An untypeable enterotoxigenic *Escherichia coli* represents one of the dominant types causing human disease

Atsushi Iguchi, Astrid von Mentzer, Taisei Kikuchi and Nicholas R. Thomson

**Abstract**

Enterotoxigenic *Escherichia coli* (ETEC) is a major cause of diarrhoea in children below 5 years of age in endemic areas, and is a primary cause of diarrhoea in travellers visiting developing countries. Epidemiological analysis of *E. coli* pathovars is traditionally carried out based on the results of serotyping. However, genomic analysis of a global ETEC collection of 362 isolates taken from patients revealed nine novel O-antigen biosynthesis gene clusters that were previously unrecognized, and have collectively been called unclassified. When put in the context of all isolates sequenced, one of the novel O-genotypes, OgN5, was found to be the second most common ETEC O-genotype causing disease, after O4, in a globally representative ETEC collection. It’s also clear that ETEC OgN5 isolates have spread globally. These novel O-genotypes have now been included in our comprehensive O-genotyping scheme, and can be detected using a PCR-based and an *in silico* typing method. This will assist in epidemiological studies, as well as in ETEC vaccine development.

**DATA SUMMARY**

1. Assembled draft genomes for 55 enterotoxigenic *Escherichia coli* isolates were deposited in GenBank/EMBL/DDBJ under the accession numbers shown in Table S1 (available with the online Supplementary Material).

2. Sequences of nine annotated O-antigen biosynthesis gene clusters and three *wzy* genes from positive control strains (OT-37 for OgN3, EHOUT43 for OgN5 and EH-OSB16 for OSB16) were deposited in GenBank/EMBL/DDBJ under the accession numbers LC177546–LC177554 and LC223608–LC223610, respectively.

**INTRODUCTION**

The first diarrhoeal illness that infants often experience in endemic areas is caused by enterotoxigenic *Escherichia coli* (ETEC) [1]. In 2010, annual mortality from illness due to ETEC was estimated at 157,000 deaths (9% of all deaths attributed to diarrhoea) and approximately 1% of all deaths in children 28 days to 5 years of age [2]. Additionally, ETEC is a primary cause of diarrhoea in travellers visiting developing countries. In our previous study, we sequenced 362 globally representative ETEC isolates collected between 1980 and 2011 from 20 countries, including isolates from children and adults in endemic areas, as well as from travellers visiting such areas [3]. The majority of the isolates were collected from patients with diarrhoea. Genome-wide analysis showed that, contrary to previous understanding, there are long-term stable associations of ETEC lineages with specific virulence factors, such as plasmid-encoded heat-labile toxin (LT) and/or heat-stable toxin (ST; including two subtypes, STh and STp), and colonization factors (CFs), and that these lineages are globally distributed.

O-serogrouping remains the gold standard for the subtyping of *E. coli* isolates, especially pathogenic *E. coli*, for taxonomical and epidemiological studies. Most of what we know about *E. coli* prevalence, outbreaks and surveillance is described in terms of the O-serogroup. Recently, sequence analyses show that phenotypic O-serogroup diversification can be correlated with differences in the gene content and genetic diversity of the O-antigen biosynthesis gene cluster (O-AGC) located on the chromosome [4]. In particular, sequences from O-antigen processing genes, such as *wzx* (encoding the O-antigen flippase), *wzy* (encoding the O-antigen polymerase), and the *wzm* and *wzt* genes (encoding components of the ABC transporter pathway) located on the O-AGCs are highly variable in sequence, and can be used as gene markers for the identification of O-serogroups via molecular approaches [4]. By applying an *in silico* BLAST analysis using a *wzx/wzy* and *wzm/wzt* sequence set...
extracted from O1 to O187 O-AGCs, we subtyped our 362 global ETEC isolates [3] into 48 O-genotypes, of which the top 6 were O6 (n=38), O25 (n=24), O27 (n=18), O114 (n=17), O115 and O159 (n=16, each). In addition to the ETEC isolates classified into 48 known O-genotypes, 55 isolates carried wzx/wzy genes that showed <50% or no sequence identity to the previously defined sequences. These ETEC isolates were categorized as O-genotype untypeable (OgUT), but carried novel O-AGCs.

