DNA Research, 2015, 22(1), 101–107
doi: 10.1093/dnares/dsu043
Advance Access Publication Date: 26 November 2014
Full Paper

Full Paper

A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster

Atsushi Iguchi[1,*], Sunao Iyoda[2], Taisei Kikuchi[3], Yoshitoshi Ogura[4,5], Keisuke Katsura[5], Makoto Ohnishi[2], Tetsuya Hayashi[4,5], and Nicholas R. Thomson[6,7]

[1]Department of Animal and Grassland Sciences, Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan, [2]Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo 162-8640, Japan, [3]Division of Parasitology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan, [4]Division of Microbiology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan, [5]Division of Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, Miyazaki 889-1692, Japan, [6]Pathogen Genomics, The Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK, and [7]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

*To whom correspondence should be addressed. Tel/Fax. +81 985-58-7507. E-mail: iguchi@med.miyazaki-u.ac.jp

Edited by Dr Katsumi Isono

Received 16 September 2014; Accepted 3 November 2014

Abstract

The O antigen constitutes the outermost part of the lipopolysaccharide layer in Gram-negative bacteria. The chemical composition and structure of the O antigen show high levels of variation even within a single species revealing itself as serological diversity. Here, we present a complete sequence set for the O-antigen biosynthesis gene clusters (O-AGCs) from all 184 recognized *Escherichia coli* O serogroups. By comparing these sequences, we identified 161 well-defined O-AGCs. Based on the *wzx/wzy* or *wzm/wzt* gene sequences, in addition to 145 singletons, 37 serogroups were placed into 16 groups. Furthermore, phylogenetic analysis of all the *E. coli* O-serogroup reference strains revealed that the nearly one-quarter of the 184 serogroups were found in the ST10 lineage, which may have a unique genetic background allowing a more successful exchange of O-AGCs. Our data provide a complete view of the genetic diversity of O-AGCs in *E. coli* showing a stronger association between host phylogenetic lineage and O-serogroup diversification than previously recognized. These data will be a valuable basis for developing a systematic molecular O-typing scheme that will allow traditional typing approaches to be linked to genomic exploration of *E. coli* diversity.

Key words: *E. coli*, O-antigen biosynthesis gene cluster, horizontal gene transfer, O serogroup, genomic diversity

1. Introduction

Cell-surface polysaccharides play an essential role in the ability of bacteria to survive and persist in the environment and in host organisms.[1] The O-antigen polysaccharide constitutes the outermost part of the lipopolysaccharide (LPS) present in the outer membrane of Gram-negative bacteria. The chemical composition and structure of the O-antigen exhibit high levels of variation even within a single species.[2–5] This observation is corroborated by the huge serological

variation of somatic O antigens. Currently, the O serogrouping, sometimes combined with H (flagellar) antigens and K (capsular polysaccharide) antigens, is a standard method for subtyping of *Escherichia coli* strains in taxonomical and epidemiological studies. In particular, identification of strains of the same O serogroup is a prerequisite to start any actions for outbreak investigations and surveillance.

Thus far, the World Health Organization Collaborating Centre for Reference and Research on *Escherichia* and *Klebsiella* based at the Statens Serum Institut (SSI) in Denmark (http://www.ssi.dk/English.aspx) has recognized 184 *E. coli* O serogroups. It is generally believed that the O serogrouping of *E. coli* strains provides valuable information for identifying pathogenic clonal groups, especially for public health surveillance. For example, O157 is a leading O serogroup associated with enterohemorrhagic *E. coli* (EHEC) and is a significant food-borne pathogen worldwide.[6,7] Other important EHEC O serogroups include O26, O103, and O111.[8] The Shiga toxin-producing *E. coli* O104:H4 was found responsible for a large human food-borne disease outbreak in Europe, 2011.[9] Another notable example is strains of serogroup O25; extended-spectrum beta lactamase (ESBL)-producing, multidrug-resistant *E. coli* O25:H4 has emerged worldwide to cause a wide variety of community and nosocomial infections.[10]

In *E. coli*, the genes required for O-antigen biosynthesis are clustered at a chromosomal locus flanked by the colanic acid biosynthesis gene cluster (*wca* genes) and the histidine biosynthesis (*his*) operon. Generally, the O-antigen biosynthesis genes fall into three classes: (i) the nucleotide sugar biosynthesis genes, (ii) the sugar transferase genes, and (iii) those for O-unit translocation and chain synthesis (*wzx/wzy* in the Wzx/Wzy-dependent pathway and *wzm/wzt* in the Wzm/Wzt-dependent ABC transporter pathway).[11] To date, >90 types of O-antigen biosynthesis gene cluster (O-AGC) sequences have been determined, with the majority derived from major human and animal pathogens.[12] Sequence comparisons of these O-AGCs indicate a great variety of genetic structures. Several studies have provided evidence to show that horizontal transfer and replacement of a part or all of the O-AGC have caused shifts in O serogroups.[13–15] Alternatively, point mutations in the glycosyltransferase genes in the O-AGC or acquisition of alternative O-antigen modification genes, which are located outside of the O-AGC, have also been shown to result in structural alterations of O antigen and concomitant change in the serotype of the isolate.[16,17]

Genes or DNA sequences specific for each O serogroup can be used as targets for the identification of O serogroups via molecular approaches, such as PCR-based and hybridization-based methods. Such systems have already been developed by several researchers to target specific O-antigen types.[12,18–20] In particular, molecular assays targeting major O serogroups are routinely used in EHEC surveillance for clinical or food sample screening. Considering the range of diseases caused by *E. coli* strains belonging to many different serogroups, a more comprehensive and detailed O-AGC information for the complete set of *E. coli* O serogroups is of significant clinical importance for generating a rational molecular typing scheme. This molecular typing scheme, which could be performed *in silico* directly on sequence data, also offers a mechanism with which to link the ever-expanding genomic data to our extensive epidemiological and biological knowledge of this pathogen, based on O-antigen typing. Moreover, these data will also provide a much better understanding of the complex mechanisms by which a huge diversity in O serogroups have arisen. Here, we present a complete sequence set for the O-AGCs from all 184 *E. coli* O serogroups, which include recently added serogroups (O182–O187), providing a complete picture of the O-AGC diversity in *E. coli*.

2. Materials and methods

2.1. Bacterial strains, culture condition, and DNA preparation

Reference strains of all 184 recognized *E. coli* O serogroups were obtained from SSI (see Supplementary Table S1). Cells were grown to the stationary phase at 37°C in Luria–Bertani medium. Genomic DNA was purified using the Wizard Genomic DNA purification kit (Promega) according to the manufacturer's instructions.

2.2. O-AGC sequences and comparative analyses

One hundred and eight *E. coli* O-AGC sequences were determined by Sanger-based capillary sequencing and/or Illumina MiSeq sequencing from PCR products covering O-AGCs (Supplementary Table S1). The O-AGC regions of the reference strains were amplified by PCR using 10 ng of genomic DNA as template with the Tks Gflex DNA polymerase (Takara Bio Inc.) by 25 amplification cycles for 10 s at 98°C and for 16 m at 69°C, and with a combination of three forward primers (TATGCCAGCGGCACCAAACG, ATACCGGCGATGAAAGCC, and GCGGGTGGGATTAAGTCTCT) designed on the *hisFI* genes and two reverse primers (GTGATGCAGGAATCCTCTGT and CCACGCTAATTACGCCATCTT) designed on the *wcaM* genes, or strain-specific primers designed based on the draft genome sequences determined using the MiSeq system from reference strains. Identification and functional annotation of the CDSs were performed based on the results of homology searches against the public, non-redundant protein database using BLASTP. The sequences reported in this article have been deposited in the GenBank database (accession no. AB811596-AB811624, AB812020-AB812085, and AB972413-AB972425). The other 76 *E. coli* O-AGC sequences were obtained from public databases. For a list of accession numbers, see Supplementary Table S1.

2.3. Phylogenetic analysis

Multilocus sequence typing (MLST) was carried out according to the protocol described on the *E. coli* MLST website (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), and the phylogenetic relationships of reference strains were analysed based on the concatenated sequences (3,423 bp) of seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) used for MLST. Multiple alignments of DNA and amino acid sequences were constructed by using the CLUSTAL W program.[21] Phylogenetic trees were constructed by using the neighbour-joining algorithm using the MEGA4 softrware.[22]

3. Results

3.1. Genetic structures of the O-AGCs from all *E. coli* O serogroups

Of the 184 known O serogroups, 76 complete O-AGC sequences were obtained from public databases. The sequence of the other 108 O-AGC was determined in this study from *E. coli* O-serogroup reference strains (Supplementary Table S1). Our analysis of these sequences confirmed several previously observed characteristics of O-AGCs in *E. coli* (Supplementary Fig. S1). In brief, O-AGCs are located between the *wca* and *his* operons. This region contains three housekeeping genes: *galF* (encoding UTP-glucose-1-phosphate uridylyltransferase), *gnd* (6-phosphogluconate dehydrogenase), and *ugd* (UDP-glucose 6-dehydrogenase), and most genes for O-antigen biosynthesis in each cluster are directly flanked by *galF* and *gnd/ugd*, while *gne* (UDP-GalNAc-4-epimerase) and *wzz* (O-antigen chain

length determination protein) located immediately outside of the region between *galF* and *gnd/ugd* (see Supplementary Fig. S1). The exceptions for this are the O-AGCs for O serogroups O14 and O57, which contain no O-antigen genes at the typical locus. However, it is known that the *E. coli* O14 reference strain Su4411-41 shows an O rough phenotype and lacks the O-AGC.[23] For O57, a further analysis is also required to investigate the presence of O-antigen structure in the LPS of the reference strain. Our data revealed that the O-AGCs located between *galF* and *gnd* ranged in size from 4.5 kbp (O155, including four genes) to 19.5 kbp (O108, including 18 genes).

3.1.1. Nucleotide sugar biosynthesis genes

Genes required for the deoxythymidine diphosphate (dTDP)-sugar biosynthesis pathway (*rmlBDAC*) to synthesize dTDP-L-rhamnose (dTDP-L-Rha), the precursor of L-Rha, were widely distributed in the O-AGCs (conserved in 56 O-serogroup O-AGCs; see Supplementary Fig. S1). The *vioAB* operon, for the biosynthesis of dTDP-*N*-acetylviosamine (dTDP-VioNAc), the precursor of VioNAc, was present in three O-serogroup O-AGCs; the *fnlABC* operon for the synthesis of uridine diphosphate (UDP)-*N*-acetyl-L-fucosamine (UDP-L-FucNAc), the precursor of L-FucNAc, was in 11 O-serogroup O-AGCs; the *fnlA-qnlBC* genes for the synthesis of UDP-*N*-acetyl-L-quinovosamine (UDP-L-QuiNAc), the precursor of L-QuiNAc, were in four O-serogroup O-AGCs; the *mnaDBCA* genes for synthesis of cytidine monophosphate (CMP)-*N*-acetylneuraminate (CMP–NeuNAc), the precursor of *N*-acetylneuraminic acid (Neu5Ac or sialic acid), were found in six O-serogroup O-AGCs (Supplementary Fig. S1). In addition, a gene set comprising seven genes putatively involved in the synthesis of di-*N*-acetyl-8-epilegionaminic acid (8eLeg5Ac7Ac) were found in three O-serogroup O-AGCs. For at least 49 O serogroups, gene sets for nucleotide sugar biosynthesis were not found in their O-AGCs (Supplementary Fig. S1), suggesting that, in these serogroups, nucleotide sugars required for O-antigen biosynthesis were synthesized by pathways encoded by the genes located outside of the O-AGCs.

3.1.2. Glycosyltransferase

Each O-AGC contained two to six genes encoding putative glycosyltransferases for synthesizing O-antigen subunits and a total of 611 glycosyltransferase genes identified in all O-AGCs. Pfam analysis revealed that at least 25 types of glycosyltransferase-related domains were found in the 611 glycosyltransferase genes (Supplementary Table S2). 'Glycosyl transferases group 1' (PF00534) and 'Glycosyl transferase family 2' (PF00535) were the most widely distributed domains, which were found in 216 and 253 genes, respectively. Except for the five genes belonging to 'Glycosyltransferase family 52' (PF07922), which were found in five of the six *mnaDBCA*-containing O-AGCs (O24, O56, O104, O131, and O171), there were no relationships between the type of glycosyltransferase-related domain and the gene set for sugar synthesis in each O-AGC.

3.1.3. O-antigen subunit translocation and chain synthesis

All O-AGCs carried either *wzx/wzy* or *wzm/wzt* gene pairs. Of the 182 O-AGCs (the above-mentioned O14 and O57 were excluded from the 184 clusters analysed in this study), 171 carried the *wzx/wzy* genes, and the other 11 carried the *wzm/wzt* genes (Supplementary Fig. S1 and Table S1). Detailed sequence comparisons of the *wzx/wzy* and *wzm/wzt* genes are described below.

3.2. Grouping the O-AGCs by sequence

On the basis of sequences and genetic structures of the entire O-AGC regions, in addition to 145 unique O-AGCs from different *E. coli* O serogroups, the O-AGCs from 37 O serogroups could be placed into 16 groups (named Gp1–Gp16) with the members of each group having identical or very similar O-AGC genes (mostly sharing ≥95% DNA sequence identity) (Fig. 1). This included nine groups with members of different serogroups but which carried identical O-AGC gene sets (Gp1–Gp9) and one group, Gp10, where two strains (O13 and O129) of the three-member group carried an identical O-AGC gene set (sharing 98.3–99.9% DNA sequence identity) (Fig. 1). The reason(s) why they belong to different O serogroups even though they have identical O-AGCs are discussed in the Discussion section. Indels or exchange of one or more genes was also shown to explain the differences between O135 and other members of Gp10 and members Gp11–Gp16, which otherwise carried highly conserved orthologous genes (summarized in Fig. 1). Simple insertions of insertion sequence (IS) elements containing one or two transposase genes were found in three groups without any gene disruption: an IS*629* insertion in O18ab of Gp12, IS*Ec11* in O164 of Gp13, and IS*1* in O62 of Gp14. IS element-associated replacement of the right-end portion of the O-AGC had occurred in three groups, Gp14, Gp15, and Gp16, resulting in the replacement (or deletion) of glycosyltransferase gene(s). Exchange of the *wzx* gene had also occurred in Gp16. These data suggest that IS elements are important drivers for generating O-antigen biosynthesis gene replacement and therefore diversity.

3.3. Diversity and specificity of the *wzx/wzy* or *wzm/wzt* genes among the *E. coli* O-AGCs

As previously proposed,[12] most *wzx/wzy* or *wzm/wzt* orthologues showed high levels of sequence diversity and their sequences were unique to each O-AGC or O-AGC group described above (Fig. 2 and Supplementary Fig. S2). DNA sequence identities of the closest pairs were <70%, except for the O96/O170 pair, the *wzx* genes of which showed 86% DNA sequence identity. Within the 16 O-AGC groups, the orthologous *wzx/wzy* or *wzm/wzt* genes also showed high sequence conservation (≥95% DNA sequence identity, but mostly ≥97% identity), except for Gp16 that shared only the *wzy* gene (Fig. 2).

