AMERICAN SOCIETY FOR MICROBIOLOGY | Infection and Immunity

Investigating the Relatedness of Enteroinvasive *Escherichia coli* to Other *E. coli* and *Shigella* Isolates by Using Comparative Genomics

Tracy H. Hazen,[a,b] Susan R. Leonard,[c] Keith A. Lampel,[c] David W. Lacher,[c] Anthony T. Maurelli,[d] David A. Rasko[a,b]

Institute for Genome Sciences[a] and Department of Microbiology and Immunology,[b] University of Maryland School of Medicine, Baltimore, Maryland, USA; Food and Drug Administration, Office of Applied Research and Safety Assessment, Center for Food Safety and Applied Nutrition, Division of Molecular Biology, Laurel, Maryland, USA[c]; Emerging Pathogens Institute and Department of Environmental and Global Health, College of Public Health and Health Professions, University of Florida, Gainesville, Florida, USA[d]

Enteroinvasive *Escherichia coli* (EIEC) is a unique pathovar that has a pathogenic mechanism nearly indistinguishable from that of *Shigella* species. In contrast to isolates of the four *Shigella* species, which are widespread and can be frequent causes of human illness, EIEC causes far fewer reported illnesses each year. In this study, we analyzed the genome sequences of 20 EIEC isolates, including 14 first described in this study. Phylogenomic analysis of the EIEC genomes demonstrated that 17 of the isolates are present in three distinct lineages that contained only EIEC genomes, compared to reference genomes from each of the *E. coli* pathovars and *Shigella* species. Comparative genomic analysis identified genes that were unique to each of the three identified EIEC lineages. While many of the EIEC lineage-specific genes have unknown functions, those with predicted functions included a colicin and putative proteins involved in transcriptional regulation or carbohydrate metabolism. *In silico* detection of the *Shigella* virulence plasmid (pINV), which is essential for the invasion of host cells, demonstrated that a form of pINV was present in nearly all EIEC genomes, but the Mxi-Spa-Ipa region of the plasmid that encodes the invasion-associated proteins was absent from several of the EIEC isolates. The comparative genomic findings in this study support the hypothesis that multiple EIEC lineages have evolved independently from multiple distinct lineages of *E. coli* via the acquisition of the *Shigella* virulence plasmid and, in some cases, the *Shigella* pathogenicity islands.

Enteroinvasive *Escherichia coli* is a unique group of disease-causing *E. coli* bacteria that have a virulence mechanism most similar to that of *Shigella* bacteria, involving the invasion of intestinal epithelial cells (1–4). In contrast, the other pathovars of *E. coli* do not invade host cells and, instead, typically associate with the surface of the host cell and secrete or translocate virulence factors onto or into the cell (1, 2, 4). While *Shigella* bacteria are among the leading agents of diarrheal illness, causing an estimated 165 million annual cases worldwide (5), EIEC bacteria are seldom identified but can occasionally be linked to small food- or waterborne outbreaks (2, 6). Although EIEC bacteria cause disease very similar to the disease caused by *Shigella* bacteria, they share biochemical and cultural traits that are intermediate between those of commensal *E. coli* and *Shigella* bacteria (7). Thus, it is not clear if the reported low incidence is due to the lack of acceptable and defined biochemical/molecular markers for this group or whether EIEC disease truly has a low incidence rate.

The virulence of *Shigella* and EIEC bacteria has been attributed to genes associated with mobile genetic elements including pathogenicity islands and a virulence plasmid, pINV (1–3, 8–11). The pINV plasmid is required for invasion of intestinal epithelial cells and encodes a type III secretion system (T3SS) and many associated effectors (8–10). Meanwhile, on the chromosome, there are several *Shigella*-specific pathogenicity islands that have been designated SHI-1/SHE (12, 13), SHI-2 (14), SHI-3 (15), SHI-O (16), and SRL (17). Within these islands are genes encoding additional virulence factors, including autotransporters (*pic* and *sigA*); factors involved in iron acquisition (*iucA* to *iucD* and *iutA*) (14, 15), O-antigen conversion (16), and antibiotic resistance (17); and the *Shigella* enterotoxin ShET1 (12, 18). While the virulence plasmids of EIEC and *Shigella* have considerable similarity (19), little is known regarding the conservation of the pathogenicity islands in EIEC and *Shigella*.

Previous studies of EIEC isolates have used primarily molecular approaches, such as phylogenetic analysis of single genes, to investigate the relatedness of EIEC to *Shigella* (20–22). In comparison to the thousands of genome sequences that have been generated for the *Shigella* species and most of the other *E. coli* pathovars, only 24 EIEC genomes have been sequenced and described to date (23–26), with the majority of those being generated in a recent study (26). The most recent study by Pettengill et al. utilized a single nucleotide polymorphism (SNP)-based approach to assess the taxonomic relationships of EIEC, *Shigella*, and *E. coli*. They also identified SNPs for use in the potential development of a screening assay for the diagnostic assessment of clinical isolates. However, that study did not provide a significant description of the differences in the total genomic content, such as virulence-associated genes of the EIEC and *Shigella* genomes compared to those of other *E. coli* bacteria.

Received 22 April 2016 **Returned for modification** 20 May 2016
Accepted 31 May 2016

Accepted manuscript posted online 6 June 2016

Citation Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA. 2016. Investigating the relatedness of enteroinvasive *Escherichia coli* to other *E. coli* and *Shigella* isolates by using comparative genomics. Infect Immun 84:2362–2371. doi:10.1128/IAI.00350-16.

Editor: B. A. McCormick, The University of Massachusetts Medical School

Address correspondence to David A. Rasko, drasko@som.umaryland.edu.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/IAI.00350-16.

TABLE 1 EIEC genomes analyzed in this study

Isolate ID	Molecular serotype	Country of isolation	Genome size (bp)	% GC	No. of contigs	Phylogroup	EIEC lineage	Accession no.
4608-58	O143:H26	USA	5,043,882	50.47	266	E	1	JTCO00000000.1
EC10010	O143:H26	USA	5,089,039	50.48	340	E	1	LSGI00000000.1
EC10032	O143:H26	USA	5,110,915	50.47	381	E	1	LSGJ00000000.1
EC10033	O143:H26	USA	5,110,277	50.47	358	E	1	LSGK00000000.1
ATM460	O143:H26	USA	5,193,883	50.43	352	E	1	LSGL00000000.1
ATM461	O143:H26	Zaire	5,395,526	50.52	416	E	1	LSGM00000000.1
53638	O124:H30	USA	5,371,790	50.99	4	A	2	AAKB00000000.2
M4163	O124:H30	USA	5,093,926	50.71	294	A	2	JTCN00000000.1
ATM456	O121:H30	South Africa	5,076,933	50.82	426	A	2	LSFZ00000000.1
EC10018	O124:H30	USA	4,902,846	50.83	233	A	2	LSGA00000000.1
EC10016	O124:H30	USA	4,995,694	50.64	452	A	2	LSGB00000000.1
ATM457	O124:H30	Bulgaria	5,098,706	50.61	326	A	2	LSGC00000000.1
1827-70	O29:H27	USA	4,803,088	50.79	35	A	None	ADUK00000000.1
ATM462	O164:H7	Bolivia	4,917,032	50.76	506	B1	3	LSGD00000000.1
ATM463	O164:H7	Bulgaria	5,127,067	50.62	562	B1	3	LSGE00000000.1
ATM465	O164:H7	Jordan	5,017,930	50.67	557	B1	3	LSGF00000000.1
ATM266	O29:H4	USA	5,019,573	50.67	502	B1	3	LSGG00000000.1
ATM459	O136:H7	Guam	5,012,551	50.55	492	B1	3	LSGH00000000.1
LT-68	O144:H25	Brazil	5,189,427	50.85	63	B1	None	ADUP00000000.1
CFSAN029787	O96:H19	Italy	5,288,947	50.53	3	B1	None	CP011416.1–CP011418.1

In this report, we provide a comprehensive assessment of the relatedness of EIEC to other *E. coli* and *Shigella* bacteria by using phylogenomic analysis and *in silico* detection of known virulence genes to highlight the similarities and differences among the virulence mechanisms of the EIEC and *Shigella* bacteria. We use phylogenomics and comparative genomics to describe the genome sequences of 20 EIEC isolates, including 14 newly sequenced genomes. This study represents the first use of whole-genome sequencing and comparative genomics to investigate the relatedness of EIEC to other *E. coli* and *Shigella* bacteria by comparing both phylogenomic relatedness and virulence factor content. Our findings demonstrate that 17 of the EIEC genomes form three phylogenomic lineages containing at least five members each, which are distinct from *Shigella* bacteria and other pathogenic *E. coli* bacteria. Gene-based comparisons identified genes that are unique to each of the EIEC lineages and can be used to develop diagnostic assays for the identification of presumptive EIEC clinical isolates. Overall, this study provides a view of the genomic diversity of EIEC isolates and provides insight into the evolution of this understudied pathovar.

