The Yersinia kristensenii O11 O-Antigen Gene Cluster was Acquired by Lateral Gene Transfer and Incorporated at a Novel Chromosomal Locus

Monica M. Cunneen and Peter R. Reeves
School of Molecular and Microbial Biosciences, University of Sydney, Sydney, New South Wales, Australia

We have sequenced the O-antigen gene clusters for the Escherichia coli O98 and Yersinia kristensenii O11 O-antigens. The basic structures of these O antigens are identical, and the sequence data indicate that Y. kristensenii O11 gained its O-antigen gene cluster by lateral gene transfer (LGT). Escherichia coli O98 has a typical O-antigen gene cluster between galF and gnd as is usual in E. coli. However, the O-antigen gene cluster of Y. kristensenii O11 is not located at the traditional Yersinia O-antigen gene cluster locus, between hemH and gsk, but at a novel chromosomal locus between arsA and cmk where it is flanked by remnant galF and gnd genes that indicate the probable source of the gene cluster. Phylogenetic analysis indicated that the source was not E. coli itself but a species in the Escherichia, Salmonella, and Klebsiella group of genera. Although other O-antigen studies imply LGT on the basis of the hypervariability of the loci and GC content, this report also identifies a potential donor and provides evidence for the mechanism involved. Remnant insertion sequence (IS) sequences flank the galF and gnd remnants and suggest that LGT of the gene cluster was IS mediated.

Introduction

O antigens are a major component of the bacterial cell wall lipopolysaccharide, a known virulence factor for mammalian infection that also acts as a protective barrier during environmental stress (Zhang et al. 1997; Lerouge and Vanderleyden 2002). O antigens exhibit extreme structural diversity, due to the hypervariability of the O-antigen gene cluster locus, with differences in sugar composition, arrangement, and glycosyl linkages within and between the repeat units that comprise this polysaccharide; for example, 186 O antigens have been identified in Escherichia coli (including Shigella strains) (Ewing et al. 1958; Brenner 1984; Ewing 1986; Lior 1994). Most reported O-antigen structures have been found in only one species but there are some exceptions. Three examples where sequence data are available are E. coli O157, Salmonella enterica O30, and Citrobacter freundii F90; E. coli O55 and S. enterica O50; and E. coli O111 and S. enterica O35 (Lindberg et al. 1981; Kenne et al. 1983; Wang and Reeves 1998, 2000; Samuel et al. 2004). In all the 3 cases, it is most probable that the gene cluster was present in a common ancestor, although lateral gene transfer (LGT) by homologous recombination between species remains a possibility as the O-antigen gene clusters for the three species involved are all located between galF and gnd (Samuel and Reeves 2003; Samuel et al. 2004).

Unlike E. coli and S. enterica, O-antigen gene clusters in Yersinia spp. are located between hemH and gsk (Skurnik et al. 1995; Zhang et al. 1996; Reeves et al. 2003; Skurnik 2003), so the situation where O antigens are shared in E. coli and Yersinia spp. must be more complex than in the above 3 cases, which involve the same locus. The O antigens of E. coli O98 and Yersinia kristensenii O11.23 and O11.24 have the same basic structure (fig 1) with the repeat unit composed of 1 residue each of N-acetyl-D-glucosamine (D-GlcNAc) and N-acetyl-D-galacturonic acid (D-GalNAcA) and 2 residues of N-acetyl-L-quinovosamine (L-QuiNAc), but in Y. kristensenii O11.23 the D-GalNAcA residue is acetylated (Marsden et al. 1994). As O-antigen acetylation is often a post O-unit synthesis modification (Clark et al. 1991; Slauch et al. 1996; Allison and Verma 2000), we have referred to Y. kristensenii O11.23 and O11.24 as Y. kristensenii O11 except for differentiation purposes. In this paper, we report the sequence of the gene clusters responsible for O antigen synthesis in E. coli O98 and Y. kristensenii O11 to better understand O-antigen gene cluster evolution and acquisition in these 2 organisms.

Materials and Methods

Bacterial Strains

Escherichia coli H510d, the E. coli O98 type strain (Ewing 1986) (lab number M1277), was kindly provided by the Institute of Medical and Veterinary Sciences, Adelaide, Australia. Yersinia kristensenii IP105, the O11.23 type strain (Bercovier et al. 1980) (lab number M2655), IP841 (serotype O11.24; lab number M2675), IP490 (serotype O12.25; lab number M2676), and 11047 (serotype O28; lab number M2674) were kindly provided by Prof. Roy Robins-Browne (University of Melbourne, Melbourne, Australia). Bacterial strains were grown overnight in nutrient broth at 37°C for E. coli and 25°C for Y. kristensenii.

Molecular Techniques

Chromosomal DNA was extracted using the Wizard chromosomal DNA purification kit (Promega, Madison, WI). Long-range polymerase chain reaction (PCR) was carried out using the Long Expand Template System according to the manufacturer (Roche Diagnostics, Dee Why, New South Wales). Short-range PCR conditions generally consisted of 94°C for 2 min followed by 30 cycles of 94°C for 30 s, 54°C for 30 s, and 72°C for 1–3 min, depending on the expected product size, and a final cycle of 72°C for 5 min, using high fidelity VENT Taq Polymerase (New England Biolabs, Ipswich, MA). Inverse PCR was carried out using the same conditions as for short-range PCR. Primers used in this study are described in table 1. Template DNA for inverse PCR was prepared by digesting 1 μg chromosomal DNA.
DNA in single restriction enzyme reactions with 10U of AvaII, EcoRI, NdeI, NcoI, or XmnI (NEB) followed by self-ligation of 4 fmol digested DNA using T4 DNA ligase (NEB). DNA fragments for cloning were end repaired with a Novagen dA-tailing kit (Merck, Kilsyth, Victoria), cloned using the pGEM-T Easy kit (Promega), and propagated in E. coli JM109.

