OXFORD
FEMS JOURNALS investing in science

FEMS Microbiology Reviews, fux002, 41, 2017, 200–217

doi: 10.1093/femsre/fux002
Advance Access Publication Date: 15 February 2017
Review Article

REVIEW ARTICLE

Genetics and evolution of *Yersinia pseudotuberculosis* O-specific polysaccharides: a novel pattern of O-antigen diversity

Johanna J. Kenyon[1,2], Monica M. Cunneen[1] and Peter R. Reeves[1,*]

[1]School of Molecular Bioscience, The University of Sydney, Sydney, NSW 2006, Australia and [2]Institute of Health and Biomedical Innovation, Queensland University of Technology. Brisbane, QLD 4001, Australia

[*]**Corresponding author:** School of Molecular Bioscience, The University of Sydney, Building G08, Sydney, NSW 2006, Australia. Tel: +61293512536; E-mail: peter.reeves@sydney.edu.au

One sentence summary: O-antigen gene clusters in the *Yersinia pseudotuberculosis* complex.
Editor: Chris Whitfield

ABSTRACT

O-antigen polysaccharide is a major immunogenic feature of the lipopolysaccharide of Gram-negative bacteria, and most species produce a large variety of forms that differ substantially from one another. There are 18 known O-antigen forms in the *Yersinia pseudotuberculosis* complex, which are typical in being composed of multiple copies of a short oligosaccharide called an O unit. The O-antigen gene clusters are located between the *hemH* and *gsk* genes, and are atypical as 15 of them are closely related, each having one of five downstream gene modules for alternative main-chain synthesis, and one of seven upstream modules for alternative side-branch sugar synthesis. As a result, many of the genes are in more than one gene cluster. The gene order in each module is such that, in general, the earlier a gene product functions in O-unit synthesis, the closer the gene is to the 5′ end for side-branch modules or the 3′ end for main-chain modules. We propose a model whereby natural selection could generate the observed pattern in gene order, a pattern that has also been observed in other species.

Keywords: *Yersinia pseudotuberculosis*; O antigen; O-specific polysaccharide; lipopolysaccharide; gene cluster; serotype

INTRODUCTION

The O-specific polysaccharide (OPS, also known as O antigen) is the most variable surface antigen for many Gram-negative bacteria, although it is a component of the otherwise quite well-conserved lipopolysaccharide (LPS), which is a hallmark structural entity (Erridge *et al.* 2002; Heinrichs, Yethon and Whitfield 2002). For many species, LPS is essential for membrane stability and cell survival (Zhang, Meredith and Kahne 2013), and is a key virulence determinant that provides resistance to phagocytosis, complement-mediated killing, antimicrobial peptides and lipophilic agents (Porat, McCabe and Brubaker 1995; Bengoechea, Najdenski and Skurnik 2004; West *et al.* 2005; Trent *et al.* 2006; Pier 2007; Plainvert *et al.* 2007; Conde-Álvarez *et al.* 2012; March *et al.* 2013). LPS molecules are characteristically composed of three structural segments: lipid A, which anchors the LPS in the outer membrane; the core oligosaccharide that contains inner-core and outer-core components; and the highly immunogenic OPS, which is a variable-length polymer of repeating oligosaccharide units (O units), each containing several different carbohydrate residues that can have acetyl or other groups attached (Erridge *et al.* 2002).

The compositions of lipid A and inner-core oligosaccharide are generally conserved within a genus, although minor structural variations occur (Frirdich and Whitfield 2005). The outer-core oligosaccharide can exist in discrete forms within a species, as in *Escherichia coli* (Whitfield, Kaniuk and Frirdich 2003), or can

Received: 26 October 2016; **Accepted:** 2 February 2017
© FEMS 2017. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in anymedium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

be quite variable in groups that lack OPS, such as *Neisseria* and *Haemophilus* (Gibson *et al.* 1993). However, in many cases, OPS contributes by far the most to cell surface diversity in Gram-negative species (Heinrichs, Yethon and Whitfield 2002). In particular, the OPS can vary considerably in composition and/or structural arrangement (Samuel and Reeves 2003). These differences form the basis of O-serotyping schemes that are often used to classify strains for epidemiological purposes. Extensive intraspecies structural diversity is widely accepted as a characteristic of O antigens, and this feature may be associated with enhancing bacterial evasion of the host immune response (Finlay and McFadden 2006).

O antigens may be synthesised and exported via one of three different processing systems (reviewed in Reeves and Cunneen 2009). In the Wzx/Wzy-dependent pathway, synthesis begins in the cytoplasm with the production of nucleotide-diphosphate (NDP)-sugar precursors. The first sugar of the O unit is transferred as a sugar phosphate to an undecaprenyl phosphate (UndP) lipid carrier on the cytoplasmic face of the inner membrane. In many *Enterobactericeae*, the first sugar is *N*-acetyl-D-glucosamine (D-Glc*p*NAc), which is attached by the WecA-initiating transferase (Lehrer *et al.* 2007). In some cases, the UndPP-D-Glc*p*NAc product is converted to UndPP-*N*-acetyl-D-galactosamine (UndPP-D-Gal*p*NAc) by the Gnu epimerase (see Cunneen *et al.* 2013), and D-Gal*p*NAc becomes the first sugar of the O unit. Glycosyltransferases (GTs) then transfer other activated sugar precursors onto the UndPP-linked substrates, and once the O unit is complete, it is translocated across the inner membrane by a Wzx flippase. O units can then be linked together by a Wzy polymerase to form a polymer, with chain length and modality determined by Wzz (Woodward *et al.* 2010; Kenyon and Reeves 2013). Both single O units and OPS are then ligated to lipid A-core by the WaaL O-antigen ligase, and the LPS products are exported to the cell surface by Lpt export machinery (Silhavy, Kahne and Walker 2010).