3.4. Phylogenetic relationships of *E. coli* O-serogroup reference strains

Based on the concatenated nucleotide sequences of seven housekeeping genes used for MLST, we determined the evolutionary relationships of all *E. coli* O-serogroup reference strains (Fig. 3). This analysis revealed that the members of five groups sharing the common O-AGCs (Gp8, Gp10, Gp11, Gp14, and Gp15) and two members (O17 and O77) of Gp9 were found in closely related lineages. However, the members of other groups (and three members of Gp9) were found in distinct evolutionary lineages. For example, O20 and O137, both carrying the Gp1 O-AGC, were found in two distinct lineages, each belonging to phylogroups A and E/D, respectively, and five serogroups (O17/O77, O44, O73, and O106) belonging to Gp9 were found in multiple lineages (A, E/D, and B1).

The systematic phylogenetic analysis of all *E. coli* O-serogroup reference strains further revealed that one-quarter of the reference strains (46/184) belonged to a single clonal group (≥99.9% sequence identity), which was represented by sequence type (ST) 10 and its very close relatives in phylogroup A (Fig. 3 and Supplementary Fig. S3). Additionally, three clonal groups containing five or more reference

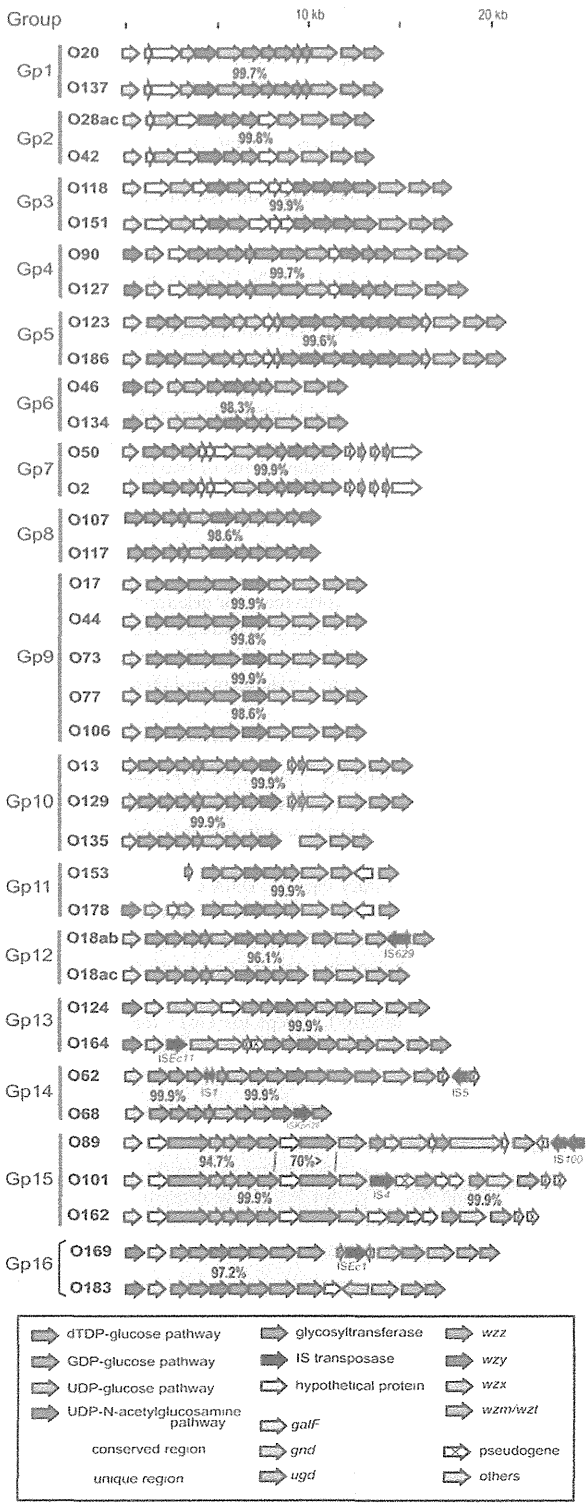

Figure 1. Sixteen *Escherichia coli* O-AGC groups identified in this study. Group members have different O serogroups in each group, but these share nearly identical or highly similar genetic organizations. Group names (Gp) are indicated at the left side. DNA sequence identities (%) between group members are indicated in each group.

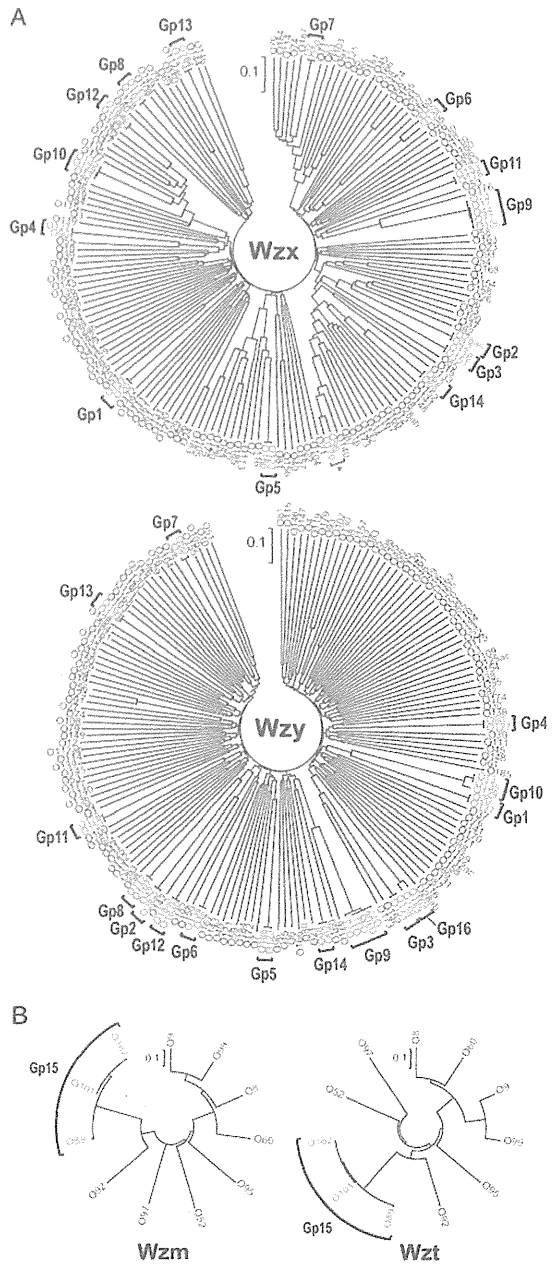

Figure 2. Phylogenetic analysis of homologues of (A) Wzx and Wzy and (B) Wzm and Wzt from *Escherichia coli* O-serogroup reference strains based on the amino acid sequences. The group names are indicated outside of trees. The pair or groups of homologues with high DNA sequence identity (≥95%, mostly ≥97%) are indicated in red. The Wzx homologues of O96 and O170, which are indicated in blue and by an asterisk, showed 86% DNA sequence identity, but in all other proteins showed low-sequence homologies to each other (<70% identity). Note that while the DNA sequence identity between the *wzx*_O46 and *wzx*_O134 in Gp6 is 99.7%, the *wzx*_O46 has a 2-bp deletion at the 3′-region, causing a frame shift.

strains were also identified in phylogroups A (ST34 and ST57) and B1 (ST300) (Fig. 3). The phylogenetic analysis also showed that the types of sugar synthesis gene sets and processing gene sets (*wzx/wzy* and *wzm/wzt*) were not limited to a specific lineage (Fig. 3).

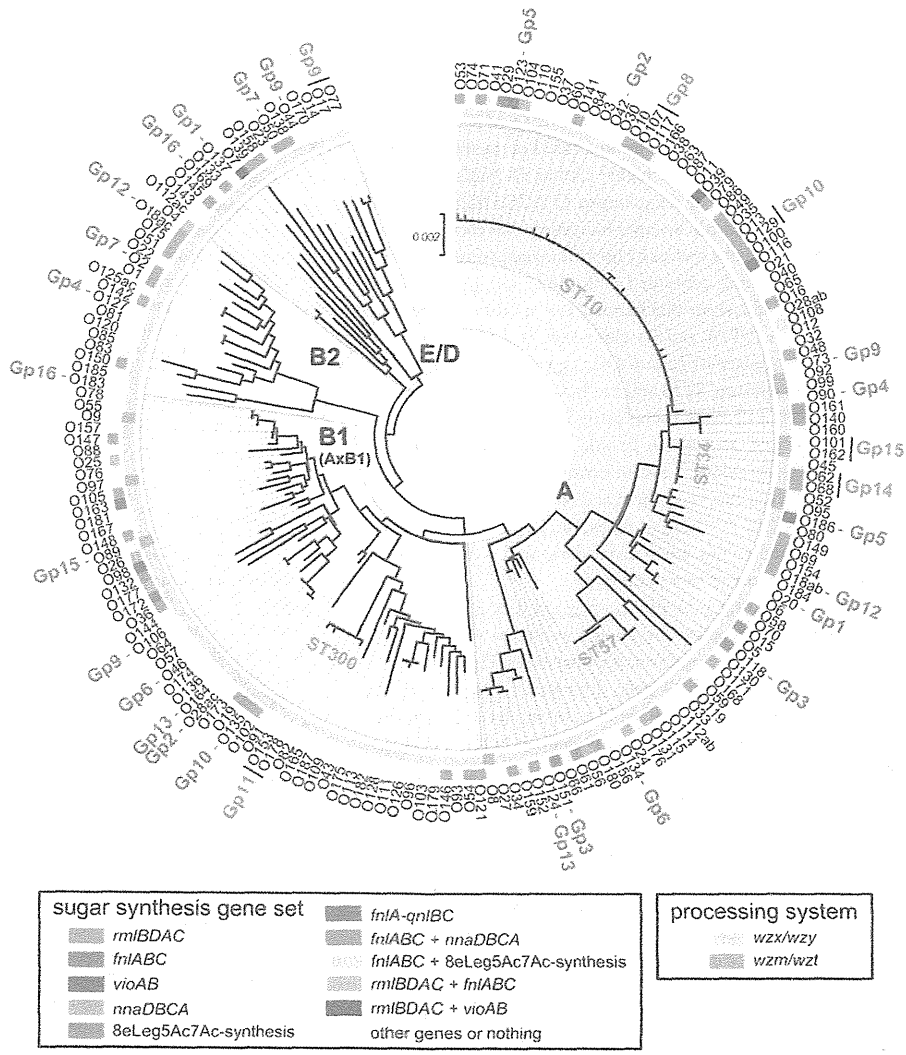

Figure 3. Correlation between the *Escherichia coli* evolutionary lineages and the distribution of O-AGCs. The phylogenetic tree was constructed based on the concatenated sequences of seven housekeeping genes from all 184 *E. coli* O-serogroup reference strains. The group names of O-AGCs (Gp1–Gp16) are indicated in the outermost region. Members in groups indicated in green were found to belong to the same or very closely related lineage, whereas members of the groups indicated in blue were found in distinct lineages. The outer circle next to the O serogroup names indicates the distribution of sugar synthesis gene sets identified in each O-AGC. The inner circle indicates the type of O-antigen processing system (*wzx/wzy* or *wzm/wzt*). Phylogenetic groups (A, B1, B2, D, and E) were determined by comparing the sequences of the strains tested with the known sequences from the ECOR collection (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli).

3.5. Relationships of the *E. coli* and *Shigella* O-AGCs

Shigella and *E. coli* belong to the same species complex[24] and many *Shigella* O antigens are known to be serologically and genetically identical or very similar to some *E. coli* O antigens, as summarized by Liu *et al.*[25] In addition to the 21 previously shown relationships, we found two additional O-AGC groups shared by *E. coli* and *Shigella*; O38 and *Shigella dysenteria* type 8 (SD8), and O169/O183 and *Shigella boydii* type 6/10 (SB6/SB10) (Supplementary Fig. S4). The O183-AGC was highly similar to the *S. boydii* types 10 cluster (sharing 98.2% DNA sequence identity). In our previous study,[26] we provisionally named a novel O serogroup for a group of Shiga toxin-producing *E. coli* strains as OSB10, which cross-reacted with *S. boydii* type 10. Sequence comparisons in this study revealed that OSB10 is not only serologically but also genetically identical to the new serogroup O183 of Gp16.

4. Discussion

Much of what we know about *E. coli* is defined at some level by O serogroups. To link genomic information to the wealth of data held in public databases, in our collective knowledge, outbreak, and disease reports and elsewhere, we endeavoured to determine whether molecular O-serogroup identification, targeting O-serogroup-specific genes (or unique sequences), was a valuable method to capture this information and maintain this important link. Not only do we show evidence supporting the effectiveness of molecular O-typing, but also we open

up the possibility of generating a molecular O-typing scheme and relate O serogroups to the underlying phylogeny of this bacterium.

By determining and comparing the sequences of O-AGCs from all known *E. coli* O serogroups, we newly defined the sequence and gene content of 145 unique O-AGCs and showed that O-AGCs from 37 O serogroups could be placed into 16 groups based on members in each group sharing nearly identical or highly similar O-AGCs. It is clear from these data that many of the grouped O-AGCs (Gp1-16) were found in distinct phylogenetic lineages indicating that these O-AGCs have been spread across this species by horizontal gene transfer. Moreover, several lineages that contained multiple O serogroups, ST10, ST34, ST57, and ST300, show that frequent exchange occurs between and within lineages. ST10 and its close relatives are particularly interesting as one-quarter of *E. coli* O-serogroup reference strains fell within this clonal group. ST10 and its clonal complex are clinically very important being recently found to include ESBL-producing *E. coli* from human and animals in Spain,[25] Italy and Denmark,[26] China,[27] and the Netherlands,[28] and in various intra-intestinal pathotypes of *E. coli*, such as enteroaggregative *E. coli*,[27,28] enterotoxigenic *E. coli*,[29,30] and EHEC.[31,32] In most cases, the O serogroups of these ST10 or ST10-related strains are unusual compared with the typical O serogroups that represent that pathotype.