MATERIALS AND METHODS

Bacterial isolates and media. The genomes of 19 EIEC isolates from humans and 1 isolate from cheese associated with human infection (M4163) were analyzed in this study, 14 of which were sequenced as part of this study. Details of the locations of isolation are included in Table 1. The sequenced EIEC isolates were cultured in Luria broth at 37°C.

Genome sequencing and assembly. Genomic DNA was extracted from overnight cultures with the DNeasy blood and tissue kit (Qiagen, Germantown, MD, USA). Sequencing libraries with insert sizes of 350 to 700 bp were prepared from genomic DNA with the Nextera DNA Sample Preparation kit (Illumina, San Diego, CA, USA) and sequenced on an Illumina MiSeq sequencer, generating paired-end 250-bp reads in sufficient quantity to provide between 86× and 261× coverage for each genome. Raw reads were trimmed, and draft genomes were assembled *de novo* with CLC Genomics Workbench v8.0.1, v8.0.2, or v8.0.3 (CLC bio, Aarhus, Denmark). Molecular serotypes were determined from the draft genomes by BLASTn analysis utilizing an in-house custom database including the *wzx*, *wzy*, and *fliC* loci with methods that have been previously described (27, 28). Assembled sequences were submitted to GenBank, and accession numbers, assembled genome information, and serotypes are included in Table 1.

Phylogenomic analysis. The genomes of the EIEC isolates analyzed in this study were compared with 37 previously sequenced reference *E. coli* and *Shigella* genomes by whole-genome phylogenomic analysis as described previously (29, 30). Briefly, the genomes were aligned by using Mugsy (31) and homologous blocks were concatenated with the bx-python toolkit (https://bitbucket.org/james_taylor/bx-python). The columns that contained one or more gaps were removed with mothur (32). The concatenated regions from each genome were used to construct a maximum-likelihood phylogeny with RAxML v7.2.8 (33) that was visualized with FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). The phylogeny was generated by using the GTR model of nucleotide substitution with the GAMMA model of rate heterogeneity and 100 bootstrap replicates.

Shigella virulence plasmid pCP301 (accession no. NC_004851.1) was aligned with the EIEC and *Shigella* genomes that had similarity with more than half of the plasmid-borne genes. The plasmid sequence and the genomes were aligned by using Mugsy (31), and 24 kb of aligned sequence from each genome was used to construct a maximum-likelihood phylogeny as described above.

LS-BSR analysis. The 20 EIEC genomes and 37 *E. coli* and *Shigella* reference genomes were compared by large-scale BLAST score ratio (LS-BSR) analysis as previously described (29, 34, 35). Briefly, the predicted protein-encoding genes of each genome that had ≥90% nucleotide sequence identity to each other were assigned to gene clusters with uclust (36). Representative sequences of each gene cluster were then compared to each genome with TBLASTN (37) with composition-based adjustment turned off, and the TBLASTN scores were used to generate a BSR indicating the detection of each gene cluster in each of the genomes analyzed. The BSR was determined by dividing the score of a gene compared to a genome by the score of the gene compared to its own sequence. The predicted protein function of each gene cluster was determined with an ergatis-based (38) in-house annotation pipeline (39).

LS-BSR was also used to detect the presence of previously identified *Shigella* virulence genes in the 29 EIEC and *Shigella* genomes. The protein-

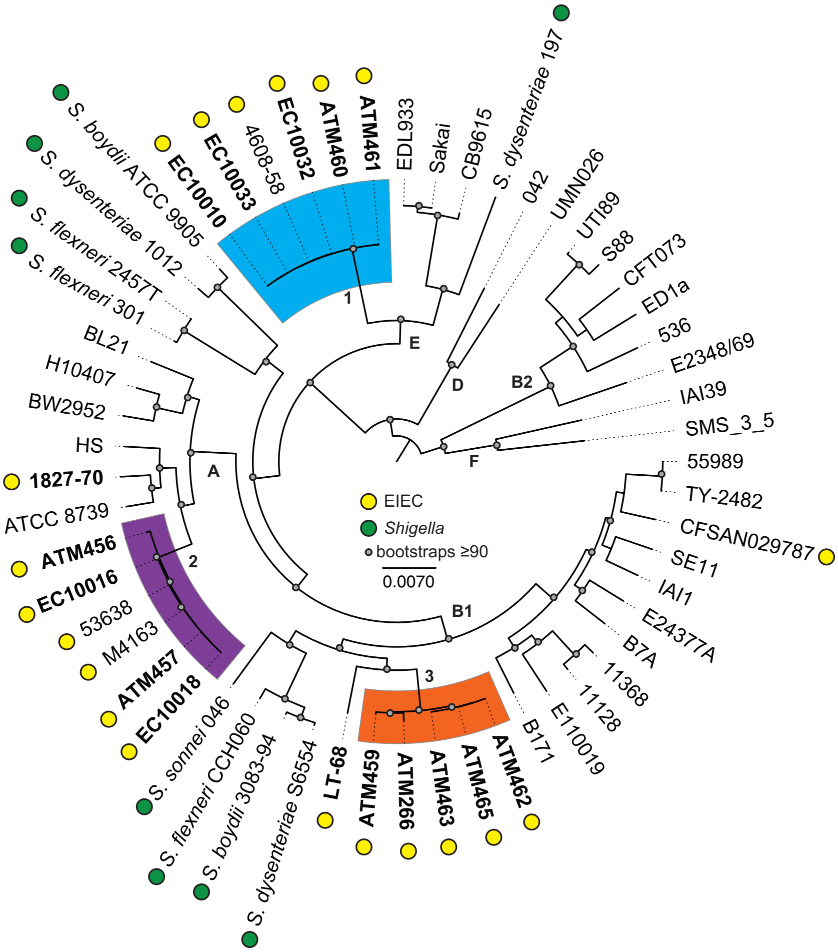

FIG 1 Phylogenomic analysis of EIEC genomes. Whole-genome phylogeny of the EIEC genomes analyzed in this study compared with a diverse collection of *E. coli* and *Shigella* genomes was performed as previously described (29). Three lineages of EIEC genomes are blue, purple, or orange and are numbered 1 to 3. The EIEC genomes are designated by yellow circles, and the *Shigella* genomes are designated by green circles. The genome sequences that were generated in this study are in bold. Bootstrap values of ≥90 are indicated by gray circles. The scale bar indicates an approximate distance of 0.007 nucleotide substitution per site. The previously designated *E. coli* phylogroups (A, B1, B2, D, E, and F) (42, 43) are indicated.

encoding genes of the *Shigella* virulence plasmid (pCP301; accession no. NC_004851.1) were detected in each of the genomes by LS-BSR analysis as described above. Protein-encoding genes of each of the *Shigella* pathogenicity islands (SHI-1/SHE [12, 13], SHI-2 [14], SHI-3 [15], SHI-O [16], and SRL [17]) were also identified in the EIEC and *Shigella* genomes. Genes not on the plasmid or in any of the pathogenicity islands, including the gene for *Shigella* enterotoxin ShET2 (accession no. CAA90899.1), the lysine decarboxylase (LCD) gene *cadA* (accession no. NP_418555.1), genes of the T2SS of *Shigella dysenteriae* 197 (accession no. CP000034.1), and those of the T6SS of enteropathogenic *E. coli* (EPEC) isolate B171 (accession no. AAJX02000009.1), were also detected by LS-BSR analysis.

RESULTS

EIEC genome sequences. The EIEC genomes analyzed in this study included 14 newly sequenced genomes (ATM456, EC10018, EC10016, ATM457, ATM462, ATM463, ATM465, ATM266, ATM459, EC10010, EC10032, EC10033, ATM460, and ATM461) and 6 previously sequenced genomes that are available in the public domain: 53638 (accession no. AAKB00000000.2), M4163 (accession no. JTCN00000000.1), 1827-70 (accession no. ADUK00000000.1), CFSAN029787 (accession no. CP011416.1 to CP011418.1), LT-68 (accession no. ADUP00000000.1), and 4608-58 (accession no. JTCO00000000.1) (Table 1). The most frequently represented serotype among these EIEC isolates was O143:H26, followed by O124:H30 and O164:H7 (Table 1). All of the O antigens were among those previously identified for EIEC isolates, with the exception of O96, which was recently linked to isolates from a foodborne outbreak in Milan, Italy (2, 6, 23, 40). These EIEC isolates represent a diverse global collection from nine countries and almost every continent, including North America, South America, Europe, Africa, and Asia (Table 1). The numbers of contigs of the EIEC genome assemblies ranged from 3 to 557 (Table 1). The average genome size was 5.09 (range, 4.80 to 5.39) Mb, and the average GC content was 50.64% (range, 50.43 to 50.99%), values that are consistent with previously sequenced *E. coli* and *Shigella* genomes (24, 41) (Table 1).

