Structure of the *Shigella dysenteriae* 7 O antigen gene cluster and identification of its antigen specific genes

Lu Feng\(^{a,b}\), Jiang Tao\(^{a}\), Hongjie Guo\(^{a,b}\), Jianguo Xu\(^{c}\), Yayue Li\(^{a}\), Fedousi Rezwan\(^{d}\), Peter Reeves\(^{d}\), Lei Wang\(^{a,b,*}\)

\(^{a}\)College of Life Sciences, Nankai University, Tianjin 300071, China
\(^{b}\)Center for Functional Genomics Research, TEDA College, Nankai University, TEDA, Tianjin 300457, China
\(^{c}\)National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China
\(^{d}\)School of Molecular and Microbial Biosciences (G08), University of Sydney, Sydney, NSW 2006, Australia

Received 13 May 2003; received in revised form 13 October 2003; accepted 15 October 2003

**Abstract**

*Shigella* strains are human pathogens. The O antigen gene cluster of *Shigella dysenteriae* O7 was sequenced and analyzed. It contains genes for synthesis of nucleotide sugars including UDP-2-acetamido-2-deoxy-D-galacturonic acid, UDP-2-acetamido-2-deoxy-D-galacturonic acid and dTDP-4-amino-4,6-dideoxy-D-glucose. Also found in the gene cluster are genes encoding O unit flippase, O antigen polymerase and sugar transferases. The *Escherichia coli* O121 O antigen, which is present in an important Shiga toxin-producing strain, has the same structure as that of *S. dysenteriae* O7, and we found that the gene clusters also had the same genes and organization. Four genes specific to *S. dysenteriae* O7 and *E. coli* O121 were identified by PCR screening against representatives of 186 *E. coli* (including *Shigella*) O serotypes. *E. coli* O121 and *S. dysenteriae* O7 isolates can be distinguished by PCR of the H antigen *fliC* gene.

© 2003 Elsevier Ltd. All rights reserved.

**Keywords:** Molecular typing; *Shigella dysenteriae* O7

1. Introduction

*Shigella* strains are important human pathogens, and a major cause of diarrhea. *Shigella* strains are in reality clones of *Escherichia coli* based on sequence similarity between their housekeeping genes [1]. *E. coli* clones are normally classified by a combination of their O and H (and sometimes K) antigens, but for *Shigella* clones only the O antigen classification system is used, as being non-motive they lack H antigen. O antigen contributes major antigenic variability to the cell surface. At least 186 O antigen forms have been recognized in *E. coli* including *Shigella* forms. Because

*E. coli* and *Shigella* species have been treated as different species, each has its own antigenic typing scheme with its own series of numbered O antigens. There are 166 in the *E. coli* scheme [2] and 33 found in *Shigella* strains after allowing for some overlap between different *Shigella* species and the fact that most *S. flexneri* forms have variants of a single O antigen structure. After also allowing for overlap between *E. coli* and *Shigella* (Table 1), there are currently 186 distinct O antigen structures in *E. coli* and *Shigella* combined.

The O antigen, which consists of many repeats of an oligosaccharide unit (O unit), is part of the lipopolysaccharide (LPS) in the outer membrane of Gram-negative bacteria. O antigens are present in both pathogenic and non-pathogenic strains of *E. coli*, but some are most commonly found in pathogenic clones [3]. The current model of biosynthesis suggested that the O unit was synthesized by sequential transfer of one sugar phosphate and then one or more sugars from respective sugar nucleotides to the carrier lipid, undecaprenyl phosphate (UndP). O units are then polymerized on UndPP into polysaccharide chains, which

**Abbreviations:** GalNAcA, 2-acetamido-2-deoxy-D-galacturonic acid; GalNAcA(N), 2-acetamido-2-deoxy-D-galacturonamide; Qui4N, 4-amino-4,6-dideoxy-D-glucose; GlcNAc, 2-acetamido-2-deoxy-D-glucose; PCR, polymerase chain reaction; UndP, undecaprenol phosphate; ml, microlitre; LPS, lipopolysaccharide; STEC, Shiga toxin-producing *Escherichia coli*.

* Corresponding author. Address: College of Life Sciences, Nankai University, Tianjin 300071, China. Tel.: +86-22-23503151; fax: +86-22-23506281.

E-mail address: wanglei@nankai.edu.cn (L. Wang).
are transferred to the independently synthesized core/lipid A to form LPS [4].