Sequencing and Analysis

Plasmid DNA templates for sequencing were prepared using the Wizard Plus Miniprep DNA Purification System (Promega). The pGEM-T Easy clones, containing E. coli O98 O-antigen gene cluster fragments, were sequenced using M13F and M13R primers (Promega). PCR templates were purified using the UltraClean Purification System (Mo Bio Laboratories, Solana Beach, CA). Sequencing was carried out at the Sydney University and Prince Alfred Macromolecular Analysis Centre in Sydney, Australia, using Applied Biosciences 377 automated DNA sequencers.

Sequences were analyzed and compiled using the PHRAP/PHRED package from the University of Washington Genome Centre and CONSED (Gordon et al. 1998). Both the forward and reverse strands were sequenced to minimize PCR and sequencing errors. Sequences were submitted to ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and potential open reading frames (Orfs) were submitted to Blast (Altschul et al. 1990). O-antigen genes were named in accordance with the Bacterial Polysaccharide Genes Database (Reeves et al. 1996). TMHMM version 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0/) was used for transmembrane segment (TMS) prediction (Sonnhhammer et al. 1998). Sequence alignments were constructed using ClustalW (Thompson et al. 1994). Sequences were screened for programmed ribosomal frameshift sites using FSFinder (http://wilab.inha.ac.kr/FSFinder/) (Moon et al. 2004).

LPS Extraction and Analysis

Lipopolysaccharide was extracted from 10 mL overnight cultures and visualized by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and silver staining as previously described (Hitchcock and Brown 1983; Lesse et al. 1990).

Table 1

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence 5′–3′</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>1523</td>
<td>ATTTGCGCGTGCAGGAGATCAAGAAATC</td>
<td>galF</td>
</tr>
<tr>
<td>1524</td>
<td>TAGTCCGGCCTGGGCTTGAATTAGGTCG</td>
<td>gnd</td>
</tr>
<tr>
<td>5249</td>
<td>TTCTGCAGGCGAGACGTA</td>
<td>adhB</td>
</tr>
<tr>
<td>5250</td>
<td>ATGCAAGCGGGCCAAGCAGGAG</td>
<td>gskB</td>
</tr>
<tr>
<td>5251</td>
<td>CCACCTGGGATATACGGCAA</td>
<td>gskB</td>
</tr>
<tr>
<td>5301</td>
<td>GCCTAGGCCCTAACCCTGA</td>
<td>cmkB</td>
</tr>
<tr>
<td>5383</td>
<td>GGTACCTCAAGCTGTTGAGCTGG</td>
<td>hemHb</td>
</tr>
<tr>
<td>5385</td>
<td>GCCTTTGCGCGTGAAACCGTG</td>
<td>hemHb</td>
</tr>
<tr>
<td>5402</td>
<td>CGGCCTGTTAATCTGAA</td>
<td>whbOb</td>
</tr>
<tr>
<td>5469</td>
<td>GACCATCCATGAGAAAC</td>
<td>whbOb</td>
</tr>
<tr>
<td>5491</td>
<td>CTGTGAACCTTGCGACTAA</td>
<td>cmk</td>
</tr>
<tr>
<td>5520</td>
<td>TGATTATATCGAGTGCGAC</td>
<td>aroA</td>
</tr>
</tbody>
</table>

PCR walking and gap-filling primer descriptions are available on request.

Primer sequences were deposited into the GenBank database under accession numbers DQ180602, DQ180603, and DQ192109.

Results and Discussion

E. coli O98 Sequence

The E. coli O98 O-antigen gene cluster was PCR amplified using primers 1523 and 1524 (table 1), and the product was digested with NcoI. The 4 fragments produced (~6, 3.5, 2.9, and 2.4 kb, were end repaired and cloned into pGEM-T-Easy. Primers were designed for walking out from the sequence obtained with one clone of each sequenced from both ends. PCR with chromosomal DNA template, was used for gap filling and verification of the complete sequence. Genes were identified on the basis of Blast results (table 2). The O-antigen gene cluster, containing 11 genes, is 11,709 bp in length (fig 2).

Y. kristensenii O11 Sequences

PCR was carried out targeting the adk–hemH (primers 5383/5249) and gsk (primers 5250/5251) regions in strain IP105, and the products were purified and sequenced. Long-range PCR between hemH and gsk, with primers designed from the new sequence, was unsuccessful, so primers based on the E. coli O98 O-antigen gene cluster were then used in short-range PCR to successfully amplify and sequence the Y. kristensenii O11 wbwU–wbwW region. Short-range PCR using hemH (5383) and gsk (5301) primers in combination with primers based on the Y. kristensenii O11 sequence already obtained was unsuccessful, with the exception of the gne–gsk combination. We concluded that the Y. kristensenii O11 O-antigen gene cluster does not reside between hemH and gsk and so inverse PCR was used to complete the sequence and extend it to the adjacent genes aroA and cmk. The O-antigen gene cluster, containing 10 genes, is 11,315 bp in length (fig 2).