Acquisition of O-antigen modification genes located on the genomes of serotype-converting bacteriophages or plasmids is also an important strategy for diversifying O-antigen structures. This mechanism has been well investigated in *Shigella flexneri*.[33,34] In *E. coli*, the O-serogroup conversion by a prophage-like element has been reported for O17 and O44,[17] which belong to Gp9 defined in this study. Another possible mechanism to generate the variation of O antigens is the mutations in the genes of the O-AGC as observed for O107 and O117,[16] which belong to Gp8. In this case, point mutations in a glycosyltransferase gene are responsible for the alteration of O-antigen structure (and thus that of O serogroup).[16] Five O-AGC groups including Gp2, Gp5, Gp7, Gp12, and Gp13 also contained differences in the amino acid sequence of their glycosyltransferases. O serogroup differences in these groups may be generated by the point mutations in glycosyltransferase genes. On the other hand, all glycosyltransferase genes in Gp1, Gp3, Gp4, Gp6, and Gp11; four strains from Gp9 (O17, O44, O73, and O77) and two from Gp10 (O13 and O129) showed 100% amino acid sequence identity. These results suggest that the serological differences between the members of these seven groups have been generated by acquisition of modification genes outside of the O-AGC as shown for O17 and O44 of Gp9.[17]

We believe that the remarkable sequence diversity observed in the *wzx/wzy* and *wzm/wzt* O-AGC genes of all known *E. coli* O serogroups appears to be sufficiently discriminative from one another to make identification of each of the known O serogroups possible. Therefore, our sequence data will serve as a valuable resource for the development of rationally designed molecular methods for O-typing as well as for detecting novel O serogroups.

In conclusion, our study provides a complete sequence set of O-AGCs of all known *E. coli* O serogroups and thus offers a full view on the genetic diversity of O-AGCs of this bacterium. In addition, the results presented suggest that horizontal gene transfer has been involved in the O serogroup diversification in *E. coli* more frequently and in a more biased or lineage-dependent fashion than previously thought.

Acknowledgements

We thank A. Akiyoshi, Y. Kato, and A. Yoshida for technical assistance.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Health Labor Sciences Research Grants from the Ministry of Health, Labor, and Welfare, Japan to A.I. (H25-Syokuhin-Wakate-018) and M.O. (H24-Shinkou-Ippan-012); Adaptable and Seamless Technology Transfer Program through Target-driven R&D (AS242Z00217P) from Japan Science and Technology Agency to A.I.; and a Scientific Research Grant on Priority Areas from the University of Miyazaki and the Program to Disseminate Tenure Tracking System from the Japanese Ministry of Education, Culture, Sports, Science, and Technology to A.I. (http://www.miyazaki-u.ac.jp/ir/english/index.html). This work was also supported by Wellcome Trust grant (098051). Funding to pay the Open Access publication charges for this article was provided by the University of Miyazaki, Japan.

References

1. Bazaka, K., Crawford, R.J., Nazarenko, E.L. and Ivanova, E.P. 2011, Bacterial extracellular polysaccharides, *Adv. Exp. Med. Biol.*, 715, 213–26.
2. Liu, B., Knirel, Y.A., Feng, L., et al. 2013, Structural diversity in *Salmonella* O antigens and its genetic basis, *FEMS Microbiol. Rev.*, 38, 56–89.
3. Stenutz, R., Weintraub, A. and Widmalm, G. 2006, The structures of *Escherichia coli* O-polysaccharide antigens, *FEMS Microbiol. Rev.*, 30, 382–403.
4. Lam, J.S., Taylor, V.L., Islam, S.T., Hao, Y. and Kocincova, D. 2011, Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide, *Front Microbiol.*, 2, 118.
5. Penner, J.L. and Aspinall, G.O. 1997, Diversity of lipopolysaccharide structures in *Campylobacter jejuni*, *J. Infect. Dis.*, 176 (Suppl. 2), S135–138.
6. Armstrong, G.L., Hollingsworth, J. and Morris, J.G. Jr. 1996, Emerging foodborne pathogens: *Escherichia coli* O157:H7 as a model of entry of a new pathogen into the food supply of the developed world, *Epidemiol Rev.*, 18, 29–51.
7. Tarr, P.I., Gordon, C.A. and Chandler, W.L. 2005, Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome, *Lancet*, 365, 1073–86.
8. Johnson, K.E., Thorpe, C.M. and Sears, C.L. 2006, The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*, *Clin. Infect. Dis.*, 43, 1587–95.
9. Buchholz, U., Bernard, H., Werber, D., et al. 2011, German outbreak of *Escherichia coli* O104:H4 associated with sprouts, *N. Engl. J. Med.*, 365, 1763–70.
10. Peirano, G. and Pitout, J.D. 2010, Molecular epidemiology of *Escherichia coli* producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4, *Int. J. Antimicrob. Agents*, 35, 316–21.
11. Samuel, G. and Reeves, P. 2003, Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly, *Carbohydr. Res.*, 338, 2503–19.
12. DebRoy, C., Roberts, E. and Fratamico, P.M. 2011, Detection of O antigens in *Escherichia coli*, *Anim. Health Res. Rev.*, 12, 169–85.
13. Leopold, S.R., Magrini, V., Holt, N.J., et al. 2009, A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis, *Proc. Natl Acad. Sci. USA*, 106, 8713–8.
14. Iguchi, A., Shirai, H., Seto, K., et al. 2011, Wide distribution of O157-antigen biosynthesis gene clusters in *Escherichia coli*, *PLoS ONE*, 6, e23250.
15. Iguchi, A., Iyoda, S. and Ohnishi, M. 2012, Molecular characterization reveals three distinct clonal groups among clinical Shiga toxin-producing *Escherichia coli* strains of serogroup O103, *J. Clin. Microbiol.*, 50, 2894–900.
16. Wang, Q., Perepelov, A.V., Wen, L., et al. 2012, Identification of the two glycosyltransferase genes responsible for the difference between *Escherichia coli* O107 and O117 O-antigens, *Glycobiology*, 22, 281–7.

17. Wang, W., Perepelov, A.V., Feng, L., et al. 2007, A group of *Escherichia coli* and *Salmonella* enterica O antigens sharing a common backbone structure, *Microbiology*, 153, 2159–67.
18. Lacher, D.W., Gangiredla, J., Jackson, S.A., Elkins, C.A. and Feng, P.C. 2014, Novel microarray design for molecular serotyping of Shiga toxin-producing *Escherichia coli* isolated from fresh produce, *Appl. Environ. Microbiol.*, 80, 4677–82.
19. Tzschoppe, M., Martin, A. and Beutin, L. 2012, A rapid procedure for the detection and isolation of enterohaemorrhagic *Escherichia coli* (EHEC) serogroup O26, O103, O111, O118, O121, O145 and O157 strains and the aggregative EHEC O104:H4 strain from ready-to-eat vegetables, *Int. J. Food Microbiol.*, 152, 19–30.
20. Wang, Q., Ruan, X., Wei, D., et al. 2010, Development of a serogroup-specific multiplex PCR assay to detect a set of *Escherichia coli* serogroups based on the identification of their O-antigen gene clusters, *Mol. Cell Probes*, 24, 286–90.
21. Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22, 4673–80.
22. Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, *Mol. Biol. Evol.*, 24, 1596–9.
23. Jensen, S.O. and Reeves, P.R. 2004, Deletion of the *Escherichia coli* O14:K7 O antigen gene cluster, *Can. J. Microbiol.*, 50, 299–302.
24. Pupo, G.M., Lan, R. and Reeves, P.R. 2000, Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics, *Proc. Natl Acad. Sci. USA*, 97, 10567–72.
25. Liu, B., Knirel, Y.A., Feng, L., et al. 2008, Structure and genetics of *Shigella* O antigens, *FEMS Microbiol. Rev.*, 32, 627–53.
26. Iguchi, A., Iyoda, S., Seto, K. and Ohnishi, M. 2011, Emergence of a novel Shiga toxin-producing *Escherichia coli* O serogroup cross-reacting with *Shigella boydii* type 10, *J. Clin. Microbiol.*, 49, 3678–80.
27. Olesen, B., Scheutz, F., Andersen, R.L., et al. 2012, Enteroaggregative *Escherichia coli* O78:H10, the cause of an outbreak of urinary tract infection, *J. Clin. Microbiol.*, 50, 3703–11.
28. Okeke, I.N., Wallace-Gadsden, F., Simons, H.R., et al. 2010, Multi-locus sequence typing of enteroaggregative *Escherichia coli* isolates from Nigerian children uncovers multiple lineages, *PLoS ONE*, 5, e14093.
29. Turner, S.M., Chaudhuri, R.R., Jiang, Z.D., et al. 2006, Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages, *J. Clin. Microbiol.*, 44, 4528–36.
30. Nada, R.A., Shaheen, H.I., Khalil, S.B., et al. 2011, Discovery and phylogenetic analysis of novel members of class b enterotoxigenic *Escherichia coli* adhesive fimbriae, *J. Clin. Microbiol.*, 49, 1403–10.
31. Monaghan, A.M., Byrne, B., McDowell, D., Carroll, A.M., McNamara, E. B. and Bolton, D.J. 2012, Characterization of farm, food, and clinical Shiga toxin-producing *Escherichia coli* (STEC) O113, *Foodborne Pathog. Dis.*, 9, 1088–96.
32. Hauser, E., Mellmann, A., Semmler, T., et al. 2013, Phylogenetic and molecular analysis of food-borne shiga toxin-producing *Escherichia coli*, *Appl. Environ. Microbiol.*, 79, 2731–40.
33. Allison, G.E. and Verma, N.K. 2000, Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*, *Trends Microbiol.*, 8, 17–23.
34. Sun, Q., Knirel, Y.A., Lan, R., et al. 2012, A novel plasmid-encoded serotype conversion mechanism through addition of phosphoethanolamine to the O-antigen of *Shigella flexneri*, *PLoS ONE*, 7, e46095.

Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution

Astrid von Mentzer[1,2], Thomas R Connor[2,3], Lothar H Wieler[4], Torsten Semmler[4], Atsushi Iguchi[5], Nicholas R Thomson[2], David A Rasko[6], Enrique Joffre[1], Jukka Corander[7], Derek Pickard[2], Gudrun Wiklund[1], Ann-Mari Svennerholm[1], Åsa Sjöling[1,8] & Gordon Dougan[2]

Enterotoxigenic *Escherichia coli* (ETEC), a major cause of infectious diarrhea, produce heat-stable and/or heat-labile enterotoxins and at least 25 different colonization factors that target the intestinal mucosa. The genes encoding the enterotoxins and most of the colonization factors are located on plasmids found across diverse *E. coli* serogroups. Whole-genome sequencing of a representative collection of ETEC isolated between 1980 and 2011 identified globally distributed lineages characterized by distinct colonization factor and enterotoxin profiles. Contrary to current notions, these relatively recently emerged lineages might harbor chromosome and plasmid combinations that optimize fitness and transmissibility. These data have implications for understanding, tracking and possibly preventing ETEC disease.

ETEC cause approximately 400 million diarrheal cases and almost 400,000 deaths per year in children aged less than 5 years in low- and middle-income countries and are also a common cause of travelers' diarrhea[1]. ETEC are defined by their ability to produce a heat-labile toxin (LT) and/or a heat-stable toxin (ST; including two subtypes, STh and STp)[2,3]. At least 25 antigenically distinct colonization factors have been described in human ETEC. Colonization factors are fimbrial or afimbrial surface structures with the potential to mediate adherence to the human intestinal mucosa[2]. The most prevalent colonization factors are CFA/I and coli surface antigens 1–6 (CS1–CS6), although in certain geographical regions CS7, CS14 and CS17 are also common[2]. Individual ETEC isolates typically carry and/or coexpress one, two or three colonization factors and/or toxin types, with combinations such as CS1 + CS3 with LT + STh, CS2 + CS3 with LT + STh, CS5 + CS6 with LT + STh, CS6 with STp, CFA/I with STh and CS7 with LT repeatedly isolated globally[2–5]. However, 20–50% of all clinical ETEC isolates, in particular, those expressing LT only, do not express any of the identified colonization factors, suggesting that additional colonization factors might exist[4,5].

In addition to the large number of identified colonization factors and colonization factor–toxin combinations, ETEC can express a wide variety of O antigens (over 100 different O antigens have been associated with clinical ETEC isolates)[2,3,6]. Together with the large number of colonization factors, this wide range of O antigens indicates that there is a substantial level of genetic diversity within this pathovar[6]. Limited sequence-based studies of ETEC phylogeny have indicated that the acquisition of colonization factor and toxin genes by non-pathogenic, commensal strains might be sufficient to cause clinical ETEC disease[7,8]. Previous studies, exploring a potential association between chromosomal backgrounds and virulence factors in ETEC, have not shown any consistent evidence of phylogenetic clustering of isolates, although a potential association between virulence profiles and genetic backgrounds was suggested in some studies[8–14]. In addition, it has been concluded that the acquisition of virulence-related genes has occurred multiple times, consistent with the key virulence genes being encoded on mobile plasmids[8,10]. These observations have led to the hypothesis that ETEC are simply any *E. coli* lineage that can acquire, express and retain plasmids harboring colonization factors and/or toxins.

In this study, we have used next-generation sequencing to develop a more complete understanding of ETEC phylogeny and evolution. To this end, we have examined a global collection of ETEC isolated from 20 different countries in Asia, Africa, and North, Central and South America between 1980 and 2011 by whole-genome sequencing, serotyping and phylogenetic analyses. To cover the breadth of ETEC diversity, we selected ETEC isolates on the basis of their colonization factor and toxin profiles, including isolates lacking known colonization factors, as well as isolates from individuals from different age groups in endemic countries and travelers. Our analyses show that ETEC are widely distributed across the *E. coli* species. There is also a clear signature of several globally distributed ETEC lineages that show consistent long-term association with a specific O antigen and virulence gene repertoire.

[1]Department of Microbiology and Immunology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [2]Wellcome Trust Sanger Institute, Hinxton, UK. [3]Organisms and Environment Division, Cardiff University School of Biosciences, Cardiff University, Cardiff, UK. [4]Centre of Infection Medicine, Institute of Microbiology and Epizootics, Freie Universität Berlin, Berlin, Germany. [5]Interdisciplinary Research Organization, University of Miyazaki, Miyazaki, Japan. [6]Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA. [7]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [8]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to A.v.M. (astrid.von.mentzer@gu.se), Å.S. (asa.sjoling@ki.se), A.-M.S. (ann-mari.svennerholm@microbio.gu.se) or G.D. (gd1@sanger.ac.uk).

Received 12 February; accepted 17 October; published online 10 November 2014; doi:10.1038/ng.3145

RESULTS

ETEC are distributed throughout the *E. coli* species

Considering the diversity of *E. coli* and to provide an accurate genome-wide phylogeny, we identified 1,429 genes representing the 'maximum common genome' (MCG) from the genomes of 47 *E. coli* isolates representing the known species diversity (Online Methods), with these genes also found in all ETEC strains sequenced in this study.