Phylogenomic analysis of EIEC. Phylogenomic analysis of the 20 EIEC genomes with 37 reference *E. coli* and *Shigella* isolate genomes demonstrated that 17 of these EIEC genomes were present in three distinct EIEC lineages (numbered 1 to 3 in Fig. 1) that contained only EIEC genomes and at least five members in each lineage (Fig. 1). There are multiple ways to delineate the new lineages; however, we chose to select the terminal node of the phy-

TABLE 2 Numbers of shared or unique genes identified by LS-BSR analysis

		No. of genomes		No. of gene clusters[a]		
Group 1	Group 2	Group 1	Group 2	All genomes	≥50% of genomes	≥1 genome
EIEC	Other genomes including *Shigella*	20	37	0	7	472
EIEC	Other genomes not including *Shigella*	20	28	0	96	687
EIEC and *Shigella*	Other genomes	29	28	0	87	1,002
Phylogroup E EIEC	Other EIEC	6	14	155	172	249
Phylogroup A EIEC	Other EIEC	7	13	16	68	458
Phylogroup B1 EIEC	Other EIEC	7	13	3	55	469
Clade 1 EIEC	Other EIEC	6	14	155	172	249
Clade 2 EIEC	Other EIEC	6	14	12	21	305
Clade 3 EIEC	Other EIEC	5	15	13	41	221

[a] There were 1,628 gene clusters identified with significant similarity (LS-BSR, ≥0.9) in all 57 *E. coli* and *Shigella* isolates. The number of gene clusters that were present in all genomes, ≥50% of the genomes, or ≥1 of the genomes of group 1 (LS-BSR, ≥0.9) and absent from all of the genomes of group 2 (LS-BSR, <0.4).

logeny that is well supported by the bootstrap values and contains only EIEC isolates. There was one EIEC lineage in each of three different *E. coli* phylogroups (A, B1, and E) (42, 43), demonstrating that there is considerable phylogenomic diversity in this global collection of EIEC isolates (Table 1; Fig. 1). EIEC lineage 1 contained six EIEC genomes and is in phylogroup E along with the O157:H7 EHEC genomes and type I *S. dysenteriae* isolate Sd197 (Fig. 1). The six EIEC isolates in lineage 1 were all serotype O143: H26, and all but one of these isolates were obtained from clinical cases in the United States (Table 1). The one geographic exception in EIEC lineage 1 was EIEC isolate ATM461 from Zaire (Table 1). EIEC lineage 2 is in phylogroup A and also contained six EIEC genomes (Fig. 1). These isolates had molecularly defined serotypes O121:H30 and O124:H30 (Table 1; note that isolate 53638 was originally serotyped as O144 with an unknown H antigen) (27, 28). These isolates were also primarily from the United States, except for two that were from either South Africa or Bulgaria (Table 1). EIEC lineage 3 was in phylogroup B1 and contained five EIEC genomes (Fig. 1). These isolates had serotypes O164:H7, O29:H4, and O136:H7 and were from five different countries (United States, Bolivia, Bulgaria, Guam, and Jordan) (Table 1).

Three additional EIEC genomes were outside these defined lineages but were in phylogroups A and B1 (Fig. 1). One of these genomes, LT-68, was closely related to EIEC lineage 3 in phylogroup B1. Meanwhile, CFSAN029787 was most closely related to enteroaggregative *E. coli* (EAEC) 55989 and O104:H4 *E. coli* TY-2482, which was the etiological agent of a severe foodborne outbreak in Germany in 2012 (44–46). The third EIEC isolate that was not in one of the three EIEC phylogenomic lineages was 1827-70, which was most closely related to the nonpathogenic *E. coli* strains ATCC 8739 and HS (Fig. 1; see Table S1 in the supplemental material).

Gene-based comparisons of EIEC, *Shigella*, and other *E. coli*. The 57 *E. coli* and *Shigella* genomes included in the phylogenomic analysis were also compared by LS-BSR analysis, which is a *de novo* method used to determine gene-based similarity in groups of genomes (34, 47) (Table 2). There were 16,418 total gene clusters that were identified in the 57 genomes, which included 1,628 gene clusters with significant similarity (LS-BSR, ≥0.9) that were present in all of the genomes (Table 2). There were no gene clusters that were present in all of the EIEC genomes that were absent (LS-BSR, < 0.4) from all of the other genomes both with and without the *Shigella* genomes (Table 2). There were also no gene clusters present in all of the EIEC and *Shigella* genomes that were absent from all of the other *E. coli* genomes (Table 2). There were only seven gene clusters identified in ≥50% of the EIEC genomes that were absent from all other *E. coli* and *Shigella* genomes (Table 2). These included the genes for a putative pyruvate kinase, a periplasmic protein, and several hypothetical proteins (see Table S2 in the supplemental material). In contrast, there were 96 gene clusters in ≥50% of the EIEC genomes when the *Shigella* genomes were not included in the comparison and 87 gene clusters that were present in ≥50% of the EIEC and *Shigella* genomes combined compared to the reference *E. coli* genomes (Table 2). Among the 87 gene clusters that were exclusive to EIEC and *Shigella* were plasmid-associated genes encoding a hypothetical toxin-antitoxin system and putative proteins hypothesized to be involved in conjugal transfer (see Table S2 in the supplemental material). Also, there were 1,002 gene clusters that were present in at least one of the EIEC or *Shigella* genomes that were not in any other *E. coli* genome (Table 2).

Comparison of only the EIEC genomes demonstrated that 3 to 155 gene clusters were present in all of the EIEC genomes of one phylogroup that were not in the EIEC genomes of the other phylogroups (Table 2). Similarly, 12 to 155 gene clusters were present in all of the EIEC genomes of one lineage and not in the EIEC genomes of the other lineages (Table 2). Among the 155 EIEC lineage-specific gene clusters that were unique to lineage 1 were putative protein-encoding genes involved in transcriptional regulation, metabolism and transport, and also a colicin (see Table S3 in the supplemental material). The number of genes exclusive to EIEC lineage 1 was 10 times greater than the number of genes that were exclusive to the genomes of lineage 2 or 3. There were only 12 gene clusters unique to lineage 2 and 13 gene clusters unique to lineage 3 (Table 2). Among the lineage 2-specific gene clusters were a putative membrane protein, the aerobactin siderophore receptor *iutA*, and hypothetical proteins (see Table S3 in the supplemental material). The lineage 3-specific gene clusters included several putative transcriptional regulators and hypothetical proteins (see Table S3 in the supplemental material).

***In silico* detection of invasion plasmid pINV.** Since EIEC isolates have a virulence mechanism similar to that of *Shigella*, we used LS-BSR to perform *in silico* detection of the previously identified *Shigella* virulence genes (1, 3, 12, 15–17, 48–50). *Shigella* virulence plasmid, pINV encodes a T3SS involved in the invasion of host cells and is a major component of the *Shigella* and EIEC virulence mechanism; thus, we detected the presence of protein-encoding genes of sequenced *Shigella* virulence plasmid pCP301

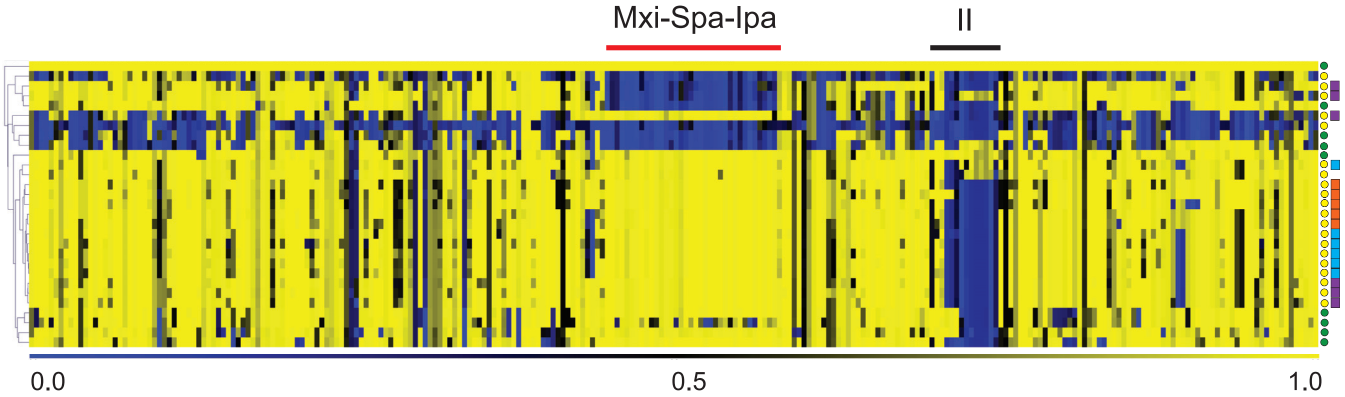

FIG 2 *In silico* detection of the protein-encoding genes of the *Shigella* virulence plasmid. The BSRs were determined for all of the protein-encoding genes by comparing the amino acid sequences encoded by invasion plasmid pCP301 (accession no. NC_004851.1) from *S. flexneri* 2a strain 301 to each of the EIEC and *Shigella* genomes by using TBLASTN (37). The heat map and a hierarchical cluster analysis were generated with MeV (83) as previously described (29). The colors of the heat map indicate the presence of each protein-encoding gene with significant similarity (yellow) or divergent similarity (black) or its absence (blue). Each column represents a different gene, and each row is a different genome. The isolates are color coded by species and lineage on the right. The EIEC genomes are designated by yellow circles, while the *Shigella* genomes are designated by green circles. Genomes of EIEC lineage 1 are indicated by blue squares, those of lineage 2 are indicated by purple squares, and those of lineage 3 are indicated by orange squares. The Mxi-Spa-Ipa region and a second variable region, designated II, are identified by the red and black lines at the top, respectively. Region II contains plasmid stability proteins, a methyltransferase, and hypothetical proteins.