Genes for O antigen synthesis are normally located in a gene cluster, which maps between galF and gnd in the genome of E. coli. A number of E. coli O antigen gene clusters have been cloned and characterized [5]. Genes involved in O antigen synthesis are classified into 3 main classes: (1) nucleotide sugar synthesis genes; (2) genes encoding sugar transferases; (3) O unit processing genes encoding flippase (wzx) and polymerase (wzy) [4]. The genes in classes 2 and 3 are often specific to individual O antigen gene clusters and have potential for use in PCR diagnosis for rapid detection of relevant strains.

The Shigella dysenteriae O7 (D7) O antigen contains four different sugars: 2-acetamido-2-deoxy-D-glucose (D-GlcNAc), 2-acetamido-2-deoxy-D-galacturonic acid (D-GalNAcA), 2-acetamido-2-deoxy-D-galacturonamide (D-GalNAcA(N)) and 4-amino-4,6-dideoxy-D-glucose (D-Qui4N) [6] (Fig. 1). Qui4N is a dideoxy-aminohexose rarely found in O antigen structures, but it is known that E. coli O7 antigen carries Qui4NAc [7]. GalNAcA(N), a rare acidic sugar, is also present in the Pseudomonas aeruginosa O6 O antigen [7]. Acidic sugars are commonly found in Shigella O antigens, while most other E. coli O antigens are composed only of neutral sugars [7]. E. coli strains having acidic sugars in their O antigens often cause dysentery-like diseases and their O antigens are often closely related to those of Shigella strains [8].

In this study, the D7 O antigen gene cluster was sequenced and analyzed. Putative genes involved in synthesis of D-GalNAcA, D-GalNAcA(N) and D-Qui4N as well as genes for O unit processing and polymerizing were identified. The genes and organization of the E. coli O121 O antigen gene cluster were found to be the same as for D7, and four genes specific for these O antigens were identified. We also sequenced the flagellin gene region of D7 and found that it has a mutated fliC gene. Comparison of the fliC gene of D7 and those of E. coli strains revealed specific DNA fragments, which can be used to discriminate between D7 and E. coli O121 strains.

### 2. Results and discussion

#### 2.1. Sequencing

The O antigen gene clusters of E. coli strains are commonly located between galF and gnd [5]. Long-range PCR primers were designed based on the sequences of galF and gnd, respectively, and used to amplify the D7 O antigen gene cluster. In order to minimize PCR errors, five individual PCR products were combined and used for the construction of a DNaseI shot gun library. A sequence of 17433 bases, including parts of galF (positions 1-738) and gnd (positions 16123–17433), was obtained. The D7 O antigen gene cluster (positions 1160–15973) contained 14 open reading frames (orfs), all with the same transcriptional direction from galF to gnd (Fig. 2).

#### 2.2. D7 O antigen genes

Orf1 and Orf2 shared 70 and 67% identity with WbpO and WbpP of P. aeruginosa O6, respectively. WbpP is an UDP-GlcNAc C4 epimerase, which converts UDP-GlcNAc

<table>
<thead>
<tr>
<th>Table 1 O antigens identical in Shigella and E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shigella O antigen</strong></td>
</tr>
<tr>
<td>Dysenteriae 2 and Boydii 15</td>
</tr>
<tr>
<td>Dysenteriae 3*</td>
</tr>
<tr>
<td>Dysenteriae 5*</td>
</tr>
<tr>
<td>Dysenteriae 7*</td>
</tr>
<tr>
<td>Dysenteriae 12*</td>
</tr>
<tr>
<td>Flexneri 1-5**</td>
</tr>
<tr>
<td>Boydii 1*</td>
</tr>
<tr>
<td>Boydii 2*</td>
</tr>
<tr>
<td>Boydii 4*</td>
</tr>
<tr>
<td>Boydii 5*</td>
</tr>
<tr>
<td>Boydii 8*</td>
</tr>
<tr>
<td>Boydii 11*</td>
</tr>
<tr>
<td>Boydii 14*</td>
</tr>
</tbody>
</table>

a Evidence for the E. coli equivalents was summarized by Ewing [39].
b S. dysenteriae O7 and E. coli O121 shown to have the same O antigen by Parolis et al. [9].
c S. flexneri 1–5 have the same basic structure, with phage encoded glucosyl and/or acetyl substitutions, and are treated as one O antigen as is done for other E. coli and S. enterica [7,39].

