS, under NIH Research Contract # N01-AI-30058.

CHAPTER 3

MOLECULAR EVOLUTION OF TYPICAL ENTEROPATHOGENIC *ESCHERICHIA COLI*

SUMMARY

Enteropathogenic *Escherichia coli* (EPEC) infections are a leading cause of infantile diarrhea in developing nations. Typical EPEC are differentiated from other types of pathogenic *E. coli* by two distinctive phenotypes – attaching effacement and localized adherence. The genes specifying these phenotypes are found on the locus of enterocyte effacement (LEE) and the EPEC adherence factor (EAF) plasmid. To describe how typical EPEC have evolved, a diverse collection of strains was characterized by performing multilocus sequence typing (MLST) and restriction fragment length polymorphism (RFLP) analysis of three virulence genes (*eae*, *bfpA*, and *perA*) to assess allelic variation. Among 129 strains representing 20 O-serogroups, 21 clonal genotypes were identified using MLST. RFLP analysis resolved 9 *eae*, 9 *bfpA*, and 4 *perA* alleles. Each *bfpA* allele was associated with only one *perA* allele class, suggesting that recombination has not played a large role in shuffling the *bfpA* and *perA* loci between separate EAF plasmids. The distribution of *eae* alleles among typical EPEC strains is more concordant with the clonal relationships than the distribution of the EAF plasmid types. These results provide further support for the hypothesis that the EPEC pathotype has evolved multiple times within *E. coli* through separate acquisitions of the LEE island and EAF plasmid.

INTRODUCTION

Enteropathogenic *E. coli* (EPEC) infections are a leading cause of infantile diarrhea in developing nations (63, 83). A key characteristic of EPEC strains is the ability to intimately attach to intestinal epithelial cells and create attaching and effacing (AE) lesions (38). The AE phenotype is specified by genes of the locus of enterocyte effacement (LEE), a ~35-kb pathogenicity island located in the bacterial chromosome (36, 93). The LEE island comprises approximately 40 genes and encodes the components of a type III secretion system, various effector molecules, and the intimin adhesin (36, 66, 151). Intimin plays a crucial role in AE lesion formation (21) and is encoded by the highly polymorphic *eae* gene (2, 160). To date, more than 25 major allelic variants of *eae* have been described (79).

Most typical EPEC strains fall into one of two phylogenetically distinct groups or clonal lineages, designated EPEC 1 and EPEC 2 (156), and differ from atypical EPEC and other types of pathogenic *E. coli* by their ability to form microcolonies on the surface of intestinal epithelial cells (6). This phenotype, termed localized adherence (LA), correlates with the presence of a large virulence plasmid called the EPEC adherence factor (EAF) plasmid (29). The EAF plasmids from different EPEC strains show considerable variation in size (~70 to 110 kb) (15, 102, 145) and, presumably, gene content. Comparison of the complete EAF plasmid sequences from two prototypical EPEC strains (O127:H6 EPEC 1 strain E2348/69 and O111:NM EPEC 2 strain B171) indicates that the EPEC 2 plasmid of B171 carries fewer genes (80 vs. 115 open reading frames) and a greater percentage of intact or partial insertion sequence elements (33% vs. 19%) than the pMAR7 plasmid of EPEC 1 strain E2348/69 (15, 145). Nevertheless, certain parts of the plasmid show a high degree of sequence conservation among typical EPEC strains (101), particularly in the region encoding the bundle-forming pilus (BFP), a type IV fimbria whose production is associated with the LA phenotype. An operon of 14 genes is necessary for expression of the BFP (31, 140), with *bfpA* encoding the major structural subunit (bundlin). Sequence comparisons of 9 *bfpA* alleles have provided compelling evidence for the action of positive selection at the molecular level (10, 11). A second locus on the EAF plasmid implicated in the full virulence of EPEC is the plasmid-encoded regulator (Per), consisting of three genes (*perA*, *perB*, and *perC*). Per has been shown to activate genes within the *bfp* operon (146) and the LEE pathogenicity island (35, 96).

Little is known about the allelic distributions of *eae*, *bfpA*, and *perA* among the EPEC 1, EPEC 2, and other clonal lineages of typical EPEC. In this study a diverse collection of 129 EPEC, including strains of the classical EPEC serotypes (117), was characterized through multilocus sequence typing (MLST) and restriction fragment polymorphism (RFLP) analysis to elucidate the extent to which horizontal transfer of the LEE island and EAF plasmid have contributed to the evolution and diversification of EPEC clones.

MATERIALS & METHODS

Strains. A collection of 95 EPEC strains was assembled based on serotype or their inclusion in one of two studies examining *bfpA* allelic variation (10, 11). These strains represent a variety of serotypes originally isolated between 1947 and 1998 from different regions around the world and were obtained from a number of sources, including the Centers for Disease Control and Prevention, Dr. Alejandro Cravioto, Prof. Helge Karch, Drs. Frits and Ida Ørskov, Dr. Phillip I. Tarr, and Dr. Luis Trabulsi (Table 5). An additional 34 *eae*$^+$ *bfpA*$^+$ strains were selected from a cohort study of childhood diarrheal disease in Guinea-Bissau, West Africa (150) (Table 5). Each strain was grown overnight at 37°C in 10 ml of Luria-Bertani (LB) broth with moderate shaking. Genomic DNA was isolated using the Puregene DNA isolation kit (Gentra Systems Inc., Minneapolis, MN). DNA concentrations were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Rockland, DE), which were diluted to 25 ng/μl for PCR.

MLST. Multilocus sequence typing (MLST) was performed on 7 conserved housekeeping genes (*aspC*, *clpX*, *fadD*, *icdA*, *lysP*, *mdh*, and *uidA*). A detailed protocol of the MLST procedure, including allelic type and sequence type (ST) assignment methods, can be found at the EcMLST website (http://www.shigatox.net/mlst). Sequences were concatenated for phylogenetic analyses.

***eae* fRFLP.** Allelic variation in *eae* was resolved using fluorescent RFLP (fRFLP) as described previously (79).

***bfpA* PCR and DNA sequencing.** PCR primers were designed to target conserved flanking regions of *bfpA* based on 9 available allelic sequences (10, 11). Each

Table 5. Summary of EPEC strains investigated.

Serogroup	# of isolates	Year(s) of isolation	Locale(s)
Classical			
O55	21	1947-1998	Brazil, Congo, Dutch Guiana, France, Germany, Guinea-Bissau, Mexico, Scotland, USA
O86	10	1950-1997	Brazil, Bulgaria, Denmark, Germany, Guinea-Bissau, UK, USA
O111	16	1947-1996	Austria, Brazil, Dutch Guiana, Germany, Mexico, Peru, Scotland, UK, USA
O114	4	1969-1997	Guinea-Bissau, UK, USA
O119	28	1960-1998	Brazil, Chile, Guinea-Bissau, Mexico, Peru, UK, USA
O127	4	1969-1997	Guinea-Bissau, UK
O142	11	<1960-1997	Brazil, Canada, Guinea-Bissau, Indonesia, Peru, Portugal, Scotland, USA
Other			
O2	1	1997	Guinea-Bissau
O33	2	1997	Guinea-Bissau
O34	1	1997	Guinea-Bissau
O49	2	1997	Guinea-Bissau
O51	1	1997	Guinea-Bissau
O73	1	1997	Guinea-Bissau
O76	1	1982	Peru
O110	1	1993	Germany
O126	3	1962-1964	Egypt, Iran, Pakistan
O128	15	1953-1991	Denmark, Germany, Pakistan, UK, USA
O157	3	1983-1998	Brazil, USA
OX9	1	1997	Guinea-Bissau
O-	3	1996-1997	Guinea-Bissau

25-µl reaction contained 2.5 µl 10X buffer II (Applied Biosystems, Foster City, CA), 2.5 µl 2 mM dNTP, 2.0 µl 25 mM $MgCl_2$, 0.5 µl 10 µM bfpA_-52F primer (5'-AGA TTA TTC CGT GAC CTA TT-3'), 0.5 µl 10 µM bfpG_9R primer (5'-TGT CCT CAC ATA TAC CTC CC-3'), 1.5 U AmpliTaq Gold (Applied Biosystems), 1 µl 25 ng/µl genomic DNA template, and 15.7 µl ddH_2O. Amplification of the approximately 700-bp fragment utilized an initial denaturing step at 94°C for 10 min, followed by 35 cycles of 92°C for 1 min, 52°C for 1 min, and 72°C for 30 s. A final step of 72°C for 5 min was used to complete any partially extended product. PCR products (5 µl) were visualized on ethidium bromide-stained 1.5% agarose gels by illumination with UV light, purified using the QIAquick PCR purification kit (QIAGEN Inc., Valencia, CA), and quantified. Cycle sequencing reactions contained 4.0 µl CEQ DTCS Quick Start premix (Beckman Coulter Inc., Fullerton, CA), 1.0 µl 20 µM bfpA_-52F or bfpG_9R primer, approximately 70 ng of *bfpA* PCR product, and ddH_2O to a final volume of 10 µl. Amplification utilized an initial denaturing step at 94°C for 1 min, followed by 35 cycles of 96°C for 30 s, 52°C for 30 s, and 60°C for 2 min. Upon completion of cycle sequencing, samples were purified with Sephadex G-50 Fine columns (Amersham Pharmacia Biotech Inc., Piscataway, NJ), dried under vacuum centrifugation (Savant Instruments Inc., Holbrook, NY), suspended in 40 µl of deionized formamide, and run on a CEQ2000XL (Beckman Coulter Inc.). Samples were analyzed using the CEQ2000XL software and then exported for further analysis with the SeqMan module of the Lasergene software (DNASTAR Inc., Madison, WI).

***bfpA* PCR and RFLP.** PCR conditions were identical to those described above for *bfpA* except primers bfpA_114F (5'-GTC TGC GTC TGA TTC CAA TA-3') and bfpA_521R (5'-TCA GCA GGA GTA ATA GC-3') were used to amplify a 408-414 bp internal fragment of the gene. Prior to digestion, each *bfpA* PCR product was purified using the QIAquick PCR purification kit. Three different restriction enzyme digests were used. Digestion with *Alu*I and with *Bfa*I was performed in separate 30-µl reactions with 10 U of enzyme, 3.0 µl 10X reaction buffer, and 26.0 µl purified PCR product; and samples were incubated overnight at 37°C, while digestion with 10 U of *Bst*NI was performed in 30-µl reactions with 3.0 µl 10X reaction buffer, 0.3 µl 100X BSA, and 25.7 µl purified PCR product followed by an overnight incubation at 60°C. All restriction enzymes were obtained from New England BioLabs Inc. (Ipswich, MA) and the reaction buffer provided with each enzyme was used. After incubation, 15 µl of the digests were visualized on ethidium bromide-stained 1.5% agarose gels by illumination with UV light.

***perA* PCR and DNA sequencing.** Primers were designed to target the conserved flanking and internal regions of *perA* based on 15 available sequences (45, 111, 146). PCR conditions are similar to those described above for *bfpA* except primers perA_-24F (5'-AAC AAA CGC GCA TGA AGG TG-3') and perB_222R (5'-TTC GCC GGT GAT GTG GTC T-3'), a 58°C annealing temperature, and a 1 min extension time were used. The resulting PCR products (approximately 1.1 kb) were purified and quantified as described above. Cycle sequencing reactions were similar to those used for *bfpA*, except that 120 ng of *perA* PCR product; primers perA_-24F, perA_539F (5'-AAA ACT GGA AAC TAG GCG ATG TCA-3'), perA_562R (5'-TGA CAT CGC CTA GTT TCC AGT TTT-3'), and perB_222R; and a 58°C annealing temperature were used.

***perA* PCR and RFLP.** PCR conditions were similar to those described for *bfpA* except primers perA_-24F and perA_562R were used at an annealing temperature of 58°C. Digestion with *Dde*I and with *Sau*96I was performed in separate 30-μl reactions with 10 U of enzyme, 3.0 μl 10X reaction buffer, 20.0 μl unpurified PCR product, and ddH_2O to volume; samples were incubated overnight at 37°C, and visualized on ethidium bromide-stained 1.5% agarose gels.

***fliC* typing.** Strains that were nonmotile or lacked flagellar serotype data were typed for the *fliC* locus. The entire *fliC* gene was amplified with primers fliC_1F (5'-ATG GCA CAA GTC ATT AAT ACC AA-3') and fliC_1497R (5'-TTA ACC CTG CAG CAG AGA CA-3') using the same PCR conditions described for *bfpA* except for an annealing temperature of 55°C and an extension time of 2 min. Amplicons (approximately 2 kb) were either digested with 5 U of *Dde*I under conditions similar to that described for *perA* or sequenced to determine the allele. H-types that were determined by *fliC* sequencing or RFLP are denoted with a lower case 'h' and are enclosed in square brackets.

Phylogenetic analyses. Sequences were aligned with the ClustalW algorithm using the MegAlign module of the Lasergene software. Neighbor-joining trees were constructed using the Kimura 2-parameter model of nucleotide substitution with the MEGA3 software (78) and the inferred phylogenies were each tested with 500 bootstrap replications. Phylogenetic network analysis was conducted with the SplitsTree 4 (54) program using the neighbor-net algorithm (18) and untransformed distances (*p* distance). The ϕ_w recombination test (17) as implemented by SplitsTree 4 was used to distinguish recurrent mutation from recombination in generating genotypic diversity.

RESULTS

MLST analysis. PCR amplification and sequencing of the 7 MLST loci in 129 EPEC strains was successful in most (90%) cases. The notable exception was *uidA*, which failed to amplify in 13 strains, including a single O114:H2 (380/69) strain and 12 strains 11 of which are serotype O55:[h51]. The 12 *uidA*-negative strains were identical to each other at the 6 remaining MLST loci. PCR amplification with primers located in the genes flanking the *uidA* locus produced a truncated amplicon suggesting that these strains belong to a clonal genotype that has lost most of the *uidA* gene (data not shown). For phylogenetic analysis, the sequenced internal fragments of the 7 housekeeping genes were concatenated to yield 3,732 nucleotides. The *uidA* locus was treated as missing data and replaced with alignment gaps in the fully concatenated sequence for the 13 *uidA*-negative strains.

MLST analysis resolved an average of 25.4 variable nucleotide sites per locus, which defined a number of alleles, ranging from 7 to 12, at the 7 MLST genes (Table 6). The distinct combinations of alleles across the 7 MLST loci were used to define 21 multilocus genotypes or sequence types (STs) among the 129 EPEC strains. Classification of the strains based on the bootstrap analysis indicates that most (77%) of the strains belong to one of four main clonal groups, designated EPEC 1, EPEC 2, EPEC 3, and EPEC 4 (Figure 4). With the exception of EPEC 3 strains, which were all O86:H34 (or nonmotile relatives), the EPEC groups based on the classification of STs included strains of various O types. There were strains representing 3 O-types in EPEC 1 (O55, O127, and O142), 5 O-types in EPEC 2 (O111, O114, O119, O126, and O128), and 2 O-types in EPEC 4 (O110 and O119). H-types (or the inferred H-type from the

Table 6. Sequence variation among alleles of 7 MLST genes from EPEC strains.

Locus	# of sites	# of variable sites	# of alleles
aspC	513	21	9
clpX	567	34	11
fadD	483	32	11
icdA	567	30	8
lysP	477	10	8
mdh	549	25	12
uidA	576	26	7
Average	533.1	25.4	9.4

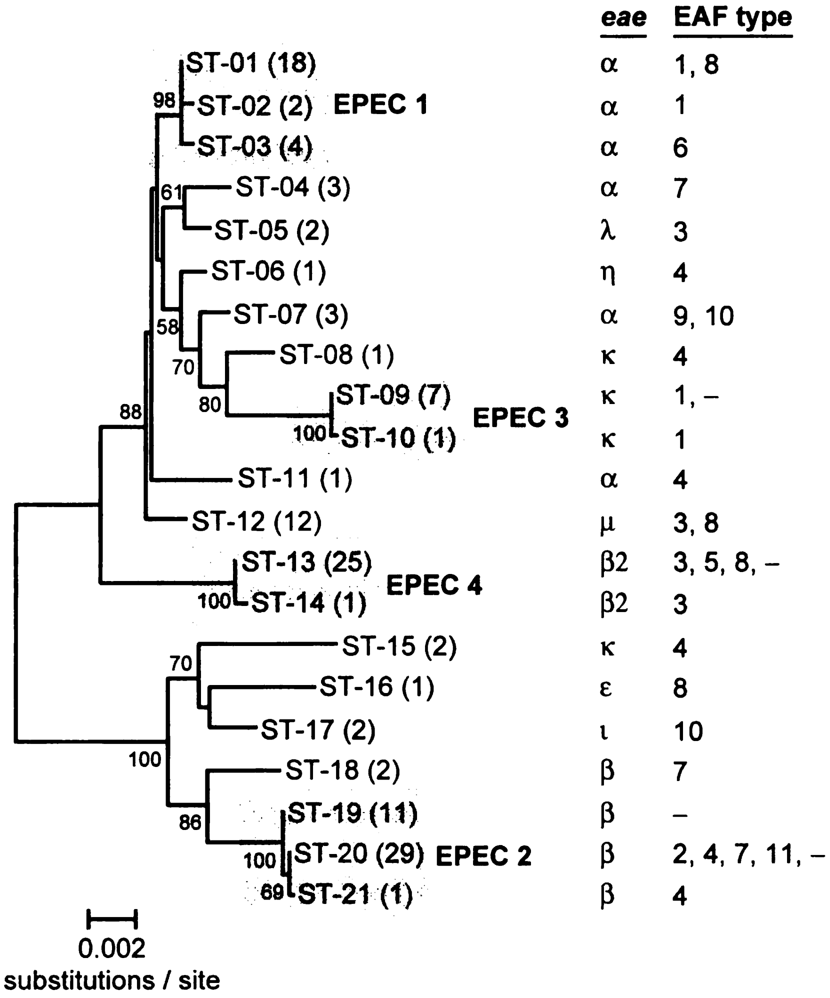

Figure 4. Phylogenetic relationships of 21 EPEC sequence types. An unrooted phylogenetic tree was constructed by the neighbor-joining algorithm based on the Kimura 2-parameter model of nucleotide substitution. The four main clonal groups are indicated by gray boxes. The sequence type (ST) and number of isolates are given at the branch tips. Bootstrap values greater than 50% based on 500 replications are given at the internal nodes. The distributions of *eae* alleles and EAF plasmid types are shown on the right (see Table 9 for plasmid type definitions).

fliC allele) were conserved among strains of each group: EPEC 1 strains were H6, EPEC 2 were H2, EPEC 3 were H34, and EPEC 4 were H6. These four clonal groups were represented among both the worldwide and Guinea-Bissau strains. The 21 STs differed on average at 1.4% and 0.2% of the nucleotide and amino acid sites, respectively. ST-20 was the most common multilocus genotype (22.5%) followed by ST-13 (19.4%), and ST-1 (14.0%) (Figure 4).

The splits network (Figure 5) reveals several parallel paths indicative of the presence of phylogenetic incompatibilities in the divergence of EPEC clones. Such incompatibilities could arise from recurrent mutation or recombination in the MLST loci. The ϕ_w test was used to detect recombination since it has been shown to discriminate between recurrent mutation and recombination in a variety of circumstances (17). In application to the concatenated sequences of the 21 STs, there were 129 informative sites and the ϕ_w test found statistically significant evidence of recombination (p <0.001). The four main clonal groups, however, are separated and intact. Three of the four groups occur at the end of long branches without evidence of multiple paths suggesting that recombination occurred early in the divergence of EPEC genotypes. With an EPEC phylogenetic framework in place, the allelic distributions of *eae*, *bfpA*, and *perA* were assessed.

Allelic variation in *eae*. The *eae* locus was subtyped by fRFLP (79) and 9 alleles (α, β, β2, ε, η, ι, κ, λ, and μ) were observed among the 129 EPEC strains (Figure 4). As shown previously (2), the EPEC 1 and EPEC 2 clonal groups possess the α and β alleles of *eae*, respectively. Furthermore, strains of the same ST had identical *eae* alleles as resolved by fRFLP. The α-*eae* allele had the widest distribution among the EPEC

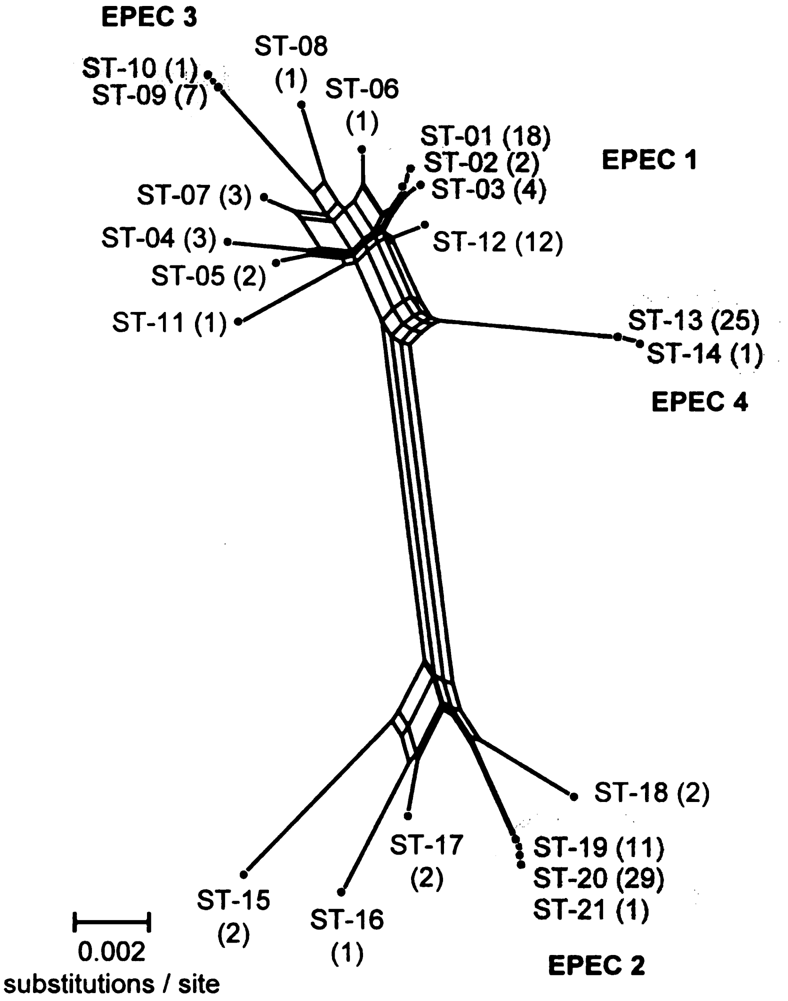

Figure 5. Phylogenetic network of 21 EPEC sequence types. Phylogenetic (splits) network is based on the neighbor-net algorithm using a *p* distance matrix. The four main clonal groups are indicated by gray ellipses. The sequence type (ST) and number of isolates are given at the branch tips.

clones, being found in strains with O51:[h49], O73:[h34], O142:[h34], and O157:[h45] serotypes in addition to the EPEC 1 group (O55:H6, O127:H6, and O142:H6). The rarest *eae* alleles were ε, η, ι, and λ, and combined, these alleles account for less than 5% of the strains examined.

Allelic variation in *bfpA*. The entire *bfpA* gene was amplified and sequenced in 15 strains with STs for which *bfpA* allelic data were not previously available (STs 4, 5, 7, 8, 11, 15, 16, 17, and 18). Comparative sequence analysis revealed the existence of a tenth allele of *bfpA*, which was designated β7.1 (Figure 6). A minor variant of this allele that differed by a single synonymous substitution was designated β7.2. Using the identified *bfpA* sequences, an RFLP-based typing system was devised to subtype *bfpA* alleles based on new PCR primers designed to target the conserved internal regions of the gene. Three restriction enzymes (*Alu*I, *Bfa*I, and *Bst*NI) were identified which, when used separately produced digestion patterns that combined could resolve 9 *bfpA* alleles (Table 7). *In silico* analysis with over 500 restriction endonucleases failed to identify an enzyme that could easily distinguish the β1 and β7 alleles. However, β1 and β7 *bfpA* strains can be easily differentiated based on their *perA* allele (see below).

