Structure and genetics of *Escherichia coli* O antigens

Bin Liu§, Axel Furevi§, Andrei V. Perepelov§, Xi Guo§, Hengchun Cao, Quan Wang, Peter R. Reeves, Yuriy A. Knirel, Lei Wang and Göran Widmalm

§These authors contributed equally to this work.

**Supporting information**

**Table of contents**

Anomalies in *E. coli* O antigen gene clusters S2 – S3

- Separated sugar biosynthetic pathway genes S2
- Unusual gene locations or transcriptional directions S2
- Mobile elements S3
- Remnant genes S3
- Non-coding regions S3

Order of GTs in the *E. coli* O antigen gene cluster S4 – S5

O antigens present in *Shigella* and *E. coli* S5

References S6 – S9

Figure S1 S10 – S21

Figure S2 S22

Table S1 S23 – S25

Table S2 S26
Anomalies in *E. coli* O antigen gene clusters

There are some patterns in O antigen gene clusters that are strongly conserved, but with occasional anomalous exceptions. These are often associated with evolutionary events and we list here a number of these anomalies that may well have resulted from evolutionary events.

Separated sugar biosynthetic pathway genes

The genes in each monosaccharide biosynthesis pathway are usually co-located with a conserved order, as in genes responsible for the synthesis of dTDP-L-Rha, GDP-L-Fuc, and UDP-L-FucNAc. We list below some exceptions.

i) dTDP-L-Rha. In *E. coli* dTDP-L-Rha synthesis involves an *rml* gene set comprising *rmlB*, *rmlD*, *rmlA* and *rmlC*, located at the 5’-end of the O antigen gene cluster. Among the 54 O antigen gene clusters include *rmlBDAC* genes; there are five exceptions (O2, O54, O71, O119 and O177), in which the *rmlC* gene is located one to seven genes downstream of the *rmlBDA* genes.

ii) dTDP-6d-L-Tal. Four O antigen gene clusters (O45, O66, O88 and O109) are found to contain the *rmlB*, *rmlA*, *rmlC* and *tll* genes involved in dTDP-6d-L-Tal synthesis. In O66 and O88 these genes are located at or near the 5’-end of the gene cluster in the above order. In O109, the *rmlC* and *tll* genes are 8 genes downstream of the *rmlBA* genes at the 3’-end of the O antigen gene cluster. The arrangement of the four genes in O45 is totally different, with *rmlB* at the 5’-end, *rmlC* and *wbhT* four genes downstream, and *rmlA* gene at the 3’-end.

iii) GDP-L-Fuc. In *E. coli* the pathways that include *Gmd*, which extends the GDP-Man pathway to generate other GDP sugars, often have a *gmm* gene. *Gmm* hydrolyses GDP-mannose and GDP-glucose to release the sugar (Frick *et al.*, 1995), and is proposed to be involved in regulation but we are not aware of any data on this. The gene set for most of the 18 GDP-L-Fuc clusters is *gmd*, *fcl*, *gmm*, *manC*, and *manB* in that order, but in O126 the *gmm* gene is missing. The O126 *gmd*, *fcl*, *manC* and *manB* genes do not cluster with those from the other O serogroups in a phylogenetic tree, indicating a different origin for GDP-L-Fuc synthesis genes in O126. There is also a glycosyltransferase gene between the *manB* and *manC* genes in 9 of the gene clusters (O11, O36, O41, O43, O84, O125, O156, O159 and O168).

Unusual gene locations or transcriptional directions

Most genes for O antigen synthesis are located in the O antigen gene cluster between *galF* and *gnd* and in the same transcriptional direction from *galF* to *gnd*. We find unusual gene locations or transcriptional direction in four gene clusters. In O59 and O155, the *wzx* and O-acetyl transferase (*wclD*) genes are found downstream of the *gnd* gene, which may be an intermediate stage in incorporation of genes into the main O antigen gene clusters. In O55, two genes (*wbdJ* and *wbdK*) responsible for the synthesis of GDP-Col are also located downstream of *gnd*, and are suggested to be added recently to replace alternative genes for extension of the GDP-Man pathway. Among all the O antigen gene clusters analyzed, we find only a *qdtC* gene in O71 transcribed in the reverse orientation to the other *E. coli* O antigen genes. The O45 gene cluster is the only one in *E. coli* with an initial transferase (IT) gene; O45 lacks GlcNAc and GalNAc residues, those that normally initiate O-unit synthesis. The O45 WbhQ, which shares 53% identical level to WbpL, the IT of *P. aeruginosa*, is proposed to transfer FucNAc-P to Und-P to initiate O45 antigen synthesis.
Mobile elements

IS and H-repeat elements are known to cause mobilization of DNA fragments, and are also likely to be involved in the formation of new O antigen forms. One or more have been found in 23 *E. coli* O antigen gene clusters. There are H-repeat elements in seven O antigen gene clusters (O2, O3, O15, O24, O40, O56, and O157).