In this study, we focused on characterizing the OgUT ETEC isolates detected in our previous study [3]. Such ETEC isolates may not be recognized by public-health surveillance, because they cannot be assigned to any known O-genotype and/or O-serogroup. To determine the relative contribution of OgUT ETEC isolates, we defined the novel O-genotypes of these ETEC isolates and estimated their overall contribution to disease by screening our global collection of ETEC.

**METHODS**

**Source sequences**

ETEC O-AGC sequences were obtained from draft genomes used in a previous study [3]. The wzx and wzy sequences were extracted from each O-AGC. The known wzx/wzy sequences from typical *E. coli* O-serogroups (171 groups, O1–O187 except for wzm/wzt-type groups) [4], *Shigella* O-serogroups (total 34 types; 13 from *Shigella dysenteriae*, 18 from *Shigella boydii*, 2 from *Shigella flexnerii* and a single type of *Shigella sonnei*) [5], *E. coli* OX serogroups (11 groups: OX6, OX9, OX10, OX13, OX18, OX19, OX21, OX25, OX28, OX38, OX43) [6] and 6 recently defined novel O genotypes from Shiga toxin-producing *E. coli* (OgN1, OgN8, OgN9, OgN10, OgN12, OgN31) [7] were used (except for wzy genes of OX25 and *S. dysenteriae* type 6, which are not found in the O-AGCs). A complete set of *flIC* sequences [8] was also used for molecular-based H-typing.

**Sequence comparisons**

Phylogenetic trees of wzx and wzy were constructed by using the neighbour-joining algorithm using MEGA5 software [9], following multiple alignments of nucleotide sequences by the CLUSTAL W program [10].

**PCR**

OgN-specific PCR primers targeting OgN3, OgN5 and OSB16 were designed based on each alignment of wzy sequences. PCR was performed as follows: the 30 µl reaction mixture contained 2 µl genomic DNA, 6 µl 5× Kapa Taq buffer, dNTP mix (final concentration 0.3 mM each), MgCl₂ (final concentration 2.5 mM), primers (final concentration 0.5 µM each) and 0.8 Unit Kapa Taq DNA polymerase (Kapa Biosystems). The thermocycling conditions were: 25 cycles of 94 °C for 30 s, 58 °C for 30 s and 72 °C for 1 min. The PCR products were visualized following agarose gel (1.5%) electrophoresis in 0.5x TBE (25 mM Tris borate, 0.5 mM EDTA) and staining with ethidium bromide (1 mg/ml). Three strains (OT-37 for OgN3, EHOUT43 for OgN5 and EH-OSB16 for OSB16) were used as positive-control strains for PCR.