The majority of the genes that direct synthesis of OPS are clustered, and the chromosomal location is usually conserved within a species, with major loci being between the *galF* and *gnd* genes in *E. coli*, *Salmonella enterica* and *Shigella*, and between *hemH* and *gsk* in many *Yersinia* (Reeves and Cunneen 2011). The reason for the genes being clustered is thought to be that it allows exchange of the whole gene cluster by recombination (Lawrence 1997). O-antigen gene clusters include genes for the synthesis of NDP-sugar precursors, glycosyl transfer, O-unit modification and O-antigen processing (Reeves and Wang 2002). However, genes for synthesis of sugars that are also required for other cellular processes are generally located elsewhere (Samuel and Reeves 2003). The *waaL* gene is also located outside the O-antigen gene cluster, usually with the genes for the outer-core oligosaccharide of the LPS (Reeves and Wang 2002). Additionally, for the many *Enterobactericeae* species that produce WecA-initiated O units, the *wecA* gene is located in the gene cluster for the enterobacterial common antigen (ECA) (Meier-Dieter *et al.* 1992).

Over the past few decades, the O antigens of several *Enterobacteriaceae* species have been studied in depth. Much of our understanding of O antigens comes from studies on *E. coli*, which produces more than 184 different serologically classified O-antigen types (DebRoy, Roberts and Fratamico 2011; Iguchi *et al.* 2015). The O antigens of *Shigella* and *S. enterica* have also been studied extensively, and have been reviewed (Liu *et al.* 2008, 2014). In most species, the majority of O-unit structures and gene clusters are not significantly related to any others in the same species, providing little information on the origins of the diversity. An exception to this is 8 of the 54 *S. enterica* serogroups that have D-galactose (D-Gal*p*) as the first sugar. These structures are clearly related, and comparisons give insights into their evolution (Reeves *et al.* 2013).

Yersinia pseudotuberculosis is a distantly related species in the same *Enterobacteriaceae* family (Paradis *et al.* 2005; Hata *et al.* 2016), and has 21 types in the O-antigen-based serotyping scheme (Tsubokura and Aleksic 1995). The O-unit structures and gene cluster sequences have been determined for isolates representing 20 of these (Reeves, Pacinelli and Wang 2003; Cunneen *et al.* 2009, 2011; De Castro *et al.* 2009, 2011, 2012; Kenyon *et al.* 2011; Beczala *et al.* 2013; Kenyon *et al.* 2016), and many of them are closely related, as for the D-Gal*p*-initiated O antigens of *S. enterica* (Reeves *et al.* 2013). The relationships between *Y. pseudotuberculosis* O antigens were last reviewed in 2003 (Reeves, Pacinelli and Wang 2003; Skurnik and Bengoechea 2003) when only 11 gene cluster sequences were available. Here we review the current data for the OPS gene clusters of *Y. pseudotuberculosis*, and provide an overview of their evolutionary relationships, including a model for generating a common pattern of OPS gene order.

BIOLOGY OF THE GENUS *YERSINIA*

The genus *Yersinia* includes 18 species, of which *Y. pseudotuberculosis, Y. pestis* and *Y. enterocolitica* are pathogenic, but as *Y. pestis* is effectively a clone of *Y. pseudotuberculosis* (Bercovier *et al.* 1980; Achtman *et al.* 1999), there are currently only two full species that are of clinical significance. The other species are commonly found in soil and water, and generally not pathogenic (McNally *et al.* 2016). It is interesting that *Y. pseudotuberculosis* and *Y. enterocolitica* are among the most divergent of the species, and are now thought to have gained pathogenicity independently, although they cause similar gastrointestinal diseases in humans and also in animals. They also share pathogenicity islands and other virulence factors, now proposed to have been gained independently, presumably initially from other genera, but perhaps reaching the second species by transfer within the genus (Reuter *et al.* 2014).

As genome sequences became available, many isolates first typed as *Y. pseudotuberculosis* were reclassified into other species that are now included in a group known as the '*Y. pseudotuberculosis* complex' (Laukkanen-Ninios *et al.* 2011). The species include *Y. pseudotuberculosis/Y. pestis*, *Y. similis* (Sprague *et al.* 2008) and the newly characterised *Y. wautersii* (previously referred to as the 'Korean group') that is proposed to have pathogenic potential (Savin *et al.* 2014).

Y. pseudotuberculosis is generally considered an enteric pathogen, but it has also been associated with other medical complications such as arthritis, erythema nodosum, desquamation, rash, pneumonia and nephritis (Carniel *et al.* 2006). Yersiniosis caused by *Y. pseudotuberculosis* is often acquired through the ingestion of contaminated food, but zoonotic transmission is possible, and outbreaks have been reported in Finland, Russia and Japan (Nakano *et al.* 1989; Jalava *et al.* 2004; Nuorti *et al.* 2004; Pärn *et al.* 2015; Timchenko *et al.* 2016). However, not all strains are capable of causing severe infections in humans (Nagano *et al.* 1997). Successful host colonisation, survival and persistence rely on a multitude of facultative virulence factors, which most notably include a high pathogenicity island, a 70-kb pYV virulence plasmid, a YPM superantigen and the LPS (Carniel 2002; Carniel *et al.* 2006). Interestingly, expression of the O-antigen component of the *Y. pseudotuberculosis* LPS is downregulated at 37°C (Ho *et al.*

2008), and this is also true for other *Yersinia* species (Bengoechea *et al.* 2002; Skurnik and Bengoechea 2003). However, the O antigen is required for virulence (Mecsas, Bilis and Falkow 2001), as well as for protection against antimicrobial chemokines such as polymyxin B (Erickson *et al.* 2016). Thus, downregulation may be delayed until the later stages of infection (Ho *et al.* 2008).