Two or more mobile elements are found in nine serogroups, including O2, O24, O40, O50, O63, O64, O156, O169 and O176. For instance, there are IS630, IS629, IS3 and H-repeat elements present in three regions in the O24 O antigen gene cluster, and IS3 and several H-repeat elements are found in three regions of the O40 gene cluster. Additionally, in the O64 O antigen gene cluster, a total of six IS1 elements are distributed at three regions. The high frequency of the mobile elements in these gene clusters indicates that lateral gene transfer occurred more often in these O antigen gene clusters.

Remnant genes

There are gene remnants in five O antigen gene clusters including O10, O16, O23, O24, O37, O40, O56, O101 and O164, which were presumably involved in reactions no longer in use. In O16, a glycosyltransferase *wbbL* at the 3’-end of the O antigen gene cluster is interrupted by an IS element. In O40, an IS3 element, discussed above, is inserted into a *wffR* pyruvyltransferase gene at the 3’-end of the gene cluster. In O164, there is a frame shift mutation in a pyruvyltransferase gene (*wfeP*) located between *wfeO* and *wfeQ*, which leads to the presence of two remnants of this gene. There are also the remnant genes in O24 that were discussed above.

Non-coding regions

In most *E. coli* O antigen gene clusters, the gaps between genes are short, but larger gaps are present in 14 O antigen gene clusters. The advantage of overlaps or short gaps between genes in an operon is probably in translational coupling in which a ribosome reinitiates before the two subunits separate and return to the cytoplasm. This can extend up to 350 nucleotides (Tian & Salis 2015) but is usually much less. Here we list only the longer gaps that potentially are regions that once had functions that cannot be obtained by searching the usual databases. In O18, a 341 bp non-coding region is found between *wekV* and *wekW*. In O37, a 523 bp non-coding region is found between *gne* and *wbeS*. In O54, a 1827 bp non-coding region is found between *wzx* and *rmlC*. In O81, a 365 bp non-coding region is found between *gne* and *wbeS*. In O82, a 323 bp non-coding region is found between *wcnV* and *wfaP*. In O92, a 1471 bp non-coding region is found between *wbgM* and *wbsH*. In O128, a 470 bp non-coding region is found between *wzy* and *gne*. In O131, a 464 bp non-coding region is found between *wetG* and *wdaK*. In O133, a 313 bp non-coding region is found between *wbdH* and *wzy*. In O139, a 326 bp non-coding region is found between *rmlC* and *wzy*. In O143, a 1939 bp non-coding region is found between *wzy* and *gne*. In O153, a 518 bp non-coding region is found between *wbuO* and *wetO*. In O163, a 293 bp non-coding region is found between *wcqV* and *wcqW*. In O178, a 518 bp non-coding region is found between *wbuO* and *wetO*.
Order of GTs in the *E. coli* O antigen gene cluster

It has often been observed that O antigen GT genes are commonly present in inverse order of their function order, with the gene for the last sugar to be added being the first GT gene in the cluster, and so on. We used the large number of *E. coli* GTs with a putative function for a quantitative analysis, and found that 68 (52%) of the 131 suitable gene clusters have all GTs in reverse order to function. The probability of being in a specific order by chance is 50% for two GTs, 17% ($\frac{1}{3} \times \frac{1}{2}$) for three GTs, but 4%, and 0.8% for four or five GTs, respectively.

There are 131 structures for which GT order can be predicted, being those with linear structures (as flipped), as the order of GT functions cannot be readily predicted for branched structures. There are 8, 59, 55 and 9 structures with two, three, four, or five GTs, respectively, with 68 (51%) of these having the predicted reverse GT order. The chances of 68 gene clusters having the predicted order by chance is extremely low, and the existence of 68 instances of a possible 131 indicates strong selection for this order. Furthermore, many of the exceptions have most genes in the predicted order, and the exceptions may reflect relatively recent evolutionary changes, indicating that these have also been under selection for the same predicted reverse GT gene order.

The prevalence of a specific pattern in gene order requires an explanation, and a recent paper proposed a selective benefit for this gene order, based on analysis of a set of 14 closely related O antigens and associated gene clusters in *Yersinia pseudotuberculosis* (Kenyon, Cunneen and Reeves 2017). It was observed (*i*) that the gene order was such that in all pairwise comparisons the genes that differed in the two clusters were generally grouped together and flanked by two blocks of genes that were shared and (*ii*) that the gene order in all clusters was related to the function order of the proteins. This order also allows the variable components of the O antigen in one strain to be replaced by any of the other O antigens by homologous recombination across the two shared ends of the gene clusters.