Fig. 1. Structure of the S. dysenteriae O7 O antigen [6].
to UDP-GalNAc [14]. Our group has previously used the name gne for UDP-GlcNAc C4 epimerase genes and as Orf2 is homologous with Gne of E. coli O113 [15] throughout its length, we designate Orf2 as gne. WbpO is responsible for converting UDP-GalNAc to UDP-GalNAcA and was complemented by its homologue wcdA in the S. enterica sv Typhi Vi gene [16]. Since GalNAcA is also present in S. dysenteriae O7 [6], the same role can be assigned to Orf1.

Putative homologues for conversion of UDP-GalNAc to UDP-GalNAcA have now been found in the gene clusters for Vi capsule, P. aeruginosa O6, Dysenteriae O7 and Staphylococcus aureus type 1 capsule [17]. It seems appropriate to give the genes a common name and propose the name gna (GalNAcA) for these genes.

Orf12 shared 55% identity with WbpS in the P. aeruginosa O6 gene cluster which was sequenced by Belanger et al. [18] and for which no role could be assigned. It was proposed that GalNAcA(N) was produced by addition of an amino group to GalNAcA but Belanger et al. were unable to find the gene responsible. We find that both Orf12 and WbpS belong to the AsnB (Asparagine synthase) family (COG2244, E value \(= 5 \times 10^{-35}\)). AsnB catalyses the conversion of aspartate to asparagine by amination. Orf12 and WbpS were also similar to members of the pfam00733 amidotransferase family (E value \(= 7 \times 10^{-46}\) and \(6 \times 10^{-50}\), respectively). Therefore, we propose that Orf12 and WbpS are the amidotransferases responsible for addition of an amino group to GalNAcA but Belanger et al. were unable to find the gene responsible. We find that both Orf12 and WbpS belong to the AsnB (Asparagine synthase) family (COG0367, E value \(= 5 \times 10^{-12}\)) of enzymes that transfer acetyl groups.

The three genes identified as responsible for converting Glucose-1-P to dTDP-Qui4NAc (Fig. 3). Orf10 belongs to a Serine O-acetyl transferase family (COG1045, E value \(= 3 \times 10^{-25}\)) of enzymes that transfer O-acetyl groups. Orf7 was similar to members of an acetyltransferase family (pfam00583, E value \(= 5 \times 10^{-12}\)) of enzymes that transfer acetyl groups. The main chain of the repeating unit of D7 is modified by the attachment of an O-acetyl group to GalNAcA(N) and an N-acetylglucosamine group to Qui4N (Fig. 1). Orf7 and Orf10 are probably involved in the addition of N-acetyl and O-acetyl groups, respectively, to the main chain of the O unit. We named orf7 as wbnH, and named orf10 as wbnJ.

The only two orfs encoding predicted membrane proteins are orf8 and orf11. Orf8 had 12 predicted transmembrane segments, and belonged to the RfbX (wzx) family (COG2244, E value \(= 4 \times 10^{-8}\)). Members of this family are membrane proteins involved in the export of O antigen and teichoic acid. Orf8 was grouped with Wzx of E. coli O55 and Shigella flexneri 2a, which showed highest identity with Orf8. This indicates that orf8 is the expected O unit flippase gene (wzx) and it was named accordingly. The other predicted membrane protein Orf11 has nine predicted transmembrane segments and a large periplasmic loop. Orf11 was grouped with Wzy proteins of E. coli O7 and Yersinia pseudotuberculosis O1b, which showed highest identity with Orf8. This indicates that orf11 is an O unit polymerase gene (wzy) and it was named accordingly.

Orf9 belongs to the Glycos_trans_2 family (pfam00535, E value \(= 1.7 \times 10^{-31}\)). Members of this family transfer nucleotide sugars such as UDP-glucose to a range of substrates. Orf13 and Orf14 belong to Glycos_transf_1 family (pfam00534, E value \(= 2 \times 10^{-25}\) and \(1.2 \times 10^{-37}\), respectively) and shared 44 and 52% identity with WbpT and WbpU of P. aeruginosa O6, respectively. Both WbpT

Fig. 2. The O antigen gene cluster of S. dysenteriae O7.

Fig. 3. Proposed pathway for the biosynthesis of the Qui4N and GalNAcA(N) in S. dysenteriae O7. The synthesis of UDP-Qui4NAc from UDP-GlcNAc via UDP-Qui4N was proposed for E. coli O7 [19]. The biosynthetic pathway for GalNAcA(N) was biochemically verified by Creuzenet [14] and Zhao [16].
and WbpU were predicted galactosyltransferases in P. aeruginosa O6 [18]. We propose that orfs 9, 13 and 14 encode three transferases, and name them as wbnL, wbnK and wbnL, respectively. The wecA gene, which is outside the O antigen gene cluster and encodes the enzyme for transferring GlcNAc-P as the first sugar in synthesis of the Enterobacterial common antigen repeat unit, has been shown to carry out the same function for several GlcNAc-containing O antigen gene clusters [20]. Therefore, the predicted three transferases in D7 gene cluster are assumed to transfer GalNAcA, GalNAcA(N) and Qui4N, while WecA acts as the GlcNAc-P transferase. However, the exact roles of the three sugar transferases need to be further investigated.