Nucleotide Sugar Biosynthesis and Transferase Genes

The gene clusters contain the genes for the biosynthesis of UDP-D-GalNAcA (gna and gne) and UDP-L-QuiNAc
**Table 2**

**BlastX Results of *E. coli* O98 and *Y. kristensenii* O11 O-Antigen Genes**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Blast Hit (Accession number)</th>
<th>Protein Description</th>
<th>% Identity and Similarity (matching/total residues) to <em>Escherichia coli</em> O98 and <em>Yersinia kristensenii</em> O11</th>
</tr>
</thead>
<tbody>
<tr>
<td>gna</td>
<td><em>Escherichia coli</em> subsp. dysenteriae Gna (AA97954)</td>
<td>UDP-N-acetyl-D-galactosamine dehydrogenase</td>
<td>99 (423/426) 79 (338/426)</td>
</tr>
<tr>
<td>gne</td>
<td><em>Escherichia coli</em> subsp. dysenteriae Gne (AA97955)</td>
<td>UDP-N-acetyl-D-glucosamine 4-epimerase</td>
<td>99 (424/426) 89 (382/426)</td>
</tr>
<tr>
<td>wzx</td>
<td><em>Vibrio vulnificus</em> MO6-24 Wzx (ABD38622)</td>
<td>Flippase</td>
<td>75 (321/427) 78 (333/425)</td>
</tr>
<tr>
<td>wzy</td>
<td><em>Vibrio vulnificus</em> MO6-24 Wzy (ABD38621)</td>
<td>UDP-N-acetyl-D-glucosamine 4-epimerase</td>
<td>85 (293/342) 87 (298/340)</td>
</tr>
<tr>
<td>wbwU</td>
<td><em>Vibrio vulnificus</em> MO6-24 hp4 (ABD38624)</td>
<td>Glycosyltransferase</td>
<td>57 (226/396) 56 (226/397)</td>
</tr>
<tr>
<td>wbwV</td>
<td><em>Syntrophomonas wolfi</em> str. Goettingen SwollDRAFT_2223 (ZP_00664310)</td>
<td>Glycosyltransferase</td>
<td>76 (301/396) 73 (290/397)</td>
</tr>
<tr>
<td>fnlA</td>
<td><em>Vibrio vulnificus</em> MO6-24 WjB (ABD38626)</td>
<td>dTDP-4-dehydrohamnose reductase</td>
<td>61 (323/400) 60 (227/373)</td>
</tr>
<tr>
<td>qnlA</td>
<td><em>Vibrio vulnificus</em> MO6-24 RmID (ABD38627)</td>
<td>UDP-N-acetyl-D-glucosamine 2-epimerase</td>
<td>64 (186/288) 63 (183/288)</td>
</tr>
<tr>
<td>qnlB</td>
<td><em>Vibrio vulnificus</em> MO6-24 WbID (ABD38628)</td>
<td>UDP-N-acetylglucosamine 2-epimerase</td>
<td>66 (265/400) 62 (244/393)</td>
</tr>
<tr>
<td>wbuC</td>
<td><em>Escherichia coli</em> O145 WbuC (AAN60464)</td>
<td>Glycosyltransferase</td>
<td>80 (322/400) 76 (299/393)</td>
</tr>
</tbody>
</table>


Three putative glycosyltransferase genes (wbwU, wbwV, and wbwW) are present, but no initial sugar phosphate transferase gene was identified (fig 1). This is probably because WecA, the initial transferase for the enterobacterial common antigen (ECA) functions as the initial GlcNac-P transferase for both ECA and O antigen. The 1st sugar of the ECA repeat unit is GlcNac and in *E. coli* it has been shown that WecA can also initiate O antigen synthesis (Meier-Dieter et al. 1992; Klena and Schmittman 1993; Alexander and Valvano 1994; Rick et al. 1994; Kido et al. 1995). It has also been shown that WecA can act as a GalNac-P transferase to initiate synthesis of the *Yersinia enterocolitica* O8 O antigen (Zhang et al. 1997). In *E. coli* and *Yersinia* spp., gene clusters for O antigens that have GlcNac or GalNac in the structure always lack an initial transferase gene, and WecA is generally assumed to be the initial transferase in all such cases. It seems clear that WecA transfers D-GlcNac-P to undecaprenol phosphate to initiate GlcNAc or GalNAc in the structure always lack an initial transferase gene, and WecA is generally assumed to be the initial transferase in all such cases. It seems clear that WecA transfers D-GlcNac-P to undecaprenol phosphate to initiate E. coli O98 and Y. kristensenii O11 O antigen synthesis, and the wbuU, wbwV, and wbwW genes must encode the other glycosyltransferases required. *E. coli* O98 and *Y. kristensenii* O11 WbwW proteins have 66% and 62% identity, respectively, to *E. coli* O26 WbuB (table 2) and similar values for other *E. coli* WbuB proteins. WbuB is the presumed N-acetyl-L-fucosamine (L-FucNAc) transferase in the synthesis of the *E. coli* O26 O-antigen which contains L-FucNAc linked to D-GlcNAc via an 1,3 linkage (Manca et al. 1996; D’Souza et al. 2002). A similar situation occurs in the *Shigella boydii* 13 and *Vibrio cholerae* O37 O-antigen gene clusters which contain the fnlA, qnlA, and qnlB gene set with a wbuB-like gene immediately downstream; wbwU and wbwV, respectively (Li et al. 2002; Feng et al. 2004). *E. coli* O98 WbwV has 44% and 40% identity, respectively, to these putative L-QuiNAc transferases. L-FucNAc and L-QuiNAc are similar sugars with similar biosynthetic
pathways and these homologous genes are in the same order in all the above strains including *E. coli* O98 and *Y. kristensenii* O11. WbwW is, therefore, proposed to transfer L-QuinAc onto D-GlcNAc via an $\alpha 1,3$ linkage, with WbwU and WbwV responsible for the other 2 glycosidic linkages (fig 1).

Frameshift Mutation in WbwW and Potential for Programmed Ribosomal Frameshifting

The *Y. kristensenii* O11,23 wbwW gene has a frameshift mutation, confirmed by sequencing both forward and reverse strands of this region twice, and the wbwW region is predicted to encode 2 Orfs (Orfs10 and 11; 240 and 167 amino acids, respectively). The *Y. kristensenii* O11,23 and O11,24 wbwW sequences are identical (data not shown), and for both strains O-antigen expression was confirmed by SDS-PAGE and silver staining of extracted LPS samples (fig 3). The wbwW is the predicted glycosyltransferase gene for one of the L-QuinAc residues, and must be functional for O-antigen expression to occur. The Pfam-A domain present in both *Y. kristensenii* and *E. coli* WbwW is Glycos_trans_1, which is located in the C-terminal domain of *E. coli* WbwW, and other glycosyltransferases that contain it (http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00534), and is in the 2nd Orf of *Y. kristensenii* WbwW. It may be that only the downstream Orf is required for function in *Y. kristensenii*, although the presence of an N-terminal region domain in all other proteins containing Glycos_trans_1 suggests a functional role for the upstream Orf. Also, there is no obvious Shine Dalgarno sequence to initiate translation of the downstream Orf although translational read-through may occur.