PCR amplification of *bfpA* was successful in all but 21 isolates. Of the *bfpA*-negative strains, O128:H2 was the most common serotype with 13 isolates. RFLP analysis of the 108 *bfpA*-positive strains showed that the α1 (n=23) and α3 (n=24) alleles were the most common. The α2 (n=15) and β5 (n=17) alleles were also frequently identified, but the β2 (n=1), β3 (n=4), β4 (n=7), and β6 (n=2) alleles were rarely observed. Fourteen strains were classified as β1/β7 by RFLP, DNA sequencing confirmed 11 as β1 and 3 as β7. The α1 and α2 alleles of *bfpA*, as well as β4 and β5,

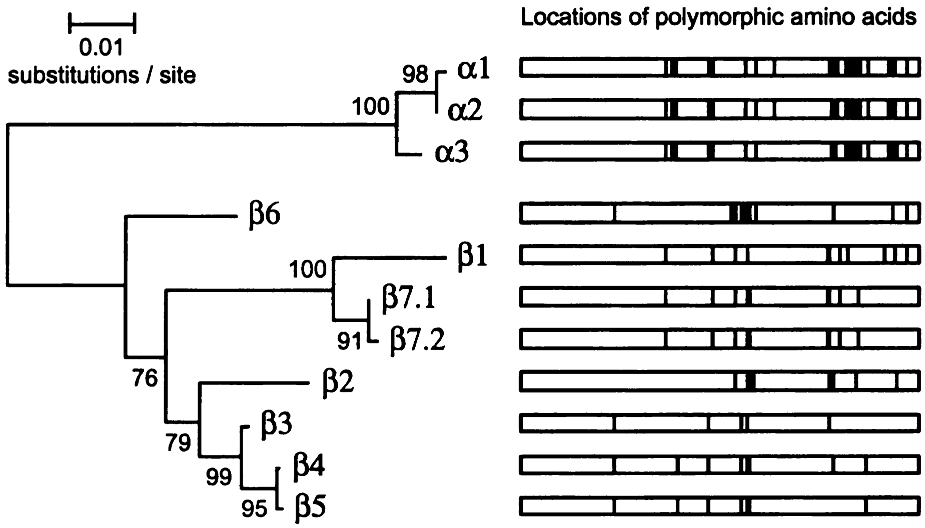

Figure 6. Eleven *bfpA* alleles cluster into two major groups. A phylogenetic tree constructed by the neighbor-joining algorithm based on the Kimura 2-parameter model of nucleotide substitution is shown on the left. Bootstrap values based on 500 replications are given at the internal nodes. To the right is a graph of the locations of the 39 polymorphic amino acid sites (195 total), which are marked as vertical lines that indicate differences from the consensus of all 11 alleles.

Table 7. Expected restriction fragment length polymorphisms of *bfpA* PCR amplicons.

bfpA	Digestion pattern (bp)[a]		
allele	*Alu*I	*Bfa*I	*Bst*NI
α1	408	408	155, 253
α2	408	75, 333	155, 253
α3	408	75, 333	408
β1	16, 54, 113, 231	39, 375	414
β2	16, 170, 228	414	414
β3	16, 54, 344	39, 375	414
β4	16, 177, 215	39, 369	408
β5	16, 54, 123, 215	39, 369	408
β6	16, 392	39, 369	408
β7	16, 54, 113, 225	39, 369	408

[a] Underlined fragments are not detectable under standard electrophoretic conditions.

differ only by one nonsynonymous nucleotide substitution. The two sets of closely related *bfpA* alleles were found in divergent EPEC lineages: EPEC 1 contains α1 and β5 whereas EPEC 2 has α2 and β4 (Figure 4). In addition, multiple *bfpA* alleles were found within the same sequence type: α3, β2, and β5 within ST-13 and α2, β1, and β4 within ST-20 (Figure 4).

Allelic variation in *perA*. The entire *perA* gene was amplified and sequenced in 33 strains representing a diverse set of EPEC STs and *bfpA* alleles. One strain, 2309-77 (O111:H2), resulted in a PCR product approximately 1 kb larger than expected. DNA sequencing revealed the presence of a 1055-bp IS element inserted into *perA* at position 414 with significant similarity to IS102 (86%) and IS903 (84%). Sequence analysis also identified 8 strains that contained one or more frameshifts within mononucleotide repeats in *perA* that presumably inactivate the gene. The variability of these frameshifts among closely related alleles indicates their relatively recent occurrence, as there has not been sufficient time for the inactivated alleles to accumulate further mutations. These frameshifts were corrected and the IS element sequence was excised *in silico* prior to allele assignment and phylogenetic analyses.

The 33 sequences yielded 20 alleles, which clustered into 4 groups based on phylogenetic sequence analysis, and in keeping with the nomenclature for *eae* and *bfpA*, these 4 allele classes were designated α, β, γ, and δ (Figure 7). As with *bfpA*, each distinct translated *perA* sequence was given an allele designation, resulting in 11 major α types, 3 β, 2 γ, and a single δ. Two of the α alleles had variants resulting from synonymous substitutions and each variant was given its own subtype designation (α1.1, α1.2, α1.3, α5.1, and α5.2).

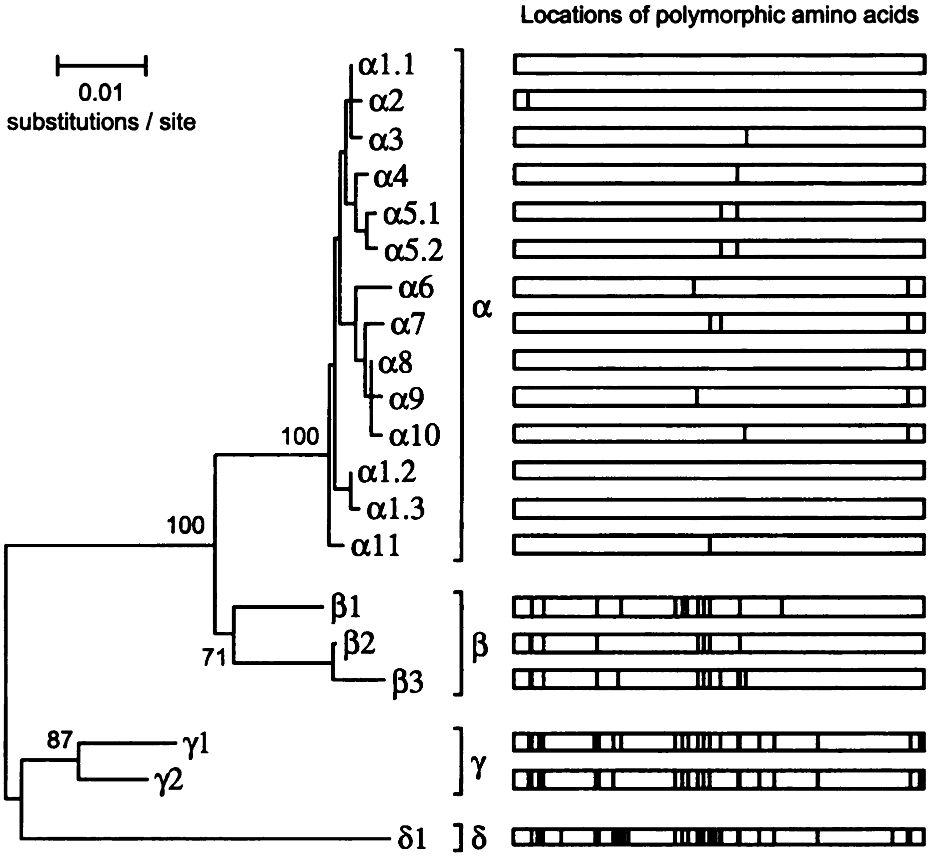

Figure 7. Twenty *perA* alleles cluster into four major groups. A phylogenetic tree constructed by the neighbor-joining algorithm based on the Kimura 2-parameter model of nucleotide substitution is shown on the left. Bootstrap values for the major groups based on 500 replications are given at the internal nodes. To the right is a graph of the locations of the 46 polymorphic amino acid sites (274 total), which are marked as vertical lines that indicate differences from the consensus of all 20 alleles.

Based on the sequence data, an RFLP method using *Dde*I and *Sau*96I was designed to detect the 4 *perA* allele classes (Table 8). PCR amplification of *perA* was successful in all but 17 isolates. As with the *bfpA*-negative strains, O128:H2 was the most common serotype among the *perA*-negative isolates (n=12). RFLP analysis of the 112 *perA*-positive strains showed that the α allele was the most common (n=74), followed by β (n=29), γ (n=8), and δ (n=1).

Association between EAF types and STs. By combining the *bfpA* and *perA* allelic data, each *bfpA* allele was associated with only one *perA* allele class, resulting in 11 distinct EAF plasmid types, which were designated EAF type 1 to 11 (Table 9). EAF plasmid types 4 and 8 appear to be the most promiscuous, being found in 5 and 4 clonal groups, respectively. Interestingly, the EAF type represented by the fully sequenced plasmid from O111:NM EPEC 2 strain B171 (145) is among the least promiscuous, being found in only one serotype (O111:H2) of a single sequence type (ST-20).

Table 8. Expected restriction fragment length polymorphisms of *perA* PCR amplicons.

perA	Digestion pattern (bp)[a]	
allele class	*Dde*I	*Sau*96I
α	13, 74, 82, 417	586
β	13, 74, 499	586
γ	82, 87, 417	174, 186, 226
δ	87, 499	226, 360

[a] Underlined fragments are not detectable under standard electrophoretic conditions.

Table 9. EAF plasmid types and distribution among EPEC clones.

EAF type	*bfpA* allele	*perA* allele	# of isolates	# of STs	# of clonal groups
1	α1	α	23	4	2
2	α2	α	15	1	1
3	α3	β	24	4	3
4	β1	α	11	6	5
5	β2	δ	1	1	1
6	β3	γ	4	1	1
7	β4	α	7	3	3
8	β5	α	18	4	4
9	β6	β	2	1	1
10	β7	β	3	2	2
11	–	γ	4	1	1
–	–	–	17	4	3

DISCUSSION

Common EPEC clones. This is the first study to take a comprehensive look at the evolution of typical EPEC by combining clonal relatedness based on MLST with the allelic distributions of three important virulence factors. Previous clonal studies of EPEC have focused primarily on two main groups, EPEC 1 and EPEC 2, which were first described and defined based on the genetic relatedness of strains as determined by multilocus enzyme electrophoresis (MLEE) (156). EPEC 1 was described to comprise strains with O55:H6, O86:H34, O127:H6, and O142:H6 serotypes, while EPEC 2 included O111:H2, O114:H2, O126:H2, and O128:H2. O119:H6 strains have also been regarded as members of EPEC 1 because they share a number of genetic traits with the group (H6 flagellar antigen, *eae*$^+$, and EAF$^+$) (147) even though MLEE places them just outside of EPEC 1 (156). However, there appear to be sufficient genetic differences to warrant the removal of O86:H34 and O119:H6 from EPEC 1 and they have been reclassified as EPEC 3 and EPEC 4, respectively (Figure 4). The data indicate that EPEC 1 strains (O55:H6, O127:H6 and O142:H6) all possess α-*eae*, whereas the κ and β2 *eae* alleles are associated with EPEC 3 and EPEC 4, respectively. O86:H34 strains have also been shown to possess cytolethal distending toxin, while EAF$^+$ O55:H6, O119:H6, O127:H6, and O142:H6 strains are negative (46). O119:H6 strains, on the other hand, are negative for *espC*, which encodes an enterotoxin, whereas EPEC 1 strains are positive (97).

Other EPEC clones. Thirty of the 129 strains (23%) that were analyzed did not belong to any of the above-mentioned major lineages, and most of these had unusual serotypes for EPEC. Of these serotypes, only O2:[h2], O49:[h10], and O51:[h49] have previously been reported to possess *eae* and/or express the AE phenotype (3, 7, 16), and only O33:[h34], O142:[h34], and O157:[h45] have been previously described as *eae*$^+$ *bfpA*$^+$, and therefore classified as typical EPEC (7, 40, 43, 90, 112, 138). Literature searches on the remaining serotypes, including O34:[h45], O73:[h34], O76:H51, O86:[h8], O142:[h21], OX9:[h7], O-:[h7], and O-:[h34] failed to find any association with typical EPEC.

An interesting finding of this study was the prevalence of the O55:[h51] clone among the strains isolated in Guinea-Bissau. Strains with this serotype have previously been described as relatively minor members of the O55 serogroup and have been reportedly isolated in only South America (8, 126). It is possible that O55:[h51] strains are simply common among children in Guinea-Bissau or their higher prevalence may be due to sampling bias. Alternatively, the abundance of these strains among the West African isolates could indicate that O55:[h51] is an emerging clone, which is increasing in frequency and spreading geographically, possibly because of a distinct combination of μ intimin and EAF type 8.

Virulence gene distribution. Recombination appears to have played a role in the initial generation of the EAF plasmid types. The highly divergent α1/α2 and β4/β5 *bfpA* alleles are all associated with α-*perA*, whereas the closely related β1 and β7 *bfpA* alleles are each found with different *perA* alleles. However, since *bfpA* and *perA* are in complete linkage disequilibrium (each *bfpA* allele is associated with only one *perA* allele class), it does not appear that recombination has played a large role in assorting allele combinations. The results also indicate varying degrees of promiscuity among the different EAF types (Table 9). Recently the complete sequence of a derivative of the wild-type EAF plasmid (pMAR7) from prototypical EPEC 1 strain E2348/69 was determined (15). In comparison to the EAF of O111:NM strain B171, the primary difference is the presence of the *tra* locus in pMAR7 (EPEC 1, EAF type 1) and its absence from pB171 (EPEC 2, EAF type 2). The *tra* genes, which are responsible for conjugal transfer in plasmids F and R100, were found to have varying degrees of conservation among the EAF plasmids from different EPEC strains (15). This finding has intriguing implications for the results presented here. In addition to EAF type 2, plasmid types 5, 6, 9, and 11 were all found in only a single ST. Like pB171, these plasmids may lack the entire *tra* locus or possess defective conjugation machinery thereby preventing their transfer to other EPEC clones. In contrast to the distribution of the EAF plasmid types, the distribution of *eae* alleles among EPEC strains is more consistent with their clonal relationships. This suggests that the EAF plasmid is more mobile than the LEE island.

***bfpA*-negative EPEC.** EAF type 11 is unusual in that it was the only plasmid type that did not contain *bfpA* according to the PCR screening. Bortolini and colleagues (14) reported a similar finding when they described EAF$^+$ O119:H2 and O128:H2 strains in which most (~13 kb) of the *bfp* operon had been deleted and replaced with an IS66-like element. Since the deletion encompasses the 3' end of *bfpA*, this could explain why plasmid type 11 is *bfpA*-negative. To confirm this, new primers were designed to target the 5' end of *bfpA* and the IS66-like element. All EAF type 11 strains yielded the expected amplicon indicating that this plasmid type possesses a similar *bfp* operon structure as that described by Bortolini and colleagues (14). Aside from the EAF type 11 strains, 17 additional isolates were *bfpA*-negative. These strains, however, were also *perA*-negative, suggesting that they did not possess the EAF plasmid. Of the EAF-negative isolates, most (82%) were part of the EPEC 2 clonal group, with the O128:H2 serotype being the most common. This finding was not unexpected as O128:H2 strains are often reported to be EAF-negative and, therefore, classified as atypical EPEC (147).

Relationship between typical and atypical EPEC. The findings presented here show that at least some atypical EPEC strains, such as those that lost both *bfpA* and *perA* (and presumably the whole EAF plasmid), evolved from typical EPEC, rather than typical EPEC evolving from atypical EPEC by acquisition of the plasmid. It has been shown that typical EPEC can lose the EAF plasmid at a surprisingly high rate during passage through adult volunteers (30, 84), so there appears to be selective pressure to lose the plasmid and convert from typical to atypical EPEC. This outcome is interesting given the recent reports of atypical EPEC in human clinical isolates, some of which belong to typical EPEC O-serogroups (4, 7, 105).

ACKNOWLEDGEMENTS

I would like to thank Lindsey Ouellette, Dr. Weihong Qi, and Dr. Hans Steinsland for technical assistance. This work appears in the second January 2007 issue of the Journal of Bacteriology (189:342-350). This project has been funded in part with Federal funds from the NIAID, NIH, DHHS, under NIH Research Contract # N01-AI-30058.

CHAPTER 4

SEQUENCE VARIATION WITHIN THE TYPE 1 FIMBRIAL PHASE SWITCH OF PATHOGENIC *ESCHERICHIA COLI*

SUMMARY

Strains of *Escherichia coli* O157:H7 do not express the type 1 fimbriae encoded by the *fim* operon because of a 16-bp deletion within the *fim* switch. This regulatory switch element is an invertible piece of DNA that is responsible for the phase variation phenotype of the fimbriae. The entire *fim* switch was sequenced in 129 strains representing 22 O-serogroups to assess the amount of variation within this element and to determine when this deletion occurred during the evolution of O157:H7. Sequence analysis of 32 *fim* switch alleles revealed that the inverted repeats, promoter sites, and the third binding site for the leucine-responsive regulatory protein (Lrp) were well conserved. In contrast, the first and second Lrp sites and the integrated host factor (IHF) binding site were more variable. Phylogenetic analyses indicated that the *fim* switch deletion occurred at a recent stage in the evolution of *E. coli* O157:H7: only β-glucuronidase-negative, sorbitol-negative (GUD-, SOR-) O157:H7 and O157:NM strains possess the deletion, whereas GUD+, SOR+ O157:H- strains and GUD+, SOR- O157:H7 contain an intact *fim* switch. These observations suggest that the deletion in the *fim* switch occurred after the loss of GUD expression.

INTRODUCTION

Type 1 fimbriae are filamentous structures expressed on the surface of most clinical *E. coli* isolates (5). These fimbriae bind to mannose-containing receptors on epithelial cells (109), have been shown to be required for colonization of the urinary tract (23, 53, 65, 132), and may play a role in the colonization of the intestinal tract (76, 77). Type 1 fimbriae are encoded by the *fim* operon, a cluster of nine genes required for biosynthesis (52, 70, 116), and are composed primarily of the structural subunit FimA (69). Expression of these fimbriae is under the control of the *fim* switch, an invertible genetic element that contains the promoter for *fimA* (1). Transcription of *fimA* only occurs when the *fim* switch is in the "on" orientation. Inversion of the *fim* switch is mediated by two site-specific recombinases, FimB and FimE, which are encoded upstream of the switch (41, 92). In addition to these recombinases, two cofactors, the integration host factor (IHF) and the leucine-responsive regulatory protein (Lrp), have positive regulatory roles in the inversion process (12, 13, 32, 34, 42, 128). It has been hypothesized that these proteins organize the structure of the nucleoprotein complex so that the site-specific recombination event can occur.

E. coli strains with the O157:H7 serotype are unable to express type 1 fimbriae because the *fim* switch appears to be permanently locked in the "off" orientation so that transcription of *fimA* does not occur (85, 127). DNA sequencing has revealed that *E. coli* O157:H7 has a 16-bp deletion within the *fim* switch, which prevents inversion of the element (85, 127). It has been hypothesized that since the end of the deletion is within the IHF binding site, it may prevent IHF from binding to and bending the DNA at this position (127). In addition, the deletion may interfere with strict spatial requirements necessary for recombination between the switch's inverted repeats (81, 127).

Feng et al. (37) have proposed a model for the stepwise evolution of *E. coli* O157:H7. Under this model, enterohemorrhagic *E. coli* (EHEC) O157:H7 strains evolved from enteropathogenic *E. coli* (EPEC) O55:H7 strains. To determine if the deletion within the *fim* switch is compatible with this model, the entire *fim* switch was sequenced in 25 O157:H7, 35 non-motile O157, and 8 O55:H7 strains. An additional 61 strains representing other types of pathogenic *E. coli* were sequenced to determine if the 16-bp deletion is unique to O157:H7 and to assess the level of variation within the switch.

MATERIALS & METHODS

Strains. A collection of 129 strains representing 22 O-serogroups was assembled (Table 10). These strains were originally isolated between 1947 and 2003 from different regions around the world. Each strain was grown overnight at 37°C in 10 ml of Luria-Bertani (LB) broth with moderate shaking. Genomic DNA was isolated using the Puregene DNA isolation kit (Gentra Systems Inc., Minneapolis, MN). DNA concentrations were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Rockland, DE), which were diluted to 25 ng/µl for PCR.

MLST. Multilocus sequence typing (MLST) was performed on 7 conserved housekeeping genes (*aspC*, *clpX*, *fadD*, *icdA*, *lysP*, *mdh*, and *uidA*) as described in Chapter 3.

***fliC* typing.** Strains that were nonmotile or lacked flagellar serotype data were typed for the *fliC* locus as described in Chapter 3.

PCR. To select for the "off" orientation, the *fim* switch was amplified in two pieces. Primers fimsw-F6 (5'-TGC CGG ATT ATG GGA AAG A-3') and off-R1 (5'-ATT TGG GGC CAT TTT GAC TC-3') were used to amplify the 5' end of the *fim* switch and primers off-F5 (5'-GTT TCT GTG GCT CGA CGC ATC T-3') and fimsw-R1 (5'-GGA CAG AGC CGA CAG AAC AA-3') amplified the 3' end. These primers were used for both PCR amplification and DNA sequencing of the *fim* switch (see below). Each 25-µl reaction contained 2.5 µl 10X buffer II (Applied Biosystems, Foster City, CA), 2.5 µl 2 mM dNTP, 2.0 µl 25 mM $MgCl_2$, 0.5 µl 10 µM forward primer, 0.5 µl 10 µM reverse primer, 1.5 units AmpliTaq Gold (Applied Biosystems), 1 µl 25 ng/µl genomic DNA template, and 15.7 µl ddH_2O. Amplification utilized an initial denaturing step at 94°C for

Table 10. Summary of strains investigated for *fim* switch sequence variation.

Serogroup	Flagellar type(s) [a]	# of isolates	*fim* switch deletion
O6	H1, H31	2	no
O26	H11, [h11]	3	no
O55	H6, H7, [h34]	12	no
O86	H34, [h34]	5	no
O91	H21	1	no
O103	[h25]	1	no
O104	H21	1	no
O110	H6	1	no
O111	H2, [h2], H8, H9, [h11]	12	no
O113	H21	1	no
O114	H2	2	no
O118	H16	1	no
O119	H6	5	no
O121	H19	1	no
O126	H2	1	no
O127	H6	1	no
O128	H2	3	no
O142	H6, [h21]	5	no
O157	H7, [h7], [h42], [h45]	65	no (36), yes (29)
O173	[h16]	2	no
O174	[h21]	1	no
O-	[h7], H18, H48	3	no

[a] Lower case H-types in square brackets were inferred from *fliC* allele.

10 min., followed by 35 cycles of 92°C for 1 min., 55°C for 1 min., and 72°C for 30 sec. A final step of 72°C for 5 min. was used for final completion of any partially extended product. PCR products (5 µl) were visualized on ethidium bromide-stained 1.5% agarose gels by illumination with UV light.

DNA sequencing. PCR products were purified using the QIAquick PCR purification kit (QIAGEN Inc., Valencia, CA) and quantified. Cycle sequencing reactions contained 4.0 µl CEQ DTCS Quick Start premix (Beckman Coulter Inc.), 1.0 µl 20 µM primer, approximately 40 ng of purified *fim* switch product, and ddH_2O to 10 µl. Amplification utilized an initial denaturing step at 94°C for 1 min., followed by 35 cycles of 96°C for 30 sec., 55°C for 30 sec., and 60°C for 2 min. Upon completion of cycle sequencing, samples were purified with Sephadex G-50 Fine columns, dried under vacuum centrifugation, suspended in 40 µl of deionized formamide, and sequenced using a Beckman CEQ2000XL DNA sequencer. Samples were analyzed using the CEQ2000XL software and then exported for further analysis with the SeqMan module of the Lasergene software (DNASTAR Inc., Madison, WI).