No function can be assigned to orf6 by searching currently available databases. We expect a gene for a glycine transferase, and it is probable that orf6 is this gene. We name orf6 wbnG.

2.3. The E. coli O121 and D7 gene clusters have the same genes and organization

It has been shown that the O antigen of E. coli O121 has the same structure as that of D7 [9]. To determine if the two strains have the same O antigen gene cluster, PCR was carried out with the E. coli O121 type strain. Pairs of primers were designed, based on the sequence of the D7 O antigen gene cluster, for each pair of adjacent genes, and used to amplify both D7 and E. coli O121. The PCR products from each primer pair were the same size for both strains (data not shown). This shows that O antigen gene clusters of E. coli O121 and D7 have the same genes in the same order.

2.4. The flagellin gene region of D7

Shigella strains are non-motile and do not produce H antigen [21]. Flagellin is the protein subunit of the flagellum that carries H antigen specificity. In E. coli, flagellin proteins of most H antigens are encoded by the fliC gene [22]. Four operons containing more than 40 genes are needed for flagellar assembly in E. coli, with fliC in one of the operons [23]. Studies of some Shigella strains have shown that they have cryptic fliC genes and the loss of flagellation is due to defects in other genes of the four operons [24,25]. We amplified the D7 fliC gene and sequenced the PCR product. Analysis of the sequence showed that the gene is 1714 bp in length and shares 98% identity to the fliC gene for the E. coli H45 antigen [22]. However the fliC of D7 has a 7 bp insertion between positions 1110 and 1111 of the fliC of E. coli H45, making the D7 fliC transcript end prematurely at position 1131. Thus, it is highly likely that the truncated fliC gene is the main reason for the loss of flagellin in D7 strains.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Base positions of genes</th>
<th>Base positions of forward/reverse primers</th>
<th>Annealing temp(°C) of PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>wbnl</td>
<td>9200–10090</td>
<td>9336–9533/10017–10034</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>9274–9292/9746–9763</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>wcys</td>
<td>10646–11836</td>
<td>11120–1137/11366–11383</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>10957–10976/11547–11564</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10707–10725/11767–11785</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>wbnK</td>
<td>13691–14848</td>
<td>14084–14101/14420–14437</td>
<td>45</td>
</tr>
<tr>
<td>wbnL</td>
<td>14849–15973</td>
<td>15225–15244/15765–15782</td>
<td>58</td>
</tr>
</tbody>
</table>

2.5. Identification of D7 and E. coli O121 specific genes

Pairs of primers based on wzy and three transferase genes (wbnL, wbnK and wbnL) were used to screen representatives of 186 known O antigens of typical E. coli and Shigella strains by PCR. Except for D7 and E. coli O121, which gave the expected PCR products, none of the other E. coli or Shigella strains gave PCR products (Table 2). Therefore, all four genes are specific for D7 and E. coli O121.

E. coli O121 strains of H antigen 7, 10 and 19 have been reported as pathogens [10–13], and serotype O121:H19 is being isolated more frequently from clinical specimens [11]. fliC genes of E. coli H antigens 7, 10, 19 and 45 have been sequenced, and specific regions of each gene have been identified [22]. As described above, the D7 fliC gene is similar to the E. coli H45 fliC gene, and the main difference between the two genes is a 7 bp insertion in D7. Sequence of 18 bp with the 7 bp insertion sequence at 3’ end (CAGTGCAGAGT) share less than 77.8% similarity with sequences of any fliC genes of E. coli. Thus, primers binding to this region paired with one of the two E. coli H45 specific primers (5’-AGC GAT GCA CTG TCT CGC-3’ and 5’- GGA CAA TTT CGC TGA TCT TGC-3’, www.mmb.usyd.edu.au/archives/) can be used to differentiate D7 and E. coli O121 strains. The two E. coli H45 specific primers are conserved in the D7 fliC sequence.