There are many cases of genes with products that contribute to a single function being in a gene clusters, and in an analysis of the benefits of such clusters Lawrence and Roth (Lawrence *et al.*, 1996) proposed that the most common benefit was facilitating replacement of one gene cluster by another by horizontal gene transfer (HGT) with the ends in shared genes (the selfish operon model). This is documented for O antigens by serotype switching, which can rescue a colonizing strain from being eliminated by an immune response. It was proposed (Kenyon, Cunneen and Reeves 2017) that this selection for the genes that differ being in a central block, could account for the order of the GT genes being related to their function order. The connection between gene order and selective advantage is illustrated in Figure S2, which shows an alignment of 5 out of 14 related *Y. pseudotuberculosis* gene clusters. Switching from one serotype can be achieved by HGT replacing one contiguous block of genes by the other – these blocks are indicated by the red lines between adjacent gene clusters in Figure S2. If these varying genes were not in a contiguous block, or if the GT genes were not in function order, then recombination across shared genes could give rise to strains with GT combinations that were non-functional or produce truncated structures that cannot be polymerized by Wzy. The selective advantage for this pattern is that it facilitates O antigen switching to the advantage of the recipient strain and the O antigen gene cluster that is gained, and for the specific gene cluster there is the advantage of reaching a new strain. This can be very powerful selection, as the new O antigen may give the recipient strain the ability to continue the infection, and immediately reach numbers in the feces that allow infection of new hosts. This is the only explanation that we know of for gene order in O antigen gene clusters being related to function order.
The selective benefit of the GT gene order thus only arises when there is a set of related gene clusters within a species or bigger taxonomic group undergoing successive HGT. The high proportion of *E. coli* O antigen gene clusters that have GTs present in inverse order relative to their function order implies that each of these gene clusters evolved at a time when related structures were present and subject to serotype switching at a frequency that made the pattern advantageous for the gene clusters. Note that the situation for the related *Y. pseudotuberculosis* O antigens is complicated by having variation in both the main-chain and a DDH sugar or related sugar side-branch, and gene function order being the same as transcription for the DDH sugar set of genes, discussed by Kenyon et al. (Kenyon, Cunneen and Reeves 2017).

**O antigens present in *Shigella* and *E. coli***

Twenty-two *Shigella* O antigens are also present in *E. coli,* and others are very similar to an *E. coli* O antigen (Knirel et al. 2016; Knirel et al. 2018), and in such cases the *Shigella* have the same or very similar O antigen gene cluster to that in the *E. coli* counterpart. This is not surprising as most *Shigella* are phylogenetically within *E. coli.* However, this situation is causing difficulty in some approaches to *in silico* serotyping, as strains of the alternate “genus” are confused with the target “genus”. Table S1 shows the current structures of relevant *Shigella* serotypes with repeats of the associated *E. coli* serotype from Table 1. These are the pairs of O antigens that may get confused, and the table (*vide infra*) shows the occasional minor differences in the structures.
References

Adeyeye A, Jansson PE, Lindberg B, Abaas S & Svenson SB (1988) Structural studies of the *Escherichia coli* O-149 O-antigen polysaccharide. *Carbohydr Res* 176: 231–236.

Cheng J, Wang Q, Wang W, Wang Y, Wang L & Feng L (2006) Characterization of *E. coli* O24 and O56 O antigen gene clusters reveals a complex evolutionary history of the O24 gene cluster. *Curr Microbiol* 53: 470-476

Dmitriev BA, Backinowsky LV, L’vov VL, Kochetkov NK & Hofman IL (1975a) Somatic antigen of *Shigella dysenteriae* type 3. Structural features of specific polysaccharide chain. *Eur J Biochem* 50: 539–547.

Dmitriev BA, Knirel YA, Kochetkov NK & Hofman IL (1975b) Selective cleavage of glycosidic linkages: studies with the polysaccharide component of *Shigella dysenteriae* type 6 lipopolysaccharide. *Carbohydr Res* 44: 77–85.

Dmitriev BA, Knirel YA, Kochetkov NK & Hofman IL (1976a) Somatic antigens of *Shigella*. Structural investigation on the O-specific polysaccharide chain of *Shigella dysenteriae* type 1 lipopolysaccharide. *Eur J Biochem* 66: 559–566.