Phylogenetic Analyses. Sequences were aligned with the ClustalW algorithm using the MegAlign module of the Lasergene software. A neighbor-joining tree of the concatenated MLST sequences was constructed using the Kimura 2-parameter model of nucleotide substitution with the MEGA3 software (78). The inferred phylogeny was tested with 500 bootstrap replications. Phylogenetic network analysis was conducted with the SplitsTree 4 (54) program using the neighbor-net algorithm (18) and untransformed distances (*p* distance).

RESULTS

Variation within the *fim* switch. PCR amplification and sequencing of the *fim* switch was successful in all cases. A total of 61 polymorphic sites are present within the *fim* switch, resulting in 32 alleles (Figure 8). Of the regions within the switch with functional importance, the –10 and –35 promoter sites, the Lrp-3 binding site, and the right inverted repeat were completely conserved across all 32 alleles. The left inverted repeat was conserved in all but a single O157:H7 strain, which contained a C → T transition. The IHF binding site and the remaining two Lrp binding sites (Lrp-1 and Lrp-2) were more variable. With the exception of the strains possessing the 16-bp deletion, the size of the switch was fairly consistent with a length of 314±1 bp. The 16-bp deletion was found to be associated only with strains expressing the O157 antigen. Examination of the flagellar data revealed that the deletion was unique to O157:H7 (n=20) and O157:[h7] (n=9) strains. However, 5 O157:H7 and 26 O157:[h7] strains possess an intact *fim* switch.

Clonal analysis. Since the presence of the *fim* switch deletion was variable among the O157:H7 and O157:[h7] isolates, MLST was performed to assess the clonal relatedness of the strains investigated. The sequenced internal fragments of the 7 housekeeping genes were concatenated to yield 3,732 nucleotides for comparative analysis. MLST analysis resolved an average of 35.7 variable nucleotide sites per locus, which defined a number of alleles ranging from 12 to 24 at the 7 housekeeping genes (Table 11). The distinct combinations of alleles across the 7 MLST loci were used to define 41 multilocus genotypes or sequence types (STs) among the 129 strains (Figure 9). Strains closely related to O157:H7 had one of 12 STs that cluster together at the end of a

```
   IRL               -10                     -35
TTGGGGCCAAA-CTGTCCATATCATAAATAAGTTACGTATTTTTTCTCAAGCATAAAAAT      [60]
      t    A     T           G     T  G-

                                 IHF
ATTAAAAAACGACAAAAAGCATCTAACTGTTTGATATGTAAATTATTTCTCTTGTAAATT     [120]
T C-T             T  C      a   t    A     C      A   cG

AATTTCACATCACCTCCGCTATATGTAAAGCTAACGTTTCTGTGGCTCGACGCATCTTCC     [180]
                     A   C               t Aa     A   A
                                            T     T

   Lrp-3                          Lrp-1
TCATTCTTCTCTCCAAAAACCACCTCATGCAATATAAAAATCTATAAATAAAGATAACAA     [240]
                  Ga     a  AT        C A  GC       Cg   TC
                                                          T

     Lrp-2
TAGAATATTAAGCCAACAAATAAACTGAAAAAGTTTGTGCGCGATGCTTTCCTCTATGAG     [300]
    gC  a  T T G G                    C
        C

         IRR
TCAAAATGGCCCCAA                                                  [315]
```

Figure 8. Nucleotide polymorphism within the *fim* switch. The consensus sequence determined from 32 *fim* switch alleles is given for the "off" orientation. Mutations from the consensus are below the sequence with substitutions found in a single isolate in lower case. Regions with known functional importance are in bold and labeled above the sequence (IRL, inverted repeat left; -10 and -35, promoter sites; IHF, integrated host factor binding site; Lrp, leucine-responsive regulatory protein binding site; IRR, inverted repeat right). The location of the O157:H7 deletion between sites 67 and 82 is in the shaded box. Nucleotide positions are given at the end of each line.

Table 11. Variation among alleles of 7 MLST genes in the *fim* switch strains.

Locus	# of sites	# of variable sites	# of alleles
aspC	513	26	12
clpX	567	39	16
fadD	483	61	17
icdA	567	29	17
lysP	477	17	14
mdh	549	35	24
uidA	576	43	18
MLST avg.	533.1	35.7	16.9

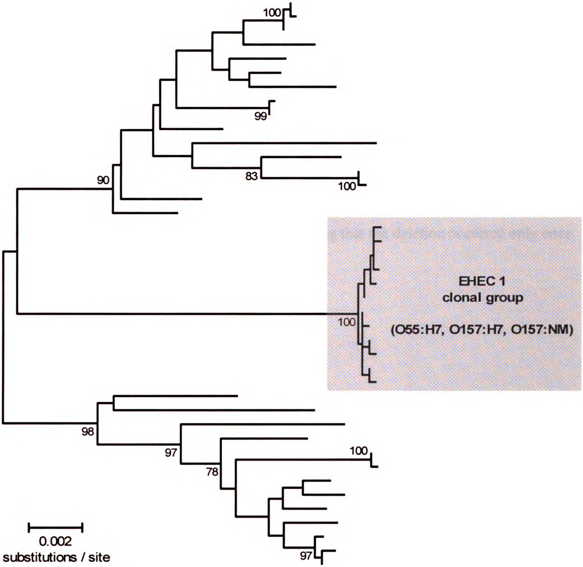

Figure 9. Phylogenetic relationships of 41 observed sequence types. An unrooted phylogenetic tree was constructed by the neighbor-joining algorithm based on the Kimura 2-parameter model of nucleotide substitution. The gray box indicates the 12 sequence types representing strains with O55:H7, O157:H7, and O157:NM serotypes belonging to the EHEC 1 clonal group. Bootstrap values greater than 75% based on 500 replications are given at the internal nodes.

long internal branch with 100% bootstrap support. In addition to O157:H7, the clonal group contains isolates with O157:[h7], O55:H7, and O-:[h7] serotypes. Reid et al. (124) have designated this cluster of closely related strains the EHEC 1 clonal group. When these STs were combined with the *fim* switch allelic data, 18 genotypes were resolved within the EHEC 1 clonal group (Figure 10). Phylogenetic splits network analysis, which does not force the data into a bifurcating tree, revealed that the genotypes with the *fim* switch deletion cluster together, strongly suggesting that the deletion occurred only once during the diversification of the clonal group.

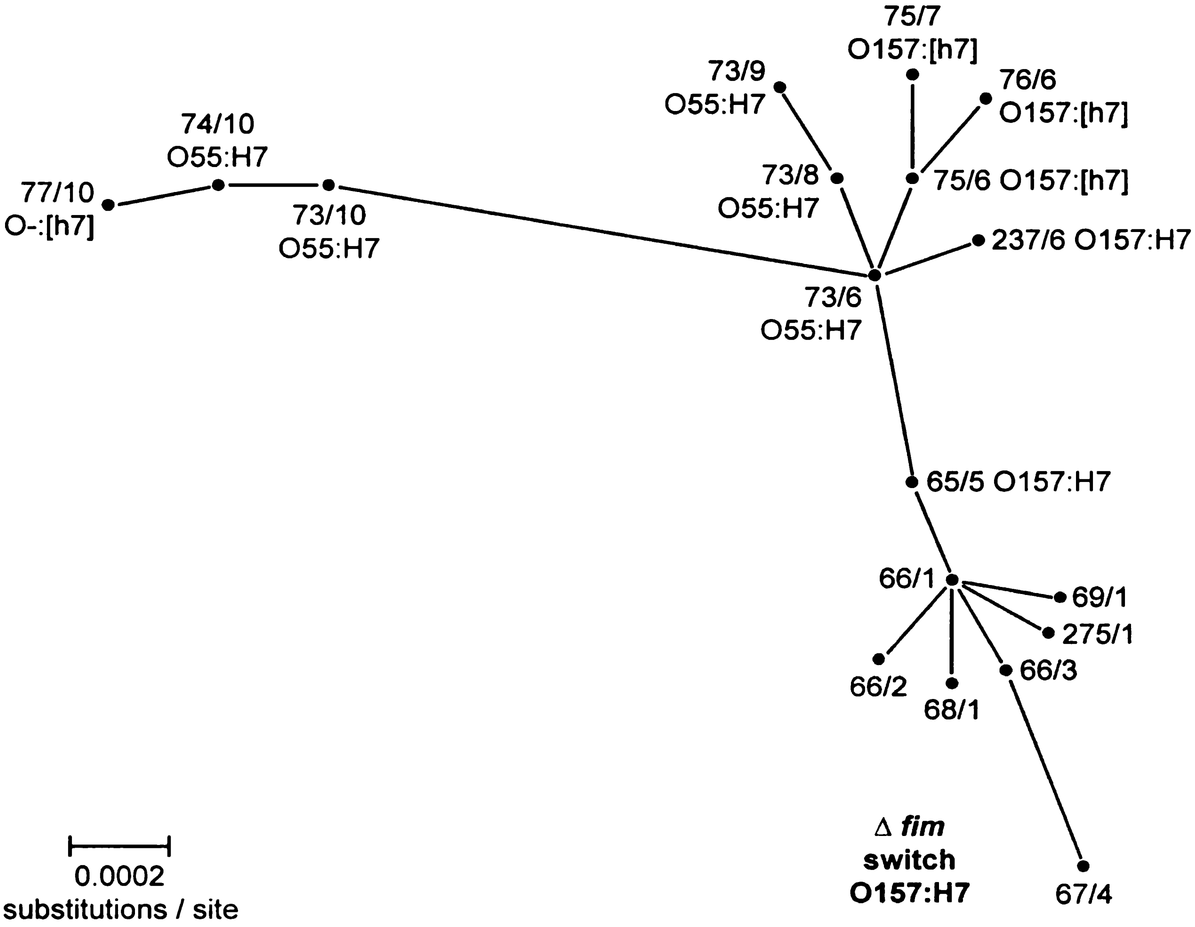

Figure 10. Phylogenetic network of 18 genotypes belonging to the EHEC 1 clonal group. The splits network is based on the neighbor-net algorithm using a *p* distance matrix. Unique combinations of sequence type (ST) and *fim* switch allele were used to define the genotypes, which are denoted as ST/*fim* switch. Serotypes common to each genotype are also shown. Genotypes with the *fim* switch deletion are enclosed in the gray box. The 9 O157:[h7] isolates with the *fim* switch deletion are all genotype 66/1.

DISCUSSION

Sequence polymorphism within the *fim* switch. The work presented here is the first to examine the level of DNA sequence variation within the *fim* switch from a diverse set of strains. Previous work on the switch has been primarily focused on identifying binding sites for cofactors (IHF and Lrp) involved in the inversion process (13, 42, 128). The results of these studies have made it possible to examine the switch for differing selective pressures by classifying the sites within the switch as either "domain" or "independent." The domain sites consisted of the two inverted repeats, both promoter sites, and the IHF and Lrp binding sites; all other sites were classified as independent. When sites within the switch were placed into one of these two categories, an intriguing pattern emerged. Substitution rates involving transitions were similar between the two categories of sites (0.0199 for domain, 0.0170 for independent). However, a greater difference between the two categories was observed for the transversion rate (0.0107 for domain, 0.0155 for independent), suggesting that there may be some selective disadvantage towards transversions within the domain sites.

Stepwise evolution of O157:H7 and the *fim* switch deletion. A stepwise evolution model has been proposed for the evolution of *E. coli* O157:H7 (37). This model is based on multilocus enzyme electrophoresis and was created using a parsimony approach. The *fim* switch deletion appears to follow the stepwise evolution model and occurred only once during the evolution of O157:H7. The *fim* switch deletion occurred relatively recently in the evolutionary history of *E. coli* O157:H7; only GUD-, SOR- O157:H7 and O157:NM strains possess the 16-bp deletion. This implies that the *fim* switch deletion occurred after the loss of GUD expression. The stepwise evolution model was updated using sequence types resolved by MLST as the foundation rather than electrophoretic types as determined by MLEE (Figure 11). While the original model had a bifurcating split after the unobserved GUD+, SOR+, Stx2+ O157:H7 phenotype (A3 in Figure 11), the updated model now has a trifurcation after this point. The discovery of a GUD+, SOR+, Stx2- O157:H7 strain provides further support to the existence of the A3 phenotype. As the contribution of O157 strains without the typical diagnostic GUD- and SOR- phenotypes to the cases of EHEC disease becomes known, a strain with the elusive A3 phenotype may finally be observed.

The biological relevance of the loss of type 1 fimbrial expression in O157:H7 remains unclear. However, an intriguing hypothesis has been put forth in a recent study by Low et al. (88). They measured the expression of 15 fimbrial gene clusters in O157:H7 and found that most (n=11) were not expressed under the variety of conditions examined. The authors concluded that the limited collection of expressed fimbriae may be an important part of O157:H7's biology. Low et al. hypothesize that O157:H7's niche at the terminal rectum of cattle is possibly due to limited adherence to other sites in the gastrointestinal tract. This hypothesis, however, remains to be tested.

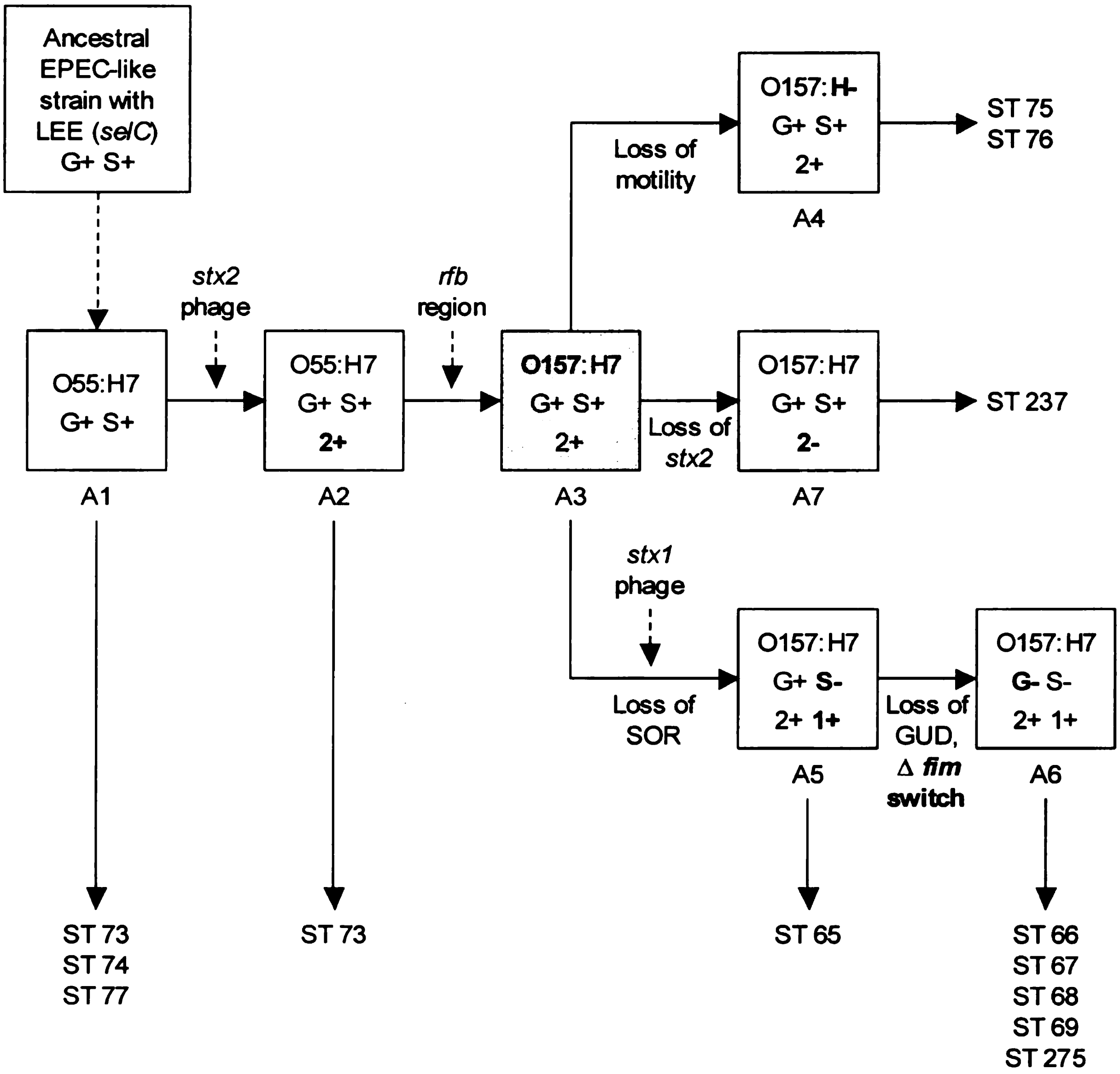

Figure 11. Revised stepwise evolution model of *E. coli* O157:H7. Phenotypes of ancestors A1–A7 are shown; changes predicted to have occurred are in bold (G, GUD; S, SOR; 1, Stx1; 2, Stx2). Sequence types (ST) observed among the strains investigated are shown. A strain with the traits of ancestor A3 (shaded square) has not been reported.

ACKNOWLEDGEMENTS

I would like to thank Lindsey Ouellette for technical support and Susan Francisco for her assistance in the sequencing of the EPEC isolates examined. This project has been funded in part with Federal funds from the NIAID, NIH, DHHS, under NIH Research Contract # N01-AI-30058.

CHAPTER 5

POSITIVE SELECTION AND RECOMBINATION IN SURFACE PROTEIN-ENCODING GENES

SUMMARY

Escherichia coli is a diverse species of Gram-negative bacteria. Most strains are nonpathogenic and do not harm their host, but some may cause a variety of harmful intestinal and extra-intestinal infections. A common theme in the pathogenesis of the different types of *E. coli* is bacterial attachment mediated by the expression of surface proteins. A wide range of surface proteins are expressed by these strains, some of which are ubiquitous while others are specific to certain pathogenic types (pathotypes). The primary objective of this study is to examine the allelic diversity of 5 surface protein-encoding genes (*bfpA*, *csgA*, *eae*, *espA*, and *fimA*) for the actions of positive selection and recombination. Allelic sequences were obtained for at least one of the 5 loci from a collection of 324 strains representing 44 O-serogroups of *E. coli* as well as the newly described *E. albertii*. Sequence analysis identified 11 *bfpA*, 12 *csgA*, 20 *eae*, 31 *espA*, and 32 *fimA* alleles in the strains investigated. Comparison to the allelic sequences of 7 conserved housekeeping loci in the same strains revealed higher levels of sequence variation in the 5 surface protein-encoding genes. Analysis of the housekeeping loci suggests that these genes are under weak negative selection with most (76%) of the codons examined evolving neutrally with no significant intragenic recombination events identified. In contrast, evidence of positive selection and/or recombination was detected in all 5 of the surface protein-encoding genes. These results support the hypothesis that surface proteins alter their three-dimensional structure or surface epitopes in order to confer an evolutionary advantage.

INTRODUCTION

Escherichia coli is a diverse species of Gram-negative bacteria, some strains of which are pathogenic. The early stages of *E. coli* pathogenesis often involve bacterial attachment mediated by the expression of surface proteins. It has been hypothesized that pathogens alter their surface proteins in order to evade detection by their host's immune system. Therefore, it is likely that natural selection is acting to generate new allelic variants. Furthermore, homologous recombination may be acting to generate new allelic variants. This could take place through the exchange of whole genes or gene segments. Several methods have been developed to test for the occurrence and boundaries of recombinational events. The most commonly used methods examine the sequences for conflicting phylogenetic signals (58, 74) or a non-random distribution of nucleotide substitutions (130).

There are two types of natural selection, positive and negative. These types of selection can be detected at the DNA sequence level by comparing the rates of synonymous (silent) and nonsynonymous (amino acid changing) nucleotide substitutions. Numerous methods have been developed for estimating the numbers of synonymous and nonsynonymous substitutions (22, 44, 57, 86, 87, 99, 103). These methods calculate the substitution rate for an entire gene by computing the average number of substitutions over a particular length of codons. More recently, methods have been developed to detect positive selection at single amino acid sites (51, 73, 106-108, 142, 143, 158, 159).

It is the goal of this research to examine the allelic diversity of genes that encode a variety of surface structures in different classes of pathogenic *E. coli* (pathotypes). Five genes encoding surface proteins involved in bacterial attachment during different types of infection or survival in the external environment were examined for evidence of positive selection and homologous recombination. The genes chosen for this analysis are: 1) *bfpA*, the major structural subunit of the bundle-forming pilus of enteropathogenic *E. coli* (EPEC); 2) *csgA*, the major structural subunit of the curli fimbriae involved in biofilm formation; 3) *eae*, the intimin protein of the attaching/effacing phenotype; 4) *espA*; the filamentous extension on the type III secretion system needle of attaching and effacing *E. coli* (AEEC); and 5) *fimA*, the major structural subunit of the type 1 fimbriae. The level of sequence polymorphism was assessed in each locus and compared to the amount of variation present within 7 genes with conserved housekeeping function. A similar approach was taken for both the selection and recombination analyses. Overall values were calculated for d_N, d_S, and phylogenetic compatibility, followed by regional and site-by-site examination of each locus.

MATERIALS AND METHODS

Strains. A collection of 324 strains representing 44 O-serogroups of *E. coli* as well as the newly described *E. albertii* was assembled (Table 12). These strains were originally isolated between 1947 and 2004 from 29 countries around the world. Each strain was grown overnight at 37°C in 10 ml of Luria-Bertani (LB) broth with moderate shaking. Genomic DNA was isolated using the Puregene DNA isolation kit (Gentra Systems Inc., Minneapolis, MN). DNA concentrations were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Rockland, DE), which were diluted to 25 ng/μl for PCR.

MLST. Multilocus sequence typing (MLST) was performed on 7 conserved housekeeping genes (*aspC*, *clpX*, *fadD*, *icdA*, *lysP*, *mdh*, and *uidA*) as described in Chapter 3.

***fliC* typing.** Strains that were nonmotile or lacked flagellar serotype data were typed for the *fliC* locus as described in Chapter 3.

PCR primer design. DNA sequences of the regions encompassing each of the genes (*bfpA*, *csgA*, *eae*, *espA*, and *fimA*) were obtained from GenBank and aligned. All primers were designed to target conserved regions, were synthesized by Integrated DNA Technologies, Inc., and were stored at a concentration of 100 μM in ddH_2O.

PCR. The *eae* and *bfpA* genes were amplified and sequenced as described in Chapters 2 and 3, respectively. For *csgA*, each 25-μl reaction contained 2.5 μl 10X buffer II (Applied Biosystems, Foster City, CA), 2.5 μl 2 mM dNTP, 2.0 μl 25 mM $MgCl_2$, 0.5 μl 10 μM csgA_-70F primer (5'-CAA ATG GCT ATT CGC GTG AC-3'), 0.5 μl 10 μM csgA_542R primer (5'-GTG CCG CAA GGA GTA ATA AC-3'), 1.5 U

Table 12. Summary of strains investigated for allelic variation in at least one of 5 surface protein-encoding genes.