It is also possible that programmed -1 ribosomal framesshifting is involved in *Y. kristensenii* O11 WbwW expression. Programmed ribosomal framesshifting involves the ribosome stalling during translation and slipping out of frame, by either +1 or -1 base, to express 2 overlapping orfs as a single polypeptide (Ton-Hoang et al. 1998; Plant

FIG. 2.—*E. coli* O98, *Y. kristensenii* O11 O-antigen, and *V. vulnificus* MO6-24 capsule gene clusters. (A) Map of the 3 gene clusters. A common gene block (middle) with different chromosomal loci (complete boxes) in each strain indicated (full arrows) though a different wzy gene is present in *V. vulnificus*. Gene remnants and strain specific polysaccharide genes are shown (dashed boxes). (B) *E. coli* O98 and *Y. kristensenii* O11 DNA and protein identity levels are indicated as is the G + C% gene content. Not to scale. Note that for *Y. kristensenii* O11 wbwW, a predicted programmed ribosomal framesshifting may be involved in expression and is discussed in the text.

FIG. 3.—SDS-PAGE and silver staining of LPS profiles from (A) *Escherichia coli* O98, (B) *Yersinia kristensenii* O11,23, and (C) *Y. kristensenii* O11,24.

<table>
<thead>
<tr>
<th>% identity</th>
<th>DNA</th>
<th>72.1</th>
<th>72.0</th>
<th>71.6</th>
<th>74.2</th>
<th>69.8</th>
<th>73.3</th>
<th>82.0</th>
<th>72.8</th>
<th>79.4</th>
<th>64.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>78.9</td>
<td>76.8</td>
<td>69.8</td>
<td>70.4</td>
<td>64.7</td>
<td>72.5</td>
<td>90.4</td>
<td>74.9</td>
<td>74.9</td>
<td>86.2</td>
<td>61.3*</td>
</tr>
<tr>
<td>G+C% content</td>
<td>DNA</td>
<td>37.0</td>
<td>35.7</td>
<td>29.9</td>
<td>24.0</td>
<td>32.2</td>
<td>33.2</td>
<td>36.7</td>
<td>33.3</td>
<td>35.5</td>
<td>41.0</td>
</tr>
<tr>
<td>Protein</td>
<td>37.8</td>
<td>38.9</td>
<td>29.7</td>
<td>25.1</td>
<td>31.2</td>
<td>32.4</td>
<td>36.1</td>
<td>32.5</td>
<td>36.7</td>
<td>37.3</td>
<td></td>
</tr>
</tbody>
</table>
et al. 2004; Makelainen and Makinen 2005; Manktelow et al. 2005). Successful -1 programmed ribosomal frameshifting requires several mRNA signals; a frameshift site, a suitable spacer region and a stable pseudoknot, or hairpin loop stem secondary structure (Blinkova et al. 1997; Hammell et al. 1999; Rettberg et al. 1999; Baranov et al. 2002b; Ferracci et al. 2004). *Y. kristensenii* O11 *wbwW* contains these frameshift signals, as detected by FSFinder (Moon et al. 2004), that would enable expression as a single polypeptide (fig 4). *E. coli* O98 *wbwW* and 12 other related genes, found by Blast that are also located downstream of *fna* or *qnl* gene sets, were analyzed by FSFinder (fig 4) in addition to the other O-antigen genes in *E. coli* O98 and *Y. kristensenii* O11. Of the 12 only the *E. coli* O172 gene has both frameshift signals in this region, and in a further test, of the 10 O-antigen genes shared between *E. coli* O98 and *Y. kristensenii* O11, only *wbwW* and *gnd* contain both a conserved frameshift site and potential secondary structure with a spacing consistent with functional frameshift sites (Dimman et al. 1991; Baranov et al. 2002a; Bekaert et al. 2003). It seems unlikely that by chance *wbwW* has the appropriate set of signals for programmed -1 ribosomal frameshifting at the frameshift mutation site, but experimental work is required to determine if ribosomal frameshifting does occur.

It will be very interesting to see if programmed ribosomal frameshifting is involved in *wbwW* expression or regulation and if so, it would be the first case associated with O-antigen expression.

O-Unit Processing Genes

Most *E. coli* and *Yersinia* O-antigens are synthesized by the Wzy dependent pathway (Valvano 2003), as shown by the presence of *wzy* and *wzx* genes. *E. coli* O98 and *Y. kristensenii* O11 Orf3 have 56% and 60% identity, respectively, to *Vibrio vulnificus* MO6-24 Wzx, whereas Orf4 has low level identity (<26%) with putative Wzy proteins (table 2). Wzx and Wzy often have low levels of similarity to the closest homologue, but usually encode 12 or 9 potential TMS, respectively (Macpherson et al. 1995; Samuel et al. 2004), and that was found to be the case here using TMHMM (Sonhammer et al. 1998). We assigned Orf3 and Orf4 as Wzx and Wzy, respectively.