O ANTIGENS IN *YERSINIA*

The O antigens of the major species, *Y. pseudotuberculosis* and *Y. enterocolitica*, have been studied the most extensively, and have very different sets of O antigens that may be synthesised via one of two different pathways. *Y. pseudotuberculosis* O antigens are synthesised by the Wzx/Wzy-dependent pathway, and genes for synthesis of these structures are clustered between conserved *hemH* and *gsk* genes. This set will be discussed in detail below. There are over 70 *Y. enterocolitica* serotypes, but only 11 have been associated with human diseases (Garzetti *et al.* 2014), and genetic analysis has focussed on the O:3, O:8 and O:9 serotypes (Skurnik and Bengoechea 2003), of which O:8 has the Wzx/Wzy-dependent pathway and the others the ABC-transporter pathway. The gene cluster at the *hemH-gsk* locus for the O:8 structure has been well documented (Zhang *et al.* 1997; Bengoechea *et al.* 2002; Skurnik and Bengoechea 2003), and has a typical arrangement for a Wzx/Wzy pathway structure (Fig. 1). However, the O:3 and O:9 gene clusters do not map to the *hemH-gsk* locus, but rather to the *galF-gnd* locus (Zhang *et al.* 1993; Skurnik *et al.* 2007), as is typical for Wzx/Wzy pathway, *E. coli* and *S. enterica* O-antigen gene clusters, and also *E. coli* group 1 capsule gene clusters. The O:3 and O:9 strains have a different gene cluster at the *hemH-gsk* locus (Fig. 1), which directs the synthesis of a structure that has been called the outer core of the LPS (Skurnik *et al.* 1995). This outer core consists of what is in effect a single O unit that is synthesised by a pathway involving Wzx, but is not polymerised after translocation. In both serotypes, LPS molecules can have either the ABC pathway O antigen or the outer core (Skurnik *et al.* 1999). The eight *Y. enterocolitica* structures reported (Knirel 2011) include six with a homopolymeric O antigen or main chain common in ABC-transporter repeat units, and the species may well have a significant number of O antigens resembling O:3 and O:9 in their biosynthesis pathway and gene cluster location.

The O-antigen gene cluster of *Y. kristensenii* serotype O:11 is the only other fully annotated sequence for the genus outside of the *Y. pseudotuberculosis* complex (Fig. 1). As for *Y. enterocolitica* O:3 and O:9, this gene cluster is an example of O-antigen genes being found outside of the *hemH-gsk* locus, in this case between *aroA* and *cmk* elsewhere in the genome. Interestingly, it is flanked by remnants of *galF* and *gnd* genes. The gene cluster is very similar to the *E. coli* O98 gene cluster, which suggested that it had been imported from an *E. coli* relative (Cunneen and Reeves 2007). The *hemH-gsk* locus in *Y. kristensenii* O:11 contains a ∼15-kb uncharacterised gene cluster adjacent to *wbcQ* and *gne* genes (Fig. 1). These genes are also found in the *Y. enterocolitica* O:3 and O:9 outer-core gene cluster at the same locus, and it is probable that the same outer-core gene cluster is present in *Y. kristensenii* O:11.

O SEROTYPES IN THE *YERSINIA PSEUDOTUBERCULOSIS* COMPLEX

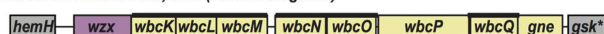
The O-serotyping scheme for *Y. pseudotuberculosis* was formally established in 1971 (Thal and Knapp 1971), and now has a total of 21 serotypes, including 6 originally classified as subtypes of either O:1, O:2, O:4 or O:5 (Tsubokura *et al.* 1984, 1993; Aleksic

hemH/gsk locus

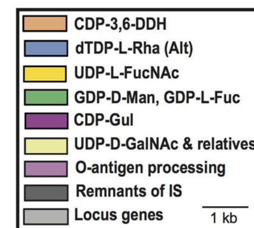
Y. mollaretii ATCC 43969

Y. enterocolitica O:8

Y. enterocolitica O:3, O:9 (*Outer core genes*)



Y. kristensenii O:11



galF/gnd locus

Y. enterocolitica O:9

Y. enterocolitica O:3

aroK/cmkA locus

Y. kristensenii O:11

Figure 1. O-antigen gene clusters of *Yersinia* spp. outside of *Y. pseudotuberculosis*. Genes are coloured according to the respective pathways of their products, and the scheme is shown on the right. Figure is drawn to scale and the scale is shown below. Bold lines bordering genes indicate that the gene encodes a GT.

et al. 1991; Tsubokura and Aleksic 1995). However, while these subtypes are now treated as types following the convention set for *S. enterica* in which variation in the main gene clusters is used to define types, the use of names such as O:1a, O:2b, etc. has been retained for historical reasons. The epidemiology and geographical distribution of *Y. pseudotuberculosis* serotypes have been summarised previously (Fukushima *et al.* 2001; Carniel *et al.* 2006).

Yersinia pestis isolates carry genes for the O:1b serotype, suggesting that it emerged from a *Y. pseudotuberculosis* O:1b progenitor (Duan *et al.* 2014). However, in *Y. pestis*, four O-antigen genes have inactivating mutations, and an O antigen is not produced (Skurnik, Peippo and Ervela 2000; Prior, Hitchen and Williamson 2001; Bogdanovich *et al.* 2003; Skurnik and Bengeochea 2003). Interestingly, genes specific to many of the *Y. pseudotuberculosis* O-antigen serotypes were identified in other species in the *Y. pseudotuberculosis* complex (Laukkanen-Ninios *et al.* 2011; Savin *et al.* 2014), which led to the proposal that the O-antigen serotyping scheme should apply to all members of this complex (Laukkanen-Ninios *et al.* 2011; De Castro *et al.* 2012).

Since the establishment of the current serotyping scheme, anomalies have been reported for serotypes O:8, O:13 and O:14. The O:8 form has been shown to be a 'rough' mutant of either serotype O:4a or O:1b (i.e. lipid A-core without attached O antigen) (Tsubokura *et al.* 1993; Kenyon *et al.* 2016). It has also been shown that an O:14 isolate has a complete copy of the O:11 gene cluster, but is also a rough mutant as the isolate did not produce O antigen (Cunneen *et al.* 2009), while O:13 isolates carried genes characteristic of either O:1a, O:1b or O:3 (Bogdanovich *et al.* 2003). The structural basis for the O:13 and O:14 epitopes has not been identified, but they are clearly not O antigens, and only the 18 validated serotypes will be discussed.

O-UNIT STRUCTURES

O-unit structures for the 18 O serotypes are shown in Fig. 2, some of which have been reviewed together previously (Bruneteau and Minka 2003; Knirel 2011). Each O unit consists of a di- to tetrasaccharide main chain that contains either D-Glc*p*NAc or D-Gal*p*NAc as the first sugar (see below), with one or two sugars present in side branches. Many O units also contain L-fucose (L-Fuc*p*), D-mannose (D-Man*p*) and D-galactose (D-Gal*p*), and there are single cases of L-quinovose (L-Qui*p*) in O:12, and 3-*O*-acetyl-*N*-acetyl-D-glucosaminuronic acid (D-Glc*p*NAcA3OAc) and 2,6-dideoxy-2-acetamidino-L-galactose (L-Fuc*p*NAm) in O:9. Most O-unit sugars have pyranose (*p*) rings; the exceptions being two furanose (*f*) structures: paratofuranose (Par*f*) and L-altrofuranose (L-Alt*f*). Abbreviations of sugar names are expanded in Table S1 (Supporting Information).