(accession no. NC_004851.1) (1, 3, 8–10, 48–50). The majority of the protein-encoding genes of *Shigella* virulence plasmid pCP301 were identified with similarity (LS-BSR, ≥0.8) in 85% (17/20) of the EIEC genomes (Fig. 2). The plasmid genes were mostly absent from LT-68, EC10018, and 1827-70. However, genes of the Mxi-Spa-Ipa region were present in EC10018, although the majority of the other plasmid genes were absent (Fig. 2). The Mxi-Spa-Ipa region of the plasmid (Fig. 2) was absent from four of the EIEC genomes (1827-70, LT-68, ATM456, and 53638). A second region of the pINV plasmid (region II in Fig. 2) was absent from almost all of the EIEC genomes. This region encodes a putative methylase, plasmid stability proteins, and numerous hypothetical proteins. Phylogenetic analysis of 24 kb of highly conserved aligned sequence from each of the EIEC and *Shigella* genomes that contained the plasmid also demonstrated the lineage-specific similarity observed in the cluster analysis of the BSR data (see Fig. S1 in the supplemental material). The plasmid sequences from the *Shigella* genomes grouped together in one part of the phylogeny, while all but three of the plasmid sequences from the EIEC genomes were present in the other part of the phylogeny (see Fig. S1). All of the plasmids from EIEC lineages 1 and 2 formed lineage-specific groups in the phylogeny (see Fig. S1). In contrast, the plasmid sequences from two of the EIEC genomes from lineage 3 were related to the plasmids from the other EIEC genomes, while the aligned plasmid regions from the other three EIEC genomes from lineage 3 exhibited greater similarity to the *Shigella* plasmids (see Fig. S1).

***In silico* detection of virulence genes.** *In silico* detection of protein-encoding genes from the *Shigella* pathogenicity islands (SHI-1/SHE [12, 13], SHI-2 [14], SHI-3 [15], SHI-O [16], and SRL [17]) in the EIEC genomes demonstrated that portions of these pathogenicity islands were present in the EIEC genomes in a mostly lineage-specific manner (Table 3). More than 50% of the genes of *Shigella* pathogenicity islands SHI-1, SHI-2, and SHI-3 were identified in all of the EIEC genomes of lineage 1 (Table 3).

The EIEC genomes contained a wide range (3 to 91%) of the total number of genes from SHI-1, but none of the genomes contained all of the SHI-1 genes (Table 3). The serine protease autotransporters (*sigA* and *pic*) (12, 13, 51–54) of SHI-1 were identified in 75% (15/20) and 15% (3/20) of the EIEC genomes, respectively. The *Shigella* enterotoxin (ShET) genes, *set1A* and *set1B*, of SHI-1 were previously identified and characterized in *S. flexneri* 2a (12, 18, 55). In the present study, the ShET1 genes were detected only in the *S. flexneri* 2a genomes and in none of the other *Shigella* genomes analyzed (Table 3). The ShET1 genes were detected in three EIEC genomes, which were all in lineage 2 (Table 3). Since the reading frames of *set1A*, *set1B*, and *pic* overlap but are transcribed in opposite directions (51, 55, 56), it is not surprising that the genomes containing ShET1 were the same in the three EIEC genomes that also contained *pic* (Table 3). ShET1 was initially identified in *S. flexneri* 2a (55); however, it has also been found along with *pic* in EAEC and uropathogenic *E. coli* (51, 57). A second version of the *Shigella* enterotoxin, designated *sen* or ShET2 (58), was identified in all but two of the EIEC genomes (Table 3). Unlike ShET1, *sen*/ShET2 is encoded on the *Shigella* virulence plasmid (49, 58), and both of the EIEC genomes that lacked *sen*/ShET2 were also missing most of the virulence plasmid genes. Of the three EIEC genomes that contained ShET1, one genome (EC10018) did not have ShET2, while the other two genomes contained both ShET1 and ShET2 (Table 3).

Only two of the EIEC genomes (ATM460 and ATM461), both in lineage 1, had 100% (7/7) of the genes of SHI-3 (Table 3). The genes for aerobactin synthesis, *iucA* to *iucD* and *iutA* (14, 15) are present in both SHI-2 and SHI-3 and are involved in iron acquisition and contribute to *Shigella* virulence (Table 3). The aerobactin synthesis genes were present in some of the other EIEC genomes that did not contain the entire SHI-3 region (Table 3). Included among these were all of the EIEC genomes in lineage 3 (Table 3). A gene in the SHI-2 region, *shiA*, is involved in reducing the host cell inflammatory response (3, 59, 60) and was absent from all of the EIEC genomes and all but four of the *Shigella* genomes analyzed (Table 3). Another putative virulence gene in the SHI-2 region, *shiD*, which provides immunity to colicin I and colicin V (3), was identified in all of the EIEC genomes in EIEC

TABLE 3 *In silico* detection of *Shigella* virulence genes and pathogenicity islands

			No. (%) of genes[c]							Presence[d] of:						
Isolate ID	Phylogroup[a]	EIEC lineage[b]	SHI-1/SHE	SHI-2	SHI-3	SRL	T2SS	T6SS	*iucA-iucD*, *iutA*	*cadA*	*sigA*	*pic*	ShET1	ShET2	*shiA*	*shiD*
EIEC																
4608-58	E	1	21 (66)	14 (61)	6 (85)	24 (41)	11 (100)	4 (17)	4 (80)	+	+	−	−	+	−	+
ATM460	E	1	21 (66)	17 (74)	7 (100)	24 (41)	11 (100)	5 (22)	5 (100)	+	+	−	−	+	−	+
ATM461	E	1	21 (66)	17 (74)	7 (100)	25 (43)	11 (100)	5 (22)	5 (100)	+	+	−	−	+	−	+
EC10010	E	1	21 (66)	15 (65)	6 (85)	24 (41)	11 (100)	4 (17)	4 (80)	+	+	−	−	+	−	+
EC10032	E	1	21 (66)	16 (70)	6 (85)	24 (41)	11 (100)	4 (17)	4 (80)	+	+	−	−	+	−	+
EC10033	E	1	21 (66)	16 (70)	6 (85)	24 (41)	11 (100)	4 (17)	4 (80)	+	+	−	−	+	−	+
53638	A	2	9 (28)	8 (35)	0 (0)	12 (21)	11 (100)	23 (100)	0 (0)	+	−	−	−	+	−	−
ATM456	A	2	24 (75)	9 (39)	1 (14)	16 (28)	10 (91)	22 (96)	0 (0)	+	+	−	−	+	−	−
ATM457	A	2	29 (91)	16 (70)	7 (100)	16 (28)	11 (100)	23 (100)	5 (100)	+	+	+	+	+	−	−
EC10016	A	2	19 (59)	9 (39)	1 (14)	12 (21)	0 (0)	22 (96)	0 (0)	−	+	+	+	+	−	−
EC10018	A	2	29 (91)	16 (70)	7 (100)	14 (24)	11 (100)	22 (96)	5 (100)	+	+	+	+	−	−	−
M4163	A	2	16 (50)	8 (35)	0 (0)	19 (33)	11 (100)	22 (96)	0 (0)	+	−	−	−	+	−	−
1827-70	A	None[e]	9 (28)	3 (13)	0 (0)	18 (31)	11 (100)	18 (78)	0 (0)	+	−	−	−	−	−	−
ATM266	B1	3	14 (44)	16 (70)	6 (85)	10 (17)	11 (100)	22 (96)	5 (100)	−	+	−	−	+	−	+
ATM459	B1	3	12 (38)	16 (70)	6 (85)	10 (17)	11 (100)	20 (87)	5 (100)	−	+	−	−	+	−	+
ATM462	B1	3	12 (38)	13 (57)	6 (85)	7 (12)	11 (100)	22 (96)	5 (100)	−	+	−	−	+	−	+
ATM463	B1	3	13 (41)	14 (61)	6 (85)	17 (29)	11 (100)	1 (4)	5 (100)	−	+	−	−	+	−	+
ATM465	B1	3	14 (44)	15 (65)	6 (85)	11 (19)	11 (100)	22 (96)	5 (100)	−	+	−	−	+	−	+
CFSAN029787	B1	None	1 (3)	8 (35)	0 (0)	2 (3)	11 (100)	22 (96)	0 (0)	+	−	−	−	+	−	−
LT-68	B1	None	7 (22)	5 (22)	0 (0)	21 (36)	11 (100)	9 (39)	0 (0)	+	−	−	−	+	−	−
Shigella																
S. boydii ATCC 9905	B1*	NA[f]	2 (6)	15 (65)	6 (85)	5 (9)	10 (91)	5 (22)	5 (100)	−	−	−	−	+	−	−
S. dysenteriae 1012	B1*	NA	12 (38)	16 (70)	7 (100)	36 (62)	0 (0)	3 (13)	5 (100)	−	−	−	−	−	−	−
S. flexneri 2a 2457T	B1*	NA	32 (100)	23 (100)	7 (100)	10 (17)	0 (0)	0 (0)	5 (100)	−	+	+	+	−	+	+
S. flexneri 2a 301	B1*	NA	32 (100)	23 (100)	7 (100)	12 (21)	0 (0)	0 (0)	5 (100)	−	+	+	+	+	+	+
S. boydii 3083-94	B1	NA	10 (31)	16 (70)	6 (85)	11 (19)	10 (91)	0 (0)	5 (100)	−	+	−	−	+	−	+
S. dysenteriae S6554	B1	NA	10 (31)	16 (70)	6 (85)	11 (19)	10 (91)	0 (0)	5 (100)	−	+	−	−	−	−	+
S. flexneri CCH060	B1	NA	2 (6)	15 (65)	6 (85)	0 (0)	10 (91)	0 (0)	5 (100)	−	−	−	−	−	−	−
S. sonnei 046	B1	NA	6 (19)	18 (78)	7 (100)	13 (22)	0 (0)	20 (87)	5 (100)	−	+	−	−	+	+	−
S. dysenteriae 197	E	NA	8 (25)	9 (39)	0 (0)	22 (37)	11 (100)	0 (0)	0 (0)	+	−	−	−	+	+	−