Species &	# of	# of	Allelic data [a]				
serogroup	strains	STs	*bfpA*	*csgA*	*eae*	*espA*	*fimA*
E. coli O2	2	2	1		2	1	
E. coli O5	2	2			2	2	
E. coli O6	2	2		2			2
E. coli O20	1	1			1	1	
E. coli O21	1	1			1	1	
E. coli O26	4	2		3	3	3	3
E. coli O33	2	1	2		2	2	
E. coli O34	3	2	1		3	1	
E. coli O49	2	1	2		2	2	
E. coli O51	2	2	1		2		
E. coli O55	36	9	21	9	36	16	12
E. coli O73	1	1	1		1	1	
E. coli O76	1	1	1	1	1	1	
E. coli O84	1	1		1	1	1	
E. coli O85	1	1			1	1	
E. coli O86	14	7	9	3	10	4	5
E. coli O88	7	1			7		
E. coli O91	1	1					1
E. coli O101	3	2			3	2	
E. coli O103	3	2		2	3	3	1
E. coli O104	1	1					1
E. coli O108	4	1			4		
E. coli O109	1	1			1	1	
E. coli O110	2	2	1	1	2	2	1

continued

Table 12, continued

Species &	# of	# of	Allelic data [a]				
serogroup	strains	STs	*bfpA*	*csgA*	*eae*	*espA*	*fimA*
E. coli O111	24	7	15	9	21	5	12
E. coli O113	1	1					1
E. coli O114	5	3	4	2	4		2
E. coli O116	1	1			1	1	
E. coli O118	1	1					1
E. coli O119	29	4	23	2	28	5	5
E. coli O121	5	3		5	2	2	1
E. coli O125	4	2			4	1	
E. coli O126	4	3	2	1	3		1
E. coli O127	5	2	4	1	4	1	1
E. coli O128	20	6	2	5	15	1	3
E. coli O142	12	5	11	3	11	3	5
E. coli O145	12	7			12	2	
E. coli O153	1	1			1		
E. coli O157	67	13	3	13	46	14	65
E. coli O173	2	1					2
E. coli O174	1	1					1
E. coli OX9	1	1	1		1	1	
E. coli O-	9	9	3	2	7	2	3
E. coli O? [b]	7	3			7		
E. albertii	12	7		5	10	7	4
S. boydii type 13	4	3			4	1	
total	324	103	108	70	269	91	133

[a] Number of strains for which allelic data was determined. Blanks indicate no allelic data was obtained because the strain was either negative or not tested for the locus.

[b] O-type not determined.

AmpliTaq Gold (Applied Biosystems), 1 µl 25 ng/µl genomic DNA template, and 15.7 µl ddH_2O. Amplification of the approximately 600-bp fragment utilized an initial denaturing step at 94°C for 10 min, followed by 35 cycles of 92°C for 1 min, 55°C for 1 min, and 72°C for 30 s. A final step of 72°C for 5 min was used to complete any partially extended product. For *espA*, an approximately 950-bp fragment was amplified under conditions similar to those described for *csgA* with the exceptions of primer sepL-F887 (5'-AGA GCC CTT CTC GGG TAT CG-3'), primer espD-R126 (5'-GGC CGT GGA TTT AAC CAG TTG TAA-3'), an annealing temperature of 53°C, and an extension time of 1 min. For *fimA*, an approximately 700-bp fragment was amplified under conditions similar to those described for *csgA* with the exceptions of primers fimA-F6 (5'-ACT GCC CAT GTC GAT TTA GAA-3') and fimA-R8 (5'-GAG CAA ACA TTG GCA GCA AC-3'). PCR products (5 µl) were visualized on ethidium bromide-stained 1.5% agarose gels by illumination with UV light, purified using the QIAquick PCR purification kit (QIAGEN Inc., Valencia, CA), and quantified.

DNA sequencing. For *csgA*, cycle sequencing reactions contained 4.0 µl CEQ DTCS Quick Start premix (Beckman Coulter Inc., Fullerton, CA), 1.0 µl 20 µM csgA_-70F or csgA_542R primer, approximately 60 ng of purified *csgA* PCR product, and ddH_2O to a final volume of 10 µl. Amplification utilized an initial denaturing step at 94°C for 1 min, followed by 35 cycles of 96°C for 30 s, 55°C for 30 s, and 60°C for 2 min. Upon completion of cycle sequencing, samples were purified with Sephadex G-50 Fine columns (Amersham Pharmacia Biotech Inc., Piscataway, NJ), dried under vacuum centrifugation (Savant Instruments Inc., Holbrook, NY), suspended in 40 µl of deionized formamide, and run on a CEQ2000XL (Beckman Coulter Inc.). For *espA*, sequencing conditions were similar to those described for *csgA* with the exceptions of primers sepL-F887, espD-R126, espA-F205c (5'-GAG GCA TCT AAR GMG TCA AC-3'), espA-F246 (5'-GGA TGC CAA GAT CGC TGA AGT T-3'), espA-R413 (5'-GCT TTT ACG GTT TGC AGG TCA C-3'), and espA-R426c (5'-AAT AGC NGC YTT CAC YGT TTG-3'); an annealing temperature of 53°C; and approximately 100 ng of purified *espA* PCR product. For *fimA*, sequencing conditions were similar to those described for *csgA* with the exception of primers fimA-F6 and fimA-R8 and approximately 70 ng of purified *fimA* PCR product. Samples were analyzed using the CEQ2000XL software and then exported for further analysis with the SeqMan module of the Lasergene software (DNASTAR Inc., Madison, WI).

Phylogenetic analyses. Sequences were aligned with the ClustalW algorithm using the MegAlign module of the Lasergene software. Neighbor-joining trees were constructed using the Kimura 2-parameter model of nucleotide substitution with the MEGA3 software (78) and the inferred phylogenies were each tested with 500 bootstrap replications.

Recombination analyses. Putative regions of recombination were identified through the construction of compatibility matrices of parsimoniously informative sites using the program Reticulate (58). The significance of each matrix was evaluated by a Monte Carlo approach in which the original matrix was compared to 1000 random matrices of the gene's informative sites. Allelic sequences were examined for evidence of gene conversion by performing Sawyer's test using the program GENECONV (Version 1.81) (131). Calculations were based on 10,000 permutations and global fragments with Bonferroni-corrected Karlin-Altschul p-values ≤ 0.05 were considered significant.

Allelic sequences were also fit to a nucleotide substitution model and potential recombinational breakpoints were identified using the genetic algorithm recombination detection (GARD) method (74) as implemented on the Datamonkey website (http://www.datamonkey.org/GARD) using the $\beta - \Gamma$ model with 3 rate classes. For loci in which GARD identified a possible recombinational breakpoint, the regions of the gene on either side of the breakpoint were analyzed separately for the action of selection on individual codons (see below).

Selection analyses. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) were estimated by the modified Nei-Gojobori method using MEGA3 (78). The proportions of polymorphic synonymous (p_S) and nonsynonymous (p_N) sites were calculated by the method of Nei and Gojobori (103). Variation in the functional constraints of different regions of each gene was examined by tabulating p_S and p_N in a sliding-window analysis of 10 or 30 codons using the program PSWIN. Allelic sequences were fit to a nucleotide substitution model using the Datamonkey website (http://www.datamonkey.org) and then the single likelihood ancestor counting (SLAC) and random effects likelihood (REL) methods were used to fit a codon model to detect selection on individual codons (72). Codons under significant positive or negative selection were identified by p-values ≤ 0.01 for SLAC and posterior probabilities ≥ 0.99 for REL.

RESULTS

MLST analysis. PCR amplification and sequencing of the 7 MLST loci in 324 strains was successful in most (92%) cases. The notable exception was *uidA*, which failed to amplify in 25 strains, including a single O114:H2 and a O76:H51 strain, all 11 O55:[h51] strains, and all 12 *E. albertii* strains. One of the *E. albertii* strains was also PCR-negative for *lysP*. The *uidA* and *lysP* loci were treated as missing data and replaced with alignment gaps in the fully concatenated sequence for these 25 strains. For phylogenetic analysis, the sequenced internal fragments of the 7 housekeeping genes were concatenated to yield 3,732 nucleotides. MLST analysis resolved an average of 78.4 variable nucleotide sites per locus, which defined a number of alleles, ranging from 30 to 45, at the 7 MLST genes (Table 13). The distinct combinations of alleles across the 7 MLST loci were used to define 103 multilocus genotypes or sequence types (STs) among the 324 strains (Figure 12).

The synonymous rate of substitution (d_S) ranged from a low of 5.52% for *uidA* to a high of 22.66% for *fadD* with an average of 11.33 synonymous substitutions per 100 synonymous sites (Table 13). The nonsynonymous rate (d_N) per 100 nonsynonymous sites was generally two orders of magnitude lower than d_S, ranging from 0.07% for *clpX* to 0.58% for *uidA*. Tests for natural selection operating on the allelic variation at each MLST locus based on the SLAC and REL methods revealed that, on average, most (76.0%) of the codons analyzed are evolving neutrally with no significant difference between the levels of synonymous and nonsynonymous substitutions (Table 14). The number of negatively selected codons ranged from a low of 2.9% for *aspC* to a high of 49.1% for *fadD* with an average of 24.0%. No codons in any of the MLST loci were

Table 13. Sequence variation among alleles of 7 MLST genes.

Locus	# of sites	# of variable sites	# of alleles	d_S × 100	d_N × 100
aspC	513	80	37	12.80	0.30
clpX	567	73	37	9.55	0.07
fadD	483	114	36	22.66	0.35
icdA	567	59	39	8.13	0.16
lysP	477	70	30	10.23	0.41
mdh	549	86	45	10.43	0.12
uidA	576	67	40	5.52	0.58
Avg.	533.1	78.4	37.7	11.33	0.28

Table 14. Summary of recombination and selection analyses for 7 MLST genes.

Locus	Overall compatibility	Global fragments[a]	GARD breakpoints	Neutral codons	Negative codons	Positive codons
aspC	94.4%	0	0	97.1%	2.9%	0.0%
clpX	77.4%	0	0	70.4%	29.6%	0.0%
fadD	81.3%	0	0	50.9%	49.1%	0.0%
icdA	70.0%	0	0	78.8%	21.2%	0.0%
lysP	92.6%	0	0	76.7%	23.3%	0.0%
mdh	94.9%	0	0	77.0%	23.0%	0.0%
uidA	83.3%	0	0	81.3%	18.8%	0.0%
Avg.	84.8%	0	0	76.0%	24.0%	0.0%

[a] Number of significant global inner fragments identified by Sawyer's test.

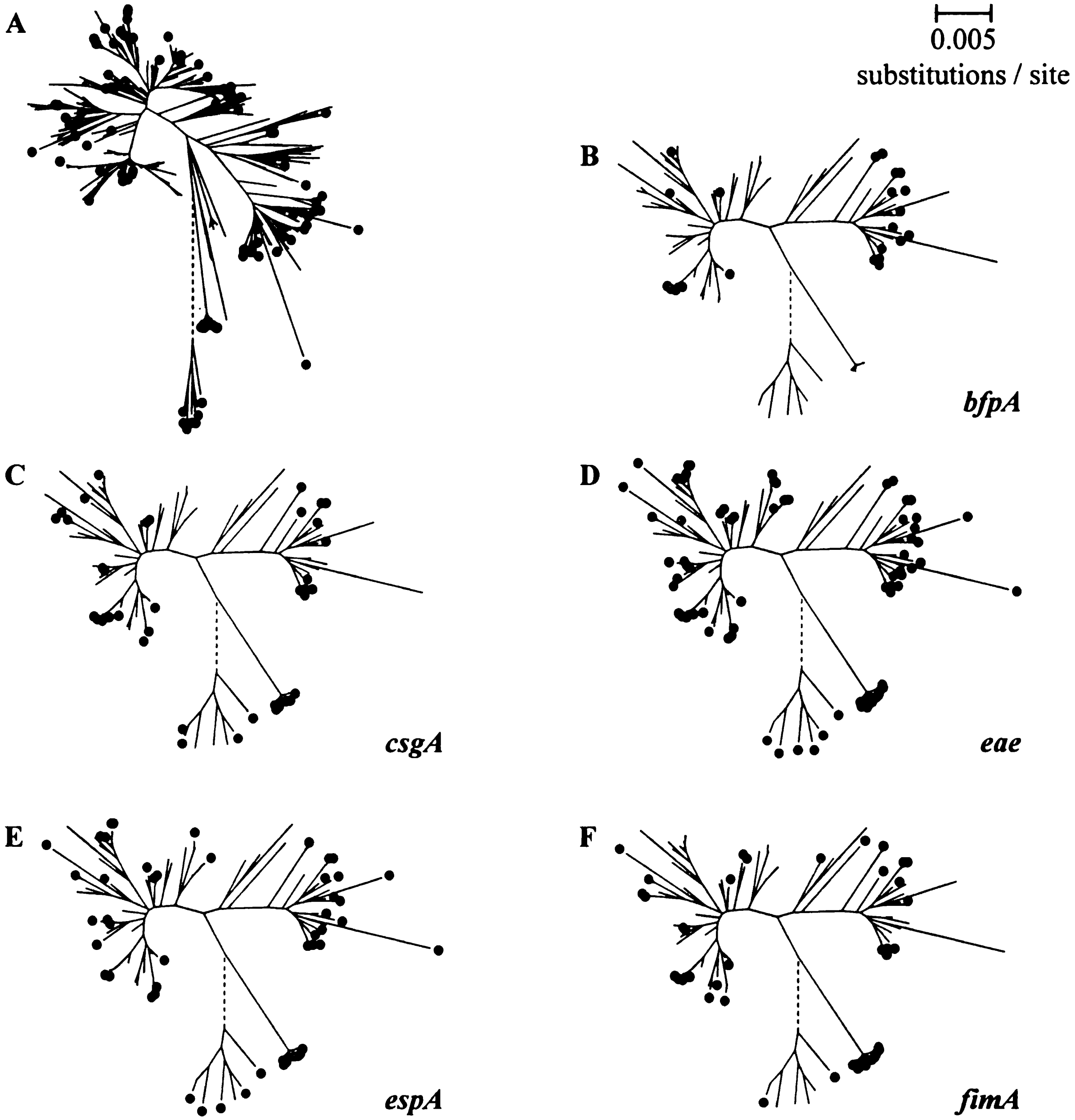

Figure 12. Distribution of allelic data among sequence types (STs). A) 500 STs showing the diversity of *E. coli* and *E. albertii.* The 103 STs investigated are marked with gray circles. The branch between *E. coli* and *E. albertii* is represented by the dashed line (length = 0.063 substitutions/site). Panels B through F show the distributions of allelic data obtained for *bfpA* (B, 22 STs), *csgA* (C, 39 STs), *eae* (D, 74 STs), *espA* (E, 50 STs), and *fimA* (F, 44 STs) among the 103 STs indicated in panel A.

found to be under significant positive selection. Thus, low values of d_N/d_S at the MLST loci reflect weak negative selection over many codons.

Compatibility analysis indicated that, on average, 84.8% of the pairwise comparison of parsimoniously informative sites in the 7 MLST genes have compatible phylogenies, suggesting that recombination or parallel mutation may have played a small role in the allelic diversification at these loci. The results of Sawyer's test and GARD appear to support the possibility of parallel mutation since neither method found evidence of significant intragenic recombination in any of the 7 MLST loci (Table 14).

Allelic variation in *bfpA*. Of the 324 strains investigated, *bfpA* allelic data was obtained for 108 isolates through a combination of DNA sequencing and RFLP analysis. Sequence analysis revealed the presence of 11 distinct alleles of *bfpA*, which are defined by 99 polymorphic nucleotide sites within the 588-bp gene. Phylogenetic compatibility analysis with 72 parsimoniously informative sites resulted in an overall compatibility of 92.6%. No globally significant inner fragments were detected by Sawyer's test; however, GARD identified a potential recombinational breakpoint at codon 150.

In comparison to the MLST loci from the same 108 strains, the synonymous rate for *bfpA* was 9.22%; slightly greater than the range (3.44 – 8.56%) and mean (5.41%) for the 7 MLST genes. The nonsynonymous rate of 6.38% for *bfpA* was more than 40 times greater than the average d_N across all 7 MLST genes (0.15%). The proportions of nonsynonymous and synonymous codon changes (p_N and p_S, respectively) were calculated within a 10-codon sliding window to determine if regions of *bfpA* are experiencing different selective pressures (Figure 13A). The level of selective constraint appears to vary along the length of *bfpA*, for most of the gene $p_S > p_N$, but

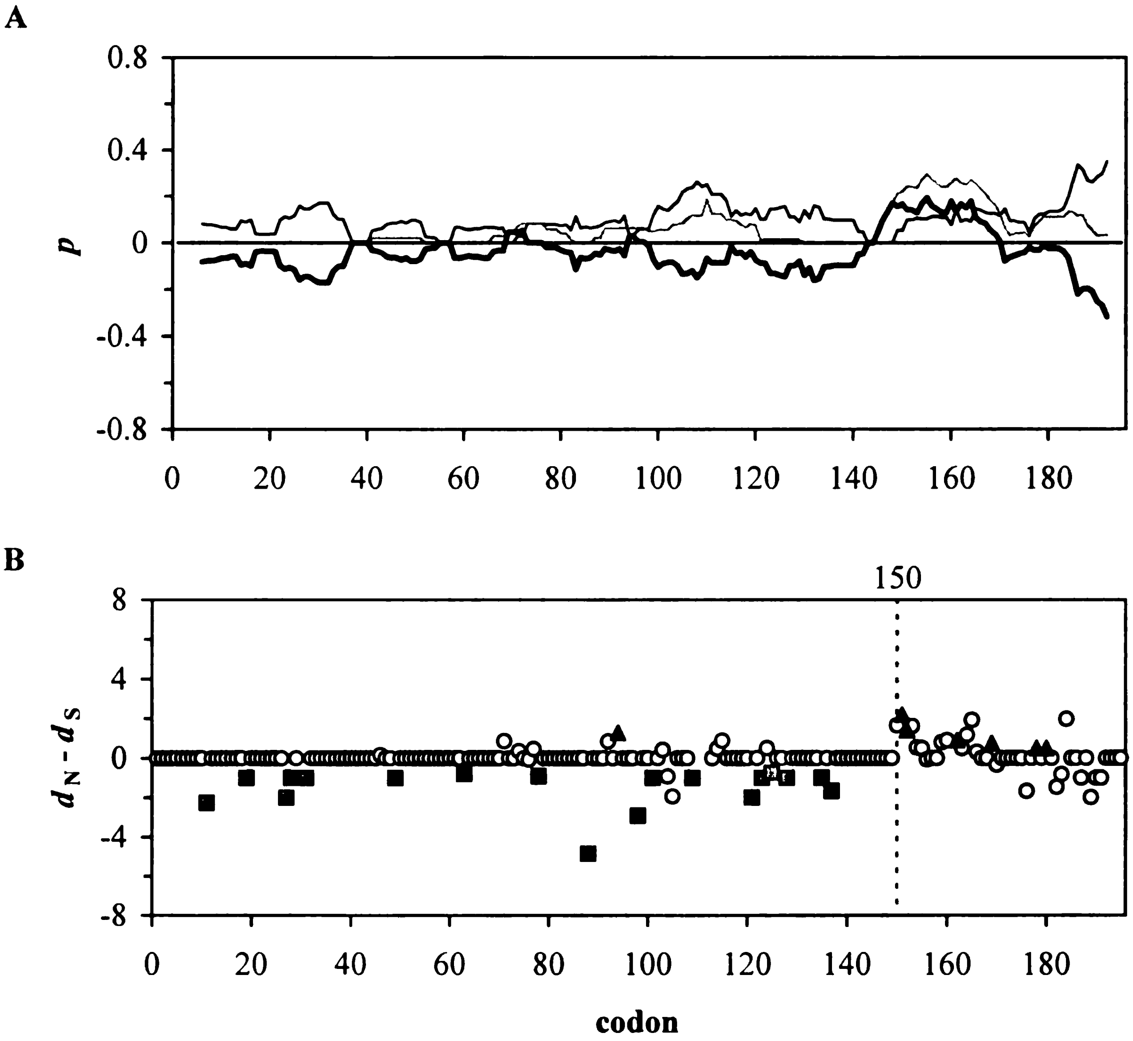

Figure 13. Selection and recombination in *bfpA*. A) Sliding window analysis depicting changes in p_S (black line), p_N (gray line), and $p_N - p_S$ (heavy black line) along the length of *bfpA* using a 10-codon window. B) Codons under significant positive (triangles, n=7) or negative (squares, n=19) selection are plotted using the $d_N - d_S$ values as determined by the SLAC method. Sites found to be significant by SLAC and REL are shaded black, while sites significant under REL only are gray. Codons evolving neutrally are indicated by the white circles. The recombination breakpoint identified by GARD at codon 150 is indicated by the dashed line.

codons 145-170 are distinct in that $p_N > p_S$. When the regions on either side of the recombinational breakpoint identified by GARD were analyzed separately for the action of selection on individual codons, 19 codons were identified as being under significant negative selection, while 7 codons appear to be under positive selection (Figure 13B).

Allelic variation in *csgA*. Of the 324 strains investigated, *csgA* allelic data was obtained for 70 isolates. Sequence analysis revealed the presence of 12 distinct alleles, which are defined by 73 polymorphic nucleotide sites within the 459-bp gene. One of the *E. albertii* strains investigated (K-1) possessed an unusual *csgA* allele. This allele, designated as allele 12, was unique in that it is 135-bp shorter than the other alleles due to a large in-frame deletion. In addition to the deletion, allele 12 also possesses a premature stop codon because of a C → T transition at position 418 in the alignment. Since some of the methods require the removal of alignment gaps, allele 12 was excluded from the recombination and selection analyses because the removal of the 135-bp from the other alleles may lead to inaccurate results. Phylogenetic compatibility analysis with 41 parsimoniously informative sites within the remaining 11 alleles resulted in an overall compatibility of 82.7%. Two globally significant inner fragments were detected by Sawyer's test, both of which were located between codons 16 and 57 and GARD identified a potential recombinational breakpoint within the same region of the gene at codon 53.

In comparison to the MLST loci from 69 of the 70 *csgA* strains (*E. albertii* strain K-1 was excluded), the synonymous rate for *csgA* was 20.02%; within the range (5.19 – 28.65%) but greater than the mean (12.65%) for the 7 MLST genes. The nonsynonymous rate of 1.54% for *csgA* was approximately 5 times greater than the average d_N across all 7 MLST genes (0.30%). Sliding window analysis revealed that $p_S > p_N$ for most of the gene, but a small region centered near codon 20 was distinct in that $p_N > p_S$. (Figure 14A) This region along with 3 others was characterized by an increase in p_N above zero. The codon selection analyses identified 42 codons under significant negative selection and 2 codons under positive selection (Figure 14B).

Allelic variation in *eae*. Of the 324 strains investigated, *eae* allelic data was obtained for 269 isolates through a combination of DNA sequencing and fRFLP analysis (79). Among these isolates, 21 of the 28 major allelic variants of *eae* were observed, and representative sequences of each allele were chosen for subsequent analyses. Sequence analysis of the 21 *eae* alleles identified 1009 polymorphic nucleotide sites within the ~2.8-kb gene. Of these polymorphic sites, 773 (76.6%) are located in the 3' end of the gene, which encodes the extracellular domains of intimin. Because of this large difference in the distribution of polymorphic sites, *eae* was divided into intracellular (periplasmic and transmembrane) and extracellular regions.

In comparison to the MLST loci from the same 269 strains, the synonymous rate for the intracellular portion of the 21 *eae* alleles was 12.57%; within the range (5.31 – 25.57%) and slightly greater than the mean (12.43%) for the 7 MLST genes. The nonsynonymous rate of 1.93% was approximately 6 times greater than the average d_N across all 7 MLST genes (0.31%). As expected, the extracellular region of *eae* was quite different from the periplasmic and transmembrane domains. With a synonymous rate of 61.39% and a nonsynonymous rate of 26.18%, the extracellular domains possess roughly 5 times as many synonymous and approximately 85 times as many nonsynonymous substitutions per site as the average values for the 7 MLST loci.