**RESULTS**

Comparative analyses revealed that chromosomal O-AGC sequences flanked by gene clusters of the colanic acid biosynthesis (*wca*) (upstream) and the histidine biosynthesis (*his*) (downstream) extracted from OgUT ETEC genomes contained nine novel O-AGCs, OgN2, OgN3, OgN4, OgN5, OgN13, OgN14, OgN15, OgN16 and OgN17 (Fig. 1), which all carried unique wzx/wzy genes, and showed <70% nucleotide sequence identities among those from known *E. coli* O-AGCs (Fig. 2). One of the O-AGCs was similar (the same gene construction and >97% sequence identity) to that of *S. boydii* type 16 (OSB16) (Fig. 2). Among the novel O-AGC sequences identified in our ETEC collection, it was apparent that the novel OgN5 was carried by more than half of ETEC OgUT isolates (29/55), followed in frequency by OgN3 and OgN13, OgN14, OgN15, OgN16 and OgN17 (Fig. 1), which all carried unique wzx/wzy genes, and showed <70% nucleotide sequence identities among those from known *E. coli* O-AGCs. OgN5 strains were isolated in five countries, Guatemala (n=12), Argentina (n=5), Egypt (n=4), Mexico (n=4) and Indonesia (n=4), between 1989 and 2003. All ETEC OgN5 isolates carried the ST-encoding gene, except for a single isolate (E1542) that carried the LT-encoding gene, and all OgN5 isolates were confirmed negative for all known CFs by dot-blot and PCR-based analyses in our previous study [3]. The *flIC*-based H-typing showed that the OgN5 isolates were classified into eight H-types: H5, H9, H10, H16, H18, H19, H32 and H39 (see Table S1). Phylogenetic analysis showed OgN5 isolates were spread across phylogenetically distinct lineages where most/some isolates grouped based on their H-types (see Fig. S1, Table S1). The major groups were OgN5: H10 and OgN5: H16 (n=6, each) belonging to a phylogenetic lineage L14 [3] of *E. coli* phylogroup A and a lineage L22 (designed in this study) of phylogroup B, respectively (see Fig. S1). Additionally, OgN5: H18 (L11, n=1), OgN5: H9 (L14, n=1), OgN5: H32 (L15, n=4), OgN5: H19 (L15, n=1), OgN5: H34 (L16, n=4), OgN5: H16 (L23, n=3) and OgN5: H5 (L24, n=3)
were observed. No significant association between lineages and geographical origins was observed (see Table S1).

Six of eight OgN3 strains carried the fliC of H45, and were confirmed positive for STh and two CFs, CFA/I (a fimbria) and CS21 (a type IV pilus) (see Table S1). All OgN3 : H45 isolates were phylogenetically grouped into L6 of phylogroup A, and originated from Central and South America, including Mexico, Guatemala and Argentina (see Fig. S1, Table S1). Six of eight OSB16 strains carried the fliC of H32, and were confirmed positive for LT, LT+STh or LT+STp, and negative for all known CFs (see Table S1). All OSB16 : H32 and OSB16 : H2 isolates were phylogenetically grouped into L13 of phylogroup A, and originated from Guatemala and Argentina (see Fig. S1, Table S1). Additionally, three OgN13, two OgN15 and single OgN2, OgN4, OgN14, OgN16 and OgN17 strains were observed in our ETEC collection (see Table S1).

To aid in the identification of the predominant novel ETEC O-AGCs, we designed specific PCR primers for identifying OgN5, OgN3 and OSB16 targeting unique sequences on wzy genes (Table 1, Fig. 4), and their specificities were confirmed by using positive-control strains and all O-serogroup reference strains (O1–O188).

DISCUSSION

The results of the current study encapsulate problems with using a limited number of phenotypic markers for typing and tracking pathogenic bacteria of importance to human or animal health. It also highlights how, with whole-genome sequencing, there are still major discoveries to be made in identifying the isolates/lineages or types that are responsible for a significant burden of reported disease and yet untypeable by traditional methods. Since the O-serogroup is still used as a general marker for the subtyping of E. coli isolates, epidemiologically there can be an unawareness of the presence, emergence or spread of O-serogroup untypeable pathogenic E. coli groups. In this study, we used extracted O-AGC sequences from 362 ETEC genomes and by comparative genomics revealed nine novel O-genotypes that were previously unrecognized by the standard methods used for serotype detection and surveillance.