[a] *E. coli* phylogroups E, A, and B1 are indicated. B1* indicates that the genome is typically within the B1 phylogroup.

[b] The phylogenomic clades that contained EIEC are designated 1 to 3.

[c] The total numbers of genes associated with the representative pathogenicity islands or other regions are 32 (SHI-1/SHE, AF200692.2), 23 (SHI-2, AF141323.1), 7 (SHI-3, AF335540.1), 58 (SRL, AF326777.3), 11 (T2SS, CP000034.1), and 23 (T6SS, AAJX02000009.1).

[d] The genes that were identified with an LS-BSR of ≥0.8 were considered present (+), while those with an LS-BSR of <0.8 but ≥0.4 were considered divergent and those with an LS-BSR of <0.4 were considered absent (−).

[e] The genome was not within one of the EIEC clades.

[f] NA, genome not applicable.

lineages 1 and 3 but in none of the genomes in EIEC lineage 2 (Table 3). The presence of *Shigella* genomic island SHI-O (16) in the EIEC genomes was also investigated. The genes from SHI-O involved in serotype conversion of *Shigella* (*gtrA*, *gtrB*, and *gtrV*) were not identified in any of the EIEC genomes.

The genes encoding LCD (*cadA*), a lysine:cadaverine antiporter (*cadB*), and a transcriptional activator (*cadC*) of the *cadBA* operon were previously determined to be absent from or disrupted in *Shigella* and EIEC isolates, although they are present in most other *E. coli* isolates (61). The *cadA* and *cadC* genes are considered antivirulence genes, since their loss enhances the pathogenicity of these strains (62–64). *In silico* detection of *cadA* in the EIEC and *Shigella* genomes in this study demonstrated that it was present in 70% (14/20) of the EIEC genomes but only one *Shigella* genome (*S. dysenteriae* Sd197) (Table 3). The genomes in EIEC lineages 1 and 2 all contained an intact *cadA* gene, with the exception of EC10016 (Table 3). However, upon closer inspection, the *cadA* gene is present in isolate EC10016 but interrupted by an insertion element (Table 3). The five EIEC genomes that were missing *cadA* were all in EIEC lineage 3 of phylogroup B1, which is the EIEC lineage that was most closely related to the *Shigella* genomes of phylogroup B1 (Table 3; Fig. 1). The other two EIEC genomes of phylogroup B1 (LT-68 and CFSAN029787) did contain *cadA* (Fig. 1). In some EIEC strains, *cadA* was identified as intact and the lack of LDC activity in these strains was due to inactivation of *cadC*, which encodes the transcriptional regulator of *cadBA* (65). While 70% of the EIEC genomes described in this study contain an intact *cadA* gene, further characterization is required to determine whether *cadA* encodes a functional enzyme in these strains or whether the expression of *cadA* is disrupted by a mutation in *cadC*.

Other putative virulence-related features that were detected in the EIEC genomes included genes with similarity to those of a T2SS (Table 3). Over 90% of the T2SS genes from *S. dysenteriae* Sd197 were identified in all of the EIEC genomes, except EC10016 (Table 3). Interestingly, while EC10016 was missing all of the T2SS genes, the other genomes belonging to the same EIEC lineage (lineage 2) did contain these genes (Fig. 1; Table 3). Further investigation is necessary to confirm that the T2SS genes are missing from this isolate or whether they are absent from the genome assembly. The T2SS is also called the general secretion pathway and has been linked to biofilm formation and virulence of EPEC (66) and toxin secretion in enterotoxigenic *E. coli* (ETEC) (67). However, the T2SS is also functional in nonpathogenic *E. coli* and is thought to contribute to survival (68). The role of T2SS in the pathogenicity of EIEC is unknown.

Another secretion system of interest for its potential contribution to virulence is the T6SS (69–72). Genes with similarity to those of the T6SS of EPEC isolate B171 were identified in nearly all of the EIEC genomes of phylogroups A and B1 but were absent from all of the EIEC genomes of phylogroup E (Table 3). Two genomes of phylogroup B1 that were missing most of the T6SS genes were ATM463 and LT-68; each had less than half of the T6SS genes required for a functional secretion system (Table 3). In contrast, the T6SS genes were absent from all of the *Shigella* genomes investigated except *S. sonnei* 046 (Table 3). While the T6SS has been demonstrated to contribute to virulence in avian pathogenic *E. coli* (69) and EAEC (71, 72), its role in the virulence of EIEC or *Shigella* has not been investigated. The T6SS has also been reported to have roles other than those directly involved in virulence, such as providing a competitive advantage in the presence of other bacteria in a community structure (73–75). Another unique feature identified in the genome of EIEC isolate LT-68 was the heat-stable enterotoxin encoded by *astA* (accession no. L11241.1), which is typically found in EAEC (76).

DISCUSSION

The genomes analyzed in this study represent a global collection of EIEC isolates that have diverse serotypes and, as demonstrated by our findings, also have considerable genomic diversity (Table 1; Fig. 1). While previous studies have demonstrated close genetic similarity among EIEC and *Shigella* isolates (20, 77), the findings from this study highlight the genomic and potential virulence diversity of this pathovar.

Using phylogenomic analysis, we demonstrated that most of the EIEC isolates were more closely related to other *E. coli* isolates than to *Shigella* isolates (Fig. 1). Meanwhile, a gene-based comparison highlighted that EIEC isolates had greater similarity to *Shigella* than to other *E. coli* isolates. Furthermore, the identification of only seven genes that were unique to EIEC demonstrated that EIEC has many genomic similarities to other *E. coli* and *Shigella* strains. Meanwhile, the detection of 87 genes that were present in EIEC and *Shigella* that were not in the other *E. coli* strains demonstrates that there is greater gene-based similarity between EIEC and *Shigella*. Further inspection of these genes indicated that many of them are likely plasmid associated and could be part of invasion plasmid pINV, which is a genetic feature that unites EIEC and *Shigella* both genotypically and phenotypically (1, 2) (see Table S2 in the supplemental material).