Dmitriev BA, L’vov VL, Kochetkov NK, Jann B & Jann K (1976b) Cell-wall lipopolysaccharide of the ‘*Shigella*-like’ *Escherichia coli* O124. Structure of the polysaccharide chain. *Eur J Biochem* 64: 491–498.

Dmitriev BA, Knirel YA, Kochetkov NK, Hofman IL & Capek K (1977a) Somatic antigens of *Shigella*. Structural investigation on the O-specific polysaccharide chain of *Shigella dysenteriae* type-2 lipopolysaccharide. *Eur J Biochem* 76: 433–440.

Dmitriev BA, L’vov VL & Kochetkov NK (1977b) Complete structure of the repeating unit of the O-specific polysaccharide chain of *Shigella dysenteriae* type 3 lipopolysaccharide. *Carbohydr Res* 56: 207–209.

Dmitriev BA, L’vov VL, Kochetkov NK & Hofman IL (1977c) Antigenic polysaccharides of bacteria. VI. The structure of O-specific polysaccharide chain of *Shigella dysenteriae* type 4 lipopolysaccharide. *Bioorg Khim* 3: 1226–1233.

Dmitriev BA, Backinowsky LV, Knirel YA & Kochetkov NK (1977d) Somatic antigens of *Shigella*. The structure of the specific polysaccharide chain of *Shigella dysenteriae* type 5 lipopolysaccharide. *Eur J Biochem* 78: 381–387.

Dmitriev BA, Knirel YA, Kochetkov NK, Jann B & Jann K (1977e) Cell-wall lipopolysaccharide of the ‘*Shigella*-like’ *Escherichia coli* O58. Structure of the polysaccharide chain. *Eur J Biochem* 79: 111–115.

Dmitriev BA, Knirel YA, Vinogradov EV, Kochetkov NK & Hofman IL (1978) Antigenic polysaccharides of bacteria. 7. Structure of polysaccharide chain of the *Shigella dysenteriae* type 9 lipopolysaccharide. *Bioorg Khim* 4: 40–46.

Dmitriev BA, Knirel YA, Shremmet OK, Shashkov AA, Kochetkov NK & Hofman IL (1979) Somatic antigens of *Shigella*. The structure of the specific polysaccharide of *Shigella Newcastle* (*Sh. flexneri* type 6) lipopolysaccharide. *Eur J Biochem* 98: 309–316.
Feng L, Han W, Wang Q, Bastin DA & Wang L (2005) Characterization of Escherichia coli O86 O-antigen gene cluster and identification of O86-specific genes. *Vet Microbiol* 106: 241-248.

Feng L, Perepelov AV, Zhao G et al. (2007) Structural and genetic evidence that the *Escherichia coli* O148 O antigen is the precursor of the *Shigella dysenteriae* type 1 O antigen and identification of a glucosyltransferase gene. *Microbiol* 153: 139–147.

Frick DN, Townsend BD & Bessman MJ (1995) A novel GDP-mannose mannosyl hydrolase homology with the *MutT* family of enzymes. *J Biol Chem* 270: 24086-24091.

Hygge Blakeman K, Weintraub A & Widmalm G (1998) Structural and genetic evidence that the *Escherichia coli* O148 O antigen is the precursor of the *Shigella dysenteriae* type 1 O antigen and identification of a glucosyltransferase gene. *Microbiol* 153: 139–147.

Jakhetia R & Verma NK (2015) Identification and molecular characterisation of a novel Mu-like bacteriophage, SfMu, of *Shigella flexneri*. *PLoS One* 10: e0124053.

Jonsson KH, Weintraub A & Widmalm G (2006) Structural studies of the O-antigenic polysaccharides from *Shigella dysenteriae* type 3 and *Escherichia coli* O124, a reinvestigation. *Carbohydr Res* 341: 2986–2989.

Katzenellenbogen E, Mulczyk M & Romanowska E (1976) Structural studies on D-specific polysaccharide of *Shigella flexneri* serotype 6. *Eur J Biochem* 61: 191–197.

Kenne L, Lindberg B, Petersson K, Katzenellenbogen E & Romanowska E (1978) Structural Studies of *Shigella flexneri* O-Antigens. *Eur. J. Biochem.* 91: 279-284.

Kenyon JJ, Cunneen MM, Reeves PR, (2017) Genetics and evolution of *Yersinia pseudotuberculosis* O-specific polysaccharides: a novel pattern of O-antigen diversity. *Microbiol Rev* 41: 200-217.

Kido N & Kobayashi H (2000) A single amino acid substitution in a mannosyltransferase, WbdA, converts the *Escherichia coli* O9 polysaccharide into O9a: generation of a new O-serotype group. *J Bacteriol* 182: 2567-2573.