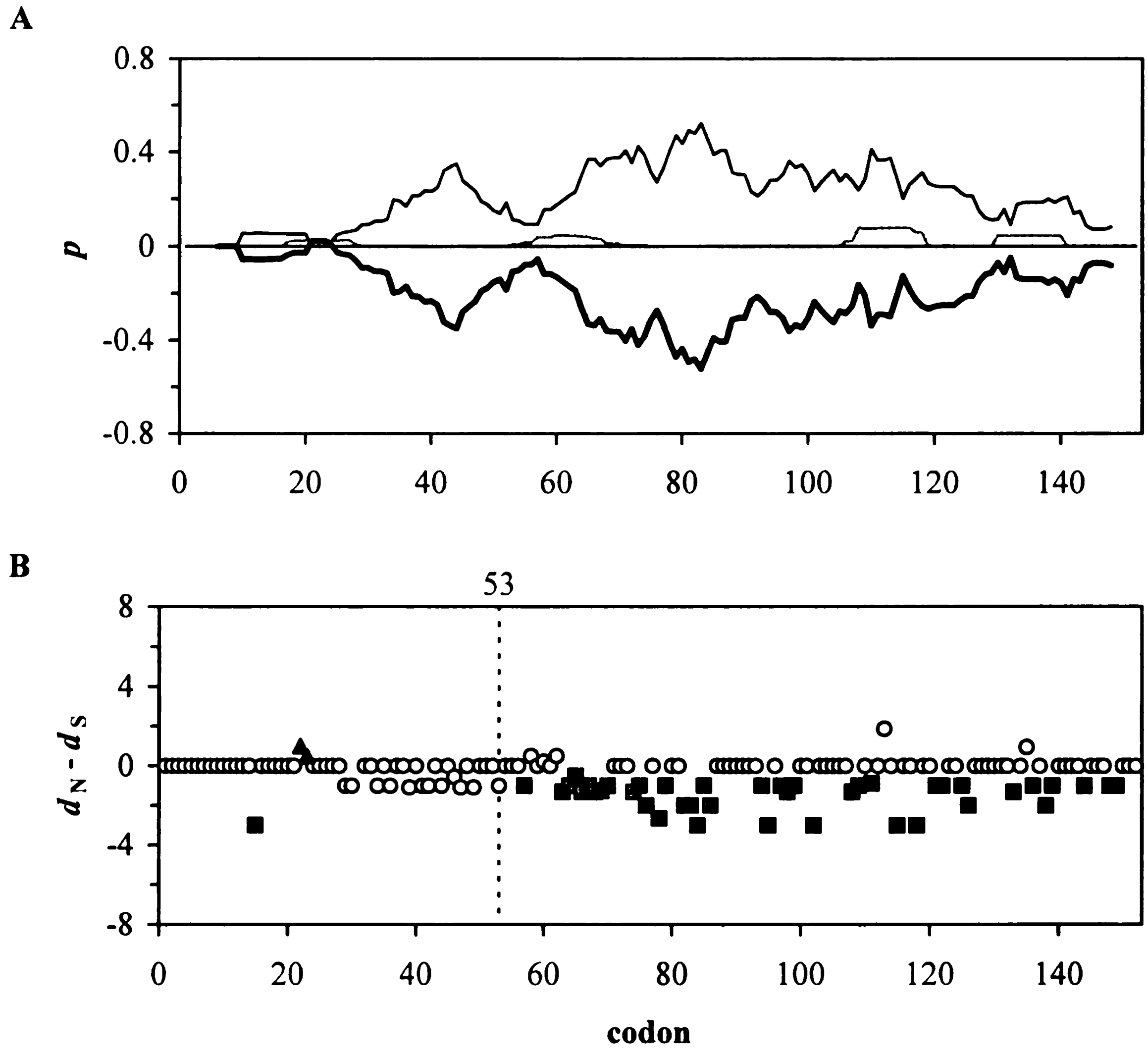

Figure 14. Selection and recombination in *csgA*. A) Sliding window analysis depicting changes in p_S (black line), p_N (gray line), and $p_N - p_S$ (heavy black line) along the length of *csgA* using a 10-codon window. B) Codons under significant positive (triangles, n=2) or negative (squares, n=42) selection are plotted using the $d_N - d_S$ values as determined by the SLAC method. Sites found to be significant by SLAC and REL are shaded black, while sites significant under REL only are gray. Codons evolving neutrally are indicated by the white circles. The recombination breakpoint identified by GARD at codon 53 is indicated by the dashed line.

As mentioned previously, 21 of the 28 major allelic variants of *eae* were observed among the strains investigated. When all 28 variants were analyzed, an additional 116 polymorphic sites were identified; all selection and recombination analyses were performed on these 28 allelic variants. Phylogenetic compatibility analysis with 860 parsimoniously informative sites from the 28 alleles resulted in an overall compatibility of 29.8%. When *eae* was divided into the intracellular and extracellular regions, the periplasmic and transmembrane portion had an observed compatibility of 71.3% while the extracellular domains had a compatibility of 30.4%. Since the compatibility analysis indicates strong support for parallel mutation and/or recombination within the extracellular domains as opposed to the intracellular region of the gene, the regional division of *eae* was maintained in the remaining recombination and selection analyses.

Intracellular region of *eae*. Eleven globally significant inner fragments were detected within the intracellular region of *eae* by Sawyer's test, all of which were located either between codons 107 and 345 or between codons 352 and 507. GARD identified two potential recombinational breakpoints, one at codon 173 and a second at codon 351. Sliding window analysis indicated that p_S is greater than p_N over the entire region (Figure 15A). A slight increase in p_N was observed over the first 175 codons, which corresponds to the periplasmic domain of intimin. Codon selection analysis identified 119 negatively selected codons and 2 positively selected codons (Figure 15B).

Extracellular region of *eae*. Sawyer's test identified 237 significant global inner fragments ranging in length from 51 to 973 bp. These fragments tended to cluster into one of three overlapping regions within the extracellular domains (codons 551 to 697, codons 648 to 885, and codons 849 to 947). GARD identified potential recombinational

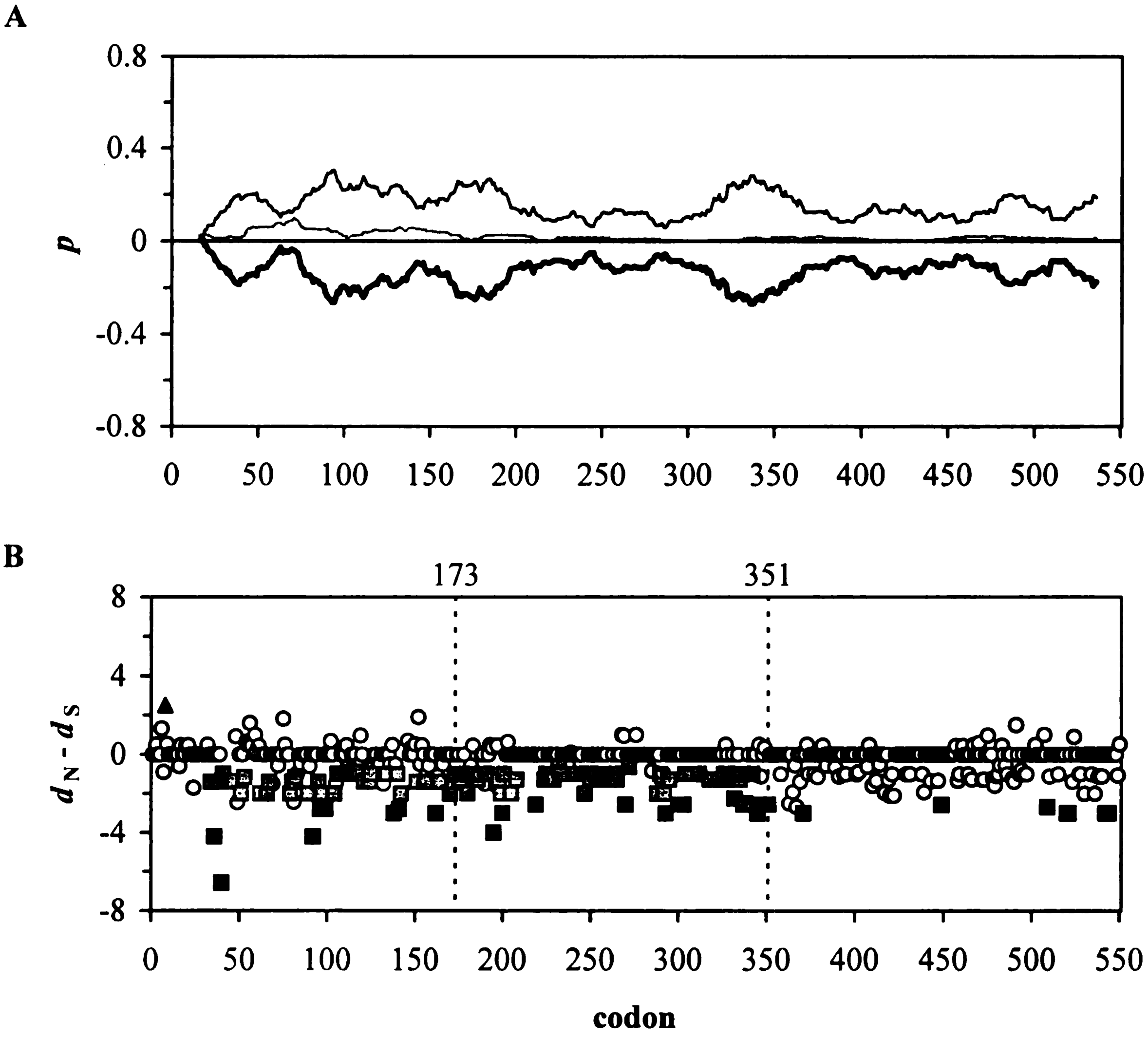

Figure 15. Selection and recombination in the periplasmic (PPD) and transmembrane (TMD) domains of *eae*. A) Sliding window analysis depicting changes in p_S (black line), p_N (gray line), and $p_N - p_S$ (heavy black line) along the length of $eae_{PPD/TMD}$ using a 30-codon window. B) Codons under significant positive (triangles, n=2) or negative (squares, n=119) selection are plotted using the $d_N - d_S$ values as determined by the SLAC method. Sites found to be significant by SLAC and REL are shaded black, while sites significant under REL only are gray. Codons evolving neutrally are indicated by the white circles. The recombination breakpoints identified by GARD at codons 173 and 351 are indicated by the dashed lines.

breakpoints at codons 598, 656, 689, 743, and 829. Sliding window analysis indicated that, similar to the intracellular region of *eae*, p_S is greater than p_N over the entire extracellular region (Figure 16A). An increase in both p_S and p_N was observed at approximately codon 675, which corresponds to the immunoglobulin-like D1 domain of intimin. Codon selection analysis identified 213 negatively selected codons and 8 positively selected codons (Figure 16B). Six of the 8 positively selected codons are located within the lectin-like D3 domain of intimin, which interacts with the translocated intimin receptor (Tir). In addition to the 8 positively selected sites, the D3 domain features a pair of cysteine residues located at positions 865 and 948 in the alignment. These two cysteine residues are under significant negative selection (Figure 16B), suggesting that they may play a role in stabilizing the affinity of the intimin-Tir interaction as suggested by the structural model of Luo et al. (89).

Allelic variation in *espA*. Of the 324 strains investigated, *espA* allelic data was obtained for 91 isolates. Sequence analysis revealed the presence of 31 distinct alleles, which are defined by 241 polymorphic nucleotide sites within the 579-bp gene. Phylogenetic compatibility analysis with 214 parsimoniously informative sites resulted in an overall compatibility of 74.5%. Neither Sawyer's test nor GARD found evidence of significant intragenic recombination, suggesting that the phylogenetic incompatibilities could be due to parallel mutation or ancient recombination events that have since been obscured by subsequent mutation.

In comparison to the MLST loci from the same 91 strains, the synonymous rate for *espA* was 39.22%, which is greater than the range (5.20 – 26.64%) and mean (13.10%) for the 7 MLST genes. The nonsynonymous rate of 11.93% for *espA* was more

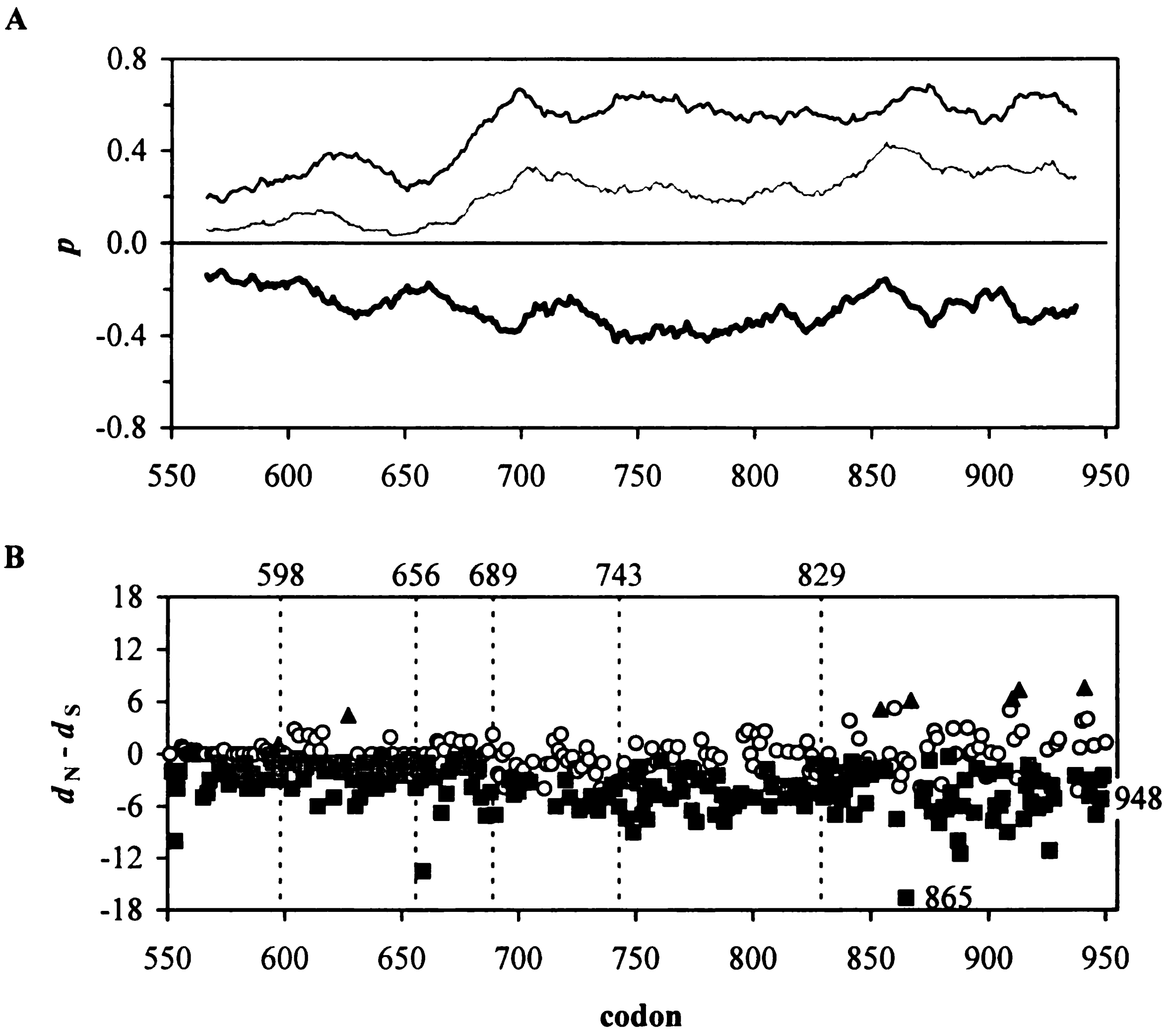

Figure 16. Selection and recombination in the extracellular domains (ECD) of *eae*. A) Sliding window analysis depicting changes in p_S (black line), p_N (gray line), and $p_N - p_S$ (heavy black line) along the length of *eae*$_{\text{ECD}}$ using a 30-codon window. B) Codons under significant positive (triangles, n=8) or negative (squares, n=213) selection are plotted using the $d_N - d_S$ values as determined by the SLAC method. Sites found to be significant by SLAC and REL are shaded black, while sites significant under REL only are gray. Codons evolving neutrally are indicated by the white circles. The recombination breakpoints identified by GARD at codons 598, 656, 689, 743, and 829 are indicated by the dashed lines. Cysteine residues are located at codons 865 and 948.

than 30 times greater than the average d_N across all 7 MLST genes (0.35%). Sliding window analysis revealed that $p_S > p_N$ for most of the gene, but a small region before codon 20 was distinct in that $p_N > p_S$. In addition, a broad region between codons 80 and 140 was characterized by a greater increase in p_N when compared to the rest of the gene (Figure 17A). When the alleles were analyzed for the action of selection on individual codons, 81 codons were found to be under significant negative selection and 6 codons are under positive selection. Of the positively selected sites, 5 are within the broad region with increased p_N identified in the sliding window analysis (Figure 17B).

Allelic variation in *fimA*. Of the 324 strains investigated, *fimA* allelic data was obtained for 133 isolates. Sequence analysis revealed the presence of 32 distinct alleles, which are defined by 156 polymorphic nucleotide sites within the 555-bp gene. Phylogenetic compatibility analysis with 134 parsimoniously informative sites resulted in an overall compatibility of 69.9%. Nineteen globally significant inner fragments were detected by Sawyer's test, all of which were located either between codons 1 and 74 or between codons 111 and 172. GARD identified a potential recombinational breakpoint within the second region of the gene at codon 112.

In comparison to the MLST loci from the same 133 strains, the synonymous rate for *fimA* was 19.99%, within the range (5.73 – 23.21%) but greater than the mean (10.51%) for the 7 MLST genes. The nonsynonymous rate of 4.84% for *fimA* was almost 20 times greater than the average d_N across all 7 MLST genes (0.26%). Sliding window analysis revealed that, with the exception of a small region centered at codon 24 where p_N is slightly greater than p_S, the difference between p_N and p_S is less than zero over the entire length of the gene (Figure 18A). However, five regions of the gene were

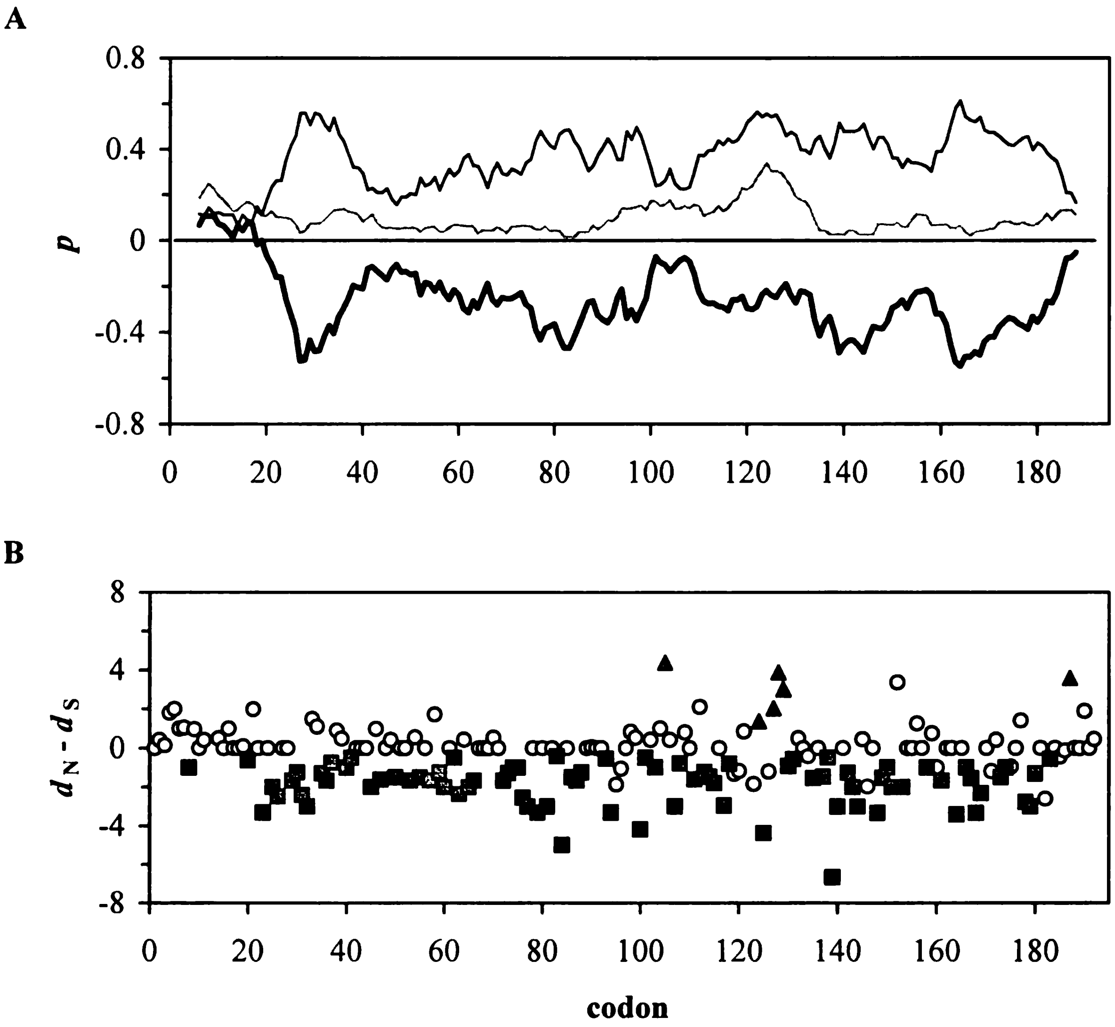

Figure 17. Selection in *espA*. A) Sliding window analysis depicting changes in p_S (black line), p_N (gray line), and $p_N - p_S$ (heavy black line) along the length of *espA* using a 10-codon window. B) Codons under significant positive (triangles, n=6) or negative (squares, n=81) selection are plotted using the $d_N - d_S$ values as determined by the SLAC method. Sites found to be significant by SLAC and REL are shaded black, while sites significant under REL only are gray. Codons evolving neutrally are indicated by the white circles. No recombination breakpoints were identified by GARD.

characterized by an increase in p_N above zero. The codon selection analysis identified 60 negatively selected codons and 2 positively selected codons (Figure 18B).

Selection and hydrophobicity. Amino acid hydrophobicity was assessed to determine if any significant differences exist in the amino acids found at codons evolving under different selective pressures (neutral, positive, or negative). For the surface protein-encoding genes, significantly more hydrophilic residues were found at sites under positive selection (χ^2=3.898, df=1, p=0.048) than expected from the observed proportions over all sites. Similarly, significantly more hydrophobic residues were found at sites under negative selection (χ^2=9.039, df=1, p=0.003) than expected. No significant difference was observed for the neutrally selected sites (χ^2=2.030, df=1, p=0.154). Analysis of the housekeeping loci used in MLST revealed no significant differences for either the negatively selected sites (χ^2=0.015, df=1, p=0.901) or neutral sites (χ^2=0.004, df=1, p=0.947).