Comparative analysis of EIEC pINV demonstrated that for most of the EIEC genomes analyzed, genes of the invasion plasmid were more similar to those of other EIEC isolates analyzed than to *Shigella* pINV (Fig. 2). This finding demonstrates that the plasmid was likely acquired early in the divergence of a particular EIEC or *Shigella* lineage and it has been stable and evolving along with the chromosome over time. However, there were exceptions for a limited number of the EIEC isolates that exhibited greater similarity to the *Shigella* plasmids than to other EIEC isolates within the same lineage (Fig. 2; see Fig. S1 in the supplemental material). This suggests that these isolates may have lost the EIEC version of the plasmid and acquired a pINV plasmid from *Shigella*. It is possible that there are multiple models of plasmid evolution, including plasmid gain and loss, in the EIEC isolates and *Shigella*. Detailed genetic studies are required to examine each of these potential pathways of evolution.

In addition to the presence of the invasion plasmid, another feature that is shared by *Shigella* and EIEC but not other *E. coli* strains is the absence of LDC activity (2). While it is present in most strains of *E. coli*, the LDC gene, *cadA*, is absent from or inactive in *Shigella* (61) and *cadA* is not expressed in EIEC strains because of mutations in the regulator gene, *cadC* (65). However, our findings demonstrate that while *cadA* was absent from nearly all of the *Shigella* isolates analyzed, it was present in 70% of the EIEC genomes, and the *cadC* gene is present in 64.3% (9/14) of the EIEC genomes. Experimental verification of LDC function in these EIEC isolates is required (Table 3). Similar to presence or absence of the other genes detected by *in silico* analysis, that of *cadA* in the EIEC genomes exhibited phylogroup specificity (Table 3). Similarly, there are few universal genes that are lacking in all EIEC or EIEC and *Shigella* isolates that are present in the *E. coli* isolates used for comparison, suggesting that there is not a single antivirulence mechanism present among all of the EIEC isolates.

The variable presence of the *Shigella* pathogenicity islands among the EIEC genomes demonstrates the phylogroup- or lineage-specific diversity of the EIEC and *Shigella* isolates analyzed in this study (Table 3). Furthermore, many of these *Shigella*-specific acquired regions appear to be stable in the EIEC isolates within a lineage, with the exception of a few isolates that are missing regions or individual genes that were present in all of the other isolates within that lineage (Table 3). There are features that are shared among the lineages identified, indicating that these lineages are more permissive to the acquisition of *Shigella* virulence genes and plasmid pINV. These findings suggest that there is not a single origin of the EIEC pathovar, but rather multiple lineages of *E. coli* have acquired common virulence factors, as has been previously suggested on the basis of single-gene analyses (20, 21).

Pettengill et al. recently used a SNP-based approach to compare the genomes of a collection of diverse *E. coli* and *Shigella* isolates (26). However, the clusters defined in their phylogeny group together isolates of the diverse pathovars, including one cluster that contains EHEC, EPEC, EAEC, EIEC, and *S. dysenteriae*. Other studies that have compared a larger collection of genomes have demonstrated these clusters could be further subdivided into lineages that in many cases contain lineage-specific genes or other genomic regions that can be used for more targeted diagnostics (29, 41). Comparative genomics provide a finer-scale level of resolution that allows discrimination between genomically related *E. coli* isolates that belong to different pathovars. As demonstrated in the present study, there were many virulence-associated regions of the EIEC and *Shigella* genomes that exhibited phylogroup or lineage specificity (Table 3), suggesting a shared ancestral lineage. Further investigation is necessary to determine whether some of these genes could be used to develop a rapid diagnostic assay that could identify whether clinical isolates belong to any of the three EIEC lineages described in this study, thus distinguishing them from other *E. coli* or *Shigella* isolates.

In conclusion, a molecular study that analyzed eight housekeeping genes (77) and recent genome sequencing analyses have suggested that *Shigella* developed through the acquisition of common virulence factors in multiple lineages of *E. coli* (41, 78, 79). The findings of the present study demonstrate that EIEC likely arose through the acquisition of mobile-element-encoded virulence genes by permissive recipient isolates occupying multiple distinct lineages of *E. coli*, in a fashion similar to that of *Shigella*. For many genes, it appears that the acquisition events occurred early in the development of a lineage since the genes exhibit lineage specificity. However, in some cases, particularly when plasmid-borne genes were analyzed, the plasmids exhibit greater sequence similarity than those of other more distantly related isolates, suggesting that the plasmids were more recently acquired. Like the evolution of *Shigella*, that of EIEC was likely driven by the acquisition of the invasion plasmid, as well as some, but not all, of the genes of the *Shigella* pathogenicity islands. As previously demonstrated for EPEC (29, 80), ETEC (30, 81), and *Shigella* (41, 82), the same pathovar can occur in evolutionarily diverse lineages. Similar to what we have previously described for the attaching and effacing *E. coli* pathovars (EPEC and EHEC) (29), *Shigella* and EIEC include diverse members that have in common virulence mechanisms driven by the acquisition of mobile genetic element-encoded virulence genes (genomic islands, plasmids, or phage). We anticipate that the decoding of additional EIEC genome sequences will reveal even greater genomic diversity beyond what has been presented in this study, most likely reflecting the diverse nature of this pathovar. This trend of the identification of significant genomic diversity is becoming the norm in the study of *E. coli* and *Shigella* genomics, as we integrate high-throughput genomics into the epidemiological studies of modern collections of strains.

ACKNOWLEDGMENTS

This project was funded in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract HHSN272200900009C, grants U19AI090873 and U19AI110820, and startup funds from the State of Maryland. S.R.L. was supported by a fellowship appointment administered by the Oak Ridge Institute for Science and Education.

REFERENCES

1. **Kaper JB, Nataro JP, Mobley HL.** 2004. Pathogenic *Escherichia coli*. Nat Rev Microbiol **2:**123–140. http://dx.doi.org/10.1038/nrmicro818.
2. **Nataro JP, Kaper JB.** 1998. Diarrheagenic *Escherichia coli*. Clin Microbiol Rev **11:**142–201.
3. **Schroeder GN, Hilbi H.** 2008. Molecular pathogenesis of *Shigella* spp.: controlling host cell signaling, invasion, and death by type III secretion. Clin Microbiol Rev **21:**134–156. http://dx.doi.org/10.1128/CMR.00032-07.
4. **DuPont HL, Formal SB, Hornick RB, Snyder MJ, Libonati JP, Sheahan DG, LaBrec EH, Kalas JP.** 1971. Pathogenesis of Escherichia coli diarrhea. N Engl J Med **285:**1–9. http://dx.doi.org/10.1056/NEJM197107012850101.
5. **Kotloff KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, Sansonetti PJ, Adak GK, Levine MM.** 1999. Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. Bull World Health Organ **77:**651–666.
6. **Escher M, Scavia G, Morabito S, Tozzoli R, Maugliani A, Cantoni S, Fracchia S, Bettati A, Casa R, Gesu GP, Torresani E, Caprioli A.** 2014. A severe foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive *Escherichia coli*, an uncommon agent. Epidemiol Infect **142:**2559–2566. http://dx.doi.org/10.1017/S0950268814000181.
7. **Silva RM, Toledo MR, Trabulsi LR.** 1980. Biochemical and cultural characteristics of invasive *Escherichia coli*. J Clin Microbiol **11:**441–444.
8. **Sinai AP, Bavoil PM.** 1993. Hyper-invasive mutants define a novel Pho-regulated invasion pathway in *Escherichia coli*. Mol Microbiol **10:**1125–1137. http://dx.doi.org/10.1111/j.1365-2958.1993.tb00982.x.
9. **Hsia RC, Small PL, Bavoil PM.** 1993. Characterization of virulence genes of enteroinvasive *Escherichia coli* by Tn*phoA* mutagenesis: identification of *invX*, a gene required for entry into HEp-2 cells. J Bacteriol **175:**4817–4823.
10. **Small PL, Falkow S.** 1988. Identification of regions on a 230-kilobase plasmid from enteroinvasive *Escherichia coli* that are required for entry into HEp-2 cells. Infect Immun **56:**225–229.
11. **Sansonetti PJ, d'Hauteville H, Ecobichon C, Pourcel C.** 1983. Molecular comparison of virulence plasmids in *Shigella* and enteroinvasive *Escherichia coli*. Ann Microbiol (Paris) **134A:**295–318.
12. **Al-Hasani K, Henderson IR, Sakellaris H, Rajakumar K, Grant T, Nataro JP, Robins-Browne R, Adler B.** 2000. The *sigA* gene which is borne on the *she* pathogenicity island of *Shigella flexneri* 2a encodes an exported cytopathic protease involved in intestinal fluid accumulation. Infect Immun **68:**2457–2463. http://dx.doi.org/10.1128/IAI.68.5.2457-2463.2000.
13. **Rajakumar K, Sasakawa C, Adler B.** 1997. Use of a novel approach, termed island probing, identifies the *Shigella flexneri she* pathogenicity island which encodes a homolog of the immunoglobulin A protease-like family of proteins. Infect Immun **65:**4606–4614.
14. **Moss JE, Cardozo TJ, Zychlinsky A, Groisman EA.** 1999. The *selC*-associated SHI-2 pathogenicity island of *Shigella flexneri*. Mol Microbiol **33:**74–83. http://dx.doi.org/10.1046/j.1365-2958.1999.01449.x.
15. **Purdy GE, Payne SM.** 2001. The SHI-3 iron transport island of *Shigella boydii* 0-1392 carries the genes for aerobactin synthesis and transport. J Bacteriol **183:**4176–4182. http://dx.doi.org/10.1128/JB.183.14.4176-4182.2001.
16. **Huan PT, Bastin DA, Whittle BL, Lindberg AA, Verma NK.** 1997. Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in *Shigella flexneri* bacteriophage SfV. Gene **195:**217–227. http://dx.doi.org/10.1016/S0378-1119(97)00143-1.
17. **Luck SN, Turner SA, Rajakumar K, Sakellaris H, Adler B.** 2001. Ferric dicitrate transport system (Fec) of *Shigella flexneri* 2a YSH6000 is encoded on a novel pathogenicity island carrying multiple antibiotic resistance genes. Infect Immun **69:**6012–6021. http://dx.doi.org/10.1128/IAI.69.10.6012-6021.2001.
18. **Noriega FR, Liao FM, Formal SB, Fasano A, Levine MM.** 1995. Prevalence of *Shigella* enterotoxin 1 among *Shigella* clinical isolates of diverse serotypes. J Infect Dis **172:**1408–1410. http://dx.doi.org/10.1093/infdis/172.5.1408.
19. **Lan R, Lumb B, Ryan D, Reeves PR.** 2001. Molecular evolution of large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*. Infect Immun **69:**6303–6309. http://dx.doi.org/10.1128/IAI.69.10.6303-6309.2001.
20. **Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR.** 2004. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. Infect Immun **72:**5080–5088. http://dx.doi.org/10.1128/IAI.72.9.5080-5088.2004.
21. **Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, Zhu Y, Yu J, Jin Q.** 2007. Revisiting the molecular evolutionary history of *Shigella* spp. J Mol Evol **64:**71–79. http://dx.doi.org/10.1007/s00239-006-0052-8.
22. **Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E.** 2003. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. J Mol Evol **57:**140–148. http://dx.doi.org/10.1007/s00239-003-2460-3.
23. **Pettengill EA, Hoffmann M, Binet R, Roberts RJ, Payne J, Allard M, Michelacci V, Minelli F, Morabito S.** 2015. Complete genome sequence of enteroinvasive *Escherichia coli* O96:H19 associated with a severe foodborne outbreak. Genome Announc **3:**e00883–15. http://dx.doi.org/10.1128/genomeA.00883-15.
24. **Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J.** 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol **190:**6881–6893. http://dx.doi.org/10.1128/JB.00619-08.
25. **Leonard SR, Lacher DW, Lampel KA.** 2015. Draft genome sequences of the enteroinvasive *Escherichia coli* strains M4163 and 4608-58. Genome Announc **3:**e01395–14. http://dx.doi.org/10.1128/genomeA.01395-14.
26. **Pettengill EA, Pettengill JB, Binet R.** 2015. Phylogenetic analyses of *Shigella* and enteroinvasive *Escherichia coli* for the identification of molecular epidemiological markers: whole-genome comparative analysis does