The ETEC phylogeny was constructed from the MCG alignments of 362 selected ETEC isolates with representative virulence profiles isolated from indigenous populations and travelers between 1980 and 2011 from different countries in Asia, Africa, and North, Central and South America. Also included were 21 available reference genomes covering commensal *E. coli*, enteropathogenic *E. coli* (EPEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC), enterohemorrhagic *E. coli* (EHEC), uropathogenic *E. coli* (UPEC), *Shigella* and previously published ETEC genomes (**Fig. 1** and **Supplementary Tables 1** and **2**). An alignment was generated covering the 1,429 genes for the 383 genomes, which was then used as a basis for the detection of recombination sites, SNP calling and phylogenetic tree construction. In total, we identified 128,214 variable sites (excluding the reference genomes) across the MCG showing the diversity captured within this pathovar. This analysis demonstrated that ETEC are clearly distributed throughout the species phylogeny, consistent with previous studies[8] (**Fig. 1**). Indeed, ETEC isolates were assigned to most recognized phylogroups of *E. coli*[15], with the majority falling within the A and B1 groups (**Fig. 1**).

Identification of several major ETEC lineages

The MCG-based phylogenetic analysis together with Bayesian analysis of the population structure (BAPS)[16] was employed to define lineages across ETEC. The BAPS analysis defined several robust ETEC lineages (L1–L21) among the 362 ETEC isolates analyzed (**Fig. 1**). The majority of the lineages encompassed isolates obtained from various countries in Asia, the Americas and Africa collected over three decades.

Lineages share specific virulence factors and are globally spread

The colonization factor and toxin profiles of all 362 ETEC isolates sequenced were mapped onto the MCG tree, and the presence of colonization factors and toxins was reconfirmed using comparative genomics approaches (Online Methods). Known colonization factors were identified in 228 ETEC isolates, whereas 134 isolates lacking an identifiable colonization factor (by phenotypic and genomic

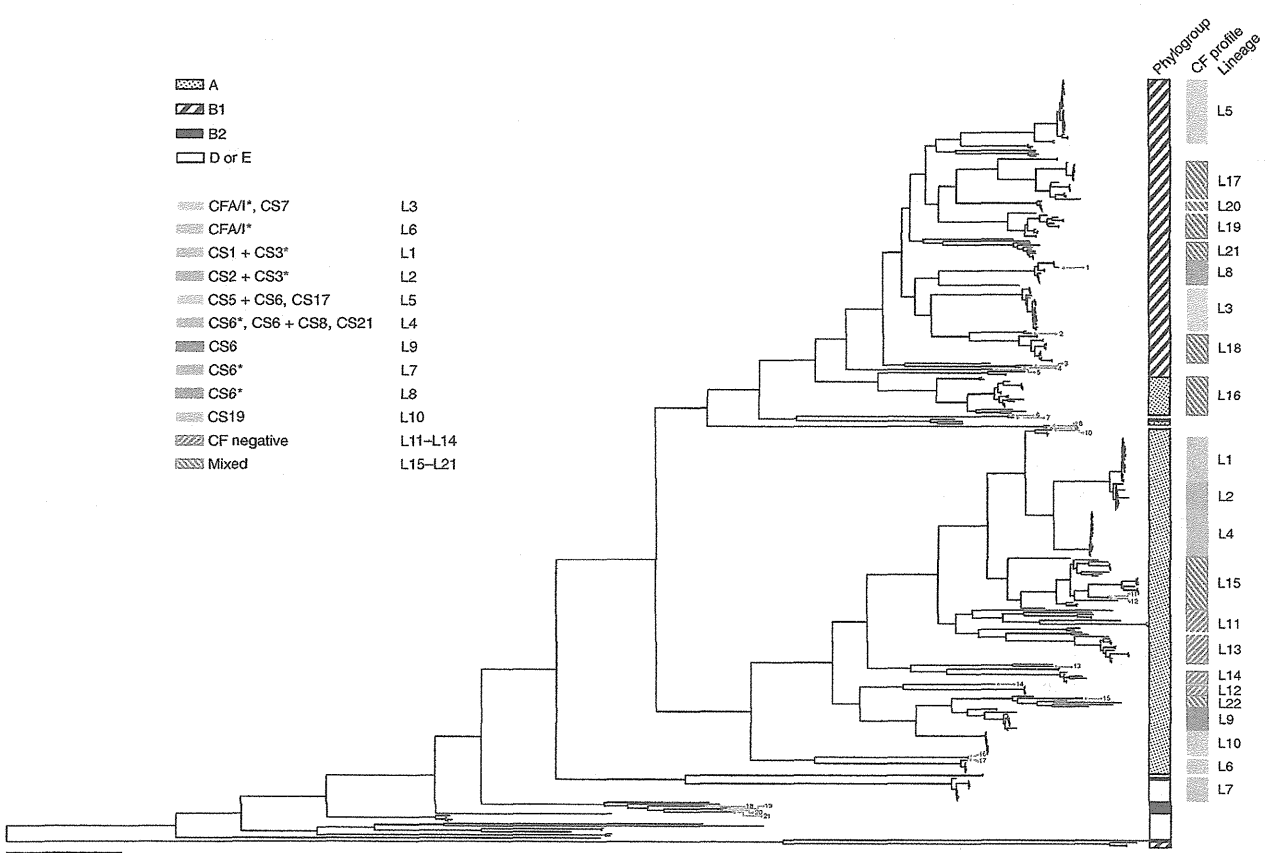

Figure 1 Population structure of ETEC isolates. Midpoint-rooted maximum-likelihood phylogenetic tree based on SNP differences across the MCG, excluding probable recombination events. The phylogenetic groups of isolates are shown in black and white to the right of the tree. Distinct lineages identified with BAPS across the data set are indicated to the right of the color-coded colonization factor profile, which match the lineages identified. An asterisk indicates isolates with or without CS21. References are indicated with red dots and arrows: 1, ETEC B7A; 2, ETEC 24377A; 3, EPEC B171; 4, EPEC E22; 5, EPEC E110019; 6, *S. sonnei* 53G; 7, *S. sonnei* Ss046; 8, *Shigella flexneri* 8401; 9, *S. flexneri* M90T; 10, ETEC H10407; 11, *E. coli* K-12 MG1655; 12, *E. coli* K12 W3110; 13, EAEC 101-1; 14, EIEC 53638; 15, *E. coli* HS; 16, EHEC EDL933; 17, EHEC Sakai; 18, UPEC CFT073; 19, UPEC 536; 20, UPEC F11; 21, UPEC UTI89. CF, colonization factor. Scale bar, 0.041 substitutions per variable site.

Table 1 Characteristics of ETEC isolates in lineages L1–L14

Lineage	Number of isolates[a]	Variable sites[b]	MLST[c]	O antigen	Colonization factor	Enterotoxin
L1	23	542	ST2353, ST4 ($n = 4$)	O6	CS1 + CS3[e]	LT + STh
L2	14	427	ST4	O6	CS2 + CS3[e]	LT + STh
L3	22	880	ST173	O78, O114, O126, O128	CFA/I[e], CS7	LT + STh, STh
L4	23	340	ST1312	O25	CS6[e], CS6 + CS8, CS21	LT, STh
L5	30	517	ST443	O115, O157	CS5 + CS6, CS17	LT + STh, LT, STh
L6	7	172	ST2332	ON3[d]	CFA/I[e]	STh
L7	12	483	ST182	O169	CS6	STp
L8	11	522	ST94	O148	CS6	STh, STp
L9	12	2,758	ST398	O27	CS6	STp
L10	12	57	ST2368	O114	CS19	LT + STp
L11	13	13,059	$n = 8$[f]	$n = 10$[f]	–	LT + STp, LT, ST
L12	5	80	ST731	O15	–	LT
L13	14	2,408	ST10, ST100, ST165, ST750, ST3860	ON49[d], O112ab, O160, O170, O179	–	LT + STp, LT
L14	6	502	ST10, ST1684, ST2705	ON5[d]	–	ST

[a]Number of isolates in L1–L5 used for Bayesian phylogenetic analysis with BEAST and in lineages L6–L14. [b]Determined on the basis of sequences with recombination sites removed. [c]MLST was determined by extracting the sequences for the *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA* and *recA* genes from the whole-genome data. [d]N, novel O antigen. [e]With or without CS21 (Longus). [f]Details on MLST and O antigens can be found in **Supplementary Table 2**.

analyses) were defined as 'colonization factor–negative' isolates. We also determined O antigen genotypes as predicted from sequence (**Fig. 1**, **Table 1** and Online Methods).

Interestingly, the isolates found in the major lineages L1–L5 also expressed the most prevalent virulence profiles described in the literature[5]. These lineages were selected for further analyses, as they comprised a large number of isolates appropriate for the subsequent studies. Lineages L1–L5 all showed a clear clustering of isolates with a specific virulence profile, i.e., a combined colonization factor and toxin profile. This clustering was also evident in several additional lineages (L6–L10) with distinct colonization factor and toxin profiles. We identified 38 colonization factor–negative isolates in 4 lineages (L11–L14). These observations indicate that ETEC is far more than a plasmid with an *E. coli* attached. The data also demonstrate that some ETEC isolates fall into distinct globally and temporally distributed lineages with specific virulence profiles (**Fig. 2**, **Table 1**, **Supplementary Figs. 1–4** and **Supplementary Table 2**). However, seven lineages (L15–L21) comprised ETEC isolates with a mix of colonization factor and toxin profiles, suggesting that in these lineages gene exchange might be common. All of the distinct ETEC lineages were represented by isolates taken from adults and children in endemic areas as well as from travelers with diarrhea (**Supplementary Table 2**).

The L1 lineage comprised ETEC with the colonization factor profile CS1 + CS3 (± CS21) (**Fig. 2** and **Table 1**). The closely related L2 lineage encompassed isolates expressing CS2 + CS3 (± CS21) and shared the O6 antigen with L1. Notably, isolates positive for CS1 + CS3 (± CS21) and CS2 + CS3 (± CS21) were not found in any other lineage (**Fig. 2** and **Table 1**). CFA/I-positive isolates were identified in two individual lineages, L3 and L6. Additional lineages, comprising isolates that shared O antigen and colonization factor and toxin profiles, were identified (**Table 1**). ETEC isolates expressing CS6 were found in four lineages (L4 and L7–L9). Ten CS19-positive isolates clustered together (L10), although these isolates were probably from a single geographically restricted outbreak.

There was also an interesting association of variation in O antigen genotype and colonization factor and toxin profile with phylogenetic lineage. On the basis of O antigen genotype, lineage L3 could be divided into four subclades: isolates encoding CFA/I with O78, O126 or O128, and five CS7 + O114 isolates. However, these isolates were related both in terms of colonization factor profile and phylogenetic origin. In contrast, isolates within the L4 lineage expressed CS6 alone or together with CS8 as well as CS21, but they all belonged to serogroup O25 (**Fig. 2** and **Table 1**). The largest ETEC lineage identified in this study was L5, which could be divided into two subclades on the basis of O antigen type. One subclade harbored O115-positive isolates that expressed either CS5 + CS6 or the distantly related CS17, whereas the second subclade harbored isolates positive for O167 and CS5 + CS6.

Colonization factor–negative isolates allocate across the ETEC tree

Four lineages (L11–L14) harboring mainly colonization factor–negative isolates were identified. In comparison to other lineages (L1–L10), which showed a substantial association between their O antigen and colonization factor and toxin profiles, the predominantly colonization factor–negative lineages showed a higher level of diversity. However, there was still structure in their phylogeny, suggesting that so-called colonization factor–negative isolates might share properties, including, for example, unidentified new colonization factors (**Fig. 1**, **Table 1**, **Supplementary Fig. 3** and **Supplementary Table 2**). Additional colonization factor–negative isolates were found across the tree, clustering together with isolates harboring less prevalent colonization factors such as CS12, CS14 and CS17 and, in some cases, CS6. The isolates in lineages L15–L21 represented 28% of our ETEC collection (**Fig. 1**, **Supplementary Fig. 3** and **Supplementary Table 2**). Hence, there are lineages with varied virulence profiles that have added to the confusion about this pathovar, but our data identify a number of persistent plasmid-chromosomal background combinations in ETEC, even in isolates without known colonization factors.

Toxin allele profiles are associated with chromosomal background

To further investigate a potential relationship between the chromosomal background, colonization factors and enterotoxins, we extracted the nucleotide sequences of the LT (*eltAB*) and ST (STh and STp genes) operons for further analyses. Sequence analysis showed that *eltAB* was more variable than the STh and STp genes, a finding in agreement with previous studies[17,18]. Mapping LT and ST allele data onto the MCG-based phylogenetic tree showed a close association of toxin alleles with the chromosomal background and the colonization factor and toxin profiles. For example, the closely related lineages L1

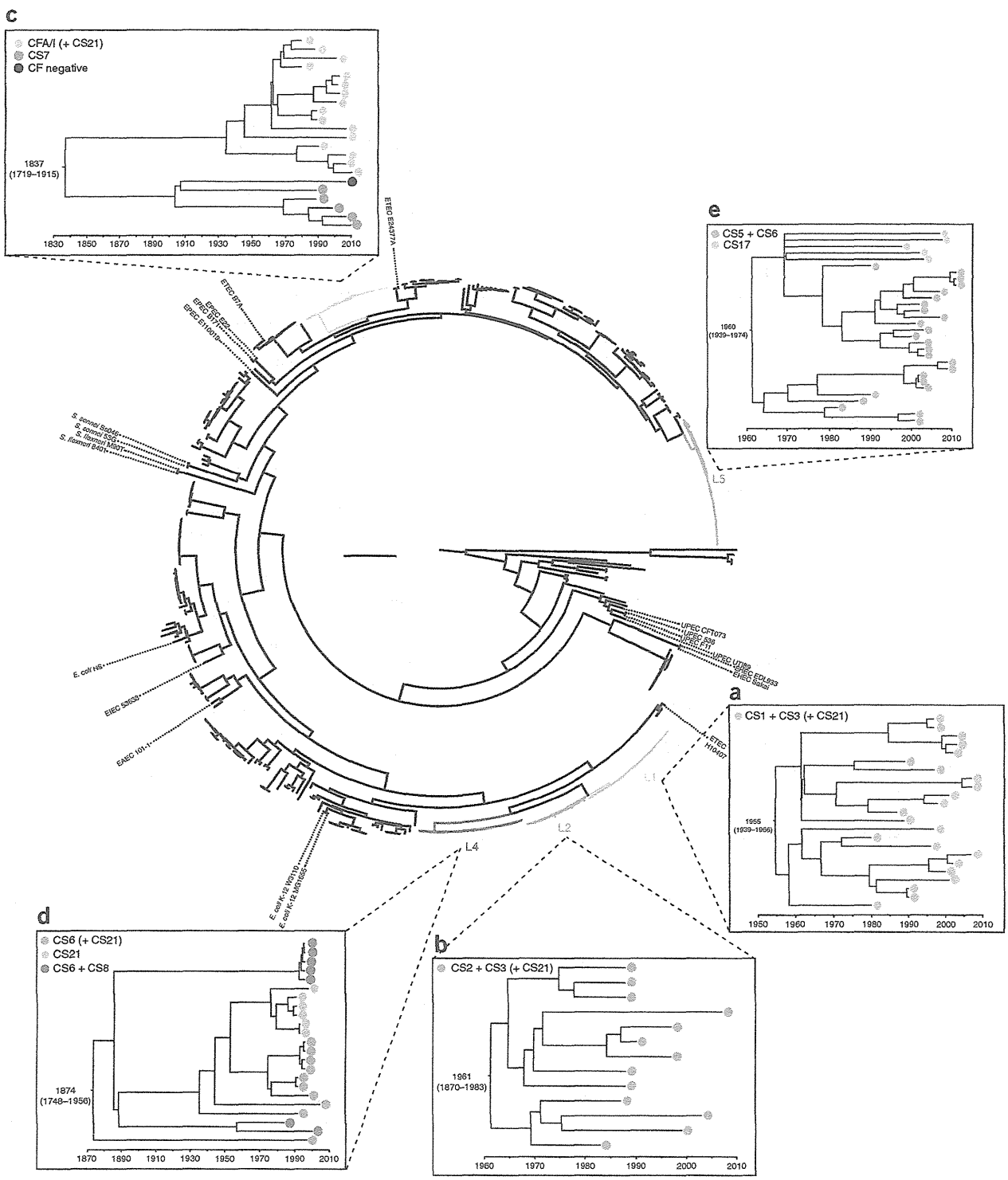

Figure 2 Most common recent ancestors of the five major lineages L1–L5. Midpoint-rooted circular phylogenetic tree of ETEC based on SNP differences, with probable recombination events removed and 21 *E. coli* references indicated. (**a**–**e**) Surrounding the circular tree are five individual SNP-based phylogenetic trees, with recombination events removed, of the major lineages: L1 (**a**), L2 (**b**), L3 (**c**), L4 (**d**) and L5 (**e**). Node heights in each tree are the median values obtained from Bayesian phylogenetic analysis with BEAST. The MRCA of each lineage is labeled with inferred dates. Scale bar, 0.08 substitutions per variable site. Two isolates have been removed from L5 and one isolate has been removed from L2 for the BEAST analyses.