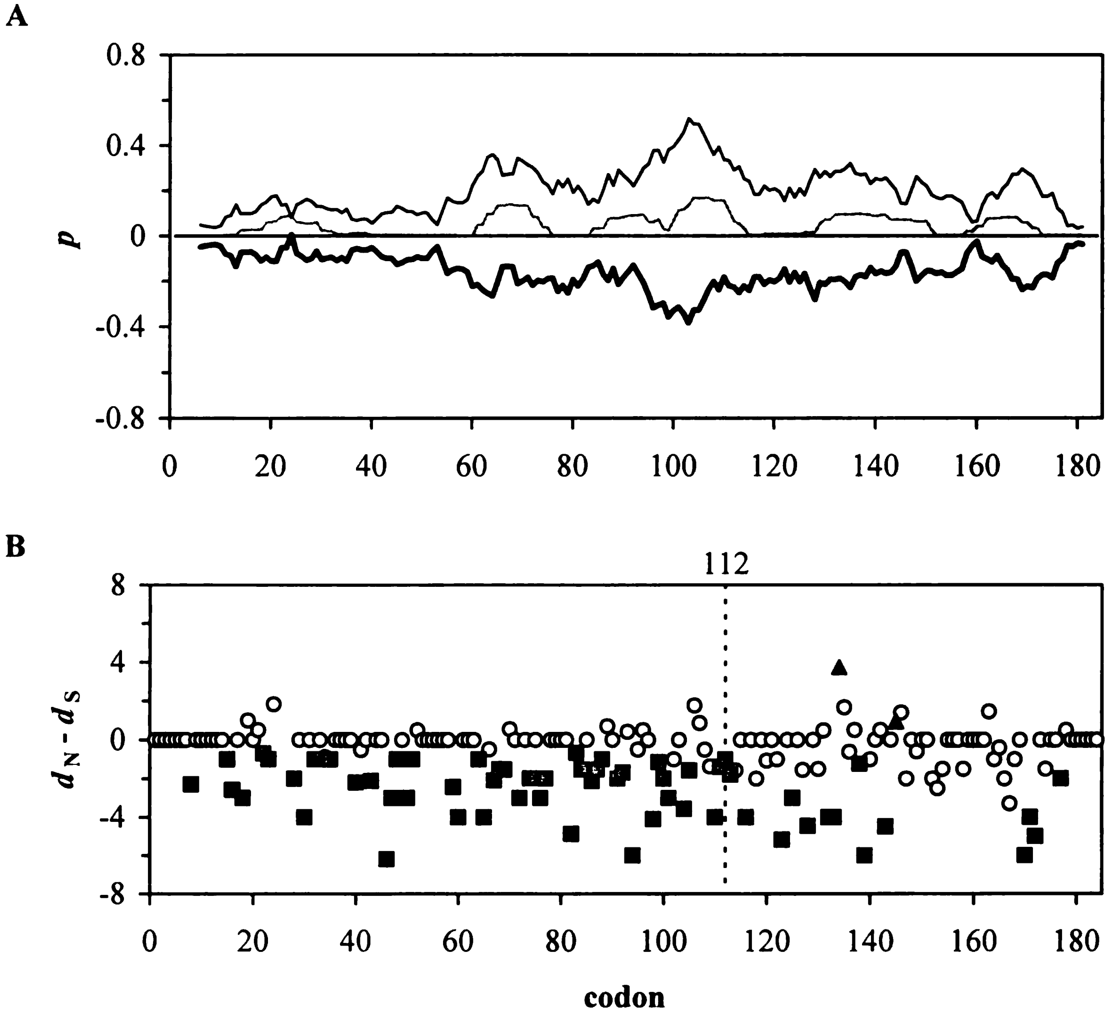

Figure 18. Selection and recombination in *fimA*. A) Sliding window analysis depicting changes in p_S (black line), p_N (gray line), and $p_N - p_S$ (heavy black line) along the length of *fimA* using a 10-codon window. B) Codons under significant positive (triangles, n=2) or negative (squares, n=60) selection are plotted using the $d_N - d_S$ values as determined by the SLAC method. Sites found to be significant by SLAC and REL are shaded black, while sites significant under REL only are gray. Codons evolving neutrally are indicated by the white circles. The recombination breakpoint identified by GARD at codon 112 is indicated by the dashed line.

DISCUSSION

In the work presented here, five genes encoding surface proteins involved in the attachment of *E. coli* during different types of infection or survival in the external environment were examined for evidence of positive selection and homologous recombination. The genes investigated were quite variable in the amount of recombination detected in the allelic sequences ranging from no recombination (*espA*) to extensive recombination (*eae*) (Table 15). With the exception of *csgA*, structural protein models have been reported for the genes analyzed allowing inferences to be drawn concerning the residues under positive selection.

bfpA. Blank et al. (11) reported a theoretical three-dimensional structure model for the bundlin encoded by the α1 *bfpA* allele. All 6 positively selected residues located in the 3' half of the gene map to the convex face of the pilin head, which is predicted to be located on the surface of the assembled pilus. The remaining positively selected residue is located along the edge of the pilin head and may also be surface-exposed. Interestingly, the two cysteine residues involved in disulfide bridge formation and stability of the mature bundlin protein appear to be evolving neutrally. However, since the codons are completely conserved across all the alleles, the lack of any synonymous substitutions may reflect a limitation of the codon selection method used here to distinguish between negative and neutral selection in codons that lack sequence variation.

csgA. For *csgA*, the pattern of codon selection was quite different from the other genes analyzed. Little variation was observed among the allelic sequences with the level of sequence polymorphism more similar to the MLST loci than the other surface protein-encoding genes. The two codons identified as being under positive selection are

Table 15. Summary of recombination and selection analyses.

Locus	Overall compatibility	Global regions [a]	GARD breakpoints	Neutral codons	Negative codons	Positive codons
bfpA	92.6%	0	1	86.4%	9.9%	3.7%
csgA	82.7%	1	1	70.9%	27.8%	1.3%
eae$_{\text{PPD/TMD}}$	71.3%	2	2	77.9%	21.8%	0.4%
eae$_{\text{ECD}}$	30.4%	3	5	41.2%	56.6%	2.1%
espA	74.5%	0	0	54.0%	42.9%	3.2%
fimA	69.9%	2	1	65.7%	33.1%	1.1%

[a] Number of regions within the gene in which the significant global inner fragments identified by Sawyer's test tended to cluster.

completely conserved among the *E. coli* alleles examined with the sequence differences at these positions due to the *E. albertii* strains. The two residues may be indicative of weak positive selection operating between the two species. However, analysis of the allelic sequences used here with the addition of *csgA* sequences from *Citrobacter* sp. (GenBank accession numbers AJ515700 and AJ515701) and *Salmonella enterica* (GenBank accession numbers NC_003197 and NC_003198) classified these two sites as evolving neutrally. The conflicting findings indicate that the selection results must be carefully inspected before conclusions about their importance can be drawn.

eae. Luo et al. (89) described the crystal structure of the extracellular domains of α intimin and focused on the interaction between the D3 domain of intimin and the translocated intimin receptor (Tir). Of particular interest is the positively charged intimin tip (residues 909-914 in the alignment used here) that directly interacts with Tir through hydrogen bonding. Five of the 6 residues within the tip have positive d_N-d_S values with the codons at positions 910 and 913 under significant positive selection according to the more conservative SLAC method. This is somewhat surprising as one might expect this region to be under negative selection so as to preserve the interaction with Tir. Variation within the tip may explain reports of intimin binding to receptors other than Tir (39).

In addition to the D3-Tir interactions, Luo et al. also described an 8-residue linker region between the transmembrane domain and the D0 domain. They hypothesized that, since it contains a glycine near each end (positions 552 and 559), the region may function as a flexible hinge because of the conformational variability of these residues. Interestingly, these two glycines are completely conserved across all 28 *eae* alleles investigated. The glycine at position 552 is under significant negative selection according to the REL method, while the codon at position 559 is classified as neutral since it lacks any sequence variation among the alleles. This supports the hypothesis that the region serves some functional importance especially given the level of sequence polymorphism observed in this part of the gene.

espA. In 2003 Neves and colleagues (104) demonstrated that there was no immunological cross-reactivity between EspA filaments from EPEC 1 (serotype O127:H6) and EHEC 1 (serotype O157:H7) strains. More recently, Crepin et al. (24) expanded upon this finding by elucidating the molecular basis of the antigenic polymorphism in these two distinct alleles of EspA. They identified a short hypervariable region of the protein located between residues 123 and 129 that when deleted did not affect filament biogenesis and function. In addition, peptide insertions into the hypervariable region were tolerated and displayed on the surface of the filament. By exchanging this surface-exposed hypervariable domain between an O127:H6 EPEC 1 strain and an O157:H7 EHEC 1 strain, they were able to swap the antigenic specificity of the EspA filaments. The hypervariable region described by Crepin et al. was easily identified in the *espA* alleles examined in this study. Of the 7 amino acids within this domain, 4 were found to be under significant positive selection, thereby adding further evidence to the immunological importance of this region. Another positively selected codon was identified at position 105 that, while not part of the dispensable hypervariable domain, may be surface-exposed given its close proximity to the region.

fimA. A 2001 study by Peek and colleagues (119) examined *fimA* sequences obtained from *E. coli* strains isolated from a broad range of host species for evidence of positive selection and recombination. They also performed a structural analysis of *fimA* based on homology to the pilin domain of *fimH*, the gene encoding the adhesive tip of the type 1 fimbriae. The authors identified 19 sites under positive selection within their *fimA* sequences by using the likelihood method of Nielsen and Yang (108). However, since this method does not allow for variation in synonymous substitution rates across sites, the validity of the Nielsen-Yang approach was recently called into question by Kosakovsky Pond and Frost (73). They discovered that if variation in both synonymous and nonsynonymous substitution rates is not taken into account, the results obtained by the Nielsen-Yang approach could be misleading. The REL method used here was developed by Kosakovsky Pond and Frost to address this issue in the Nielsen-Yang approach by allowing both d_S and d_N to vary across sites independently.

Reanalysis of the Peek et al. data using the methods employed here (recombination detection by GARD, followed by codon selection analysis with both SLAC and REL) identified 9 codons as having a positive value for d_N-d_S. All of these sites were described by Peek et al. as being under positive selection; however, none of the values were significant in the reanalysis. This suggests that site-to-site variation in the synonymous substitution rate could be responsible for the positive selection results reported by Peek et al.

ACKNOWLEDGEMENTS

I would like to thank Lindsey Ouellette for performing MLST on many of the non-EPEC strains. I would also like to thank Susan Francisco, Suzanne Henson, and Sara Kienzle for their assistance in the *fimA*, *espA*, and *csgA* sequencing projects, respectively. This project has been funded in part with Federal funds from the NIAID, NIH, DHHS, under NIH Research Contract # N01-AI-30058.

CHAPTER 6

SUMMARY AND SYNTHESIS

The overall purpose of the research presented in this dissertation is to examine the allelic diversity of genes that encode a variety of surface structures in different classes of pathogenic *E. coli*. Much of the work presented thus far has been primarily focused on the evolution of particular virulence factors. By combining this allelic data with a phylogenetic framework obtained from MLST analysis, insights into strain evolution may be achieved.

Compatibility analysis combined with phylogenetic networks. Phylogenetic incompatibilities can arise from either recombination or parallel mutation, both of which can obscure the evolutionary history of genetic information passed by vertical transmission within bacterial populations. To address this issue, phylogenetic compatibility analysis was performed as described in Chapter 5, with the modification that the least compatibles sites were sequentially removed until only the set of sites with complete compatibility was achieved (Table 16). This set of 100% compatible sites should more accurately represent the phylogenetic relationships of the loci under investigation.

For the two chromosomally encoded loci (*csgA* and *fimA*) this modified compatibility analysis was also performed on the sequence types resolved by MLST. Phylogenetic network analysis (see Chapter 3) was then performed using the compatible *csgA* or *fimA* sites with the MLST sites from the associated STs to identify potential lateral transfer events involving either of these two genes into a new chromosomal background. This type of analysis is the first of its kind to be performed with either of these two genes. The *E. coli csgA* sequences cluster into one of three groups with *E. albertii* as an outgroup (Figure 19). Of particular interest is the parallel path involving *E.*

Table 16. Summary of combined compatibility analyses.

Locus [a]	# of sites	# of BI sites [b]	# of BC sites [c]	# of BS sites [d]	Fraction compatible [e]	Fraction for network analysis [f]
csgA	456	33	25	28	75.8%	11.6%
csgA STs	3156	316	225	35	71.2%	8.2%
eae$_{\text{PPD-D0}}$	1962	234	81	164	34.6%	12.5%
eae$_{\text{D1-D3}}$	816	184	43	54	23.4%	11.9%
espA	570	137	77	24	56.2%	17.7%
fimA	549	97	51	16	52.6%	12.2%
fimA STs	3156	305	218	60	71.5%	8.8%

[a] ST, sequence type; PPD-D0, the portion of *eae* that encodes the periplasmic, transmembrane and D0 domains; D1-D3, the portion of *eae* that encodes the D1, D2, and D3 domains.

[b] BI, binary informative: sites possessing two nucleotides, each of which is found in at least two alleles.

[c] BC, binary compatible: the set of BI sites that are 100% compatible with each other.

[d] BS, binary singleton: sites possessing two nucleotides, one of which is found in only a single allele.

[e] The percentage of BI sites that are 100% compatible.

[f] The percentage of the total sites that are 100% compatible or binary singletons.

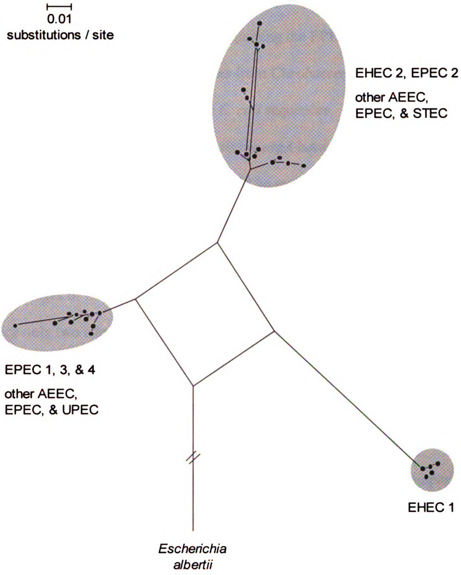

Figure 19. Phylogenetic network for *csgA* and its associated sequence types (STs). Three main clusters of *E. coli* are indicated by the gray ovals. Clonal groups and other relevant pathotypes found within each cluster are given. Multiple paths are indicative of phylogenetic incompatibilities between the STs and *csgA*. AEEC, attaching and effacing *E. coli*; EHEC, enterohemorrhagic *E. coli*; EPEC, enteropathogenic *E. coli*; STEC, Shiga toxin-producing *E. coli*; UPEC, uropathogenic *E. coli*.

albertii. MLST analysis places the EHEC 1 clonal group as more basal, while the *csgA* data places *E. albertii* closer to the cluster containing the EPEC 1, 3, and 4 clonal groups. In addition, the *csgA* analysis with sequences from *Citrobacter* and *Salmonella* places the *E. albertii* allele within the diversity of the *E. coli* sequences. Taken together, these findings suggest a lateral transfer event involving *csgA* between *E. coli* and *E. albertii* since the sequences are more similar than would be expected based on the MLST data. The results of the *fimA* comparisons are even more intriguing since numerous transfer events are apparent (Figure 20). With the exception of the EHEC 2 group, each of the two major lineages of EHEC and EPEC has experienced a lateral transfer event involving *fimA*. Since only two sites within *fimA* were found to be under significant positive selection, recombination and not point mutation appears to be the driving force in generating allelic diversity within clonal groups.

A slightly different approach was used for *eae* and *espA* since both are part of the LEE pathogenicity island, a known mobile genetic element. Instead of comparing each of these genes to their associated sequence types, they were compared to each other to identify potential recombination events resulting in the creation of new combinations of *eae* and *espA* alleles. The mosaic nature of *eae* has been previously described (95, 144), but over 20 additional major allelic variants of *eae* have been reported since these studies, so an updated analysis was warranted. In the previous analyses of *eae* presented here, the gene was partitioned on the basis of domain location (intracellular vs. extracellular). This division was also seen in the compatibility analysis, but the boundary was somewhat different. Closer inspection of the compatibility analysis indicates that the D0 domain is slightly more compatible with the periplasmic and transmembrane domains (41.4%) than

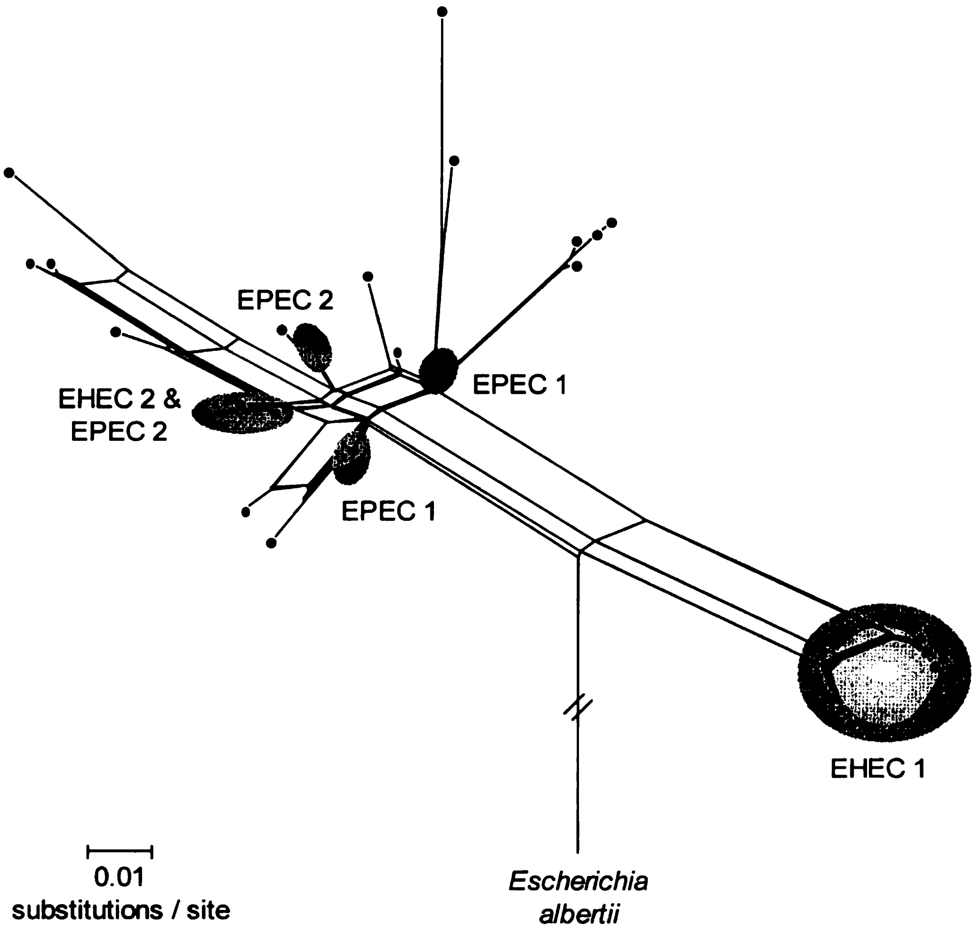

Figure 20. Phylogenetic network for *fimA* and its associated sequence types (STs). The locations of the two major clonal lineages of the enterohemorrhagic and enteropathogenic *E. coli* are indicated by the gray ovals. Multiple paths are indicative of phylogenetic incompatibilities between the STs and *fimA*. EHEC, enterohemorrhagic *E. coli*; EPEC, enteropathogenic *E. coli*.

it is with the other extracellular domains (37.1%). The resulting set of 100% compatible sites from each of these two regions revealed two very different evolutionary histories of *eae* (Figure 21). With the exception of θ-*eae*, the 5' region sequences from the alleles fell into one of three groups with the *Citrobacter rodentium* sequence as an outgroup.

Analysis of the 100% compatible sites from the 3' region revealed a different set of relationships with most of the sequences belonging to one of four major allelic families. These allelic families have been designated α, β, γ, and ε after the first reported member of each group. The extent of recombination between the two regions of *eae* is quite apparent in the network analysis (Figure 22). Since the 5' region is primarily intracellular, it is likely to be more representative of the evolutionary history of *eae* in the absence of selective pressure from the immune system. Each of the three groups contains members from different allele families; for example, Group 2 is comprised of members of the β, γ, and ε allele families. This indicates that recombination has played a large role in generating antigenic variation. Comparison of the set of 100% compatible *espA* sites to both *eae* regions further supports the hypothesis of extracellular domain exchange within *eae*. On the basis of location within the LEE island alone, one might expect *espA* to be more compatible with the 3' *eae* region given the shorter distance between them. However, the data indicate the contrary, *espA* is much more compatible with the 5' *eae* region (88.6%) than the 3' region (58.3%) (Figure 23). The three major groups of alleles identified in the 5' *eae* region are intact in the *espA* network analysis. In contrast, the relationships within each of the four allele families from the 3' *eae* region are broken with members of each family dispersed over the network. By replacing the extracellular

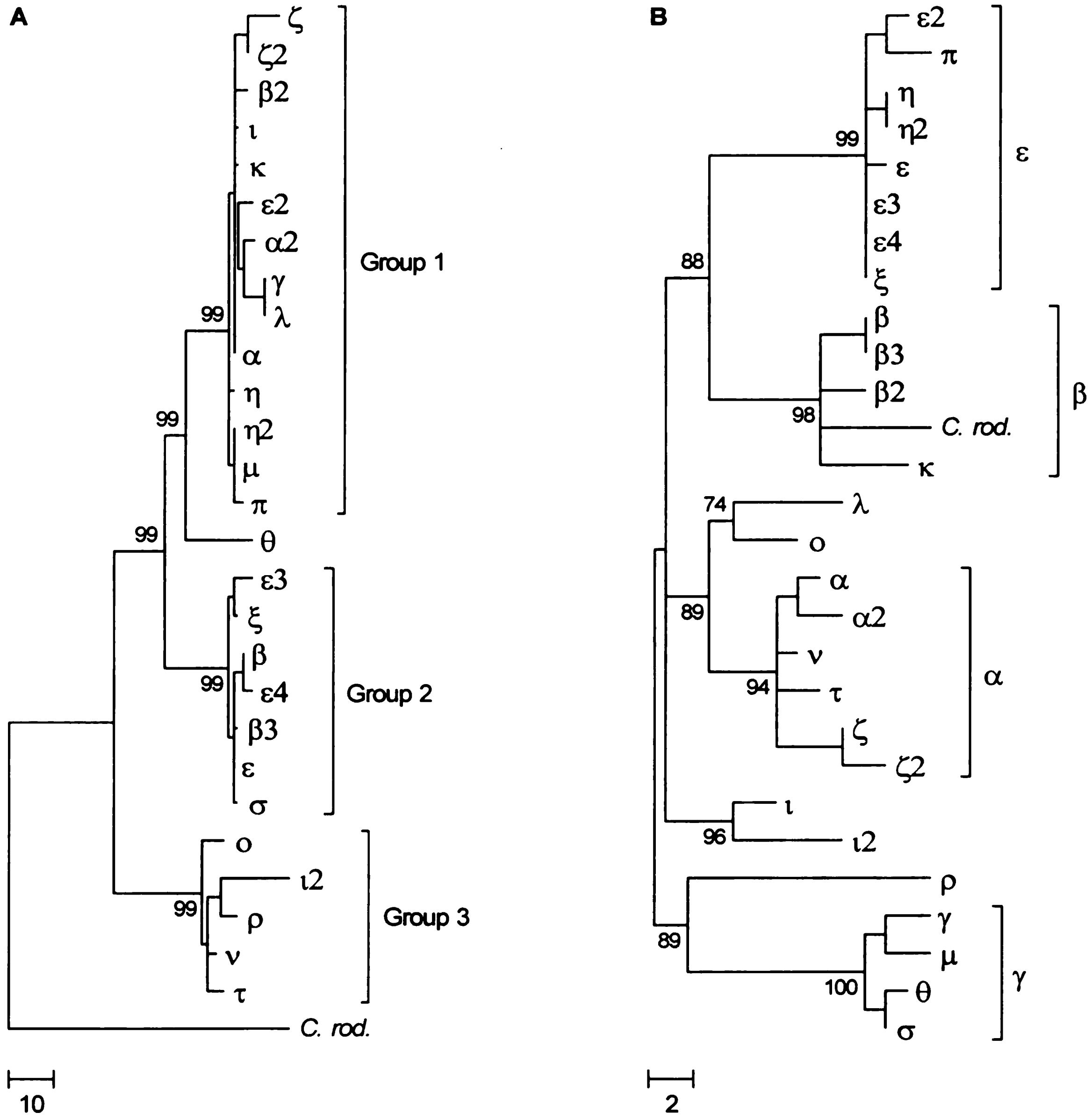

Figure 21. Conflicting evolutionary histories for two regions of *eae*. These minimum evolution trees were constructed using the number of nucleotide differences within the set of 100% compatible and binary singleton sites for the periplasmic, transmembrane, and D0 domains (A) and for the D1, D2, and D3 domains (B) of *eae*. Group or allele family designations are indicated by the square brackets. Bootstrap values for relevant clusters based on 500 replications are given at the internal nodes. *C. rod.*, *Citrobacter rodentium*