not support distinct genera designation. Front Microbiol **6**:1573. http://dx.doi.org/10.3389/fmicb.2015.01573.

27. **Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson NR.** 2015. A complete view of the genetic diversity of the Escherichia coli O-antigen biosynthesis gene cluster. DNA Res **22:**101–107. http://dx.doi.org/10.1093/dnares/dsu043.
28. **Wang L, Rothemund D, Curd H, Reeves PR.** 2003. Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. J Bacteriol **185:** 2936–2943. http://dx.doi.org/10.1128/JB.185.9.2936-2943.2003.
29. **Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA.** 2013. Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. Proc Natl Acad Sci U S A **110:**12810–12815. http://dx.doi.org/10.1073/pnas.1306836110.
30. **Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, Rasko DA.** 2011. A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. Infect Immun **79:**950–960. http://dx.doi.org/10.1128/IAI.00932-10.
31. **Angiuoli SV, Salzberg SL.** 2011. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics **27:**334–342. http://dx.doi.org/10.1093/bioinformatics/btq665.
32. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75:**7537–7541. http://dx.doi.org/10.1128/AEM.01541-09.
33. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.
34. **Sahl JW, Caporaso JG, Rasko DA, Keim P.** 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. Peer J **2:**e332. http://dx.doi.org/10.7717/peerj.332.
35. **Sahl JW, Gillece JD, Schupp JM, Waddell VG, Driebe EM, Engelthaler DM, Keim P.** 2013. Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in *Acinetobacter*. PLoS One **8:**e54287. http://dx.doi.org/10.1371/journal.pone.0054287.
36. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.
37. **Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF.** 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol **4:**41. http://dx.doi.org/10.1186/1741-7007-4-41.
38. **Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV.** 2010. Ergatis: a web interface and scalable software system for bioinformatics workflows. Bioinformatics **26:**1488–1492. http://dx.doi.org/10.1093/bioinformatics/btq167.
39. **Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, White O, Wortman J, Mahurkar A, Giglio MG.** 2011. The IGS standard operating procedure for automated prokaryotic annotation. Stand Genomic Sci **4:**244–251. http://dx.doi.org/10.4056/sigs.1223234.
40. **Matsushita S, Yamada S, Kai A, Kudoh Y.** 1993. Invasive strains of *Escherichia coli* belonging to serotype O121:NM. J Clin Microbiol **31:** 3034–3035.
41. **Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B, Breiman RF, Gilmour M, Nataro JP, Rasko DA.** 2015. Defining the phylogenomics of *Shigella* species: a pathway to diagnostics. J Clin Microbiol **53:**951–960. http://dx.doi.org/10.1128/JCM.03527-14.
42. **Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E, Lortholary O, Clermont O, Denamur E, Picard B, Nassif X, Brisse S.** 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics **9:**560. http://dx.doi.org/10.1186/1471-2164-9-560.
43. **Tenaillon O, Skurnik D, Picard B, Denamur E.** 2010. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol **8:**207–217. http://dx.doi.org/10.1038/nrmicro2298.
44. **Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Yang H, Wang J, Xu J, Pallen MJ, Aepfelbacher M, Yang R.** 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N Engl J Med **365:**718–724. http://dx.doi.org/10.1056/NEJMoa1107643.
45. **Scheutz F, Nielsen EM, Frimodt-Moller J, Boisen N, Morabito S, Tozzoli R, Nataro JP, Caprioli A.** 2011. Characteristics of the enteroaggregative Shiga toxin/verotoxin-producing *Escherichia coli* O104:H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, May to June 2011. Euro Surveill **16:**19889. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19889.
46. **Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Moller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK.** 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med **365:**709–717. http://dx.doi.org/10.1056/NEJMoa1106920.
47. **Rasko DA, Myers GS, Ravel J.** 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics **6:**2. http://dx.doi.org/10.1186/1471-2105-6-2.
48. **Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J.** 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res **30:**4432–4441. http://dx.doi.org/10.1093/nar/gkf566.
49. **Venkatesan MM, Goldberg MB, Rose DJ, Grotbeck EJ, Burland V, Blattner FR.** 2001. Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. Infect Immun **69:**3271–3285. http://dx.doi.org/10.1128/IAI.69.5.3271-3285.2001.
50. **Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, Sansonetti P, Parsot C.** 2000. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. Mol Microbiol **38:**760–771. http://dx.doi.org/10.1046/j.1365-2958.2000.02179.x.
51. **Henderson IR, Czeczulin J, Eslava C, Noriega F, Nataro JP.** 1999. Characterization of Pic, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect Immun **67:**5587–5596.
52. **Zhang J, Qian L, Wu Y, Cai X, Li X, Cheng X, Qu D.** 2013. Deletion of *pic* results in decreased virulence for a clinical isolate of *Shigella flexneri* 2a from China. BMC Microbiol **13:**31. http://dx.doi.org/10.1186/1471-2180-13-31.
53. **Al-Hasani K, Navarro-Garcia F, Huerta J, Sakellaris H, Adler B.** 2009. The immunogenic SigA enterotoxin of *Shigella flexneri* 2a binds to HEp-2 cells and induces fodrin redistribution in intoxicated epithelial cells. PLoS One **4:**e8223. http://dx.doi.org/10.1371/journal.pone.0008223.
54. **Chua EG, Al-Hasani K, Scanlon M, Adler B, Sakellaris H.** 2015. Determinants of proteolysis and cell-binding for the *Shigella flexneri* cytotoxin, SigA. Curr Microbiol **71:**613–617. http://dx.doi.org/10.1007/s00284-015-0893-8.
55. **Fasano A, Noriega FR, Maneval DR, Jr, Chanasongcram S, Russell R, Guandalini S, Levine MM.** 1995. *Shigella* enterotoxin 1: an enterotoxin of *Shigella flexneri* 2a active in rabbit small intestine *in vivo* and *in vitro*. J Clin Invest **95:**2853–2861. http://dx.doi.org/10.1172/JCI117991.
56. **Behrens M, Sheikh J, Nataro JP.** 2002. Regulation of the overlapping *pic*/*set* locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect Immun **70:**2915–2925. http://dx.doi.org/10.1128/IAI.70.6.2915-2925.2002.
57. **Navarro-Garcia F, Gutierrez-Jimenez J, Garcia-Tovar C, Castro LA, Salazar-Gonzalez H, Cordova V.** 2010. Pic, an autotransporter protein secreted by different pathogens in the *Enterobacteriaceae* family, is a potent mucus secretagogue. Infect Immun **78:**4101–4109. http://dx.doi.org/10.1128/IAI.00523-10.
58. **Nataro JP, Seriwatana J, Fasano A, Maneval DR, Guers LD, Noriega F, Dubovsky F, Levine MM, Morris JG, Jr.** 1995. Identification and cloning of a novel plasmid-encoded enterotoxin of enteroinvasive *Escherichia coli* and *Shigella* strains. Infect Immun **63:**4721–4728.
59. **Ingersoll MA, Zychlinsky A.** 2006. ShiA abrogates the innate T-cell response to *Shigella flexneri* infection. Infect Immun **74:**2317–2327. http://dx.doi.org/10.1128/IAI.74.4.2317-2327.2006.
60. **Ingersoll MA, Moss JE, Weinrauch Y, Fisher PE, Groisman EA, Zychlinsky A.** 2003. The ShiA protein encoded by the *Shigella flexneri* SHI-2