(CS1 + CS3 (± CS21)) and L2 (CS2 + CS3 (± CS21)) were positive for LT + STh and shared the same LT and STh allelic variants. Interestingly, this LT allele was not found in any other isolates of this data set, whereas the ST variant (STa3/4) was found elsewhere in the tree (L5). Two other lineages, L3 (CFA/I) and L5 (CS5 + CS6), comprised LT + STh and STh-only isolates. Isolates in both L3 and L5 had the same translated LT variant, LT2 (ref. 17). These results demonstrate that not only the colonization factor profile but also the toxin allele profile is associated with chromosomal background, further supporting the idea that specific chromosomal-plasmid combinations are present in stable ETEC lineages.

Specific plasmid incompatibility groups are found in each lineage

Using *in silico* PCR with specific primers for the most common plasmid incompatibility groups, we could estimate the number of different replicons present in each ETEC isolate and relate this number to the virulence factors present in each of the lineages. Sequences corresponding to plasmid incompatibility groups could be identified in 317 plasmids from the 362 ETEC isolates sequenced. Isolates in L1 (CS1 + CS3) contained 2–5 different incompatibility groups, FII, FIIY, FrepB, FIB and I1 (**Supplementary Figs. 1** and **2**, and **Supplementary Table 2**), corresponding to the pCoo plasmid encoding CS1 and incompatibility groups I1 (CS1) and FII (CS3) identified in the ETEC E24377A (CS1 + CS3) strain[19,20]. FIB, known to be associated with virulence traits[20], was identified in all isolates within the L3 (CFA/I) lineage. Additional incompatibility groups were found in the majority of the isolates in lineage L3; among these were FII, FIIY and FrepB (**Supplementary Figs. 1** and **2**, and **Supplementary Table 2**). In the fully annotated genome of the reference strain H10407 (CFA/I), four plasmids have previously been identified: two larger plasmids with FII, one of which harbors the CFA/I operon, and two smaller plasmids with ColE1 (ref. 21). Although the number of incompatibility groups varied within a lineage, a pattern of conserved incompatibility groups in each lineage was apparent.

Five selected ETEC lineages emerged during the last century

Our findings demonstrate a clear association between the O antigen and virulence profiles, which correlates well with the chromosomal genotype of the ETEC lineages. To estimate the time of emergence of the selected lineages, we reconstructed an evolutionary history of these lineages using BEAST. A maximum-clade-credibility (MCC) tree was generated for each lineage, and this tree was used to estimate the most recent common ancestor (MRCA). The substitution rates (number of substitutions per site per year) for the five selected lineages were largely consistent and were estimated to be 1.0×10^{-6} (L1), 1.0×10^{-6} (L2), 3.7×10^{-7} (L3), 4.0×10^{-7} (L4) and 1.1×10^{-6} (L5). The substitution rates in L1, L2 and L5 were similar to the mutation rate of 1.57×10^{-6} estimated in *Streptococcus pneumoniae*[22]. Lineages L3 and L4 had slightly lower substitution rates, which were in the same range as that of *Clostridium difficile*[23]. Estimates for L1–L5 correspond to an accumulation of approximately 2–5.5 SNPs per genome per year.

In a time-dependent reconstruction of the identified lineages, L1–L5 were estimated to have emerged between 51 and 174 years ago (**Fig. 2**). Hence, these lineages emerged between the 1840s and 1970s. This result further supports the idea that the lineages are stable and implies a tighter and longer-term coupling of the chromosome and plasmid than has previously been appreciated.

DISCUSSION

ETEC infections are a major cause of diarrhea in low- and middle-income countries, and, although diarrhea incidence may be declining slightly[24], the total burden in children in these countries remains high. Additional research is needed to improve prevention as well as treatment; hence, a more detailed understanding of what actually constitutes a naturally occurring ETEC strain is vital. We have applied high-resolution genomic analysis to determine the phylogenetic structure of a globally representative ETEC collection in the context of the whole *E. coli* species. ETEC isolates were found across *E. coli* phylogroups A, B1, B2, D and E in the MCG-based phylogenetic tree, in agreement with previous multilocus sequence type (MLST)-based studies (using internal sequences of housekeeping genes: *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA* and *recA*)[8,25], demonstrating that ETEC are indeed genetically diverse (**Fig. 1**). Despite this diversity, we unexpectedly found that many ETEC isolates formed discrete lineages. Combining the colonization factor and toxin profiles of the ETEC isolates with whole-genome data, we demonstrated the existence of clearly identifiable ETEC lineages. These lineages were not only phylogenetically related but also shared consistent plasmid-encoded virulence profiles. For example, the closely related L1 and L2 lineages harbored CS1 + CS3 (± CS21) and CS2 + CS3 (± CS21) profiles, respectively, and shared specific LT and ST alleles, and they all expressed the O6 antigen (**Fig. 2**, **Table 1** and **Supplementary Figs. 1** and **2**). This close clustering of isolates from different parts of the world collected over a period of 30 years may suggest that the acquisition of plasmid-encoded virulence factors occurred once and was then followed by a clonal expansion of isolates carrying the same virulence profile. It is clear that this event has not only occurred in a single case: we observed multiple distantly related lineages where the same pattern was evident. For instance, lineage L5 mainly comprised isolates positive for CS5 + CS6 that were not found elsewhere in the data set. In contrast, CFA/I-positive isolates were found in two lineages that had different O antigens but shared the same LT allele. Isolates expressing CS6 alone or in combination with CS8 or CS21 were identified in four lineages (L4 and L7–L9) spread across the MCG-based phylogenetic tree as well as in additional lineages together with colonization factor–negative isolates (**Figs. 1** and **2**, and **Supplementary Figs. 1** and **2**). CS6 is known to be a diverse colonization factor, as several variants of the genes encoding the structural subunits, *cssA* and *cssB*, have been identified[26]. CS6-positive isolates clustered on the basis of different variants of the CS6 structural subunits (data not shown). Screening for plasmid incompatibility groups provides further evidence that a specific plasmid and chromosomal background is stably maintained in a population (**Supplementary Figs. 1** and **2**). In summary, ETEC isolates with distinct virulence profiles can form monophyletic lineages, for example, CS1 + CS3 (± CS21), CS2 + CS3 (± CS21) and CS5 + CS6 isolates, or arise in isolates with different chromosomal backgrounds, for example, CFA/I and CS6 isolates.

It has previously been suggested that the acquisition of colonization factor and toxin genes is sufficient for the emergence of pathogenic ETEC isolates[8]. Previous studies had indicated a low frequency of phylogenetic clustering of ETEC isolates[9,10] and that the acquisition of virulence-related genes occurred at multiple times, and no common clonal lineages with distinct virulence profiles had been identified[8,12,27]. Indeed, a portion of the isolates in our ETEC collection form lineages with isolates of mixed O antigens and virulence profiles; however, these isolates are mostly represented by colonization factor–negative isolates and by ETEC expressing less prevalent colonization factors. It is possible that there are additional lineages that are yet to be defined among these less frequent isolates or that these isolates represent ETEC that have recently acquired virulence plasmids.

Our data clearly demonstrate that ETEC harbor identifiable lineages, with the majority containing consistent, definable virulence profiles. This finding implies that, in these lineages, the virulence determinants were acquired once and the clades subsequently spread,

in some instances, around the world. This narrative is markedly different from the one that could be expected on the basis of previous work on this pathovar, and it is clear from our analysis that the ETEC population contains a number of stable clones with specific virulence profiles. Similar patterns have been identified in other pathogens. Detailed analyses of the five major lineages (L1–L5) using Bayesian-based analysis suggest that these lineages emerged between 51 and 174 years ago (**Fig. 2**). This observation is consistent with the time period of the spread of several other major pathogens, including *Vibrio cholerae*[28], *Staphylococcus aureus*[29], *Shigella sonnei*[30] and invasive *Salmonella* Typhimurium[31]. The precise cause of the emergence and spread of these organisms is likely multifactorial, but factors such as international travel have certainly contributed to the spread of these pathogens around the world.

In summary, the data presented here show that ETEC-mediated disease is actually a set of overlapping global epidemics of individual ETEC lineages, which have been stable over substantial periods of time in endemic areas. Furthermore, these globally distributed lineages with consistent colonization factor profiles seem to cause disease in children and adults in endemic areas and travelers to the same extent. It is of particular interest that our data suggest that plasmid acquisition is a major vehicle driving the emergence of different ETEC clades. This suggests that the development of a vaccine based on the most prevalent colonization factors could be protective against a large proportion of ETEC diarrhea cases. New ETEC vaccines are under development that are based on the major colonization factors identified in ETEC causing diarrhea[5,32]. Hence, our study has key implications for how ETEC disease is understood and tracked, and may help in disease prevention.

URLs. Source code for SNP calling, https://github.com/sanger-pathogens/snp_sites.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Primary accession codes for the Illumina sequence reads of all 362 ETEC isolates sequenced in this study are included in **Supplementary Table 3.**

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust (grant 098051), the Swedish Research Council (grants 2012-3464 and 2011-3435), the Swedish Strategic Foundation (grant SB12-0072) and the European Research Council (grant 239784).

AUTHOR CONTRIBUTIONS

A.v.M., G.D., A.-M.S., Å.S., T.R.C. and D.A.R. contributed to the design of the study and data interpretation. A.v.M. and G.W. extracted DNA. A.v.M. screened the sequence data and performed the majority of the bioinformatics analyses with input from T.R.C. and N.R.T. A.v.M. interpreted and analyzed the results from the recombination detection and BAPS analyses, executed by J.C. T.S. and L.H.W. identified the MCG and determined sequence types from whole-genome data. A.I. performed the BLASTN analysis to identify O antigen genotypes in all ETEC isolates included. E.J. analyzed the genes encoding toxins. D.P. was responsible for forwarding extracted DNA samples to the sequencing pipeline at the Wellcome Trust Sanger Institute. G.D., Å.S. and A.-M.S. supervised the work. All authors contributed to the writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. World Health Organization. *Diarrhoeal Diseases (Updated February 2009)* (World Health Organization, Geneva, 2009).
2. Gaastra, W. & Svennerholm, A.M. Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC). *Trends Microbiol.* **4**, 444–452 (1996).
3. Qadri, F., Svennerholm, A.M., Faruque, A.S. & Sack, R.B. Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. *Clin. Microbiol. Rev.* **18**, 465–483 (2005).
4. Isidean, S.D., Riddle, M.S., Savarino, S.J. & Porter, C.K. A systematic review of ETEC epidemiology focusing on colonization factor and toxin expression. *Vaccine* **29**, 6167–6178 (2011).
5. Svennerholm, A.M. & Lundgren, A. Recent progress toward an enterotoxigenic *Escherichia coli* vaccine. *Expert. Rev. Vaccines* **11**, 495–507 (2012).
6. Wolf, M.K. Occurrence, distribution, and associations of O and H serogroups, colonization factor antigens, and toxins of enterotoxigenic *Escherichia coli*. *Clin. Microbiol. Rev.* **10**, 569–584 (1997).
7. Smith, H.W. The exploitation of transmissible plasmids to study the pathogenesis of *E. coli* diarrhoea. *Proc. R. Soc. Med.* **66**, 272–273 (1973).
8. Turner, S.M. *et al.* Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. *J. Clin. Microbiol.* **44**, 4528–4536 (2006).
9. Steinsland, H., Lacher, D.W., Sommerfelt, H. & Whittam, T.S. Ancestral lineages of human enterotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* **48**, 2916–2924 (2010).
10. Escobar-Páramo, P. *et al.* A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol. Biol. Evol.* **21**, 1085–1094 (2004).
11. Regua-Mangia, A.H. *et al.* Genotypic and phenotypic characterization of enterotoxigenic *Escherichia coli* (ETEC) strains isolated in Rio de Janeiro city, Brazil. *FEMS Immunol. Med. Microbiol.* **40**, 155–162 (2004).
12. Steinsland, H., Valentiner-Branth, P., Aaby, P., Mølbak, K. & Sommerfelt, H. Clonal relatedness of enterotoxigenic *Escherichia coli* strains isolated from a cohort of young children in Guinea-Bissau. *J. Clin. Microbiol.* **42**, 3100–3107 (2004).
13. Valvatne, H., Steinsland, H. & Sommerfelt, H. Clonal clustering and colonization factors among thermolabile and porcine thermostable enterotoxin-producing *Escherichia coli*. *APMIS* **110**, 665–672 (2002).
14. Sahl, J.W. & Rasko, D.A. Analysis of global transcriptional profiles of enterotoxigenic *Escherichia coli* isolate E24377A. *Infect. Immun.* **80**, 1232–1242 (2012).
15. Herzer, P.J., Inouye, S., Inouye, M. & Whittam, T.S. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *J. Bacteriol.* **172**, 6175–6181 (1990).
16. Corander, J., Marttinen, P., Sirén, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**, 539 (2008).
17. Lasaro, M.A., Mathias-Santos, C., Rodrigues, J.F. & Ferreira, L.C.S. Functional and immunological characterization of a natural polymorphic variant of a heat-labile toxin (LT-I) produced by enterotoxigenic *Escherichia coli* (ETEC). *FEMS Immunol. Med. Microbiol.* **55**, 93–99 (2009).
18. Rodrigues, J. *et al.* Clonal structure and virulence factors in strains of *Escherichia coli* of the classic serogroup O55. *Infect. Immun.* **64**, 2680–2686 (1996).
19. Froehlich, B., Parkhill, J., Sanders, M., Quail, M.A. & Scott, J.R. The pCoo plasmid of enterotoxigenic *Escherichia coli* is a mosaic cointegrate. *J. Bacteriol.* **187**, 6509–6516 (2005).
20. Johnson, T.J. & Nolan, L.K. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* **73**, 750–774 (2009).
21. Crossman, L.C. *et al.* A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407. *J. Bacteriol.* **192**, 5822–5831 (2010).
22. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
23. Didelot, X. *et al.* Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol.* **13**, R118 (2012).
24. Fischer Walker, C.L., Perin, J., Aryee, M.J., Boschi-Pinto, C. & Black, R.E. Diarrhea incidence in low- and middle-income countries in 1990 and 2010: a systematic review. *BMC Public Health* **12**, 220 (2012).
25. Sahl, J.W. *et al.* A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. *Infect. Immun.* **79**, 950–960 (2011).
26. Sabui, S. *et al.* Allelic variation in colonization factor CS6 of enterotoxigenic *Escherichia coli* isolated from patients with acute diarrhoea and controls. *J. Med. Microbiol.* **59**, 770–779 (2010).
27. Pupo, G.M., Karaolis, D.K., Lan, R. & Reeves, P.R. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Infect. Immun.* **65**, 2685–2692 (1997).
28. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
29. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
30. Holt, K.E. *et al. Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44**, 1056–1059 (2012).
31. Okoro, C.K. *et al.* Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. *Nat. Genet.* **44**, 1215–1221 (2012).
32. Darsley, M.J. *et al.* The oral, live attenuated enterotoxigenic *Escherichia coli* vaccine ACE527 reduces the incidence and severity of diarrhea in a human challenge model of diarrheal disease. *Clin. Vaccine Immunol.* **19**, 1921–1931 (2012).