Thirteen O-unit structures include a 3,6-dideoxyhexose (DDH) side-branch sugar. Although these sugars are considered rare in nature, six different forms are found in *Yersinia pseudotuberculosis*: Par*f* (O:1a, O:1b, O:1c, 15), paratopyranose (Par*p*; O:3), tyvelose (Tyv*p*; O:4a, O:4b), abequose (Abe*p*; O:2a, O:2b, O:2c), ascarylose (Asc*p*; O:5a) and L-colitose (L-Col*p*; O:6, O:7, O:10). The O:6 and O:12 O units include a related sugar known as yersiniose(A) (Yer(A)*p*), which has a 2-carbon addition to the hexose base (Gorshkova *et al.* 1983; De Castro *et al.* 2012), and the L-Alt*f* residue in the O:5b and O:11 O units is a 6-deoxy sugar also related to the DDH sugars (Korchagina, Gorshkova and Ovodov *et al.* 1982; Cunneen *et al.* 2009). The only O unit that does not contain a DDH or related sugar is that of serotype O:9 (Beczala *et al.* 2013).

Many of the O units fall into one of five different groups based on common main-chain structures (Fig. 2), each of which is associated with two or three different side-branch options that are generally present in more than one group. The O:12 O unit has some similarity to the O:1b/O:11 group, and the O:6, O:7 and O:10 O units form a further group, defined by the presence of L-Col*p* side branches. The O:9 O unit shares no similarity to any other *Y. pseudotuberculosis* O units, and forms a separate group. Most structures in Fig. 2 are unique to the *Y. pseudotuberculosis* complex, the exceptions being O:10, which is closely related to the *Escherichia coli* O111 and *Salmonella enterica* O35 structures (Kenyon *et al.* 2011), and the O:2c/O:4a main chain that is identical to that of the complete *S. enterica* O:18 (K) O antigen (Vinogradov, Nossovaa and Radziejewska-Lebrechtb 2004).

Initiation of O-unit synthesis

As for most other *Enterobacteriaceae*, a *wecA*-initiating transferase gene is located in the ECA gene cluster (Pacinelli, Wang and Reeves 2002), and since D-Glc*p*NAc and D-Gal*p*NAc are present in each O unit, initiation of O-unit synthesis is inferred to be by WecA. A *gnu* gene (previously annotated as *gne* or *gne2*), which is required for the reversible conversion of UndPP-D-Glc*p*NAc to UndPP-D-Gal*p*NAc, is present in the gene clusters of strains with O units containing D-Gal*p*NAc instead of D-Glc*p*NAc (reviewed in Cunneen *et al.* 2013). Where there is a second D-Gal*p*NAc sugar (O:6 and O:7), the gene cluster also includes a *gne* gene (previously annotated as *gne1*) for epimerisation of UDP-D-Glc*p*NAc to UDP-D-Gal*p*NAc (Cunneen *et al.* 2011).

GENE CLUSTERS FOR OPS BIOSYNTHESIS

All *Y. pseudotuberculosis* OPS gene clusters are located between conserved genes *hemH* and *gsk*, and are 14–29 kb in length (Fig. 3; GenBank accession numbers in Table 1). Between *hemH* and the first gene, there is always a JUMPStart (Just Upstream of Many Polysaccharide gene Starts) sequence, which includes a proposed promoter region (Hobbs and Reeves 1994; Bailey, Hughes and Koronakis 1997). Proteins encoded within each gene cluster generally fall into three functional categories, being enzymes for NDP-sugar synthesis, glycosyl transfer and OPS processing. As expected, sugar biosynthesis genes are only present in a gene cluster when the corresponding O unit includes sugars that require those gene(s) for synthesis.

The biosynthesis pathways for the NDP-linked precursors are shown in Fig. 4. Of particular interest is that there are two different pathways for DDH sugar synthesis. GDP-L-Col*p* is synthesised via GDP-D-Man*p* (Alam, Beyer and Liu 2004), whereas all other DDH sugars are synthesised as CDP-linked DDH sugars (Matsuhashi *et al.* 1966; Thorson *et al.* 1994; Chen, Guo and Liu 1998; and reviewed in Samuel and Reeves 2003). These sugars are linked as side branches to the O-unit main chain by specific GTs that are encoded by genes located close to those for synthesis of DDH or other related sugars. There are two or three additional GT genes in each OPS gene cluster, with 32 different GTs in the set of 18 serotypes. GT types are distinguished by name and defined by a cut-off value of 85% amino acid identity, though most GT sequences belonging to a type are >90% identical. The function of only one GT, WbyM, has been experimentally confirmed (Kondakova *et al.* 2012), though the linkage specificities of all GTs have been predicted (Table 2).

The *wzx, wzy* and *wzz* OPS processing genes are present in all gene clusters, indicating that all *Y. pseudotuberculosis* serotypes use the Wzx/Wzy-dependent pathway for O-antigen

O:12
β-Yer(A)p
1
↓
3
3)-α-D-Gal p-(1→4)-α-L-Quip-(1→3)-β-D-GlcpNAc-(1→

O:1b
β-Parf
1
↓
3
2)-β-D-Manp-(1→4)-α-D-Manp-(1→3)-α-L-Fucp-(1→3)-β-D-GlcpNAc-(1→

O:11
α-L-Altf
1
↓
3
2)-β-D-Manp-(1→4)-α-D-Manp-(1→3)-α-L-Fucp-(1→3)-β-D-GlcpNAc-(1→

O:5a
α-Ascp
1
↓
3
2)-α-L-Fucp-(1→3)-α-D-Manp-(1→4)-α-L-Fucp-(1→3)-β-D-GalpNAc-(1→

O:5b
α-L-Altf
1
↓
3
2)-α-L-Fucp-(1→3)-α-D-Manp-(1→4)-α-L-Fucp-(1→3)-β-D-GalpNAc-(1→