Knirel YA, Dashunin VV, Shashkov AS, Kochetkov NK, Dmitriev BA & Hofman IL (1988) Somatic antigens of *Shigella*: structure of the O-specific polysaccharide chain of the *Shigella dysenteriae* type 7 lipopolysaccharide. *Carbohydr Res* 179: 51–60.

Knirel YA, Qian C, Shashkov AS, Sizova OV, Zdorovenko EL, Naumenko OI, Senchenkova SN, Perepelov AV, Liu B (2016) Structural relationships between genetically closely related O-antigens of *Escherichia coli* and *Shigella spp*. *Biochemistry (Moscow)* 81: 600-608

Knirel YA, Qian C, Shashkov AS, Sizova OV, Zdorovenko EL, Naumenko OI, Senchenkova SN, Perepelov AV, Liu B (2018) Erratum to: Structural relationships between genetically closely related O-antigens of *Escherichia coli* and *Shigella spp*. *Biochemistry (Moscow)* 83: 1422-1423.

Landersjo C, Weintraub A & Widmalm G (1996) Structure determination of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* (EIEC) O143 by component analysis and NMR spectroscopy. *Carbohydr Res* 291: 209–216.

Lawrence, J. G. and J. R. Roth (1996) "Selfish operons: horizontal transfer may drive the evolution of gene clusters." *Genetics* 143: 1843-1860.

Linnerborg M, Wollin R & Widmalm G (1997) Structural studies of the O-antigenic polysaccharide from *Escherichia coli* O167. *Eur J Biochem* 246: 565–573.

Linnerborg M, Weintraub A & Widmalm G (1999a) Structural studies of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* O164 cross-reacting with *Shigella dysenteriae* type 3. *Eur J Biochem* 266: 460–466.
Linnerborg M, Weintraub A & Widmalm G (1999b) Structural studies utilizing $^{13}$C-enrichment of the O-antigen polysaccharide from the enterotoxigenic Escherichia coli O159 cross-reacting with Shigella dysenteriae type 4. Eur J Biochem 266: 246–251.

Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, Reeves PR & Wang L. Structure and genetics of Shigella O antigens. FEMS Microbiol Rev 32: 627-653.

Liu B, Perepelov AV, Guo D, Shevelev SD, Senchenkova SN, Feng L, Shashkov AS, Wang L & Knirel YA (2010) Structural and genetic relationships between the O-antigens of Escherichia coli O118 and O151. FEMS Immunol Med Microbiol 60: 199-207.

L’vov VL, Dashunin VM, Ramos EL, Shashkov AS, Dmitriev BA & Kochetkov NK (1983) Somatic antigens of Shigella: the structure of the polysaccharide chain of Shigella boydii type 2 lipopolysaccharide. Carbohydr Res 124: 141–149.

L’vov VL, Lapina EB, Belaya OF, Pluzhnikova GN, Shashkov AS & Dmitriev BA (1986) Somatic antigens of Shigella and Escherichia coli. Structure elucidation of O-specific polysaccharide from Shigella boydii type 12, and investigation of its serological properties in comparison with O-specific polysaccharide from E. coli O7. Bioorg Khim 12: 1240–1252.

Olsson U, Lycknert K, Stenutz R, Weintraub A & Widmalm G (2005) Structural analysis of the O-antigen polysaccharide from Escherichia coli O152. Carbohydr Res 340: 167–171.

Parolis H, Parolis LA & Olivieri G (1997) Structural studies on the Shigella-like Escherichia coli O121 O-specific polysaccharide. Carbohydr Res 303: 319–325.

Perepelov AV, Wang Q, Senchenkova SN, Shevelev SD, Zhao G, Shashkov AS, Feng L, Knirel YA & Wang L (2006a) Structure of a teichoic acid-like O-polysaccharide of Escherichia coli O29. Carbohydr Res 341: 2176–2180.

Perepelov AV, Han W, Senchenkova SN, Shashkov AS, Feng L, Knirel YA & Wang L (2006b) Structure of the O-polysaccharide of Escherichia coli O150 containing a regioisomer of N-acetylmuramic acid, 2-acetamido-4-O-[(S)-1-carboxyethyl]-2-deoxy-D-glucose. Carbohydr Res 342: 648–652.

Perepelov AV, Liu B, Senchenkova SN, Shashkov AS, Feng L, Knirel YA & Wang L (2007a) Close relation of the O-polysaccharide structure of Escherichia coli O168 and revised structure of the O-polysaccharide of Shigella dysenteriae type 4. Carbohydr Res 342: 2676–2681.