ONLINE METHODS

Isolate selection. In total, 362 human ETEC isolates from the Gothenburg University (UG) ETEC collection, comprising more than 3,500 ETEC isolates, were selected on the basis of virulence factor profile, origin and year of isolation to represent a broad collection of ETEC isolated worldwide. This collection included attempts to cover the most prevalent colonization factor and toxin profiles as well as isolates with rare colonization factors or lacking identifiable colonization factors. In addition, isolates were from as many geographical locations as possible collected over a long time period. Isolates from different patient groups were also included, i.e., from children and adults in ETEC-endemic areas as well as from travelers and soldiers visiting such areas. Most isolates were from subjects suffering from diarrhea, both hospitalized and treated as outpatients; some isolates were derived from asymptomatic carriers. The isolates were from several countries in Asia, Africa, and North, Central and South America collected between 1980 and 2011 (**Supplementary Table 2**). ETEC isolates were identified by culture on MacConkey agar followed by analysis of LT and ST toxin expression using GM1 ELISAs and in some cases PCR[33]. Different colonization factors were identified by dot-blot analyses and in some cases by multiplex PCR[34]. Isolates had been kept in glycerol stocks at −70 °C, and each isolate had been passaged as few times as possible. The strains were collected with informed consent from patients or the parents of children. Permission to use the ETEC strain collection was granted by the Regional Ethical Board of Gothenburg, Sweden (Ethics Committee Reference 088-10).

Bacterial culture and DNA preparation. All isolates were initially cultured on horse blood agar plates overnight at 37 °C to detect possible contamination. Pure ETEC cultures were used for DNA extraction with the Wizard Genomic DNA kit (Promega) according to the manufacturer's instructions. DNA quantity was measured by NanoDrop spectrophotometer (NanoDrop Technologies). A DNA concentration of at least 72 ng/μl was used for each isolate in Illumina sequencing.

Genomic library preparation and DNA sequencing. Paired-end sequencing was performed on the Illumina HiSeq 2000 platform. Libraries were constructed, according to the protocols of Quail *et al.*[35], with a fragment length of 75 or 100 bases in pools of uniquely tagged isolates. In total, 362 isolates were sequenced with a read length of 75 or 100 bases. The difference in read length did not interfere with assembly. Index tagging sequences were used to assign reads to individual samples for further analysis.

***De novo* sequence assembly.** Paired-end Illumina sequence data from each isolate were *de novo* assembled using the Velvet pipeline developed in house at the Wellcome Trust Sanger Institute. The pipeline consisted of the following steps. First, Velvet Optimiser fragmented the reads into shorter sequences (66–90% of original read length), the best parameter set was chosen on the basis of N50 values and Velvet was run to produce a set of contigs. Second, contigs of less than 300 bases in length were removed. Third, the scaffolding software SSPACE (v 2.0) was run to scaffold the assembly from the previous step. Fourth, gaps were filled by GapFiller (v 1.10). Fifth, reads were mapped to the scaffolds from the third step, and statistics and graphs were generated for this assembly. In total, four samples were removed from the data set because of contamination or incomplete assembly. The best assembly for each isolate was chosen on the basis of a combination of contig length, contig number and N50 value.

Determination of the maximum common genome and SNP detection. The genomes of 47 fully sequenced *E. coli* strains available in GenBank (June 2012) were used. We extracted a set of 1,429 genes that were termed the maximum common genome (MCG), i.e., all genes that were present in each of the 47 genomes. Genes with at least 70% identity on the protein level were considered to form the MCG. This was determined by hierarchical clustering using CD-HIT[36]. The protein sequences of the core genes were concatenated for each of the 47 genomes, and a maximum-parsimony phylogeny was created. Extracted core genes for all 362 ETEC genomes were compared, and SNPs were detected using an in-house script (source code can be found on GitHub; see URLs).

Recombination analysis. The BratNextGen method[37] was used to detect recombination events with the concatenated MCG alignment for 362 ETEC isolates and 21 *E. coli* references. Estimation of recombination was performed with the default settings as in refs. 37–39, using 20 iterations of the estimation algorithm, which was assessed to be sufficient as changes in the hidden Markov model parameters were already negligible over the last 70% of the iterations. The significance of a recombining region was determined as in ref. 37 using a permutation test with 100 permutations executed in parallel on a cluster computer with the threshold of 5% to conclude significance for each region. Every significant recombination was then masked as missing data in the MCG alignment to provide robust input data for phylogenetic and population genetic analyses.

Phylogenetic analysis. In total, 1,429 genes constituting the MCG of 47 well-annotated *E. coli* and *Shigella* strains were extracted. All phylogenetic trees were generated by FastTree (v2.1.4) for all variable sites using a general time-reversible model with gamma correction for among-site rate variation for ten initial trees and the maximum-likelihood algorithm. Phylogenetic trees are based on the alignment of the 1,429 genes, with recombination sites masked. In addition to the 362 sequenced ETEC isolates, 21 *E. coli* reference strains were included (**Supplementary Table 1**).

Population genetic analysis. To estimate the population structure, we used the BAPS v6.0 software[16,40], in particular, its module hierBAPS[41], which fits lineages to genome data using nested clustering. BAPS has been shown to efficiently estimate bacterial population structure from both limited core genome variation[42–44] and whole-genome sequence data[28,38,39,45]. Two nested levels of molecular variation were fitted to the MCG alignment with estimated recombinations masked as missing data. The estimation used 10 independent runs of the stochastic optimization algorithm with the a priori upper bound of the number of clusters varying over the interval 50–150 across the runs. The estimated mode of the posterior distribution had 8 and 39 clusters at levels 1 and 2 of the hierarchy, respectively. All clusters were significantly supported when compared against alternative partitions (posterior probability for any cluster of at least 100-fold higher than for the alternative).

Bayesian phylogeny and estimating dates of emergence of lineages. Estimation of rates of evolution and the age of individual clusters was performed using Bayesian inference, BEAST (v5.8.8)[46], on SNP alignments. Various combinations of a population size change model and molecular clock model were compared to find the model best fitting the data. In all cases, the Bayes factor showed strong support (Bayes factor >> 200) for the use of a skyline[47] model of population size change and a relaxed uncorrelated lognormal clock, which allows evolutionary rates to change among the branches of the tree[28], and a Hasegawa, Kishino and Yano (HKY) substitution model with gamma-distributed rate heterogeneity among sites was used[48]. In all cases, three independent chains were run for 500 million steps each with sampling every 10,000 steps. The three chains were combined with LogCombiner[46], with the first 50 million steps removed from each as a burn-in. MCC trees were created and annotated using TreeAnnotator and were viewed in FigTree[46]. We report estimates as median values within 95% highest probability density (HPD) intervals and report posterior probability values as support for identified ancestral node age.

Sequence-based determination of O antigens. On the basis of a sequence data set for two O antigen–processing gene sets (*wzx/wzy* and *wzm/wzt*) for all 184 *E. coli* O antigens (A.I., S. Iyoda, T. Kikuchi, Y. Ogura and K. Katsura, unpublished data) and the results of BLASTN analysis, we identified the ETEC O antigen genotype (defined by ≥97% sequence identity and ≥97% aligned length coverage of a query sequence, in both genes from a set) from the assembled draft genomes. Moderately diversified sequences (defined by 70–97% sequence identity and ≥70% aligned length coverage) were classified into '-like' categories. For isolates with previously unknown O antigen genotypes, *wzx/wzy* and *wzm/wzt* sequences were extracted from the regions of the draft genomes encoding O antigen biosynthesis genes and classified into groups that were categorized as novel *E. coli* O antigen genotypes (defined by ≥97% sequence identity and ≥97% aligned length coverage).

Screening for plasmid incompatibility groups. In total, 362 ETEC isolates were screened for known plasmid incompatibility groups using *in silico* PCR with primers for the following incompatibility groups: FII, FIIK, FIIS, FIIY, FIA, BIB, FrepB, HI1, I1, K, N, P, X1, X4 and Y[49].

33. Sjöling, A., Wiklund, G., Savarino, S.J., Cohen, D.I. & Svennerholm, A.M. Comparative analyses of phenotypic and genotypic methods for detection of enterotoxigenic *Escherichia coli* toxins and colonization factors. *J. Clin. Microbiol.* **45**, 3295–3301 (2007).
34. Rodas, C. *et al.* Development of multiplex PCR assays for detection of enterotoxigenic *Escherichia coli* colonization factors and toxins. *J. Clin. Microbiol.* **47**, 1218–1220 (2009).
35. Quail, M.A., Swerdlow, H. & Turner, D.J. Improved protocols for the Illumina Genome Analyzer sequencing system. *Curr. Protoc. Hum. Genet.* **Unit 18.2** 1–27 (2009).
36. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
37. Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012).
38. McNally, A., Cheng, L., Harris, S.R. & Corander, J. The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome Biol. Evol.* **5**, 699–710 (2013).
39. Castillo-Ramírez, S. *et al.* Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. *Genome Biol.* **13**, R126 (2012).
40. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLOS Comput. Biol.* **5**, e1000455 (2009).
41. Cheng, L., Connor, T.R., Sirén, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
42. Corander, J., Connor, T.R., O'Dwyer, C.A., Kroll, J.S. & Hanage, W.P. Population structure in the *Neisseria*, and the biological significance of fuzzy species. *J. R. Soc. Interface* **9**, 1208–1215 (2012).
43. Willems, R.J.L. *et al.* Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. *MBio* **3**, e00151–12 (2012).
44. Hanage, W.P., Fraser, C., Tang, J., Connor, T.R. & Corander, J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science* **324**, 1454–1457 (2009).
45. Croucher, N.J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45**, 656–663 (2013).
46. Drummond, A.J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
47. Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
48. Drummond, A.J., Ho, S.Y., Phillips, M.J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
49. Carattoli, A. *et al.* Identification of plasmids by PCR-based replicon typing. *J. Microbiol. Methods* **63**, 219–228 (2005).

Journal of Food Protection, Vol. 77, No. 8, 2014, Pages 1269–1274
doi:10.4315/0362-028X.JFP-13-506

Identification of O Serotypes, Genotypes, and Virulotypes of Shiga Toxin–Producing *Escherichia coli* Isolates, Including Non-O157 from Beef Cattle in Japan

HIROHISA MEKATA,[1] ATSUSHI IGUCHI,[2] KIMIKO KAWANO,[3] YUMI KIRINO,[1] IKUO KOBAYASHI,[4] AND NAOAKI MISAWA[3]*

[1]*Project for Zoonoses Education and Research, Faculty of Agriculture, and* [3]*Center for Animal Disease Control, University of Miyazaki, 1-1 Gakuen-Kibanadai-Nishi, Miyazaki, 889-2192, Japan;* [2]*Interdisciplinary Research Organization, University of Miyazaki, 5200 Kiyotake, Miyazaki, 889-1692, Japan; and* [4]*Sumiyoshi Livestock Science Station, Field Science Center, Faculty of Agriculture, University of Miyazaki, 10100-1 Shimanouchi, Miyazaki 880-0121, Japan*

MS 13-506: Received 25 November 2013/Accepted 3 April 2014

ABSTRACT

Bovines are recognized as an important reservoir of Shiga toxin–producing *Escherichia coli* (STEC). Although STEC strains are significant foodborne pathogens, not all of the STEC held by cattle are pathogenic, and which type of STEC that will become epidemic in humans is unpredictable. Information about the prevalence of serotype and virulence gene distribution in beef cattle is insufficient to develop monitoring and controlling activities for a food safety and security program. Thus, this study investigated the prevalence of O157 and non-O157 STEC in Japanese beef cattle and characterized the isolates by the type of O antigen and several virulence markers to help predict the pathogenicity. In this study, 64.2% (176 of 274) of enrichment cultures of fecal samples collected from an abattoir and farms were stx_1 and/or stx_2 positive by PCR. STEC strains were isolated from 22.1% (39 of 176) of the positive fecal samples, and these isolates represented 17 types of O antigen (O1, O2 or O50, O5, O8, O55, O84, O91, O109, O113, O136, O150, O156, O157, O163, O168, O174, and O177). Two selective media targeting major STEC groups, cefixime-tellurite sorbitol MacConkey agar and CHROMagar O26/O157, allowed isolation of a variety of STEC strains. The most frequently isolated STEC was O113 (8 of 39), which has previously been reported as a cause of foodborne infections. Although most of the O113 STEC isolated from infected patients possessed the enterohemolysin (*hlyA*) gene, none of the O113 STEC cattle isolates possessed the *hlyA* gene. The second most common isolate was O157 (6 of 39), and all these isolates contained common virulence factors, including *eae*, *tir*, lpf_1, lpf_2, and *hlyA*. This study shows the prevalence of O157 and non-O157 STEC in Japanese beef cattle and the relationship of O antigen and virulotypes of the isolates. This information may improve identification of the source of infection, developing surveillance programs or the current understanding of virulence factors of STEC infections.