O:15
β-Parf
1
↓
3
2)-α-L-Fucp-(1→3)-α-D-Manp-(1→4)-α-L-Fucp-(1→3)-β-D-GalpNAc-(1→

O:2c
α-Abep
1
↓
3
6)-α-D-Manp-(1→2)-α-D-Manp-(1→2)-β-D-Manp-(1→3)-β-D-GalpNAc-(1→

O:4a
α-Tyvp
1
↓
3
6)-α-D-Manp-(1→2)-α-D-Manp-(1→2)-β-D-Manp-(1→3)-β-D-GalpNAc-(1→

O:9
α-D-Galp
1
↓
3
3)-α-D-GlcpNAcA3OAc-(1→4)-α-L-Fucp NAm-(1→3)-α-D-GlcpNAc-(1→

O:1c
β-Parf
1
↓
3
2)-α-D-Manp-(1→4)-α-L-Fucp-(1→3)-β-D-GalpNAc-(1→

O:2b
α-Abep
1
↓
3
2)-α-D-Manp-(1→4)-α-L-Fucp-(1→3)-β-D-GalpNAc-(1→

O:3
β-Parp
1
↓
4
2)-α-D-Manp-(1→4)-α-L-Fucp-(1→3)-β-D-GalpNAc-(1→

O:1a
β-Parf-(1→3)-β-D-6dManHepp
1
↓
3
3)-α-D-Galp-(1→3)-β-D-GlcpNAc-(1→

O:2a
α-Abep-(1→3)-β-D-6dManHepp
1
↓
3
3)-α-D-Galp-(1→3)-β-D-GlcpNAc-(1→

O:4b
α-Tyvp-(1→3)-β-D-6dManHepp
1
↓
3
3)-α-D-Galp-(1→3)-β-D-GlcpNAc-(1→

O:10
α-L-Colp
1
↓
3
4)-α-D-Glcp-(1→4)-α-D-Galp-(1→3)-β-D-GalpNAc-(1→
6
↑
1
α-L-Colp

O:7
α-L-Colp α-D-Glcp
1 1
↓ ↓
3 3
6)-β-D-Glcp-(1→3)-α-D-GalpNAc-(1→3)-β-D-GalpNAc-(1→

O:6
α-L-Colp-(1→2)-β-Yer(A)p
1
↓
3
3)-β-D-GlcpNAc-(1→6)-α-D-GalpNAc-(1→3)-β-D-GalpNAc-(1→

Figure 2. *Yersinia pseudotuberculosis* O-unit structures. Serotype names are shown above each O unit, and structures are grouped according to common main-chain features. References for structures are in Table 1.

biosynthesis. There are eight different *wzx* sequence types in the set, numbered starting from O:1a (Fig. 3). The *wzx1* gene is further subdivided into three related subtypes (*1a*, *1b*, and *1c*) that share 85% or more DNA sequence identity. There are 11 *wzy* sequence types (also numbered starting from O:1a), and gene clusters with the same sequence type have the same linkage between their O units (see Fig. S1, Supporting Information). The *wzz* gene is usually the last gene in the cluster, and the function of the O:2a protein has been demonstrated experimentally (Kenyon and Reeves 2013). That study also indicated very little polymerisation in the absence of Wzz, suggesting the O:2a OPS had a much stronger propensity for ligation in the absence of Wzz than for others that have been used in *wzz* mutagenesis studies (Kenyon and Reeves 2013). However, it is not known if the other *Y. pseudotuberculosis* Wzy sequence types have the same functional dependence on Wzz for polymerisation.

Figure 3. *Yersinia pseudotuberculosis* O-antigen gene clusters. Gene clusters are drawn to scale using published sequences (GenBank accession numbers in Table 1). Gene clusters are numerically ordered by serotype name indicated on the left. Genes are coloured according to the respective pathways of their products, and the scheme is shown on the right. Figure is drawn to scale and the scale is shown below. All genes are transcribed from left to right. Bold lines bordering genes indicate that the gene encodes a GT.

Grouping by shared gene modules

The *Y. pseudotuberculosis* O-antigen genes involved in synthesis and linkage of sugar precursors are generally arranged in modules. Fifteen of the gene clusters contain a module at the 5' end (left end in Fig. 5) for synthesis of a CDP-pathway DDH or related sugar, which mostly is added last to the O-unit and becomes either a single sugar side branch or second in a two-sugar side branch on polymerisation. There are seven forms of these side-branch gene modules, designated S1–S7 (Fig. 5). Fourteen of the gene clusters with a side-branch module have also one of a range of modules (M1–M5) at the 3′ end for synthesis of different O-unit main chains (Fig. 6). These correspond to the five sets of related structures shown in Fig. 2 and named in Fig. S1. M5 O units have a two-sugar side branch, but as the 6dManHep sugar pathway genes are present with two GT genes in the location corresponding to the main-chain module genes of other modules, it is treated as part of the main chain.

The O:6 gene cluster has a typical *ddh* side-branch gene module (S1, shared with O:12), but is otherwise grouped with O:7 and O:10 due to shared presence of GDP-L-Col*p* synthesis genes (Fig. 3; Cunneen *et al.* 2011). Finally, the O:9 gene cluster is not obviously related to any of the others, having only *wzz* shared with any other serotype (Fig. 3; Beczala *et al.* 2013). We will now discuss the relationships of these gene clusters starting with those that have shared 5′ side-branch gene modules, while more detail of the S and M modules is given in the Supporting Information.

Side-branch gene modules

Genes present in more than one of the S modules are highly conserved, with an average of >98.5% nucleotide sequence identity between any two gene clusters. These include the *ddhD*, *ddhA*, *ddhB* and *ddhC* genes that direct the synthesis of CDP-4-keto-3,6-dideoxy-glucose (Matsuhashi *et al.* 1966; and reviewed in Samuel and Reeves 2003), the common precursor of four different CDP-DDH sugars, and also of the related octose sugar, Yer(A)*p* (Fig. 4A). The specific sugar produced by each module is determined by the DDH-specific gene(s) present immediately downstream of *ddhC*. Synthesis of hexose L-Alt is predicted to involve *ddhA* and *ddhB*, but the reduction of C6 by DdhC does not occur, and in S7 the *ddhC* gene is replaced by *altA* and *altB* genes for completion of L-Alt synthesis (Cunneen *et al.* 2009).