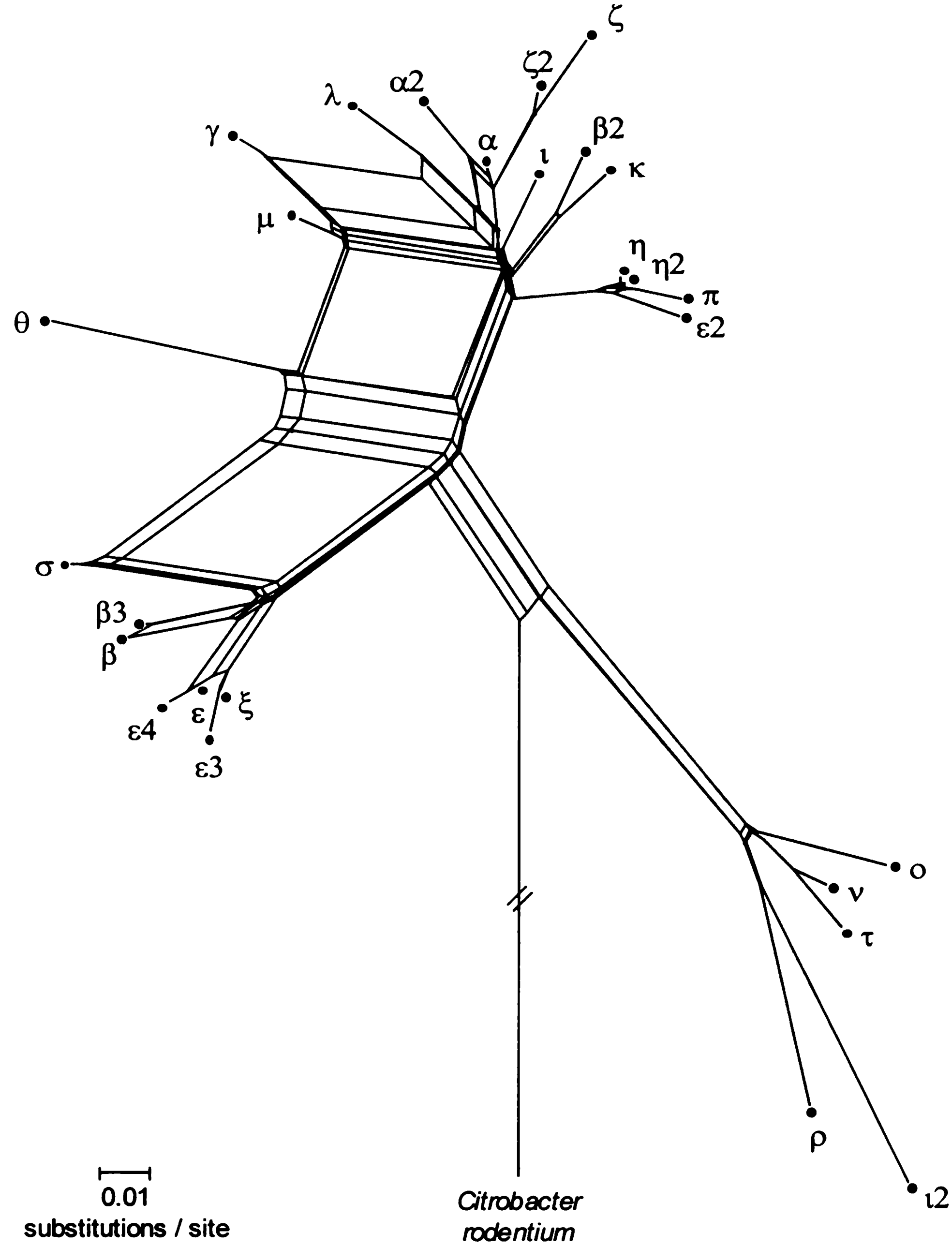

Figure 22. Phylogenetic network for *eae*. Extensive recombination between the 5' (periplasmic, transmembrane, and D0 domains) and 3' (D1, D2, and D3 domains) regions of *eae* is indicated by the multiple paths within the network.

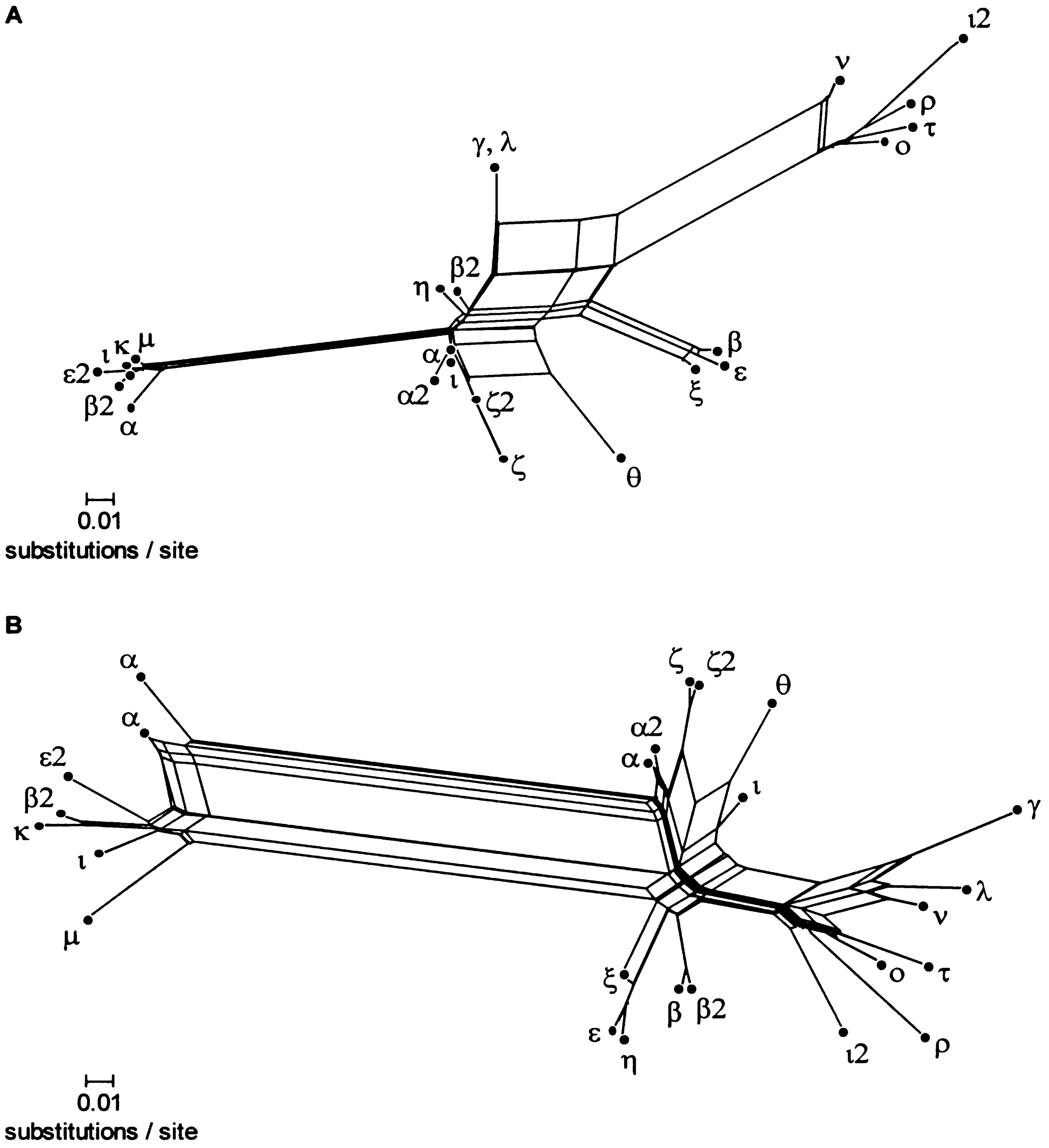

Figure 23. Phylogenetic networks for *espA* combined with *eae*. Phylogenetic incompatibilities identified between *espA* and the 5' *eae* region (A) and 3' *eae* region (B) are indicated by the multiple paths within the networks. Combinations of *eae* and *espA* are indicated by the black circles and labeled with their corresponding *eae* allele. Not all *eae* alleles are represented since they were not observed among the strains investigated.

domains from one allele family with those from another, a strain could gain a selective advantage in expressing new surface epitopes in an immunologically naïve host.

Inferred evolution of EAF plasmids and EPEC clones. The results of the EPEC study demonstrate the highly promiscuous nature of the EAF plasmid, so a different approach was used to investigate strain evolution. Based on the MLST and EAF plasmid type data, ancestral or primitive clonal types within the 4 main EPEC groups can be inferred under the parsimony principle, that is, positing a simple evolution model based on minimizing the number of evolutionary genetic events. The principal events in the evolutionary change of an ancestral EPEC clone are EAF plasmid recombination, plasmid replacement, and plasmid loss. Recombination of the EAF plasmid is defined as a change in *bfpA* allele since each *bfpA* allele is associated with a single *perA* allele class. Plasmid replacement, presumably resulting from the horizontal transfer of an EAF plasmid, is believed to have occurred when both *bfpA* and *perA* differ from the primitive condition. Plasmid loss is inferred when an isolate is PCR negative for both *bfpA* and *perA*.

Under the parsimony principle, the types of genetic events underlying the evolution of each of the EPEC groups have been deduced (Table 17). The EPEC 3 group is the most homogenous with only plasmid loss being inferred. Two of the three possible plasmid changes were found in the EPEC 1 and EPEC 4 clonal groups. For EPEC 1, the inferred ancestral type possessed EAF type 1 (α1-*bfpA*, α-*perA*). Plasmid loss was not observed, but plasmid recombination (EAF type 8: β5-*bfpA*, α-*perA*) and replacement (EAF type 6: β3-*bfpA*, γ-*perA*) have occurred. For EPEC 4, plasmid recombination was not observed, but plasmid loss and replacement were detected. EAF

Table 17. EAF plasmid changes within four common EPEC clonal groups.

Clonal group	Recombination (change in *bfpA*)	Plasmid replacement (change in *bfpA* & *perA*)	Plasmid loss
EPEC 1	EAF 1 to EAF 8	EAF 1 to EAF 6	not observed
EPEC 2	EAF 4 to EAF 2 or EAF 7	EAF 4 to EAF 11	observed
EPEC 3	not observed	not observed	observed
EPEC 4	not observed	EAF 3 to EAF 5 or EAF 8	observed

type 3 (α3-*bfpA*, β-*perA*) was replaced with EAF type 5 (β2-*bfpA*, δ-*perA*) and with EAF type 8 (β5-*bfpA*, α-*perA*). EPEC 2 is the most variable clonal group. With an inferred ancestral state of EAF type 4 (β1-*bfpA*, α-*perA*), all three types of plasmid changes were observed. Two different recombination events involving EAF type 2 (α2-*bfpA*, α-*perA*) and EAF type 7 (β4-*bfpA*, α-*perA*), plasmid replacement with EAF type 11 (*bfpA*$^-$, γ-*perA*), and plasmid loss have shaped the diversity of this clonal group.

Future considerations. The evolution of EPEC appears to be a dynamic process involving repeated acquisition of the LEE island and transfer of the EAF plasmid. The work presented here is the first to classify EAF plasmids into types based on *bfpA* and *perA* allelic data, and 11 distinct plasmid types were identified among the EPEC strains investigated. Nevertheless, it remains unclear what level of conservation exists among plasmids of the same EAF type. Given the number of IS elements present within the two fully sequenced EAF plasmids (15, 145), there could be considerable variation within each plasmid type and further characterization of the EAF types is warranted.

Three insertion sites of the LEE island have been identified within the AEEC, all of which are either within or near tRNA genes (*selC*, *pheU*, and *pheV*) (129, 137, 157). These studies have primarily focused on strains with the α, β, γ, ε, ζ, and θ, *eae* alleles, but a survey of the remaining alleles has not been performed. If additional insertion sites are uncovered, further insight into the number of transfer events of the LEE island into *E. coli* may be gained. The further characterization of pathogenic strains will improve our understanding of the processes that underlie microbial evolution. The identification of unique genetic determinants in these strains may then be used to facilitate the detection of specific epidemic clones during outbreaks of disease.

REFERENCES

1. **Abraham, J. M., C. S. Freitag, J. R. Clements, and B. I. Eisenstein.** 1985. An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. Proc Natl Acad Sci U S A **82:**5724-7.

2. **Adu-Bobie, J., G. Frankel, C. Bain, A. G. Goncalves, L. R. Trabulsi, G. Douce, S. Knutton, and G. Dougan.** 1998. Detection of intimins α, β, γ, and δ, four intimin derivatives expressed by attaching and effacing microbial pathogens. J Clin Microbiol **36:**662-8.

3. **Albert, M. J., K. Alam, M. Ansaruzzaman, J. Montanaro, M. Islam, S. M. Faruque, K. Haider, K. Bettelheim, and S. Tzipori.** 1991. Localized adherence and attaching-effacing properties of nonenteropathogenic serotypes of *Escherichia coli*. Infect Immun **59:**1864-8.

4. **Alikhani, M. Y., A. Mirsalehian, and M. M. Aslani.** 2006. Detection of typical and atypical enteropathogenic *Escherichia coli* (EPEC) in Iranian children with and without diarrhoea. J Med Microbiol **55:**1159-63.

5. **Beachey, E. H.** 1980. Bacterial adherence. Chapman and Hall, London ; New York.

6. **Bieber, D., S. W. Ramer, C. Y. Wu, W. J. Murray, T. Tobe, R. Fernandez, and G. K. Schoolnik.** 1998. Type IV pili, transient bacterial aggregates, and virulence of enteropathogenic *Escherichia coli*. Science **280:**2114-8.

7. **Blanco, M., J. E. Blanco, G. Dahbi, M. P. Alonso, A. Mora, M. A. Coira, C. Madrid, A. Juarez, M. I. Bernardez, E. A. Gonzalez, and J. Blanco.** 2006. Identification of two new intimin types in atypical enteropathogenic *Escherichia coli*. Int Microbiol **9:**103-10.

8. **Blanco, M., J. E. Blanco, G. Dahbi, A. Mora, M. P. Alonso, G. Varela, M. P. Gadea, F. Schelotto, E. A. Gonzalez, and J. Blanco.** 2006. Typing of intimin (*eae*) genes from enteropathogenic *Escherichia coli* (EPEC) isolated from children with diarrhoea in Montevideo, Uruguay: identification of two novel intimin variants (μB and ξR/β2B). J Med Microbiol **55:**1165-74.

9. **Blanco, M., J. E. Blanco, A. Mora, J. Rey, J. M. Alonso, M. Hermoso, J. Hermoso, M. P. Alonso, G. Dahbi, E. A. Gonzalez, M. I. Bernardez, and J. Blanco.** 2003. Serotypes, virulence genes, and intimin types of Shiga toxin (verotoxin)-producing *Escherichia coli* isolates from healthy sheep in Spain. J Clin Microbiol **41:**1351-6.

10. **Blank, T. E., D. W. Lacher, I. C. Scaletsky, H. Zhong, T. S. Whittam, and M. S. Donnenberg.** 2003. Enteropathogenic *Escherichia coli* O157 strains from Brazil. Emerg Infect Dis **9:**113-5.

11. **Blank, T. E., H. Zhong, A. L. Bell, T. S. Whittam, and M. S. Donnenberg.** 2000. Molecular variation among type IV pilin (*bfpA*) genes from diverse enteropathogenic *Escherichia coli* strains. Infect Immun **68:**7028-7038.

12. **Blomfield, I. C., P. J. Calie, K. J. Eberhardt, M. S. McClain, and B. I. Eisenstein.** 1993. Lrp stimulates phase variation of type 1 fimbriation in *Escherichia coli* K-12. J Bacteriol **175:**27-36.

13. **Blomfield, I. C., D. H. Kulasekara, and B. I. Eisenstein.** 1997. Integration host factor stimulates both FimB- and FimE-mediated site-specific DNA inversion that controls phase variation of type 1 fimbriae expression in *Escherichia coli.* Mol Microbiol **23:**705-17.

14. **Bortolini, M. R., L. R. Trabulsi, R. Keller, G. Frankel, and V. Sperandio.** 1999. Lack of expression of bundle-forming pili in some clinical isolates of enteropathogenic *Escherichia coli* (EPEC) is due to a conserved large deletion in the *bfp* operon. FEMS Microbiol Lett **179:**169-74.

15. **Brinkley, C., V. Burland, R. Keller, D. J. Rose, A. T. Boutin, S. A. Klink, F. R. Blattner, and J. B. Kaper.** 2006. Nucleotide sequence analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid pMAR7. Infect Immun **74:**5408-13.

16. **Broes, A., R. Drolet, M. Jacques, J. M. Fairbrother, and W. M. Johnson.** 1988. Natural infection with an attaching and effacing *Escherichia coli* in a diarrheic puppy. Can J Vet Res **52:**280-2.

17. **Bruen, T. C., H. Philippe, and D. Bryant.** 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics **172:**2665-81.

18. **Bryant, D., and V. Moulton.** 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol **21:**255-65.

19. **Bush, R. M.** 2001. Predicting adaptive evolution. Nat Rev Genet **2:**387-92.

20. **Castillo, A., L. E. Eguiarte, and V. Souza.** 2005. A genomic population genetics analysis of the pathogenic enterocyte effacement island in *Escherichia coli:* the search for the unit of selection. Proc Natl Acad Sci U S A **102:**1542-7.

21. **Cleary, J., L. C. Lai, R. K. Shaw, A. Straatman-Iwanowska, M. S. Donnenberg, G. Frankel, and S. Knutton.** 2004. Enteropathogenic *Escherichia coli* (EPEC) adhesion to intestinal epithelial cells: role of bundle-forming pili (BFP), EspA filaments and intimin. Microbiology **150:**527-38.

22. **Comeron, J. M.** 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol **41:**1152-9.

23. **Connell, I., W. Agace, P. Klemm, M. Schembri, S. Marild, and C. Svanborg.** 1996. Type 1 fimbrial expression enhances *Escherichia coli* virulence for the urinary tract. Proc Natl Acad Sci U S A **93:**9827-32.

24. **Crepin, V. F., R. Shaw, S. Knutton, and G. Frankel.** 2005. Molecular basis of antigenic polymorphism of EspA filaments: development of a peptide display technology. J Mol Biol **350:**42-52.

25. **Daniell, S. J., E. Kocsis, E. Morris, S. Knutton, F. P. Booy, and G. Frankel.** 2003. 3D structure of EspA filaments from enteropathogenic *Escherichia coli.* Mol Microbiol **49:**301-8.

26. **Deng, W., Y. Li, P. R. Hardwidge, E. A. Frey, R. A. Pfuetzner, S. Lee, S. Gruenheid, N. C. Strynakda, J. L. Puente, and B. B. Finlay.** 2005. Regulation of type III secretion hierarchy of translocators and effectors in attaching and effacing bacterial pathogens. Infect Immun **73:**2135-46.

27. **Denno, D. M., J. R. Stapp, D. R. Boster, X. Qin, C. R. Clausen, K. H. Del Beccaro, D. L. Swerdlow, C. R. Braden, and P. I. Tarr.** 2005. Etiology of diarrhea in pediatric outpatient settings. Pediatr Infect Dis J **24:**142-8.

28. **Donnenberg, M. S.** 2002. Introduction. *In* M. S. Donnenberg (ed.), *Escherichia coli:* Virulence mechanisms of a versatile pathogen. Academic Press, New York, NY.

29. **Donnenberg, M. S., J. A. Giron, J. P. Nataro, and J. B. Kaper.** 1992. A plasmid-encoded type IV fimbrial gene of enteropathogenic *Escherichia coli* associated with localized adherence. Mol Microbiol **6:**3427-37.

30. **Donnenberg, M. S., C. O. Tacket, S. P. James, G. Losonsky, J. P. Nataro, S. S. Wasserman, J. B. Kaper, and M. M. Levine.** 1993. Role of the *eaeA* gene in experimental enteropathogenic *Escherichia coli* infection. J Clin Invest **92:**1412-7.

31. **Donnenberg, M. S., H. Z. Zhang, and K. D. Stone.** 1997. Biogenesis of the bundle-forming pilus of enteropathogenic *Escherichia coli*: reconstitution of fimbriae in recombinant *E. coli* and role of DsbA in pilin stability--a review. Gene **192:**33-8.

32. **Dorman, C. J., and C. F. Higgins.** 1987. Fimbrial phase variation in *Escherichia coli*: dependence on integration host factor and homologies with other site-specific recombinases. J Bacteriol **169:**3840-3.

33. **Drouin, G., F. Prat, M. Ell, and G. D. Clarke.** 1999. Detecting and characterizing gene conversions between multigene family members. Mol Biol Evol **16:**1369-90.

34. **Eisenstein, B. I., D. S. Sweet, V. Vaughn, and D. I. Friedman.** 1987. Integration host factor is required for the DNA inversion that controls phase variation in *Escherichia coli*. Proc Natl Acad Sci U S A **84:**6506-10.

35. **Elliott, S. J., V. Sperandio, J. A. Giron, S. Shin, J. L. Mellies, L. Wainwright, S. W. Hutcheson, T. K. McDaniel, and J. B. Kaper.** 2000. The locus of enterocyte effacement (LEE)-encoded regulator controls expression of both LEE- and non-LEE-encoded virulence factors in enteropathogenic and enterohemorrhagic *Escherichia coli*. Infect Immun **68:**6115-6126.

36. **Elliott, S. J., L. A. Wainwright, T. K. McDaniel, K. G. Jarvis, Y. K. Deng, L. C. Lai, B. P. McNamara, M. S. Donnenberg, and J. B. Kaper.** 1998. The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic *Escherichia coli* E2348/69. Mol Microbiol **28:**1-4.

37. **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:**1750-3.

38. **Finlay, B. B., I. Rosenshine, M. S. Donnenberg, and J. B. Kaper.** 1992. Cytoskeletal composition of attaching and effacing lesions associated with enteropathogenic *Escherichia coli* adherence to HeLa cells. Infect Immun **60:**2541-3.

39. **Frankel, G., A. D. Phillips, L. R. Trabulsi, S. Knutton, G. Dougan, and S. Matthews.** 2001. Intimin and the host cell--is it bound to end in Tir(s)? Trends Microbiol **9:**214-8.

40. **Franzolin, M. R., R. C. Alves, R. Keller, T. A. Gomes, L. Beutin, M. L. Barreto, C. Milroy, A. Strina, H. Ribeiro, and L. R. Trabulsi.** 2005. Prevalence of diarrheagenic *Escherichia coli* in children with diarrhea in Salvador, Bahia, Brazil. Mem Inst Oswaldo Cruz **100:**359-63.

41. **Gally, D. L., J. Leathart, and I. C. Blomfield.** 1996. Interaction of FimB and FimE with the *fim* switch that controls the phase variation of type 1 fimbriae in *Escherichia coli* K-12. Mol Microbiol **21:**725-38.

42. **Gally, D. L., T. J. Rucker, and I. C. Blomfield.** 1994. The leucine-responsive regulatory protein binds to the *fim* switch to control phase variation of type 1 fimbrial expression in *Escherichia coli* K-12. J Bacteriol **176:**5665-72.

43. **Ghilardi, A. C., T. A. Gomes, W. P. Elias, and L. R. Trabulsi.** 2003. Virulence factors of *Escherichia coli* strains belonging to serogroups O127 and O142. Epidemiol Infect **131:**815-21.

44. **Goldman, N., and Z. Yang.** 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:**725-36.

45. **Gomez-Duarte, O. G., and J. B. Kaper.** 1995. A plasmid-encoded regulatory region activates chromosomal *eaeA* expression in enteropathogenic *Escherichia coli.* Infect Immun **63:**1767-76.