E. coli O98 and Y. kristensenii O11 Gene Cluster Relationships

The orfs in both clusters were in the same transcriptional direction as the adjacent neighboring genes, with most genes in common and in the same order (fig 2). Both gene clusters have low G + C% content compared with representative genus chromosomes and potentially indicate acquisition from a low G + C% content bacterium via LGT (Reeves 1991, 1993). However, the best evidence for LGT is in the flanking genes. The *E. coli* O98 O-antigen gene cluster is flanked by *galF* and *gnd* as are almost all *E. coli* and *Salmonella* O-antigen gene clusters. Remnants of these genes flank the *Y. kristensenii* O11 O-antigen gene cluster, but only the first half of *galF*, usually distal from O-antigen genes, is present and the second, usually proximal, half is missing, although the promoter region for the gene cluster is not disrupted. It appears that the *Y. kristensenii* O11 gene cluster was acquired from a species in which the O-antigen genes are between *galF* and *gnd*. The *gnd* fragment at the end of the cluster is also truncated and lacks the sequence on which our primer 1524 is based, but nonetheless the *galF–gnd* primer pair, 1523/1524, when used on our *Y. kristensenii* O11 strain, gave a PCR product showing that it had adjacent *galF* and *gnd* genes, as in other *Yersinia* spp. (data

---

**FIG. 4.—Part of ClustalW alignment of putative L-FucNAc and L-QuiNAc transferase genes.** For *Y. kristensenii* O11, the predicted frameshift (gray) and secondary sites (white; only the 1st is shown with the 2nd sequence located further downstream). For *Y. kristensenii* O11, Orf11 start codon (green) and Orf10 stop codon (red) with the T-rich frameshift region (circled) and spacer region (underlined) are shown. Other sequences are similarly shaded. Species abbreviations; Ec, *Escherichia coli*; LiC, Leptospira interrogans sv. Copenhageni; LiH, *L. interrogans* sv. Hardjo; LiP, *L. interrogans* sv. Pomona; Lb, *Leptospira borgpetersenii*; Sb, *S. boydii*; Vc, *Vibrio cholerae*; Vv, *Vibrio vulnificus*; and Yk, *Y. kristensenii*. Gene names with base positions are indicated at the start of each sequence. Organisms grouped according to relevant sugar residue; (A) L-FucNAc, (B) L-QuiNAc, (C) L-RhaNAc, (D) L-NeuNAc, (E) L-FucNAc, (F) L-QuiNAc, (G) L-RhaNAc, (H) L-NeuNAc. GenBank accession numbers (from top to bottom); AY369140, AY647260, AY545992, AF529080, DQ180602, DQ180603, AY0369140, AF906573, DQ360502, AF144879, AF078135, AF316500, and NC_005823.
not shown). The O-antigen associated galF and gnd remnants are clearly additional to these. They are also more closely related to genes from E. coli and S. enterica strains than to those from Yersinia spp. The polypeptide based on the galF remnant has ~61% identity to E. coli, S. enterica, and Klebsiella pneumoniae GalF, being the closest hits in a Blast search. Phylogenetic analysis demonstrates its relationships within the Enterobacteriaceae (fig 5), which show that it is related to the Escherichia, Salmonella, and Klebsiella group of genera. The common gene order indicates that the 2 gene clusters have a common ancestor, and the relationships of the galF gene suggest that it was near the Escherichia, Salmonella, and Klebsiella group of genera. The sequence divergence of the O-antigen genes and housekeeping genes is quite similar; so on that basis, the ancestral gene cluster could have been in the Yersinia, Escherichia common ancestor, but that would not explain how the Y. kristensenii gene cluster is bracketed by E. coli-like galF and gnd genes. It is also surprising that these gene clusters have such low G + C% content if they have been in the Yersinia, Escherichia lineage for as long as under either of the above hypotheses. The alternative hypothesis that they diverged in a low G + C% content species also does not explain the presence of remnant galF and gnd genes that are related to the Escherichia, Salmonella, and Klebsiella group of genera. With regard to the low G + C% content, this seems to be so common that one has to question the usual explanation of LGT from a low G + C% content species, although we are not aware of a convincing alternative explanation.

Insertion sequence (IS) remnants are present in Y. kristensenii O11 at the junctions of aroA and cmk, with the remnant galF and gnd sequences, respectively. The partial IS upstream encodes a protein with 54% amino acid identity to the 3’ end of Pseudomonas aeruginosa transposase IstB and that downstream of the cluster encodes a protein with 55% amino acid identity to a Yersinia pestis pCD1 transposase Lcr5 (Perry et al. 1986; Larbig et al. 2002). It is likely that these IS were involved in the transfer of the O antigen gene cluster to Yersinia.

The galF and gnd remnants transferred with the Y. kristensenii O11 cluster may not have shared sufficient identity to the galF–gnd chromosomal region for homologous recombination to occur at this site. The aroA and cmk genes are not thought to have been transferred with the gene cluster as they are closely related to those of other Yersinia spp. and mark the site of chromosomal integration.

WbuC

Directly downstream of E. coli O98 wbbW is wbuC, which is absent in Y. kristensenii O11. The block of genes qnlA, qnlB, qnlC, wbuC, and wbuB have been found in several gene clusters for O antigens that, where known, have L-FucNAc, or a derivative, as the 2nd sugar and GlcNAc as the initiating sugar, and a similar situation exists for L-QuinAc with fnlA, qnlA, qnlB, and wbuC genes. The first 3 genes encode synthesis of UDP-L-FucNAc or UDP-L-QuinAc, respectively, that have similar biosynthetic pathways with fnlB and fnlC being homologues of qnlA and qnlB, respectively. The next gene is wbbW, or homologue, encoding the transferase believed responsible for the α(1-3) linkage of L-QuinAc or L-FucNac to GlcNAc. In E. coli there is usually an additional gene wbuC, that has no known function, but its conservation implies that it must have a function. It is interesting that E. coli O98 resembles other E. coli in having wbuC but it is absent in Y. kristensenii O11. It was presumably transferred to Y. kristensenii as part of the segment from galF to gnd, but there is no remnant present. It seems that even if imported into another species wbuC provides no useful function outside of E. coli.