Perepelov AV, Liu B, Senchenkova SN, Shevelev SD, Wang V, Shashkov AS, Feng L, Wang L & Knirel YA (2007b) The structure of the glycerophosphate-containing O-specific polysaccharide from Escherichia coli O130. Bioorg Khim 33: 64–68.

Perepelov AV, Liu B, Senchenkova SN, Shashkov AS, Feng L, Knirel YA & Wang L (2008a) Structure of the O-polysaccharide of Escherichia coli O112ab containing L-iduronic acid. Carbohydr Res 343:571–575.

Perepelov AV, Weintraub A, Liu B, Senchenkova SN, Shashkov AS, Feng L, Wang L, Widmalm G & Knirel YA (2008b) The O-polysaccharide of Escherichia coli O112ac has the same structure as that of Shigella dysenteriae type 2 but is devoid of O-acetylation; a revision of the S. dysenteriae type 2 O-polysaccharide structure. Carbohydr Res 343: 977–981.

Perepelov AV, Senchenkova SN, Shashkov AS, Liu B, Feng L, Wang L & Knirel YA (2008c) Antigenic polysaccharides of bacteria. 40. Revision of the structures of the O-specific polysaccharides of Shigella dysenteriae types 3, 9 and Shigella boydii type 4 by NMR spectroscopy. Bioorg Khim 34: 121–128.
Perepelov AV, Senchenkova SN, Shashkov AS, Knirel YA, Liu B, Feng L & Wang L (2008d) Antigenic polysaccharides of bacteria. 41. Structures of the O-specific polysaccharides of *Shigella dysenteriae* types 4 and 5: a revision based on NMR spectroscopy data. *Bioorg Khim* **34**: 460–467.

Perepelov AV, Liu B, Senchenkova SN, Shashkov AS, Knirel YA, Feng L & Wang L (2009) A revised structure of the O-specific polysaccharide of *Shigella dysenteriae* type 10. *Bioorg Khim* **35**: 142–144.

Perepelov AV, Shevelev SD, Liu B, Senchenkova SN, Shashkov AS, Feng L, Knirel, YA & Wang L (2010) Structures of the O-antigens of *Escherichia coli* O13, O129 and O135 related to the O-antigens of *Shigella flexneri*. *Carbohydr Res* **345**: 1594-1599.

Reeves PR & Wang L (2002) Genomic organization of LPS-specific loci. Curr Top Microbiol Immunol **264**: 109-135.

Senchenkova SN, Feng L, Yang J, Shashkov AS, Cheng J, Liu D, Knirel YA, Reeves PR, Jin Q, Ye Q, Wang L (2005) Structural and genetic characterization of the *Shigella boydii* type 10 and type 6 O antigens. *J Bacteriol* **187**: 2551-2554.

Sturm S, Jann B, Jann K, Fortnagel P & Timmis KN (1986) Genetic and biochemical analysis of *Shigella dysenteriae* 1 O antigen polysaccharide biosynthesis in *Escherichia coli* K-12: 9 kb plasmid of *S. dysenteriae* 1 determines addition of a galactose residue to the lipopolysaccharide core. *Microb Pathog* **1**: 299–306.

Sugiyama T, Kido N, Kato Y, Koide N, Yoshida T & Yokochi T (1998) Generation of *Escherichia coli* O9a serotype, a subtype of *E. coli* O9, by transfer of the wb* gene cluster of *Klebsiella* O3 into *E. coli* via recombination. *J Bacteriol* **180**: 2775-2778.

Tian T & Salis HM (2015) A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons. *Nucleic Acids Res* **43**: 7137-7151.

Yang B, Senchenkova SN, Naumenko OI, Shashkov AS, Liu B, Knirel YA (2018) Structural and genetic relatedness of the O antigens of *Escherichia coli* O50 and O2. *Carbohydr Res* **15**: 8-11.

Wang Q, Perepelov AV, Wen L, Shashkov AS, Wang X, Guo X, Knirel YA & Wang L (2012) Identification of the two glycosyltransferase genes responsible for the difference between *Escherichia coli* O107 and O117 O-antigens. *Glycobiology* **22**: 281-287.

Wang W, Perepelov AV, Feng L, et al. (2007) A group of *Escherichia coli* and *Salmonella enterica* O antigens sharing a common backbone structure. *Microbiology* **153**: 2159-2167.