46. **Guth, B. E., R. Giraldi, T. A. Gomes, and L. R. Marques.** 1994. Survey of cytotoxin production among *Escherichia coli* strains characterized as enteropathogenic (EPEC) by serotyping and presence of EPEC adherence factor (EAF) sequences. Can J Microbiol **40:**341-4.

47. **Hammar, M., A. Arnqvist, Z. Bian, A. Olsen, and S. Normark.** 1995. Expression of two *csg* operons is required for production of fibronectin- and congo red-binding curli polymers in *Escherichia coli* K-12. Mol Microbiol **18:**661-70.

48. **Hammar, M., Z. Bian, and S. Normark.** 1996. Nucleator-dependent intercellular assembly of adhesive curli organelles in *Escherichia coli.* Proc Natl Acad Sci U S A **93:**6562-6.

49. **Hartland, E. L., S. J. Daniell, R. M. Delahay, B. C. Neves, T. Wallis, R. K. Shaw, C. Hale, S. Knutton, and G. Frankel.** 2000. The type III protein translocation system of enteropathogenic *Escherichia coli* involves EspA-EspB protein interactions. Mol Microbiol **35:**1483-92.

50. **Hein, J.** 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math Biosci **98:**185-200.

51. **Huelsenbeck, J. P., and K. A. Dyer.** 2004. Bayesian estimation of positively selected sites. J Mol Evol **58:**661-72.

52. **Hull, R. A., R. E. Gill, P. Hsu, B. H. Minshew, and S. Falkow.** 1981. Construction and expression of recombinant plasmids encoding type 1 or D-mannose-resistant pili from a urinary tract infection *Escherichia coli* isolate. Infect Immun **33:**933-8.

53. **Hultgren, S. J., T. N. Porter, A. J. Schaeffer, and J. L. Duncan.** 1985. Role of type 1 pili and effects of phase variation on lower urinary tract infections produced by *Escherichia coli.* Infect Immun **50:**370-7.

54. **Huson, D. H., and D. Bryant.** 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol **23:**254-67.

55. **Huys, G., M. Cnockaert, J. M. Janda, and J. Swings.** 2003. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. Int J Syst Evol Microbiol **53:**807-10.

56. **Hyma, K. E., D. W. Lacher, A. M. Nelson, A. C. Bumbaugh, J. M. Janda, N. A. Strockbine, V. B. Young, and T. S. Whittam.** 2005. Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. J Bacteriol **187:**619-28.

57. **Ina, Y.** 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol **40:**190-226.

58. **Jakobsen, I. B., and S. Easteal.** 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput Appl Biosci **12:**291-5.

59. **Jakobsen, I. B., S. R. Wilson, and S. Easteal.** 1997. The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol Biol Evol **14:**474-84.

60. **Jenkins, C., A. J. Lawson, T. Cheasty, G. A. Willshaw, P. Wright, G. Dougan, G. Frankel, and H. R. Smith.** 2003. Subtyping intimin genes from enteropathogenic *Escherichia coli* associated with outbreaks and sporadic cases in the United Kingdom and Eire. Mol Cell Probes **17:**149-56.

61. **Jores, J., K. Zehmke, J. Eichberg, L. Rumer, and L. H. Wieler.** 2003. Description of a novel intimin variant (type zeta) in the bovine O84:NM verotoxin-producing *Escherichia coli* strain 537/89 and the diagnostic value of intimin typing. Exp Biol Med (Maywood) **228:**370-6.

62. **Kaper, J. B.** 1998. EPEC delivers the goods. Trends Microbiol **6:**169-72; discussion 172-3.

63. **Kaper, J. B.** 1994. Molecular pathogenesis of enteropathogenic *Escherichia coli*, p. 173-195. *In* V. L. Miller, J. B. Kaper, D. A. Portnoy, and R. R. Isberg (ed.), Molecular Genetics of Bacterial Pathogenesis. American Society for Microbiology, Washington, D.C.

64. **Kaper, J. B., and A. D. O'Brien.** 1998. *Escherichia coli* O157:H7 and other shiga toxin-producing *E. coli* strains. ASM Press, Washington, DC.

65. **Keith, B. R., L. Maurer, P. A. Spears, and P. E. Orndorff.** 1986. Receptor-binding function of type 1 pili effects bladder colonization by a clinical isolate of *Escherichia coli*. Infect Immun **53:**693-6.

66. **Kenny, B.** 2002. Mechanism of action of EPEC type III effector molecules. Int J Med Microbiol **291:**469-77.

67. **Kenny, B., R. DeVinney, M. Stein, D. J. Reinscheid, E. A. Frey, and B. B. Finlay.** 1997. Enteropathogenic *E. coli* (EPEC) transfers its receptor for intimate adherence into mammalian cells. Cell **91:**511-20.

68. **Kimura, M.** 1979. The neutral theory of molecular evolution. Sci Am **241:**98-100, 102, 108 passim.

69. **Klemm, P.** 1984. The *fimA* gene encoding the type-1 fimbrial subunit of *Escherichia coli*. Nucleotide sequence and primary structure of the protein. Eur J Biochem **143:**395-9.

70. **Klemm, P., and G. Christiansen.** 1987. Three *fim* genes required for the regulation of length and mediation of adhesion of *Escherichia coli* type 1 fimbriae. Mol Gen Genet **208:**439-45.

71. **Knutton, S., I. Rosenshine, M. J. Pallen, I. Nisan, B. C. Neves, C. Bain, C. Wolff, G. Dougan, and G. Frankel.** 1998. A novel EspA-associated surface organelle of enteropathogenic *Escherichia coli* involved in protein translocation into epithelial cells. Embo J **17:**2166-76.

72. **Kosakovsky Pond, S. L., and S. D. Frost.** 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics **21:**2531-3.

73. **Kosakovsky Pond, S. L., and S. D. Frost.** 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol **22:**1208-22.

74. **Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost.** 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol **23:**1891-901.

75. **Koster, F., J. Levin, L. Walker, K. S. Tung, R. H. Gilman, M. M. Rahaman, M. A. Majid, S. Islam, and R. C. Williams, Jr.** 1978. Hemolytic-uremic syndrome after shigellosis. Relation to endotoxemia and circulating immune complexes. N Engl J Med **298:**927-33.

76. **Krogfelt, K. A.** 1991. Bacterial adhesion: genetics, biogenesis, and role in pathogenesis of fimbrial adhesins of *Escherichia coli*. Rev Infect Dis **13:**721-35.

77. **Krogfelt, K. A., B. A. McCormick, R. L. Burghoff, D. C. Laux, and P. S. Cohen.** 1991. Expression of *Escherichia coli* F-18 type 1 fimbriae in the streptomycin-treated mouse large intestine. Infect Immun **59:**1567-8.

78. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5:**150-63.

79. **Lacher, D. W., H. Steinsland, and T. S. Whittam.** 2006. Allelic subtyping of the intimin locus (*eae*) of pathogenic *Escherichia coli* by fluorescent RFLP. FEMS Microbiol Lett **261:**80-7.

80. **Lazzaro, B. P., B. K. Sceurman, S. L. Carney, and A. G. Clark.** 2002. fRFLP and fAFLP: medium-throughput genotyping by fluorescently post-labeling restriction digestion. Biotechniques **33:**539-40, 542, 545-6.

81. **Leathart, J. B., and D. L. Gally.** 1998. Regulation of type 1 fimbrial expression in uropathogenic *Escherichia coli*: heterogeneity of expression through sequence changes in the *fim* switch region. Mol Microbiol **28:**371-81.

82. **Levine, M. M.** 1987. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. J Infect Dis **155:**377-89.

83. **Levine, M. M., and R. Edelman.** 1984. Enteropathogenic *Escherichia coli* of classic serotypes associated with infant diarrhea: epidemiology and pathogenesis. Epidemiol Rev **6:**31-51.

84. **Levine, M. M., J. P. Nataro, H. Karch, M. M. Baldini, J. B. Kaper, R. E. Black, M. L. Clements, and A. D. O'Brien.** 1985. The diarrheal response of humans to some classic serotypes of enteropathogenic *Escherichia coli* is dependent on a plasmid encoding an enteroadhesiveness factor. J Infect Dis **152:**550-9.

85. **Li, B., W. H. Koch, and T. A. Cebula.** 1997. Detection and characterization of the *fimA* gene of *Escherichia coli* O157:H7. Mol Cell Probes **11:**397-406.

86. **Li, W. H.** 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol **36:**96-9.

87. **Li, W. H., C. I. Wu, and C. C. Luo.** 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol **2:**150-74.

88. **Low, A. S., N. Holden, T. Rosser, A. J. Roe, C. Constantinidou, J. L. Hobman, D. G. Smith, J. C. Low, and D. L. Gally.** 2006. Analysis of fimbrial gene clusters and their expression in enterohaemorrhagic *Escherichia coli* O157:H7. Environ Microbiol **8:**1033-47.

89. **Luo, Y., E. A. Frey, R. A. Pfuetzner, A. L. Creagh, D. G. Knoechel, C. A. Haynes, B. B. Finlay, and N. C. Strynadka.** 2000. Crystal structure of enteropathogenic *Escherichia coli* intimin-receptor complex. Nature **405:**1073-7.

90. **Makino, S., H. Asakura, T. Shirahata, T. Ikeda, K. Takeshi, K. Arai, M. Nagasawa, T. Abe, and T. Sadamoto.** 1999. Molecular epidemiological study of a mass outbreak caused by enteropathogenic *Escherichia coli* O157:H45. Microbiol Immunol **43:**381-4.

91. **Maynard Smith, J., and N. H. Smith.** 1998. Detecting recombination from gene trees. Mol Biol Evol **15:**590-9.

92. **McClain, M. S., I. C. Blomfield, and B. I. Eisenstein.** 1991. Roles of *fimB* and *fimE* in site-specific DNA inversion associated with phase variation of type 1 fimbriae in *Escherichia coli.* J Bacteriol **173:**5308-14.

93. **McDaniel, T. K., K. G. Jarvis, M. S. Donnenberg, and J. B. Kaper.** 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. Proc Natl Acad Sci U S A **92:**1664-8.

94. **McDaniel, T. K., and J. B. Kaper.** 1997. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. Mol Microbiol **23:**399-407.

95. **McGraw, E. A., J. Li, R. K. Selander, and T. S. Whittam.** 1999. Molecular evolution and mosaic structure of α, β, and γ intimins of pathogenic *Escherichia coli.* Mol Biol Evol **16:**12-22.

96. **Mellies, J. L., S. J. Elliott, V. Sperandio, M. S. Donnenberg, and J. B. Kaper.** 1999. The Per regulon of enteropathogenic *Escherichia coli*: identification of a regulatory cascade and a novel transcriptional activator, the locus of enterocyte effacement (LEE)-encoded regulator (Ler). Mol Microbiol **33:**296-306.

97. **Mellies, J. L., F. Navarro-Garcia, I. Okeke, J. Frederickson, J. P. Nataro, and J. B. Kaper.** 2001. *espC* pathogenicity island of enteropathogenic *Escherichia coli* encodes an enterotoxin. Infect Immun **69:**315-24.

98. **Menotti-Raymond, M., W. T. Starmer, and D. T. Sullivan.** 1991. Characterization of the structure and evolution of the Adh region of *Drosophila hydei*. Genetics **127:**355-66.

99. **Miyata, T.,. and T. Yasunaga.** 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol **16:**23-36.

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

101. **Nataro, J. P., K. O. Maher, P. Mackie, and J. B. Kaper.** 1987. Characterization of plasmids encoding the adherence factor of enteropathogenic *Escherichia coli*. Infect Immun **55:**2370-7.

102. **Nataro, J. P., I. C. Scaletsky, J. B. Kaper, M. M. Levine, and L. R. Trabulsi.** 1985. Plasmid-mediated factors conferring diffuse and localized adherence of enteropathogenic *Escherichia coli*. Infect Immun **48:**378-83.

103. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3:**418-26.

104. **Neves, B. C., R. K. Shaw, G. Frankel, and S. Knutton.** 2003. Polymorphisms within EspA filaments of enteropathogenic and enterohemorrhagic *Escherichia coli*. Infect Immun **71:**2262-5.

105. **Nguyen, R. N., L. S. Taylor, M. Tauschek, and R. M. Robins-Browne.** 2006. Atypical enteropathogenic *Escherichia coli* infection and prolonged diarrhea in children. Emerg Infect Dis **12:**597-603.

106. **Nielsen, R.** 2002. Mapping mutations on phylogenies. Syst Biol **51:**729-39.

107. **Nielsen, R., and J. P. Huelsenbeck.** 2002. Detecting positively selected amino acid sites using posterior predictive P-values. Pac Symp Biocomput:576-88.

108. **Nielsen, R., and Z. Yang.** 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:**929-36.

109. **Ofek, I., and E. H. Beachey.** 1978. Mannose binding and epithelial cell adherence of *Escherichia coli*. Infect Immun **22:**247-54.

110. **Ohta, T., and C. J. Basten.** 1992. Gene conversion generates hypervariability at the variable regions of kallikreins and their inhibitors. Mol Phylogenet Evol **1:**87-90.

111. **Okeke, I. N., J. A. Borneman, S. Shin, J. L. Mellies, L. E. Quinn, and J. B. Kaper.** 2001. Comparative sequence analysis of the plasmid-encoded regulator of enteropathogenic *Escherichia coli* strains. Infect Immun **69:**5553-64.

112. **Okeke, I. N., A. Lamikanra, H. Steinruck, and J. B. Kaper.** 2000. Characterization of *Escherichia coli* strains from cases of childhood diarrhea in provincial southwestern Nigeria. J. Clin. Microbiol. **38:**7-12.

113. **Olsen, A., A. Arnqvist, M. Hammar, and S. Normark.** 1993. Environmental regulation of curli production in *Escherichia coli.* Infect Agents Dis **2:**272-4.

114. **Olsen, A., A. Jonsson, and S. Normark.** 1989. Fibronectin binding mediated by a novel class of surface organelles on *Escherichia coli.* Nature **338:**652-5.

115. **Orden, J. A., M. Yuste, D. Cid, T. Piacesi, S. Martinez, J. A. Ruiz-Santa-Quiteria, and R. De la Fuente.** 2003. Typing of the *eae* and *espB* genes of attaching and effacing *Escherichia coli* isolates from ruminants. Vet Microbiol **96:**203-15.

116. **Orndorff, P. E., and S. Falkow.** 1984. Organization and expression of genes responsible for type 1 piliation in *Escherichia coli.* J Bacteriol **159:**736-44.

117. **Orskov, F., T. S. Whittam, A. Cravioto, and I. Orskov.** 1990. Clonal relationships among classic enteropathogenic *Escherichia coli* (EPEC) belong to different O groups. J Infect Dis **162:**76-81.

118. **Oswald, E., H. Schmidt, S. Morabito, H. Karch, O. Marches, and A. Caprioli.** 2000. Typing of intimin genes in human and animal enterohemorrhagic and enteropathogenic *Escherichia coli*: characterization of a new intimin variant. Infect Immun **68:**64-71.

119. **Peek, A. S., V. Souza, L. E. Eguiarte, and B. S. Gaut.** 2001. The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (*fimA*) from *Escherichia coli.* J Mol Evol **52:**193-204.

120. **Perna, N. T., G. F. Mayhew, G. Posfai, S. Elliott, M. S. Donnenberg, J. B. Kaper, and F. R. Blattner.** 1998. Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. Infect Immun **66:**3810-7.

121. **Prigent-Combaret, C., G. Prensier, T. T. Le Thi, O. Vidal, P. Lejeune, and C. Dorel.** 2000. Developmental pathway for biofilm formation in curli-producing *Escherichia coli* strains: role of flagella, curli and colanic acid. Environ Microbiol **2:**450-64.

122. **Ramachandran, V., K. Brett, M. A. Hornitzky, M. Dowton, K. A. Bettelheim, M. J. Walker, and S. P. Djordjevic.** 2003. Distribution of intimin subtypes among *Escherichia coli* isolates from ruminant and human sources. J Clin Microbiol **41:**5022-32.

123. **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:**2719-22.

124. **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:**64-7.

125. **Riley, L. W., R. S. Remis, S. D. Helgerson, H. B. McGee, J. G. Wells, B. R. Davis, R. J. Hebert, E. S. Olcott, L. M. Johnson, N. T. Hargrett, P. A. Blake, and M. L. Cohen.** 1983. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N Engl J Med **308:**681-5.

126. **Rodrigues, J., I. C. Scaletsky, L. C. Campos, T. A. Gomes, T. S. Whittam, and L. R. Trabulsi.** 1996. Clonal structure and virulence factors in strains of *Escherichia coli* of the classic serogroup O55. Infect Immun **64:**2680-6.

127. **Roe, A. J., C. Currie, D. G. Smith, and D. L. Gally.** 2001. Analysis of type 1 fimbriae expression in verotoxigenic *Escherichia coli*: a comparison between serotypes O157 and O26. Microbiology **147:**145-52.

128. **Roesch, P. L., and I. C. Blomfield.** 1998. Leucine alters the interaction of the leucine-responsive regulatory protein (Lrp) with the *fim* switch to stimulate site-specific recombination in *Escherichia coli.* Mol Microbiol **27:**751-61.

129. **Rumer, L., J. Jores, P. Kirsch, Y. Cavignac, K. Zehmke, and L. H. Wieler.** 2003. Dissemination of *pheU-* and *pheV*-located genomic islands among enteropathogenic (EPEC) and enterohemorrhagic (EHEC) *E. coli* and their possible role in the horizontal transfer of the locus of enterocyte effacement (LEE). Int J Med Microbiol **292:**463-75.

130. **Sawyer, S.** 1989. Statistical tests for detecting gene conversion. Mol Biol Evol **6:**526-38.

131. **Sawyer, S. A.** 1999. GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

132. **Schaeffer, A. J., W. R. Schwan, S. J. Hultgren, and J. L. Duncan.** 1987. Relationship of type 1 pilus expression in *Escherichia coli* to ascending urinary tract infections in mice. Infect Immun **55:**373-80.

133. **Schmidt, H., H. Russmann, and H. Karch.** 1993. Virulence determinants in nontoxinogenic *Escherichia coli* O157 strains that cause infantile diarrhea. Infect Immun **61:**4894-8.

134. **Shaw, R. K., S. Daniell, F. Ebel, G. Frankel, and S. Knutton.** 2001. EspA filament-mediated protein translocation into red blood cells. Cell Microbiol **3:**213-22.

135. **Sinclair, J. F., E. A. Dean-Nystrom, and A. D. O'Brien.** 2006. The established intimin receptor Tir and the putative eucaryotic intimin receptors nucleolin and β1 integrin localize at or near the site of enterohemorrhagic *Escherichia coli* O157:H7 adherence to enterocytes in vivo. Infect Immun **74:**1255-65.

136. **Sjobring, U., G. Pohl, and A. Olsen.** 1994. Plasminogen, absorbed by *Escherichia coli* expressing curli or by *Salmonella enteritidis* expressing thin aggregative fimbriae, can be activated by simultaneously captured tissue-type plasminogen activator (t-PA). Mol Microbiol **14:**443-52.

137. **Sperandio, V., J. B. Kaper, M. R. Bortolini, B. C. Neves, R. Keller, and L. R. Trabulsi.** 1998. Characterization of the locus of enterocyte effacement (LEE) in different enteropathogenic *Escherichia coli* (EPEC) and Shiga-toxin producing *Escherichia coli* (STEC) serotypes. FEMS Microbiol Lett **164:**133-9.

138. **Stephan, R., N. Borel, C. Zweifel, M. Blanco, and J. E. Blanco.** 2004. First isolation and further characterization of enteropathogenic *Escherichia coli* (EPEC) O157:H45 strains from cattle. BMC Microbiol **4:**10.

139. **Stephens, J. C.** 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol Biol Evol **2:**539-56.

140. **Stone, K. D., H. Z. Zhang, L. K. Carlson, and M. S. Donnenberg.** 1996. A cluster of fourteen genes from enteropathogenic *Escherichia coli* is sufficient for the biogenesis of a type IV pilus. Mol Microbiol **20:**325-37.

141. **Strom, M. S., and S. Lory.** 1993. Structure-function and biogenesis of the type IV pili. Annu Rev Microbiol **47:**565-96.

142. **Suzuki, Y.** 2004. New methods for detecting positive selection at single amino acid sites. J Mol Evol **59:**11-9.

143. **Suzuki, Y., and T. Gojobori.** 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol **16:**1315-28.

144. **Tarr, C. L., and T. S. Whittam.** 2002. Molecular evolution of the intimin gene in O111 clones of pathogenic *Escherichia coli.* J Bacteriol **184:**479-87.

145. **Tobe, T., T. Hayashi, C. G. Han, G. K. Schoolnik, E. Ohtsubo, and C. Sasakawa.** 1999. Complete DNA sequence and structural analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid. Infect Immun **67:**5455-62.

146. **Tobe, T., G. K. Schoolnik, I. Sohel, V. H. Bustamante, and J. L. Puente.** 1996. Cloning and characterization of *bfpTVW,* genes required for the transcriptional activation of *bfpA* in enteropathogenic *Escherichia coli.* Mol Microbiol **21:**963-75.

147. **Trabulsi, L. R., R. Keller, and T. A. Tardelli Gomes.** 2002. Typical and atypical enteropathogenic *Escherichia coli.* Emerg Infect Dis **8:**508-13.

148. **Uhlich, G. A., J. E. Keen, and R. O. Elder.** 2001. Mutations in the *csgD* promoter associated with variations in curli expression in certain strains of *Escherichia coli* O157:H7. Appl Environ Microbiol **67:**2367-70.

149. **Uhlich, G. A., J. E. Keen, and R. O. Elder.** 2002. Variations in the *csgD* promoter of *Escherichia coli* O157:H7 associated with increased virulence in mice and increased invasion of HEp-2 cells. Infect Immun **70:**395-9.

150. **Valentiner-Branth, P., H. Steinsland, T. K. Fischer, M. Perch, F. Scheutz, F. Dias, P. Aaby, K. Molbak, and H. Sommerfelt.** 2003. Cohort study of Guinean children: incidence, pathogenicity, conferred protection, and attributable risk for enteropathogens during the first 2 years of life. J Clin Microbiol **41:**4238-45.

151. **Vallance, B. A., and B. B. Finlay.** 2000. Exploitation of host cells by enteropathogenic *Escherichia coli.* Proc Natl Acad Sci U S A **97:**8799-806.

152. **Vidal, O., R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman, and P. Lejeune.** 1998. Isolation of an *Escherichia coli* K-12 mutant strain able to form biofilms on inert surfaces: involvement of a new *ompR* allele that increases curli expression. J Bacteriol **180:**2442-9.

153. **Wales, A. D., M. J. Woodward, and G. R. Pearson.** 2005. Attaching-effacing bacteria in animals. J Comp Pathol **132:**1-26.

154. **Weiller, G. F.** 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol **15:**326-35.

155. **Weissman, S. J., S. Chattopadhyay, P. Aprikian, M. Obata-Yasuoka, Y. Yarova-Yarovaya, A. Stapleton, W. Ba-Thein, D. Dykhuizen, J. R. Johnson, and E. V. Sokurenko.** 2006. Clonal analysis reveals high rate of structural mutations in fimbrial adhesins of extraintestinal pathogenic *Escherichia coli*. Mol Microbiol **59:**975-88.

156. **Whittam, T. S., and E. A. McGraw.** 1996. Clonal analysis of EPEC serogroups. Reviews in Microbiology **27 (Suppl. #1):**7-16.

157. **Wieler, L. H., T. K. McDaniel, T. S. Whittam, and J. B. Kaper.** 1997. Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of the strains. FEMS Microbiol Lett **156:**49-53.

158. **Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431-49.

159. **Yang, Z., and W. J. Swanson.** 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol **19:**49-57.

160. **Zhang, W. L., B. Kohler, E. Oswald, L. Beutin, H. Karch, S. Morabito, A. Caprioli, S. Suerbaum, and H. Schmidt.** 2002. Genetic diversity of intimin genes of attaching and effacing *Escherichia coli* strains. J Clin Microbiol **40:**4486-92.