A Related V. vulnificus Gene Cluster

Nine of the 10 genes in the Y. kristensenii O11 O-antigen gene cluster are also found in the group 1 capsule gene cluster of V. vulnificus MO6-24 (fig 2 and table 2), in the same order, but with an unrelated wzy gene in MO6-24 (GenBank accession number DQ360502) (fig 2). In V. vulnificus MO6-24 the wza, wzb, and wzc genes are upstream of gna and the wbbTUY-ugd genes are downstream of wbbW (Chatzidaki-Livanis et al. 2006).

The capsule structure of V. vulnificus MO6-24 is known (Reddy et al. 1992), and 3 of the 4 repeat unit sugar residues and 2 glycosyl linkages are shared with E. coli O98
Polysaccharide Gene Cluster Loci of E. coli and Yersinia spp.

Yersinia Wzy dependent O-antigen gene clusters generally map between hemH and gsk. This applies to 11 Y. pseudotuberculosis, Y. enterocolitica O8, and Y. aldovae O-antigens (GenBank accession number AJ871364) and also the Y. enterocolitica O3 and O9 outer core gene clusters that are effectively single O-units (Skurnik et al. 1995, 2000; Zhang et al. 1996; Reeves et al. 2003; Skurnik 2003). Analysis of a Y. mollaretii genome showed that putative O-antigen genes, including wzy and wzy, are also present at this locus (GenBank accession number AALD00000000), with an extensive ~43-kb gene cluster that unusually also contains wzm–wzt genes. For Y. intermedia putative O antigen genes are present at this locus but no wzy–wzy or wzm–wzt genes (GenBank accession number AALF00000000).

The Y. kristensenii O11 O-antigen gene cluster is atypical in not being between hemH and gsk. As noted above there is another gne gene upstream of gsk. The Y. enterocolitica O3 outer core, Y. aldovae and Y. enterocolitica O8 O-antigen gene clusters all have gne near or adjacent to gsk and in these instances wbcQ is upstream of gne. PCR with Y. enterocolitica O3 wbcQ (5402) and gsk (5251) based primers was successful and long-range PCR, with hemH (5385) and wbcQ (5469) primers produced a 15-kb product. The 2.6-kb wbcQ–gsk PCR product was sequenced and the 2 orfs found had similarity to Y. enterocolitica O3 outer core wbcQ and gne genes in addition to gsk. The ~17-kb segment between hemH and gsk, including wbcQ and gne genes, indicates the presence in Y. kristensenii O11,23 of another polysaccharide gene cluster in this region (fig 6), although structural studies (Marsden et al. 1994) did not indicate the presence of a 2nd polysaccharide structure. This suggests that the gene cluster between hemH and gsk is either nonfunctional or not expressed under laboratory conditions. It is quite likely that it is the native ancestral O-antigen gene cluster of this strain that has become redundant. In a comparable situation, E. coli Sonnei, with a plasmid-borne O-antigen gene cluster derived from

Table 3
DNA Identity Levels (%) for O Antigen and Housekeeping Genes between Strains

<table>
<thead>
<tr>
<th>O-Antigen Gene</th>
<th>E. coli O98 versus Y. kristensenii O11</th>
<th>E. coli O98 versus V. vulnificus MO6-24</th>
<th>Y. kristensenii O11 versus V. vulnificus MO6-24</th>
</tr>
</thead>
<tbody>
<tr>
<td>gna</td>
<td>72.1</td>
<td>67.2</td>
<td>66.4</td>
</tr>
<tr>
<td>gne</td>
<td>72.0</td>
<td>70.1</td>
<td>71.6</td>
</tr>
<tr>
<td>wzy</td>
<td>71.6</td>
<td>64.5</td>
<td>66.5</td>
</tr>
<tr>
<td>wzy</td>
<td>74.2</td>
<td>63.1 (127 bases*)</td>
<td>71.9 (60 bases*)</td>
</tr>
<tr>
<td>wbyW</td>
<td>65.8</td>
<td>63.4</td>
<td>62.8</td>
</tr>
<tr>
<td>qnlB</td>
<td>73.3</td>
<td>65.8</td>
<td>66.8</td>
</tr>
<tr>
<td>qnlA</td>
<td>82.0</td>
<td>76.7</td>
<td>78.2</td>
</tr>
<tr>
<td>qnlB</td>
<td>72.8</td>
<td>65.5</td>
<td>66.8</td>
</tr>
<tr>
<td>qnlB</td>
<td>79.4</td>
<td>72.9</td>
<td>75.4</td>
</tr>
<tr>
<td>wbyW</td>
<td>64.1</td>
<td>56.8</td>
<td>58.8</td>
</tr>
<tr>
<td>Range (average)</td>
<td>64.7–82 (73.1)</td>
<td>63.4–76.7 (67.0)</td>
<td>58.8–75.4 (68.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Housekeeping Gene</th>
<th>E. coli K12 versus Y. pseudotuberculosis</th>
<th>E. coli K12 versus V. vulnificus YJ016</th>
<th>Y. pseudotuberculosis versus V. vulnificus YJ0162</th>
</tr>
</thead>
<tbody>
<tr>
<td>msh</td>
<td>78</td>
<td>70.5</td>
<td>68</td>
</tr>
<tr>
<td>ask</td>
<td>77.8</td>
<td>71.8</td>
<td>70.4</td>
</tr>
<tr>
<td>gycB</td>
<td>78.7</td>
<td>73.0</td>
<td>72.2</td>
</tr>
<tr>
<td>mctG</td>
<td>77.4</td>
<td>68.8</td>
<td>68.5</td>
</tr>
<tr>
<td>purA</td>
<td>80.9</td>
<td>73.3</td>
<td>73.3</td>
</tr>
<tr>
<td>recA</td>
<td>80.0</td>
<td>77.0</td>
<td>75.9</td>
</tr>
<tr>
<td>Range (average)</td>
<td>77.4–80.9 (78.8)</td>
<td>68.8–77.0 (72.4)</td>
<td>68–75.9 (71.4)</td>
</tr>
</tbody>
</table>

*Not included in average.

and Y. kristensenii O11 (fig 1). The 3rd shared glycosyl-
transerase, WbwW, is presumably responsible for the re-
lated QuiNac (α1-3) GlcNAc and QuiNac (α1-3) GalNAc
linkages in E. coli O98 and V. vulnificus MO6-24, respec-
tively. Unfortunately the comparison does not help us allo-
 cate WbuU or WbuV to specific linkages.