In recent surveillance studies, O6 has been shown to be the most common ETEC O-serogroup (mostly O6 : H16 serotype) associated with diarrhoeal patients in Egypt [11], Bolivia [12], Bangladesh [13], China [14] and Japan [15]. The O6-serogroup is mainly found in isolates with the CF profile CS1+CS2 and CS1+CS3, which are three of the most prevalent CFs identified in clinical isolates [16]. In our ETEC collection, the most prevalent O-serogroup was O6. O25 is also a well-known ETEC O-serogroup frequently isolated from ETEC patients, such as in Bosnia and Herzegovina [17], Bangladesh [13], China [14] and Japan [15]. Interestingly, in our representative ETEC collection, the novel O-genotype OgN5 was the second most common ETEC after O6 and before O25. Moreover, ETEC...
OgN5 isolates included those collected from children and adults in endemic regions such as Argentina, Egypt, Guatemala and Indonesia, as well as from travellers to Guatemala and Mexico, between 1989 and 2003. These data indicate that OgN5-positive ETEC isolates have spread both globally and over time.
Novel O-genotype ETEC isolates were found in several distinct phylogenetic lineages, and sequences of \textit{wzx}/\textit{wzy} from O-AGCs were highly conserved within each O-genotype, indicating that these O-AGCs have been spread across this species by horizontal gene transfer. Phylogenetic analysis in our previous study [3] showed that most ETEC isolates clustered based on their O-serogroups, such as O6 isolates belonging to L1+L2 (37/38), O25 belonging to L4 (22/24), O27 belonging to L9+L17 (18/18), O114 belonging to L4+L10 (14/17) and O115 belonging to L5 (16/16). In this study, ETEC OgN3 and OSB16 were mostly clustered in L16 and L13, respectively. However, OgN5 isolates were found in several distinct lineages, suggesting frequent exchanges of the O-AGC occur between lineages. Horizontal gene transfer-promoting sequences, such as transposase genes and repeat sequences, were not found in and around the OgN5 O-AGC.

Recently, we have been able to access whole-genome sequencing data due to the appearance of next-generation sequence technologies. A publicly available \textit{in silico}-based serotyping web tool, SerotypeFinder, of the Center for Genomic Epidemiology (https://cge.cbs.dtu.dk/services/SerotypeFinder) [8], which has a database consisting of \textit{E. coli} O-antigen associated genes (\textit{wzx}, \textit{wzy}, \textit{wzm} and \textit{wzt}) from O1 to O187 and H-antigen associated genes (\textit{fliC} and its homologues) from H1 to H56 can be used to estimate \textit{E. coli} O : H serotypes of sequenced isolates. If it is possible to organize and update the database with new \textit{wzx} and \textit{wzy} sequences, it becomes possible to meet the requirements for broader O-genotypes, including OgN5, OgN3 and OSB16 in the \textit{in silico} typing.

In conclusion, by using detailed whole genome sequence-based analyses, it is clear that there is a hidden diversity of ETEC O-genotypes for which there is no prior information regarding their contribution to the burden of disease. It should be noted that even from our global collection, ETEC isolates from Africa are underrepresented. Hence, it is possible and perhaps likely that additional novel ETEC O-types

\[
\begin{array}{|c|c|c|c|}
\hline
\text{O genotype} & \text{Primer name} & \text{Sequence (5'→3')} & \text{Target gene} & \text{Size (bp)} \\
\hline
\text{OgN3} & \text{OgN3_PCR_F} & GCTTGGCATCGTGGGGATA & \text{wzy} & 189 \\
& \text{OgN3_PCR_R} & TGCTACAACTAGCCCGCTA & & \\
\text{OgN5} & \text{OgN5_PCR_F} & GGTAGTAGGAAGCGCTATCG & \text{wzy} & 650 \\
& \text{OgN5_PCR_R} & CCAATTCCAGCGATGGAG & & \\
\text{OSB16} & \text{OgSB16_PCR_F} & AACGGGCTGGAAACTGCA & \text{wzy} & 717 \\
& \text{OgSB16_PCR_R} & ATTTCCACATCAATCCAGGA & & \\
\hline
\end{array}
\]
and their associated lineages will continue to be discovered. These novel O-genotypes identified in this study have now been included in our comprehensive O-genotyping scheme and can be detected using the PCR-based method described in this study and by the in silico typing method.

Funding information
This research was partially supported by Health Labor Sciences Research Grants from the Ministry of Health, Labor, and Welfare, Japan (H27-Syokuhin-Wakate-018), and by the Research Program on Emerging and Reemerging Infectious Diseases from the Japan Agency for Medical Research and Development (AMED) (15fk0108008h0001).

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data bibliography


References


Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as ‘excellent’ or ‘very good’.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.