Zhao G, Perepelov AV, Senchenkova SN, Shashkov AS, Feng L, Li X, Knirel YA & Wang L (2007) Structural relation of the antigenic polysaccharides of *Escherichia coli* O40, *Shigella dysenteriae* type 9, and *E. coli* K47. *Carbohydr Res* **342**:1275–1279.
Figure S1. The O antigen clusters of 178 *E. coli* serotypes (including eight subgroups). Open arrows represent the location and orientation of putative genes.
**Figure S2.** Relationship of gene order and gene function in *Y. pseudotuberculosis.*

Top: the columns show the genes in five related O antigen gene clusters, with their genes in map order. The initial transferase (IT) gene is not in the gene cluster. The gaps in the alignment are to keep homologous genes in the same row. The yellow cells have genes involved in synthesis of dinucleotide sugars used by the GTs, while the green cells have the GT genes, Wzx flippase genes, Wzy polymerase genes and the Wzz co-polymerase gene, which together build and process the O units; these O unit assembly pathway genes generally act in a known order because the O units are built as unbranched linear structures that are acted on by Wzx and then by Wzy/Wzz. Arrows on the right side show the transcription/translation order and the function order of the genes. The genes are all transcribed in map order from bottom to top. However, function order of gene products is mostly from top to bottom for the main-chain pathway green genes, so in the reverse order to transcription order: this is very clear for the two or three GT genes and also applies to most of the yellow sugar pathway genes in relation to where that sugar is added to the structure. However, function order for a small set of genes at the 5’-end that code for synthesis of a CDP-DDH sugar is from bottom to top; thus, in the same order as transcription. Between the two sets of pathway genes there are genes for the DDH sugar GT genes that join the two components to complete the O units, and also the Wzx flippase genes; the gene products of the latter flip the O units to the periplasm, in effect the final steps for both main-chain and DDH-sugar pathways. The subsequently completed O antigen is then facing the periplasm for transfer to lipid A-core by the ligase. The red bars between adjacent columns show the DNA segments that have to be transferred by recombination for serotype switching between the two gene clusters to occur.

Bottom: The O units are shown below as linear oligosaccharides prior to polymerisation, with the sugars in blue boxes.
Table S1. O antigen structures of *Shigella* and corresponding *E. coli* serogroups.\(^1\)

| Shigella | E. coli | Structure | References |
|----------|---------|-----------|------------|
| B1       | O149    | (β1→3) → 3)-dGlcNAc46SPyr-(β1→3)-LRha-(β1→4)-dGlcNAc-(β1→<br> | Liu et al. (2008) |
|          |         | | Adeyev et al. (1988) |
| B3\(^\#\) | O167    | Gro-(2-N-6) (β1→3)-dGal-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Liu et al. (2008) |
|          |         | | Linnerborg et al. (1997) |
| B4       | O53     | (α1→2) → 3)-LRha-(β1→4)-dGlcA-(β1→3)-LRha-(β1→4)-dGlcNAc-(β1→<br> | Knirel et al. 2016 |
|          |         | | Knirel et al. 2018 |
| B5       | O79     | (β1→3) → 2)-dGal-(β1→4)-dGal6Ac-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Knirel et al. 2016 |
|          |         | | Knirel et al. 2018 |
| B8\(^\#\) | O143    | Gro-(2-N-6) (β1→3)-dGal-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | L’vov et al. (1983) |
|          |         | | Landersjo et al. (1996) |
| B10      | O183    | (β1→3) → 4)-dRibf-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Knirel et al. 2016 |
|          |         | | Knirel et al. 2018 |
| B11      | O105    | (β1→3) → 4)-dRibf-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Knirel et al. 2016 |
|          |         | | Knirel et al. 2018 |
| B12      | O7      | (β1→3) → 4)-dRibf-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | L’vov et al. (1984) |
|          |         | | L’vov et al. (1986) |
| B15\(^\#\) | O112ab  | (β1→3) → 4)-dRibf-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Liu et al. (2008) |
|          |         | | Perepelov et al. (2008a) |
| D1       | O148    | (β1→3) → 4)-dRibf-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Dmitriev et al. (1976a) |
|          |         | | Sturm et al. (1986) |
|          |         | | Peng et al. (2007) |
| D2       | O112ac\(^\#\) | (β1→3) → 4)-dRibf-(β1→4)-dGlcA-(β1→3)-dMan-(β1→4)-dGlcA-(β1→3)-dGlcNAc-(β1→<br> | Dmitriev et al. (1977a) |
|          |         | | Perepelov et al. (2008b) |
| D3 | O124 | DGlc4RLac-(β1→6)-dGlc-(α1→) |
|-----|------|-------------------------------|
|     |      | -3)-dGal-(β1→6)-dGalf-(β1→3)-dGalNAC-(β1→) |
|     |      | DGlc-(β1→6)-dGlc-(α1→4) |
|     |      | -3)-dGal-(β1→6)-dGalf-(β1→3)-dGalNAC-(β1→) |