Table 1. *Yersinia pseudotuberculosis* O-antigen gene cluster sequences and O-unit structures.

Serotype	O-unit structure reference	OPS gene cluster GenBank accession number
O:1a	Kondakova *et al.* (2012)	AF461768
O:1b	Kondakova *et al.* (2009d)	AJ251712 (AJ251713[a])
O:1c	De Castro *et al.* (2011)	GU120200
O:2a	Kondakova *et al.* (2008b)	AF461770
O:2b	Kondakova *et al.* (2009a)	GU120201
O:2c	Kondakova *et al.* (2008a)	KJ504353
O:3	Kondakova *et al.* (2008a)	KJ504354
O:4a	Kondakova *et al.* (2009b)	KJ504355
O:4b	Kondakova *et al.* (2009c)	AF461769
O:5a	Gorshkova, Korchahina and Ovodov (1983)	KJ504356
O:5b	Korchagina, Gorshkova and Ovodov *et al.* (1982)	KJ504357
O:6	Gorshkova *et al.* (1983)	HQ456392
O:7	Kotandrova *et al.* (1989)	HQ456391
O:8	Kenyon *et al.* (2016)	KM454907
O:9	Beczala *et al.* (2013)	AJ539157[b]
O:10	Kenyon *et al.* (2011)	HQ396160
O:11	Cunneen *et al.* (2009)	FJ798742
O:12	De Castro *et al.* (2012)	JX454603[b]
O:13	-	-
O:14	-	FJ798743
O:15	De Castro *et al.* (2009)	AM849474

[a] *Y. pestis* EV76.
[b] *Y. similis* strain.

L-Alt occurs only in the furanose form (L-Alt*f*), and Par is also usually furanose (Par*f*) in this species. In each case, there is a *wbyH* gene, predicted to encode a mutase (Reeves, Pacinelli and Wang 2003) shared by S5 and S7 as the last of the S-module pathway genes. In cases where *wbyH* is absent i.e. S4, a furanose side-branch sugar is not present in the O-unit structure.

In most of these gene clusters, *wzx* is located immediately downstream of the S module, followed by the GT gene for the CDP-linked DDH or related sugar. The exceptions are O:6 and O:3, both of which are discussed below. The *wzx*, GT and *wzy* genes sit between the side-branch and main-chain gene modules. Five *wzx* sequence types are present among the gene clusters with S modules, many in more than one gene cluster, and the *wzx* type commonly correlates with presence of one or two specific side-branch sugars, and in all but two cases, the S module can be extended to include the *wzx* gene (Fig. 5, dotted lines). However, the DDH GT gene and the *wzy* sequence form both correlate better with the M module.

Main-chain gene modules

Modules M1–M4 are related (Fig. 6) with many sugar pathway and GT genes shared by two or more. These modules differ mainly in the types and numbers of GT genes present, which reflects the differences in the structure linkages. The M1 structures have D-Glc*p*NAc as the first sugar, whereas the M2–M4 structures have D-Gal*p*NAc. In the latter case, the gene clusters include a *gnu* gene for conversion of UndPP-D-Glc*p*NAc to UndPP-D-Gal*p*NAc (Cunneen *et al.* 2013), and also a *wbyQ* gene predicted to catalyse the α-L-Fuc*p*-(1→3)-D-Gal*p*NAc linkage (Reeves, Pacinelli and Wang 2003). The *gnu* and *wbyQ* genes replace the *wbyL* gene for the α-L-Fuc*p*-(1→3)-D-Glc*p*NAc linkage present in M1, and together these account for the difference in the first sugar. The remaining sugars in the main chains are all D-Man*p* or L-Fuc*p*, apart from D-Gal*p* and L-Qui*p* residues in O:12 (M1a).

Finally, module M5 is involved in the formation of a common trisaccharide for O:1a, O:2a and O:4b. The synthesis genes for one sugar of this trisaccharide, 6dManHep*p*, are present along with two GTs (Ho *et al.* 2008). The synthesis genes for UDP-Gal and UDP-GlcNAc reside elsewhere, and have been discussed. The 6dManHep*p* of this trisaccharide, although a part of a common main chain, forms part of the final side branch of the O unit, with either a Abe, Tyv or Par reside (synthesised by S2, S3 or S5). In fact, all of the repeat units have one or two branches when polymerised, but most are linear until Wzy uses the last or second last sugar as the acceptor during polymerisation thereby creating a side branch.

Each M module (including M1a and M1b) has a predominant *wzy* polymerase gene, which correlates with the combination of donor sugar (D-Glc*p*NAc or D-Gal*p*NAc), the specific acceptor sugar and the linkage to be generated by that Wzy. The exception is in module M2, in which O:3 has a different acceptor sugar and also a different *wzy* gene.