Shiga toxin–producing *Escherichia coli* (STEC), alternatively referred to as verocytotoxigenic *E. coli*, is a significant foodborne pathogen causing diarrhea, hemorrhagic colitis, and hemolytic uremic syndrome *(17, 20)*. More than 100 types of O antigens from STEC isolates have thus far been identified, and O157 STEC has been the main serotype causing STEC outbreaks in the world *(9, 21, 38)*. Recently, several non-O157 STEC have also been identified as the cause of severe illnesses *(6, 19)*. In Japan, infections caused by non-O157 serotypes have increased annually and accounted for 47% of all STEC infections in 2012 *(28)*.

The pathogenicity of STEC in humans depends on a number of virulence factors, such as capacity to produce the Shiga toxin and attach to target cells *(13, 24)*. Shiga toxin is the most important virulence factor in STEC and can be differentiated into two major groups: stx_1 and stx_2. Several variants (stx_{1a}, stx_{1c}, stx_{1d}, and stx_{2a} to stx_{2g}) have been identified, and a close association has been found between the severity of disease and the presence of the variants, such as stx_{2a} and/or stx_{2c} *(4, 14, 30)*. In addition to these, STEC strains have a number of other virulence factors. One of the major virulence factors, intimin, is associated with attachment to the host's intestinal mucosa conferred by the gene *eaeA*, and its receptor is encoded by the *tir* gene *(35, 44, 46)*. Other virulence genes associated with attachment to cells include the long polar fimbriae (*lpf*) and the STEC auto-agglutinating adhesion (*saa*) gene *(5, 43)*. Additionally, the enterohemolysin (*hlyA*) and subtilase (*subAB*) virulence factors were found to associate with hemorrhagic enterocolitis *(33, 39)*.

Bovines are recognized as a primary source of STEC, and many outbreaks are linked to bovine food sources *(1, 3, 12, 23)*. In addition, consumption of water or vegetables contaminated with bovine feces is frequently involved in outbreaks *(18, 29)*. Despite an increase in the number of STEC outbreaks, there is a lack of data on prevalence at the farm level, distribution of serotypes, and virulence factors. To

* Author for correspondence. Tel and Fax: (81)-985-58-7284; E-mail: a0d901u@cc.miyazaki-u.ac.jp.

TABLE 1. *Target and primer sequences used for the detection of STEC virulence genes*

Target gene	Primer	Primer sequence	Reference
stx_1	stx1F	5′-ATAAATCGCCATTCGYTGACTAC-3′	*31*
	stx1R	5′-AGAACGYCCACTGAGATCATC-3′	
stx_2	stx2F	5′-GGCACTGTCTGAAACTGCTCC-3′	*31*
	stx2R	5′-TCGCCASTTATCTGACATTCTG-3′	
eae	eaeAF	5′-GACCCGGCACAAGCATAAGC-3′	*31*
	eaeAR	5′-CCACCTGCAGCAACAAGAGG-3′	
vtx_{1a}	vtx1a-F1	5′-CCTTTCCAGGTACAACAGCGGTT-3′	*45*
	vtx1a-R2	5′-GGAAACTCATCAGATGCCATTCTGG-3′	
vtx_{1c}	vtx1c-F1	5′-CCTTTCCTGGTACAACTGCGGTT-3′	*45*
	vtx1c-R1	5′-CAAGTGTTGTACGAAATCCCCTCTGA-3′	
vtx_{1d}	vtx1d-F1	5′-CAGTTAATGCGATTGCTAAGGAGTTTACC-3′	*45*
	vtx1d-R1	5′-CTCTTCCTCTGGTTCTAACCCCATGATA-3′	
vtx_{2a}	vtx2a-F2	5′-GCGATACTGRGBACTGTGGCC-3′	*45*
	vtx2a-R2	5′-GGCCACCTTCACTGTGAATGTG-3′	
	vtx2a-R3	5′-CCGKCAACCTTCACTGTAAATGTG-3′	
vtx_{2b}	vtx2b-F1	5′-AAATATGAAGAAGATATTTGTAGCGGC-3′	*45*
	vtx2b-R1	5′-CAGCAAATCCTGAACCTGACG-3′	
vtx_{2c}	vtx2c-F1	5′-GAAAGTCACAGTTTTTATATACAACGGGTA-3′	*45*
	vtx2c-R2	5′-CCGGCCACYTTTACTGTGAATGTA-3′	
vtx_{2d}	vtx2d-F1	5′-AAARTCACAGTCTTTATATACAACGGGTG-3′	*45*
	vtx2d-R1	5′-TTYCCGGCCACTTTTACTGTG-3′	
	vtx2d-R2	5′-GCCTGATGCACAGGTACTGGAC-3′	
vtx_{2e}	vtx2e-F1	5′-CGGAGTATCGGGGAGAGGC-3′	*45*
	vtx2e-R2	5′-CTTCCTGACACCTTCACAGTAAAGGT-3′	
vtx_{2f}	vtx2f-F1	5′-TGGGCGTCATTCACTGGTTG-3′	*45*
	vtx2f-R1	5′-TAATGGCCGCCCTGTCTCC-3′	
vtx_{2g}	vtx2g-F1	5′-CACCGGGTAGTTATATTTCTGTGGATATC-3′	*45*
	vtx2g-R1	5′-GATGGCAATTCAGAATAACCGCT-3′	
hlyA	hlyAF	5′-GCATCATCAAGCGTACGTTCC-3′	*31*
	hlyAR	5′-AATGAGCCAAGCTGGTTAAGCT-3′	
saa	SAADF	5′-CGTGATGAACAGGCTATTGC-3′	*32*
	SAADR	5′-ATGGACATGCCTGTGGCAAC-3′	
tir	TIR-F	5′-CATTACCTTCACAAACCGAC-3′	*22*
	TIR-R	5′-CCCCGTTAATCCTCCCAT-3′	
lpf_1	lpfO141-F	5′-CTGCGCATTGCCGTAAC-3′	*41*
	lpfO141-R	5′-ATTTACAGGCGAGATCGTG-3′	
lpf_2	lpfA-F	5′-ATGAAGCGTAATATTATAG-3′	*11*
	lpfA-R	5′-TTATTTCTTATATTCGAC-3′	
subAB	SubAF	5′-GTACGGACTAACAGGGAACTG-3′	*33*
	SubAR	5′-ATCGTCATATGCACCTCCG-3′	

establish monitoring programs and control methods of STEC contamination in food or the environment, we need to understand the prevalence of STEC, especially non-O157 STEC in cattle and its pathogenesis in humans. Therefore, this study investigated the relationship between the types of O antigen and the presence of virulence genes among the isolates.

MATERIALS AND METHODS

Sample collection. A total of 274 fresh cattle fecal samples were obtained from June 2012 to February 2013. Of these 274 cattle, 140 were from an abattoir, and the other 134 were from five individual farms located in different areas. Most samples were removed from animals, and some were obtained after defecation with attention to contamination. All the samples were transported to the laboratory within 1 h and stored at 4°C for no more than a 24 h prior to analysis.

Bacterial cultures and detection of *stx*-positive *E. coli*. Each fecal sample (1 g) was mixed by vortexing in 10 ml of modified *E. coli* broth (Nissui, Tokyo, Japan) and incubated overnight at 42°C. A 100-μl aliquot of each incubated broth was directly boiled for 10 min to extract the genomic DNA and centrifuged for 3 min (10,000 × *g* at 4°C). The supernatant was used as template DNA, and the remainder of the enrichment samples were stored at 4°C until the result of *stx* screening was obtained. The extracted DNAs were screened for the presence of stx_1 and/or stx_2 genes by PCR using a thermal cycler (Veriti 96-Well Thermal Cycler, Applied Biosystems, Foster City, CA). The primer sets and target genes are listed in Table 1 *(11, 22, 31–33, 41, 45)*. To neutralize the PCR inhibitor included in the template DNA, Ampdirect Plus (Shimadzu Biotech Co., Tsukuba, Japan) was used for the PCR. The PCR amplification was carried out in a reaction mixture containing 10 μl of 2× Ampdirect Plus, 1.0 μl of primer (100 pmol each), 0.1 μl of BIOTAQ HS DNA Polymerase (Bioline, London, UK), 1.0 μl of template DNA sample, and PCR-grade water to increase the volume to 20 μl. The reaction mixture was the same for both primer sets. Positive and blank control samples were included in each set of reactions. The cycling

TABLE 2. *The types and number of agar plates used for isolation of each O type STEC in this study*

Type of O antigen	No. of isolates	DHL	CT-sMAC	CHROMagar O26/O157
O1B	5	3	4	4
O2 or O50[a]	1	1	0	0
O5	1	0	0	1
O8	1	0	0	1
O55	1	0	1	0
O84	1	1	1	1
O91	1	1	1	0
O109	5	2	2	3
O113	8	1	8	5
O136	2	0	2	1
O150	1	0	0	1
O156	1	0	1	1
O157	6	1	5	6
O163	1	1	1	0
O168	1	0	1	1
O174	1	0	1	1
O177	1	0	0	1
OUT	1	0	1	1
Total	39	11	29	28

[a] O2 or O50: indistinctive.

conditions had an initial denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 30 s, 60°C for 1 min, and 72°C for 1 min. A final extension was performed at 72°C for 7 min. The PCR products were resolved by electrophoresis through a 2% agarose gel and were visualized under UV light (AE-6932GXCF, ATTO Corporation, Tokyo, Japan) by GelRed Nucleic Acid Gel Stain (Biotium, Hayward, CA). The product size was estimated using a 100-bp DNA ladder (TaKaRa Bio Company, Otsu, Japan).

Isolation of *stx*-positive *E. coli*. Then, 500 μl of each sample of stx_1 or/and stx_2 PCR-positive enrichment culture was centrifuged (10,000 × *g*, 5 min at 4°C). Because hydrochloric acid treatment is effective for isolating STEC from other bacteria, the resultant pellets were diluted in 100 μl of saline and treated with 100 μl of 0.5% NaCl and 1/8 N HCl for 30 s *(15, 16)*. Finally, these were plated onto desoxycholate hydrogen sulfide lactose (DHL) agar (Nissui), CHROMagar O26/O157 (CHROMagar Microbiology, Paris, France), and sorbitol MacConkey agar (Nissui) supplemented with cefixime-tellurite supplement (CT-sMAC, Merck, Darmstadt, Germany). After overnight incubation at 37°C, 4 to 6 colonies with differing colony color or morphology were taken from each plate and screened for the presence of *stx* genes by PCR amplification, as described above. *stx*-positive colonies were streaked onto Luria-Bertani (LB) agar (Nacalai Tesque, Kyoto, Japan) and incubated overnight at 37°C. Each streaked STEC isolate was individually cultured in 2 ml of LB broth (Nacalai Tesque) overnight at 37°C. Genomic DNA was extracted from a cell pellet by boiling for 10 min. All STEC isolates were preserved in LB broth with 10% glycerol at −80°C and were routinely recultured during the study.

Serotyping and genotyping. Serotypes of the O antigen were identified by the 50 types of antisera (Denka Seiken, Tokyo, Japan). Genotypes of O antigen were identified by multiplex PCR, which detects the specific sequence (mostly *wzx, wzy, wzt,* and *wzm* genes) of each O type of *E. coli*. This typing method was developed by Dr. Iguchi Interdisciplinary Research Organization at the University of Miyazaki (available at: http://www.cc.miyazaki-u.ac.jp/iguchi/iguchi_lab/O-genotyping.html).

Detection of virulence-related genes. The template DNA from each STEC isolate was tested by PCR for the presence of a range of virulence genes, including *eae, hlyA, saa, tir,* lpf_1, lpf_2, and *subAB* (Table 1). In the initial analysis, sequence analysis using nucleotide BLAST was performed to confirm the specificity of the PCR products. After this analysis, each PCR product size was checked by electrophoresis, and the specificity of the analyses was confirmed.

Determination of *stx* variants. Each STEC Shiga toxin gene was subtyped by PCR, according to the previous report *(45)*. The annealing temperatures and PCR conditions were slightly modified for a thermal cycler and *Taq* polymerase. Briefly, annealing temperatures of stx_1 and stx_2 variants were 66°C, except for vtx_{2c} (69°C) and vtx_{2d} (69°C).

RESULTS

Of the 274 fecal samples analyzed, 176 (64.2%) were stx_1 and/or stx_2 positive, and stx_1, stx_2, and the combination of both were detected from 42 (15.3%), 54 (19.7%), and 80 (29.1%) samples, respectively. STEC strains were isolated from 39 (22.1%) *stx*-positive samples. Of these STEC samples, 11, 29, and 28 isolates were isolated from DHL, CT-sMAC, and CHROMagar O26/O157 plates, respectively (Table 2). By the O-genotyping methods, 38 isolates were obtained representing 17 types of O antigen (O1, O2 or O50, O5, O8, O55, O84, O91, O109, O113, O136, O150, O156, O157, O163, O168, O174, and O177), and only one isolate could not be assigned (Table 3). On the other hand, 16 isolates were typeable (O1, O55, O91, O136, O157, and O168) and 23 were not typeable by the O serotyping method because only 50 types of O antisera were available. The most frequently occurring types of O antigen were O113 (20.5%), O157 (15.3%), O109 (12.8%), and O1 (12.8%).

The stx_{1d}, stx_{2b}, stx_{2e}, and stx_{2f} gene variants were not detected; however, stx_{1a}, stx_{1c}, stx_{2a}, stx_{2c}, stx_{2d}, and stx_{2g} were present in 15 (38.4%), 7 (17.9%), 18 (46.1%), 8 (20.5%), 8 (20.5%), and 1 (2.5%) of the isolates, respectively. Interestingly, the same *stx* gene variants were detected within strains from the same type of O antigen (Table 3). For example, all O1 and O109 strains possessed vtx_{2a}. Likewise, all O113, O136, and O157 isolates possessed vtx_{2d}, vtx_{1a}, and vtx_{2c}, respectively. The intimin gene (*eaeA*) was present in 38.4% (15 of 39) of isolates: serotypes O5, O84, O109, O150, O156, O157, and O177. The enterohemolysin gene (*hlyA*) was present in 64.1% (25 of 39). Other virulence-related genes, including *saa, tir,* lpf_1, lpf_2, and *subAB* were detected. Although O157 STEC strains possessed common virulence-related genes, non-O157 STEC tended to possess different profiles of virulence-related genes.