The relationships between the V. vulnificus MO6-24
gene cluster and those sequenced in this study were anal-
alyzed by using housekeeping genes taken from the V. vul-
nificus YJ016 genome (GenBank accession number
BA000037). The DNA identity levels observed suggest that
E. coli O98 and Y. kristensenii O11 O-antigen genes are
more closely related than either is to those in V. vulnificus
MO6-24, and housekeeping gene divergence indicated
a similar genera relationship (table 3). Furthermore, different
wzy genes were identified that correlate with the differ-
ent repeat unit linkages observed.

The V. vulnificus MO6-24 capsule and E. coli O98 and
Y. kristensenii O11 O-antigen gene clusters are similar and
share a conserved gene order. However, given the lower
levels of DNA identity, additional capsule genes and a dif-
f erent wzy gene, the evolution of the V. vulnificus MO6-24
gene cluster is more complex with a distinct evolutionary
history different from those of E. coli O98 and Y. kristen-
senii O11 although a common gene block is conserved.
Plesiomonas shigelloides, has lost its original chromosomal O-antigen gene cluster by deletion (Lai et al. 1998) presumably after gaining the plasmid-borne gene cluster.

It is interesting that ABC-transporter dependent O-antigen gene clusters in Yersinia spp. are usually in the galF–gnd region, for example, the Y. enterocolitica O9 gene cluster. There are also putative O-antigen gene clusters upstream of galF–gnd in Y. bercovieri and Y. frederiksenii (GenBank accession numbers NZ_AA0C0000000 and NZ_AA0E0000000, respectively) but not in Y. intermedia or Y. mollaretii (discussed above). Both gene clusters include wzm and wzt and thus also code for ABC-transporter dependent O antigens.

The galF–gnd region is a hotspot for polysaccharide gene clusters in E. coli and close relatives with the colanic acid gene cluster upstream of galF, E. coli, and S. enterica Wzy dependent O-antigen and E. coli group I capsule gene clusters between galF and gnd, and the Klebsiella and E. coli (serotypes O8 and O9) ABC-transporter dependent gene clusters are downstream of gnd (Stevenson et al. 1996; Samuel and Reeves 2003; Samuel et al. 2004; Whitfield 2006). Generally in Yersinia spp., Wzy dependent gene clusters are between hemH and gsk and ABC-transporter dependent gene clusters are upstream of galF–gnd, with only one of the sites being occupied in the available genome sequences.

Y. kristensenii O11, O12, and O28 may have related gene clusters as the O28 O antigen shares D-GalNAcA with O11 and the O12 O-antigen contains L-FucNAc, which is very similar to the L-QuiNAc of O11 (L’Vov et al. 1990, 1992; Perry and MacLean 2000). We examined the aroA–cmk region in these strains using long-range PCR, with ar–oA (5520) and cmk (5491) primers. This gave the expected 13-kb PCR product for Y. kristensenii O11 but revealed that for O28, O12,25, and O12,26 aroA and cmk were adjacent and thus not associated with an O-antigen gene cluster. The aroA and cmk are also adjacent in the genomes of Y. pseudotuberculosis, Y. bercovieri, Y. frederiksenii, Y. intermedia, Y. mollaretii, and Y. pestis (GenBank accession number AE009952) but separated by 1 gene, ycaL, in E. coli K12 and S. enterica LT2 (GenBank accession number AE008742). Thus there is no evidence for other O-antigen gene clusters between aroA and cmk. We conclude that the aroA–cmk region is a new O-antigen gene cluster locus, in which the presence of galF and gnd remnants allows one to see that it arose by insertion of a gene cluster from a species where these genes flank the O antigen gene cluster. This is the first reported instance of the establishment of a new O-antigen locus within a species. Yersinia spp and E. coli have extensive synteny (ENTERIX; http://globin.bx.psu.edu/enterix/enteric/enteric.html), but the Wzy dependent O-antigen gene clusters are at different loci. The situation in Y. kristensenii O11 shows how new loci can arise. Although it is not known why the Yersinia and Escherichia genera have different loci for Wzy dependent O-antigen gene clusters, this study indicates that new loci can emerge and may be mediated by IS activity during acquisition of new gene clusters in LGT events. Such events are significant as instances of O-antigen gene IS-related LGT events have been responsible, for example, for the emergence of the pathogenic V. cholerae O139 strain via a gene block transfer (Bik et al. 1995).

Final Remarks

The O-antigen gene clusters of E. coli O98 and Y. kristensenii O11 have provided information not only on the relationship between these O-antigen genes but also evidence in Y. kristensenii O11 for the establishment of a new and novel O-antigen gene cluster locus via LGT. Another gene cluster is also present in the traditional Yersinia O-antigen locus between hemH and gsk in Y. kristensenii O11 that probably relates to the ancestral O antigen gene cluster of this strain. It was not sequenced in this study as it does not appear to be expressed, but sequencing may provide information on the original O antigen of this strain.