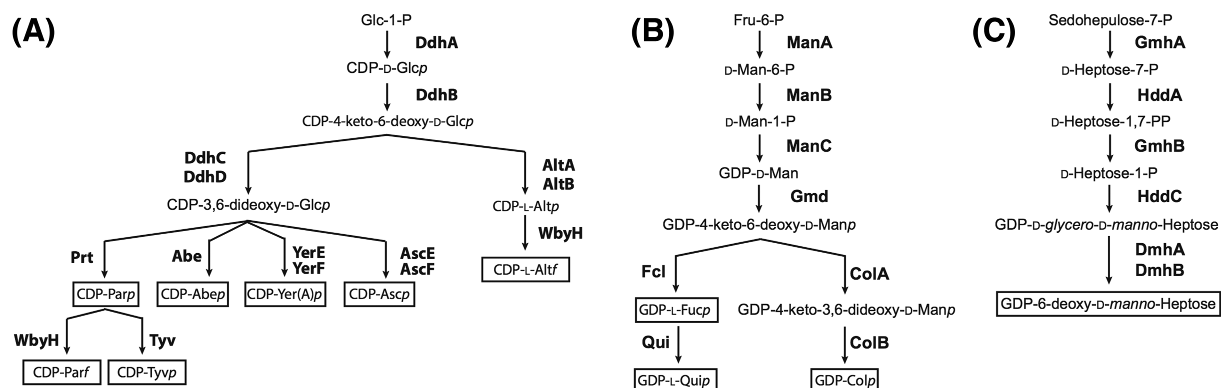


Figure 4. Synthesis pathways of activated sugar precursors that are incorporated into *Y. pseudotuberculosis* O units. Pathways with experimental data are the CDP-DDH (Chen, Guo and Liu 1998 and reviewed in Samuel and Reeves 2003), GDP-L-Col*p* (Alam, Beyer and Liu 2004), GDP-L-Fuc*p* and GDP-D-Man*p* (reviewed in Samuel and Reeves 2003) and GDP-6dManHep*p* (Butty *et al.* 2009) synthesis pathways. Predicted pathways are for CDP-L-Alt*f* (Cunneen *et al.* 2009) and GDP-L-Qui*p* (De Castro *et al.* 2012) synthesis. Substrates and products are shown, and enzymes are in bold face type. Boxed sugars are those incorporated into O units shown in Fig. 1. Sugar abbreviations are expanded in the text.

Table 2. Predicted GT functions.

GTs	Predicted linkage function	Serotypes
WbyA	α-Abe-(1→3)-D-6dManHep*p*	O:2a
	α-Tyv-(1→3)-D-6dManHep*p*	O:4b
WbyB	β-D-6dManHep*p*-(1→4)-D-Gal*p*	O:1a, O:2a, O:4b
WbyC	α-D-Gal*p*-(1→3)-D-Glc*p*NAc	O:1a, O:2a, O:4b
WbyD	α-Abe*p*-(1→3)-D-Man*p*	O:2c
	α-Tyv*p*-(1→3)-D-Man*p*	O:4a
WbyE	α-Abe*p*-(1→3)-D-Man*p*	O:2b
WbyI	β-Par*f*-(1→3)-D-Man*p*	O:1b
	α-L-Alt*f*-(1→3)-D-Man*p*	O:11
WbyJ	β-D-Man*p*-(1→4)-D-Man*p*	O:1b, O:11
WbyK	α-D-Man*p*-(1→3)-L-Fuc*p*	O:1b, O:11, O:1c, O:2b, O:3
WbyL	α-L-Fuc*p*-(1→3)-D-Glc*p*NAc	O:1b, O:11, O:12
WbyM[a]	β-Par*f*-(1→3)-D-6dManHep*p*	O:1a
WbyN/WbyO[b]	β-D-Man*p*-(1→3)-D-Gal*p*NAc	O:2c, O:4a
	α-D-Man*p*-(1→2)-D-Man*p*	
WbyP	β-Par*p*-(1→4)-L-Fuc*p*	O:3
WbyQ	α-L-Fuc*p*-(1→3)-D-Gal*p*NAc	O:2b, O:1c, O:3 O:5a, O:5b, O:15
WbyS	α-L-Alt*f*-(1→3)-L-Fuc*p*	O:5b
	α-Asc*p*-(1→3)-L-Fuc*p*	O:5a
	β-Par*f*-(1→3)-L-Fuc*p*	O:15
WbyT	α-L-Fuc*p*-(1→3)-D-Man*p*	O:5a, O:5b, O:15
WbyU	α-D-Man*p*-(1→4)-L-Fuc*p*	O:5a, O:5b, O:15
WbyV	α-L-Col*p*-(1→3)-D-Glc*p*	O:7
WbyX	α-D-Gal*p*NAc-(1→3)-D-Gal*p*NAc	O:6, O:7
WbyW	β-D-Glc*p*-(1→3)-D-Gal*p*NAc	O:7
WbzA	β-D-Glc*p*NAc-(1→6)-D-Gal*p*NAc	O:6
WbzB	α-L-Col*p*-(1→2)-D-Yer(A)*p*	O:6
WbzC	β-Yer(A)*p*-(1→3)-D-Gal*p*NAc	O:6
WbzD	β-Par*f*-(1→3)-D-Man*p*	O:1c
WbzE	α-L-Col*p*-(1→3)-D-Glc*p*	O:10
	α-L-Col*p*-(1→6)-D-Glc*p*	
WbzF	α-Glc*p*-(1→4)-D-Gal*p*	O:10
WbzG	α-Gal*p*-(1→3)-D-Gal*p*NAc	O:10
WbzH	β-Yer(A)*p*-(1→4)-D-Gal*p*	O:12
WbzI	α-Gal*p*-(1→4)-D-Qui*p*	O:12

[a]Confirmed experimentally (Kondakova *et al.* 2012).
[b]Functions cannot be clearly differentiated.

The junction between S and M modules

OPS diversity in *Y. pseudotuberculosis* mostly involves different combinations of S and M gene modules. The *wzx* gene is located primarily at the junction of these modules (Figs 5 and 6), and the sequence reveals a pattern such that the 5′ end (base positions 1 to 575) and the 3′ end (remaining sequence) groups with the type of S or M module, respectively. This is best observed in phylogenetic trees of both ends of the gene, in which distinct phylogenetic clades correspond to particular S or M modules, respectively (Fig. 7). Thus, it appears that *wzx* marks the junction between S and M modules, and it is tempting to speculate that the two ends of Wzx may interact with the two ends of the O-unit substrate.

PATTERNS IN THE GENE CLUSTERS

The M module gene order

There are some very clear patterns in the gene order within the gene clusters, and an overview of this is presented for the M modules in Fig. 8. The genes (black italic name labels) for each module are shown in each column in inverse order of their location in the gene cluster starting with *wzz* at the top. The GT-gene cells (green) are present in function order, and used as anchors for the comparison by assigning GT1 to GT3 each to a row, which requires that some intervening cells be blank to accommodate different numbers of genes. The sugars and linkages generated by each GT are shown in red in the GT cell. The addition of D-Glc*p*NAc to the UndPP carrier by the initial sugar transferase is shown in a row of striated green cells before GT1, because although WecA acts as the initial sugar transferase, the *wecA* gene is outside of the gene cluster.