pathogenicity island attenuates inflammation. Cell Microbiol **5:**797–807. http://dx.doi.org/10.1046/j.1462-5822.2003.00320.x.

61. **Maurelli AT, Fernandez RE, Bloch CA, Rode CK, Fasano A.** 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc Natl Acad Sci U S A **95:**3943–3948. http://dx.doi.org/10.1073/pnas.95.7.3943.
62. **Jores J, Torres AG, Wagner S, Tutt CB, Kaper JB, Wieler LH.** 2006. Identification and characterization of "pathoadaptive mutations" of the cadBA operon in several intestinal *Escherichia coli*. Int J Med Microbiol **296:**547–552. http://dx.doi.org/10.1016/j.ijmm.2006.07.002.
63. **Bliven KA, Maurelli AT.** 2012. Antivirulence genes: insights into pathogen evolution through gene loss. Infect Immun **80:**4061–4070. http://dx.doi.org/10.1128/IAI.00740-12.
64. **Day WA, Jr, Fernandez RE, Maurelli AT.** 2001. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. Infect Immun **69:**7471–7480. http://dx.doi.org/10.1128/IAI.69.12.7471-7480.2001.
65. **Casalino M, Latella MC, Prosseda G, Colonna B.** 2003. CadC is the preferential target of a convergent evolution driving enteroinvasive *Escherichia coli* toward a lysine decarboxylase-defective phenotype. Infect Immun **71:**5472–5479. http://dx.doi.org/10.1128/IAI.71.10.5472-5479.2003.
66. **Baldi DL, Higginson EE, Hocking DM, Praszkier J, Cavaliere R, James CE, Bennett-Wood V, Azzopardi KI, Turnbull L, Lithgow T, Robins-Browne RM, Whitchurch CB, Tauschek M.** 2012. The type II secretion system and its ubiquitous lipoprotein substrate, SslE, are required for biofilm formation and virulence of enteropathogenic *Escherichia coli*. Infect Immun **80:**2042–2052. http://dx.doi.org/10.1128/IAI.06160-11.
67. **Mudrak B, Kuehn MJ.** 2010. Specificity of the type II secretion systems of enterotoxigenic *Escherichia coli* and *Vibrio cholerae* for heat-labile enterotoxin and cholera toxin. J Bacteriol **192:**1902–1911. http://dx.doi.org/10.1128/JB.01542-09.
68. **Decanio MS, Landick R, Haft RJ.** 2013. The non-pathogenic *Escherichia coli* strain W secretes SslE via the virulence-associated type II secretion system beta. BMC Microbiol **13:**130. http://dx.doi.org/10.1186/1471-2180-13-130.
69. **Ma J, Bao Y, Sun M, Dong W, Pan Z, Zhang W, Lu C, Yao H.** 2014. Two functional type VI secretion systems in avian pathogenic *Escherichia coli* are involved in different pathogenic pathways. Infect Immun **82:**3867–3879. http://dx.doi.org/10.1128/IAI.01769-14.
70. **Shrivastava S, Mande SS.** 2008. Identification and functional characterization of gene components of type VI secretion system in bacterial genomes. PLoS One **3:**e2955. http://dx.doi.org/10.1371/journal.pone.0002955.
71. **Dudley EG, Thomson NR, Parkhill J, Morin NP, Nataro JP.** 2006. Proteomic and microarray characterization of the AggR regulon identifies a *pheU* pathogenicity island in enteroaggregative *Escherichia coli*. Mol Microbiol **61:**1267–1282. http://dx.doi.org/10.1111/j.1365-2958.2006.05281.x.
72. **Aschtgen MS, Bernard CS, De Bentzmann S, Lloubes R, Cascales E.** 2008. SciN is an outer membrane lipoprotein required for type VI secretion in enteroaggregative *Escherichia coli*. J Bacteriol **190:**7523–7531. http://dx.doi.org/10.1128/JB.00945-08.
73. **Jani AJ, Cotter PA.** 2010. Type VI secretion: not just for pathogenesis anymore. Cell Host Microbe **8:**2–6. http://dx.doi.org/10.1016/j.chom.2010.06.012.
74. **Russell AB, Hood RD, Bui NK, LeRoux M, Vollmer W, Mougous JD.** 2011. Type VI secretion delivers bacteriolytic effectors to target cells. Nature **475:**343–347. http://dx.doi.org/10.1038/nature10244.
75. **MacIntyre DL, Miyata ST, Kitaoka M, Pukatzki S.** 2010. The *Vibrio cholerae* type VI secretion system displays antimicrobial properties. Proc Natl Acad Sci U S A **107:**19520–19524. http://dx.doi.org/10.1073/pnas.1012931107.
76. **Savarino SJ, Fasano A, Watson J, Martin BM, Levine MM, Guandalini S, Guerry P.** 1993. Enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 represents another subfamily of *E. coli* heat-stable toxin. Proc Natl Acad Sci U S A **90:**3093–3097. http://dx.doi.org/10.1073/pnas.90.7.3093.
77. **Pupo GM, Lan R, Reeves PR.** 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A **97:**10567–10572. http://dx.doi.org/10.1073/pnas.180094797.
78. **Connor TR, Barker CR, Baker KS, Weill FX, Talukder KA, Smith AM, Baker S, Gouali M, Pham Thanh D, Jahan Azmi I, Dias da Silveira W, Semmler T, Wieler LH, Jenkins C, Cravioto A, Faruque SM, Parkhill J, Wook Kim D, Keddy KH, Thomson NR.** 2015. Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. eLife **4:**e07335. http://dx.doi.org/10.7554/eLife.07335.
79. **Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, Choi SY, Kim SH, da Silveira WD, Pickard DJ, Farrar JJ, Parkhill J, Dougan G, Thomson NR.** 2012. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet **44:**1056–1059. http://dx.doi.org/10.1038/ng.2369.
80. **Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A, Mandomando I, Ochieng JB, Ramamurthy T, Tamboura B, Qureshi S, Quadri F, Zaidi A, Kotloff KL, Levine MM, Barry EM, Kaper JB, Rasko DA, Nataro JP.** 2016. Genomic diversity of EPEC associated with clinical presentations of differing severity. Nat Microbiol **1:**15014. http://dx.doi.org/10.1038/nmicrobiol.2015.14.
81. **Sahl JW, Sistrunk JR, Fraser CM, Hine E, Baby N, Begum Y, Luo Q, Sheikh A, Qadri F, Fleckenstein JM, Rasko DA.** 2015. Examination of the enterotoxigenic *Escherichia coli* population structure during human infection. mBio **6:**e00501. http://dx.doi.org/10.1128/mBio.00501-15.
82. **The HC, Thanh DP, Holt KE, Thomson NR, Baker S.** 2016. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. Nat Rev Microbiol **14:**235–250. http://dx.doi.org/10.1038/nrmicro.2016.10.
83. **Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J.** 2006. TM4 microarray software suite. Methods Enzymol **411:**134–193. http://dx.doi.org/10.1016/S0076-6879(06)11009-5.