DISCUSSION

To prescreen samples before subjecting them to cultural procedures, PCR analyses, generally targeting the *stx* genes, are employed in many of the livestock fecal prevalence studies *(2, 8, 42)*. This study detected the *stx* genes in 64.2% of cattle feces. Previous studies have reported STEC

TABLE 3. *O antigen,* stx *variants, and virulence markers of STEC isolated from Japanese beef cattle*[a]

Genotype	Serotype[b]	No. of isolates (abattoir, farm)	vtx_1		vtx_2				*eae*	*tir*	*saa*	lpf_1	lpf_2	*hlyA*	*subAB*
			a	c	a	c	d	g							
O1B	O1	3 (2, 1)	−	−	+	−	−	−	−	−	−	−	−	+	−
O1B	O1	2 (1, 1)	−	−	+	−	−	−	−	−	−	−	−	−	−
O2 or O50	UT	1 (0, 1)	−	−	+	−	−	−	−	−	−	−	−	+	−
O5	UT	1 (0, 1)	+	−	−	−	−	−	+	−	−	−	−	+	−
O8	UT	1 (0, 1)	−	−	+	−	−	−	−	−	+	−	−	+	+
O55	O55	1 (1, 0)	+	−	−	−	−	−	−	−	+	−	−	−	−
O84	UT	1 (0, 1)	+	+	−	−	−	−	+	−	−	−	−	+	−
O91	O91	1 (1, 0)	+	−	−	−	−	−	−	−	+	−	−	+	+
O109	UT	1 (1, 0)	−	−	+	−	−	−	+	+	−	−	−	+	−
O109	UT	1 (0, 1)	−	−	+	−	−	−	+	+	−	+	+	+	−
O109	UT	1 (0, 1)	−	−	+	−	−	−	+	−	−	−	−	+	−
O109	UT	1 (0, 1)	−	−	+	−	−	−	−	−	−	−	−	−	−
O109	UT	1 (1, 0)	−	−	+	−	−	+	−	−	−	−	+	+	−
O113	UT	4 (2, 2)	+	+	−	−	+	−	−	−	−	−	+	−	−
O113	UT	2 (0, 2)	−	−	−	−	+	−	−	−	−	−	−	−	−
O113	UT	1 (0, 1)	−	+	−	−	+	−	−	−	−	−	+	−	−
O113	UT	1 (0, 1)	−	−	−	−	+	−	−	+	−	+	+	−	−
O136	O136	1 (0, 1)	+	−	−	−	−	−	−	−	−	−	+	+	−
O136	O136	1 (0, 1)	+	−	−	−	−	−	−	−	−	−	−	+	−
O150	UT	1 (0, 1)	+	−	+	−	−	−	+	−	−	−	−	+	−
O156	UT	1 (1, 0)	+	+	−	−	−	−	+	−	−	−	−	+	−
O157	O157	4 (1, 3)	−	−	+	+	−	−	+	+	−	+	+	+	−
O157	O157	2 (0, 2)	−	−	−	+	−	−	+	+	−	+	+	+	−
O163	UT	1 (0, 1)	+	−	−	−	−	−	−	−	+	−	−	+	+
O168	O168	1 (0, 1)	+	−	−	−	−	−	−	−	−	−	−	−	−
O174	UT	1 (0, 1)	−	−	−	+	−	−	−	−	−	−	−	−	−
O177	UT	1 (1, 0)	−	−	−	+	−	−	+	+	−	−	+	+	−
UT	UT	1 (0, 1)	+	−	+	−	−	−	+	−	−	−	−	+	−

[a] The stx_{1d}, stx_{2b}, stx_{2e}, and stx_{2f} gene variants were not detected. +, positive by PCR; −, negative by PCR.
[b] UT, untypeable.

prevalence rates in cattle of approximately 40 to 70% *(25, 37)*. The differences in these findings may be explained by differences in feed, seasonal peak, age, or detecting methods. Generally, PCR methods disclose a large number of positive samples, but only a portion of those samples yield cultures. In this study, CT-sMAC, DHL, and CHROMagar O26/O157 plates were used to isolate STEC. The CT-sMAC and CHROMagar O26/O157 plates isolated about three-fourths of all the isolates, and DHL plates isolated only two isolates (O2 or O50 and O109; data not shown). Therefore, this study indicates that CT-sMAC and CHROMagar O26/O157 plates are useful in isolating non-O157 STEC. Although 176 broths (64.2%) were *stx* positive by PCR, STEC isolates were cultivated from only 39 samples (22%). Although this result was considerably higher than the 7% previously reported, more than three-fourths of PCR-positive samples were negative by culture-based methods *(7)*.

The STEC O antigen of the isolates was determined by the O serotyping and genotyping methods. Although serotyping offers a very precise and reliable method for differentiating STEC isolates, only a few reference centers in the world are able to provide a full serotyping service. In addition, some field strains of *E. coli* have O antigens that do not react with the respective available antisera or react with more than two antisera. The genotyping method overcomes these problems, and combining both methods identifies the type of O antigen more precisely. In this study, only one isolate (1 of 39) was untyped by using both typing methods. This result was relatively small in number compared with other studies using only the serotyping method *(25, 26, 36)*.

Seventeen types of O antigen were isolated from cattle feces: O1, O2 or O50, O5, O8, O55, O84, O91, O109, O113, O136, O150, O156, O157, O163, O168, O174, and O177. The O113 STEC was the most prevalent O serotype, and other studies have also reported its isolation from cattle samples *(25, 26, 36)*. However, this occurrence has frequently been reported in clinical cases in humans *(10, 27, 34)*. Virulence gene analysis showed that stx_{2d} was common to all O113 STEC isolates, and this result coincided with previous work *(27)*. Although previous studies have shown that O113 STEC isolated from patients with hemolytic uremic syndrome or gastrointestinal illness possessed *hlyA* genes, all the O113 STEC isolates in this study did not contain *hlyA (27, 34)*. Therefore, *hlyA* might be the important virulence factor of STEC O113 against humans. Bacterial adherence and subsequent colonization on the intestinal epithelium cells also contribute to the infection and pathogenicity of STEC. As previously

reported for bovine isolates, more than half of our strains (38.4%, 15 of 39) were *eaeA* (intimin) negative *(25)*. Additionally, *saa* (another adhesion factor) was present in only four isolates. The third most prevalent types of O antigen in this study were O109 and O1. Although some O109 isolates possessed *hlyA* and *eaeA*, this serotype has been mainly isolated from healthy humans and cattle as a carrier. To the best of our knowledge, O109 STEC has not been reported as a cause of human infections. O1 was reported to be the source of some instances of diarrhea in human subjects *(40)*.

We investigated the prevalence of STEC in cattle and identified O serotypes and some virulence factors of isolates. Although many beef cattle possessed STEC, and they are recognized as an important reservoir of STEC, most of the cattle isolates are asymptomatic. This study revealed that many of cattle isolates did not harbor multivirulence factors, such as *hlyA* and some cell adhesion factors. The difference in prevalence of O antigen and virulence-related markers between humans and cattle may promote the identification of the important virulence factors of non-O157 STEC. Furthermore, accumulation of this kind of data may improve the field surveillance or monitoring programs and help find the source of infection or predict the prevalence of STEC in humans.

ACKNOWLEDGMENTS

The authors thank Y. Suehiro and K. Kumamoto (Miyakonojo Meat Inspection Center, Miyazaki Prefecture, Japan) for providing the samples. The work described in this report was funded by the Project for Zoonoses Education and Research.

REFERENCES

1. Armstrong, G. L., J. Hollingsworth, and J. G. Morris. 1996. Emerging foodborne pathogens: *Escherichia coli* O157:H7 as a model of entry of a new pathogen into the food supply of the developed world. *Epidemiol. Rev.* 18:29–51.
2. Barlow, R. S., and G. E. Mellor. 2010. Prevalence of enterohemorrhagic *Escherichia coli* serotypes in Australian beef cattle. *Foodborne Pathog. Dis.* 7:1239–1245.
3. Bettelheim, K. A., A. Kuzevski, R. A. Gilbert, D. O. Krause, and C. S. McSweeney. 2005. The diversity of *Escherichia coli* serotypes and biotypes in cattle faeces. *J. Appl. Microbiol.* 98:699–709.
4. Beutin, L., S. Kaulfuss, T. Cheasty, B. Brandenburg, S. Zimmermann, K. Gleier, G. A. Willshaw, and H. R. Smith. 2002. Characteristics and association with disease of two major subclones of Shiga toxin (Verocytotoxin)-producing strains of *Escherichia coli* (STEC) O157 that are present among isolates from patients in Germany. *Diagn. Microbiol. Infect. Dis.* 44:337–346.
5. Bono, J. L., J. E. Keen, M. L. Clawson, L. M. Durso, M. P. Heaton, and W. W. Laegreid. 2007. Association of *Escherichia coli* O157:H7 *tir* polymorphisms with human infection. *BMC Infect. Dis.* 7:98.
6. Brooks, J. T., E. G. Sowers, J. G. Wells, K. D. Greene, P. M. Griffin, R. M. Hoekstra, and N. A. Strockbine. 2005. Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983–2002. *J. Infect. Dis.* 192:1422–1429.
7. Caprioli, A., S. Morabito, H. Brugère, and E. Oswald. 2005. Enterohaemorrhagic *Escherichia coli*: emerging issues on virulence and modes of transmission. *Vet. Res.* 36:289–311.
8. Cobbold, R. N., D. H. Rice, M. Szymanski, D. R. Call, and D. D. Hancock. 2004. Comparison of Shiga-toxigenic *Escherichia coli* prevalences among dairy, feedlot, and cow-calf herds in Washington State. *Appl. Environ. Microbiol.* 70:4375–4378.
9. Currie, A., J. MacDonald, A. Ellis, J. Siushansian, L. Chui, M. Charlebois, M. Peermohamed, D. Everett, M. Fehr and L. K. Ng. 2007. Outbreak of *Escherichia coli* O157:H7 infections associated with consumption of beef donair. *J. Food Prot.* 70:1483–1488.
10. dos Santos, L. F., K. Irino, T. M. Vaz, and B. E. Guth. 2010. Set of virulence genes and genetic relatedness of O113:H21 *Escherichia coli* strains isolated from the animal reservoir and human infections in Brazil. *J. Med. Microbiol.* 59:634–640.
11. Doughty, S., J. Sloan, V. Bennett-Wood, M. Robertson, R. M. Robins-Browne, and E. L. Hartland. 2002. Identification of a novel fimbrial gene cluster related to long polar fimbriae in locus of enterocyte effacement-negative strains of enterohemorrhagic *Escherichia coli*. *Infect. Immun.* 70:6761–6769.
12. Etcheverría, A. I., and N. L. Padola. 2013. Shiga toxin-producing *Escherichia coli*: factors involved in virulence and cattle colonization. *Virulence* 4:366–372.
13. Farfan, M. J., and A. G. Torres. 2012. Molecular mechanisms that mediate colonization of Shiga toxin-producing *Escherichia coli* strains. *Infect. Immun.* 80:903–913.
14. Friedrich, A. W., M. Bielaszewska, W. L. Zhang, M. Pulz, T. Kuczius, A. Ammon, and H. Karch. 2002. *Escherichia coli* harboring Shiga toxin 2 gene variants: frequency and association with clinical symptoms. *J. Infect. Dis.* 185:74–84.
15. Fukushima, H., and M. Gomyoda. 1999. Hydrochloric acid treatment for rapid recovery of Shiga toxin-producing *Escherichia coli* O26, O111 and O157 from faeces, food and environmental samples. *Zentbl. Bakteriol.* 289:285–299.
16. Fukushima, H., K. Hoshina, and M. Gomyoda. 2000. Selective isolation of *eae*-positive strains of Shiga toxin-producing *Escherichia coli*. *J. Clin. Microbiol.* 38:1684–1687.
17. Griffin, P. M., and R. V. Tauxe. 1991. The epidemiology of infections caused by *Escherichia coli* O157:H7, other enterohemorrhagic *E. coli*, and the associated hemolytic uremic syndrome. *Epidemiol. Rev.* 13:60–98.
18. Hilborn, E. D., J. H. Mermin, P. A. Mshar, J. L. Hadler, A. Voetsch, C. Wojtkunski, M. Swartz, R. Mshar, M. A. Lambert-Fair, J. A. Farrar, M. K. Glynn, and L. Slutsker. 1999. A multistate outbreak of *Escherichia coli* O157:H7 infections associated with consumption of mesclun lettuce. *Arch. Intern. Med.* 159:1758–1764.
19. Johnson, K. E., C. M. Thorpe, and C. L. Sears. 2006. The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*. *Clin. Infect. Dis.* 43:1587–1595.
20. Karmali, M. A., M. Petric, C. Lim, P. C. Fleming, G. S. Arbus, and H. Lior. 1985. The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*. *J. Infect. Dis.* 151:775–782.
21. King, L. A., A. Mailles, P. Mariani-Kurkdjian, C. Vernozy-Rozand, M. P. Montet, F. Grimont, N. Pihier, H. Devalk, F. Perret, E. Bingen, E. Espié, and V. Vaillant. 2009. Community-wide outbreak of *Escherichia coli* O157:H7 associated with consumption of frozen beef burgers. *Epidemiol. Infect.* 137:889–896.
22. Kobayashi, H., J. Shimada, M. Nakazawa, T. Morozumi, T. Pohjanvirta, S. Pelkonen, and K. Yamamoto. 2001. Prevalence and characteristics of Shiga toxin-producing *Escherichia coli* from healthy cattle in Japan. *Appl. Environ. Microbiol.* 67:484–489.
23. Locking, M. E., S. J. O'Brien, W. J. Reilly, E. M. Wright, D. M. Campbell, J. E. Coia, L. M. Browning, and C. N. Ramsay. 2001. Risk factors for sporadic cases of *Escherichia coli* O157 infection: the importance of contact with animal excreta. *Epidemiol. Infect.* 127:215–220.
24. Mayer, C. L., C. S. Leibowitz, S. Kurosawa, and D. J. Stearns-Kurosawa. 2012. Shiga toxins and the pathophysiology of hemolytic uremic syndrome in humans and animals. *Toxins* 4:1261–1287.
25. Monaghan, A., B. Byrne, S. Fanning, T. Sweeney, D. McDowell, and D. J. Bolton. 2011. Serotypes and virulence profiles of non-O157 Shiga toxin-producing *Escherichia coli* isolates from bovine farms. *Appl. Environ. Microbiol.* 77:8662–8668.
26. Monaghan, A., B. Byrne, S. Fanning, T. Sweeney, D. McDowell, and D. J. Bolton. 2012. Serotypes and virulotypes of non-O157