2.6. General conclusions

In this study, the O antigen gene cluster of D7 was sequenced and characterized. Analyzing the sequence revealed 14 putative genes (Table 3): gna and gne for the synthesis of GalNAcA; wbpS for the synthesis of GalNAcA(N) from GalNAcA; rmlB, rmlA and vioA for the synthesis of Qui4N; wbnL, wbnK and wbnL for transferring sugars; wbnH and wbnI for transferring acetyl groups; a flippase gene (wzx) and an O unit polymerase gene (wzy); and wbnG, a putative glycine transferase.

The O antigen of E. coli O121 has the same structure as S. dysenteriae O7, and we found that E. coli O121 has a gene cluster of the same genes in the same order as S. dysenteriae O7. A similar situation applied to two S. boydii O antigens that have the same structures as
typical E. coli serotypes [26], and it seems that this may apply generally. Four O-antigen genes specific for D7 and E. coli O121 were identified, and primers based on those genes may be useful for screening of D7 and E. coli O121 in a PCR-based method. D7 strains carry a mutated flic gene similar to that of the E. coli flic gene specific to the D7 mutated strain. The gene can be used to discriminate between D7 and E. coli O121 in the same PCR-based test.

3. Materials and methods

3.1. Bacterial strains and plasmids

All plasmids used in this study were maintained in E. coli K-12 strain DH5a, which was purchased from Beijing Dingguo Biotechnology Development Center (Beijing, P. R. China). The S. dysenteriae O7 and E. coli O121 type strains were obtained from the Institute of Medical and Veterinary Science, Adelaide, Australia. Other Shigella serotypes were obtained from the Institute of Epidemiology and Microbiology, Chinese Academy of Preventive Medicine, Beijing, P. R. China and other E. coli serotypes were as previously described [27].

3.2. Construction of DNasel shot gun bank

Chromosomal DNA was prepared as previously described [28]. Primers *1523 (5'-ATT GTG CTT GCA GGG ATC AAA GAA ATC-3') and *1524 (5'-TAG TCG CGC TGN GCC TGG ATG TTC GC-3'), which were designed based on the galF and gnd genes, respectively, were used in long-range PCR to amplify the O-antigen gene cluster using the Expand Long Template PCR System from Roche. The PCR cycles used were as follows: denaturation at 94 °C for 10 s, annealing at 61 °C for 30 s and extension at 72 °C for 15 min. The PCR products were digested by DNasel and the resulting DNA fragments were cloned into pGEM-T easy to produce a bank using the method described previously [29].

3.3. Sequencing and analysis

The flic gene was amplified using primers *1575 (5'-GGG TGG AAA CCC AAT ACG-3') and *1576 (5'-GCG CAT CAG GCA ATT TGG GC-3'), as described previously [22]. The PCR product was sequenced by the Sydney University and Prince Alfred Hospital Molecular Analysis Center using an ABI 3700 DNA sequencer, and the sequence assembled and edited using programs phred, phrap and consed [30].

Sequencing of the O-antigen gene cluster was carried out by Shanghai Sangon Biological Engineering Technology and Service Co. Ltd. (Shanghai, P. R. China) using an ABI PRISM 377-96 automated DNA sequencer. Sequence data were assembled using the Staden package [31]. The program Artemis [32] was used to do annotation. The program Blockmaker [33] was used for searching conserved motifs. Blast and PSI-blast [34] were used for searching databases.
including GenBank, COG and Pfam protein motif databases [35,36]. The algorithm described by Eisenberg [37] was used to identify potential transmembrane segments. Sequence alignment and comparison were done using the program ClustalW [38].

3.4. Specificity assay by PCR

Chromosomal DNA prepared from 186 strains to represent the O antigens of E. coli in the broad sense, was examined to determine its quality by PCR amplification of the mdh gene (coding for malate dehydrogenase and present as a house-keeping gene in E. coli) using primers described previously [29]. A total of 28 pools of DNA were made, each containing DNA from 6 to 10 strains. Pools were screened using primers based on specific genes of S. dysenteriae O7. PCR was carried out in a total volume of 25 μL, of which 10 μL was run on an agarose gel to check for amplified DNA.

3.5. Nucleotide sequence accession number

The DNA sequences of the S. dysenteriae O7 O antigen gene cluster and the flic gene have been deposited in GenBank under the accession numbers AY380835 and AY380836.

Acknowledgements

We thank all the individuals and institutes for kindly supplying strains. This study was supported by the Chinese National Science Fund for Distinguished Young Scholars, NSFC General Program, Chinese National Programs for High Technology Research and Development and funding from the Science and Technology Committee of Tianjin Government to LW.

References