The most probable explanation for a given species having a single locus for all or most of its O antigens of a given class is that recombination will generally lead to replacement of one O antigen with another, rather than gain of a new O antigen. There are well-documented cases where this has happened, for example, V. cholerae O139 Bengal and E. coli O157:H7 (Bik et al. 1995; Wick et al. 2005). The new O antigen perhaps confers advantages by being not affected by host immunity to the previous O antigen. This clearly applies to V. cholerae O139 Bengal, which could infect adults immune to V. cholerae O1. The benefit may be in loss of the old O antigen while retaining the benefits of a complete LPS, rather than the specific structure of the new O antigen. This concept is supported by the finding that E. coli Sonnei with its O antigen on a plasmid has a deletion inactivating the original O antigen at the galF/gnd site (Lai et al. 1998). There will be selection for such new O-antigens to translocate to the usual locus for that species, such that when it is transferred to another strain it too can replace the pre-existing O antigen. This is similar to the selection for co-location of genes in the selfish operon hypothesis (Lawrence and Roth 1996). Presumably after some such event in either the E. coli or Yersinia lineage, the new locus somehow gained ground eventually displacing the other locus completely.

Acknowledgments

This work was supported by a grant from the Australian Research Council.

Literature Cited

Alexander DC, Valvano MA. 1994. Role of the rfe gene in the biosynthesis of the Escherichia coli O7-specific


glitserofatsoderzhashchei O-spetsificheskoi polisakharid-
noi tsipi iz lipopolisakharida Yersinia kristensenii, shtamn
L’Vov VL, Gur’yanova SV, Rodionov AV, Gorshkova RP,
1992. Structure of the repeating unit of the O-specific polysaccharide of the lipopolysaccharide of Yersinia
kristensenii strain 490 (O:12,25). Carbohydr Res. 228:
415–422.
Macpherson DF, Manning PA, Morona R. 1995. Genetic analysis
at the programmed -1 ribosomal framshefing site of
Cocksfoot mottle virus RNA in vivo. Nucleic Acids Res. 33:
2239–2247.
L’Vov VL, Gur’yanova SV, Rodionov AV, Gorshkova RP.
1364 Cunneen and Reeves
Macpherson DF, Manning PA, Morona R. 1995. Genetic analysis
at the programmed -1 ribosomal framshefing site of
Cocksfoot mottle virus RNA in vivo. Nucleic Acids Res. 33:
2239–2247.
of the Escherichia coli O26 O-antigen polysaccharide.
Manktelow E, Shigemoto K, Brierley I. 2005. Characterization of
the frameshift signal of Edr, a mammalian example of
programmed -1 ribosomal frameshifting. Nucleic Acids Res.
33:1553–1563.
Serological and structural relationships between Escherichia coli O98 and
Yersinia enterocolitica O:11,23 and O:11, 24 lipopolysaccha-
the essential gene glmM of the K-12 gene cluster
the essential gene glmM encoding phosphoglucomutase in
genes expressed via -1 and +1 frameshifts. Nucleic Acids
Res. 32:4884–4892.
Pepper RD, Harmon PA, Bowmer WS, Straley SC. 1986. A low-
Ca2+ response operon encodes the V antigen of Yersinia
Pepper MB, MacLean LL. 2000. Structural identification of the
lipopolysaccharide O-antigen produced by Yersinia enter-
-1 ribosomal frameshift signal can function as a cis-acting
mRNA destabilizing element. Nucleic Acids Res.
32:784–790.
Reddy GP, Hayat U, Abeygunawardana C, Fox C, Wright AC,
Maneval DRJ, Bush CA, Morris JGJ. 1992. Purification and
determination of the structure of capsular polysaccharide of
3:30–40.
Reeves P. 1993. Evolution of Salmonella O antigen variation by
interspecific gene transfer on a large scale. Trends Genet.
9:17–22.
Reeves PR, Hobbs M, Valmova MA, et al. 1996. Bacterial
polysaccharide synthesis and gene nomenclature. Trends
Microbiol. 4:495–503.
A three-way junction and constituent stem-loops as the
stimulator for programmed -1 frameshifting in bacterial
Rick PD, Hubbard GL, Barr K. 1994. Role of the rfe gene in the
synthesis of the O8 antigen in Escherichia coli K-12. J
the Escherichia coli O157, O111, and O55 O-antigen gene
clusters with those of Salmonella enterica and Citrobacter
186:6536–6543.
and pathways involved in nucleotide sugar precursor synthesis
Skurnik M. 2003. Molecular genetics, biochemistry and bi-
ological role of Yersinia lipopolysaccharide. Adv Exp Med
Skurnik M, Peippo A, Ervela E. 2000. Characterization of the O-
antigen gene clusters of Yersinia pseudotuberculosis and the
cryptic O-antigen gene cluster of Yersinia pestis shows that
the plague bacillus is most closely related to and has evolved from
locus of Yersinia enterocolitica serotype O:3 involved in
lipopolysaccharide outer core biosynthesis. Mol Microbiol.
17:575–594.
characterization of the oafA locus responsible for acetylation of
178:5904–5909.
model for predicting transmembrane helices in protein
Organization of the Escherichia coli K-12 gene cluster
responsible for production of the extracellular polysaccharide
Thompson JD, Higgins DG, Gibson TJ. 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.
Ton-Hoang B, Polard P, Chandler M. 1998. Efficient trans-
Front Biosci. 8:s452–s471.
Wang L, Reeves PR. 1998. Organization of Escherichia coli
O157 O antigen gene cluster and identification of its specific
Wang L, Reeves PR. 2000. The Escherichia coli O111 and
Salmonella enterica O35 gene clusters: gene clusters encod-
ing the same colistose-containing O antigen are highly
Whitfield C. 2006. Biosynthesis and assembly of capsular
75:39–68.
Wick LM, Qi WW, Lacher DW, Whittam TS. 2005. Evolution of
Zhang L, Radziejewska-Lebrecht J, Krajewska-Pietrasik D,
Toivanen P, Skurnik M. 1997. Molecular and chemical
characterization of the lipopolysaccharide O-antigen and its
role in the virulence of Yersinia enterocolitica serotype O:8.

Y. kristensenii and E. coli O Antigen Genes

Accepted March 15, 2007