| D4 | O168' | tFucp2Ac3Ac4Ac-(α1→3) |
|----|-------|------------------------|
|     |       | -4)-dGlcNAC6Ac-(α1→4)-dGlcA-(α1→3)-tFuc-(α1→3)-dGlcNAC-(β1→) |
|     | O159  | tFuc-(α1→4) |
|     |       | -3)-dGlcNAC-(β1→4)-dGalA-(α1→3)-tFuc-(α1→3)-dGlcNAC-(β1→) |

| D5 | O58   | tRha2Ac3RLac-(α1→3) |
|----|------|----------------------|
|     |      | -4)-dMan-(β1→4)-dMan-(α1→3)-dGlcNAC-(β1→) |

| D6(SR-form) vi | O130 | Gro-(2-P-4)-dGalNAC3Ac-(β1→3) |
|----------------|------|-------------------------------|
|                |      | dGal-(α1→6)-dGlc-(β1→3)-dGalNAC-(β1→) |
|                |      | Gro-(2-P-4)-dGalNAC-(β1→3) |
|                |      | -4)-dGal-(α1→6)-dGlc-(β1→3)-dGalNAC-(β1→) |

| D7 vii | O121 viii | -3)-dQui4NGlyAc-(β1→4)-dGalNAcAN3Ac-(α1→4)-dGalNAcA-(α1→3)-dGlcNAC-(α1→) |

| D9 viii | O40   | -2)-dGal34RPyr-(β1→4)-dMan-(β1→4)-dGal2Ac-(α1→3)-dGlcNAC-(β1→) |
|--------|------|-------------------------------------------------------------------|
|        |      | -2)-dGal-(β1→4)-dMan-(β1→4)-dGal-(α1→3)-dGlcNAC-(β1→) |

| D11 iv | O29   | DGlc-(α1→6)-dGal2Ac-(α1→3) |
|--------|------|-----------------------------|
|        |      | -3)-dGro-(1-P-6)-dGlc-(β1→4)-tFucNAC-(α1→3)-dGlcNAC-(β1→) |

| D12   | O152  | tRha-(β1→4) |
|-------|------|-------------|
|       |      | -3)-dGlcNAC-(α1→P-6)-dGlc-(α1→2)-dGlc-(β1→3)-dGlcNAC-(β1→) |

| D13   | O150  | DGlc-(β1→2) |
|-------|------|-------------|
|       |      | -3)-dGlcNAC4SLac-(β1→2)-tRha-(α1→2)-4GlcNAC-(α1→3)-tLrha-(α1→3)-4GlcNAC-(β1→) |
|   | Structure                                                                 | Reference(s)                          |
|---|---------------------------------------------------------------------------|---------------------------------------|
| F1-5 | \(-\alpha_1 \rightarrow 2\)-L-Rha-(-\alpha_1 \rightarrow 3)-L-Rha-(-\alpha_1 \rightarrow 3)-\text{D}GlcNAc6Ac(-\beta_1 \rightarrow \text{nGlc}(-\alpha_1 \rightarrow 2)) | Kenne et al. (1978) Liu et al. (2008) |
| O13  | \(-\alpha_1 \rightarrow 2\)-L-Rha-(-\alpha_1 \rightarrow 3)-L-Rha-(-\alpha_1 \rightarrow 3)-\text{D}GlcNAc6Ac(-\beta_1 \rightarrow \text{nGlc}(-\alpha_1 \rightarrow 2)) | Perepelov et al. (2010)               |
| F6v  | \(-\alpha_1 \rightarrow 2\)-L-Rha3Ac-(-\alpha_1 \rightarrow 4)-\text{D}GalA(-\beta_1 \rightarrow \text{D}GalNAc(-\beta_1 \rightarrow \text{nGlc}(-\alpha_1 \rightarrow 2)) | Katzenellenbogen et al. (1976) Dmitriev et al. (1979) Blakeman et al. (1998) |

i The reference to the article, in which the correct *Shigella* O polysaccharide structure was reported for the first time, is indicated in bold face. ii One of two possible structures of the biological O unit. iii The structure of the biological O unit was established. iv The *E. coli* O polysaccharide lacks O-acetyl groups. v The *E. coli* O polysaccharide lacks O-acetyl groups on the Fuc residue. vi In the *S. dysenteriae* O antigen, the degree of O-acetylation is c. 55%. vii In the *E. coli* O polysaccharide, the degree of O-acetylation is c. 60%. viii The O-acetyl group was not confirmed upon reinvestigation (Zhao et al., 2007; Perepelov et al., 2008e).
Table S2. Homology groups of glycosyl transferases in *E. coli* O antigen gene clusters.