The GT genes and sugar pathway genes have different patterns. The GTs are all in the order of the sugars added to the growing O unit. The sugar pathway genes are generally placed above all of the GT genes that use that sugar. For example, the *manB* gene is the first pathway gene in modules M1–M4, which require GDP-D-Man or the derivatives GDP-L-Fuc and GDP-L-Qui. The *gnu* gene is the next pathway gene and is required for the three modules that use GalNAc as the first sugar. It is just above the GT1 genes that add a sugar to either GlcNAc or GalNAc. The *manC* gene in M4 is immediately above the *wbyO* GT gene that adds the first Man residue in M4, while the *manC* gene in M2–M3 is between the GT1 and GT2 genes, as are the *fcl* and *gmd* genes needed for GDP-L-Fuc and GDP-L-Qui.

We can also see that each step in the parallel pathways in Fig. 8 increases the level of diversity. Reading from the 3’ end of the gene cluster, the first gene is *wzz* (on the right in Figs 5 and 6, and on top in Fig. 8), which is present in all 14 gene clusters, and is followed by two options: being the *manB* gene or 6dManHep gene set. For those with *manB*, there are three options for linkage of the second sugar, and four options for the linkage of the third sugar. Finally, this gives a total of six different O units, each with its own *wzy* gene. The selection is proposed to keep the varying genes in one central block, in any recombination event, but the effect is also to keep genes generally in function order, because each addition to the O unit has the possibility of adding to diversity. This accounts nicely for the common observation of GT genes being generally in inverse order for function.

Note that there is no selection for order of genes within a block, such as that for the genes of the 6dManHep pathway, which in *Y. pseudotuberculosis* are all either present or absent, and together add a single structural component to the O unit. It is not surprising that in this case the gene order is not related to gene function (e.g. Fig. 4C).

Another phenomenon of relevance discussed below is that the genes which differ in any pairwise alignment are in one block, with shared genes on either side (see Fig. 9). A similar situation has been documented in other species.

S gene modules have the opposite gene order to M gene modules

The S gene modules are at the 5′ end of the gene cluster and have a similar pattern, but these genes are generally in the same order as their product function as shown in Fig. 3, which is the opposite of the M module gene order. The *ddhDABC* genes for the shared DDH intermediate are at the 5′ end, followed by the genes to complete synthesis of the several DDH or related sugars found in this species. The individual genes are also generally in function order, the exception being the *ddhD* gene, which is located at the beginning of the gene cluster, although DdhD acts last in the DDH precursor pathway. In all but O:3, the S modules are followed by the *wzx* genes for O-unit translocation, which are

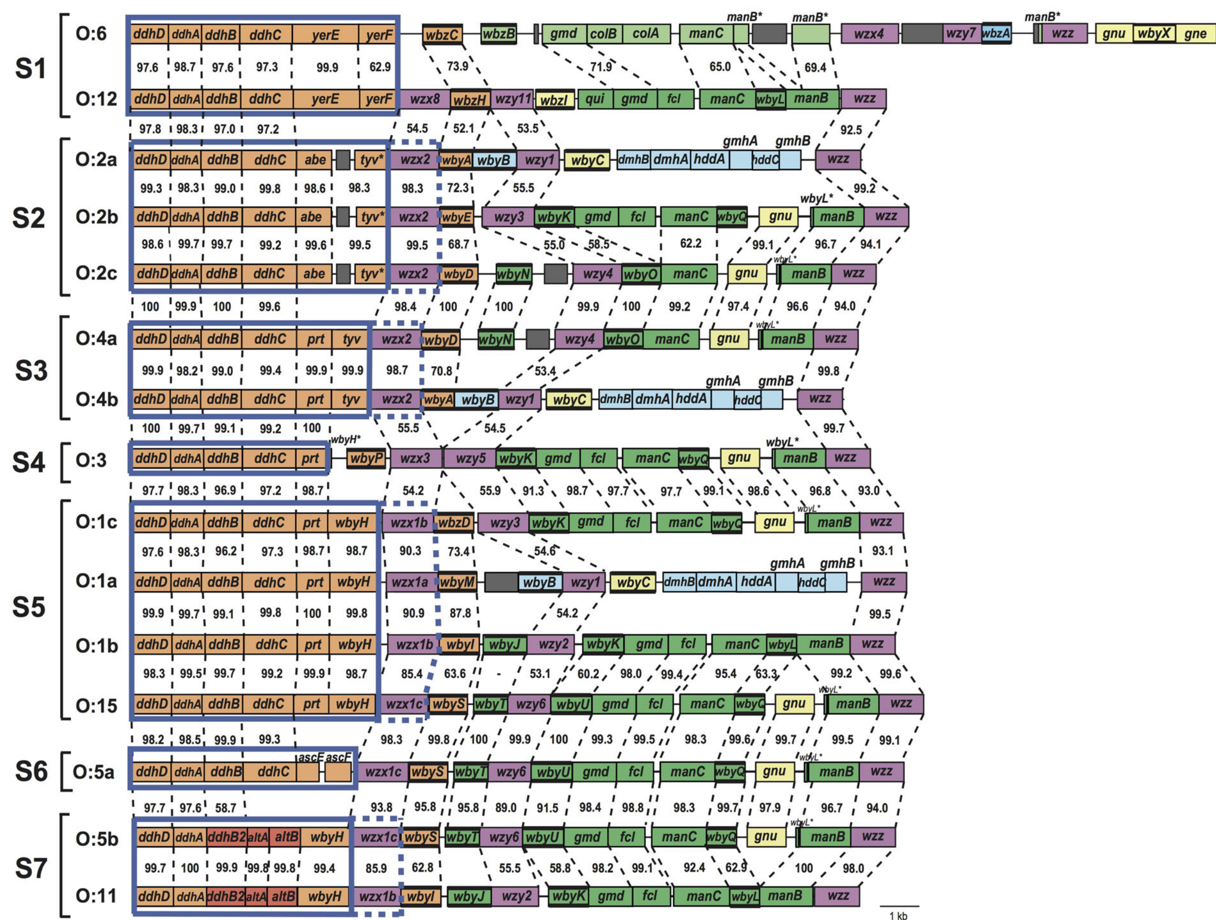


Figure 5. Gene modules for O-unit side-branch synthesis in *Y. pseudotuberculosis* O-antigen gene clusters. Gene cluster details as in legend for Fig. 3. Boxes highlight the gene modules required for the synthesis of side-branch sugars. Names of modules and serotypes are shown on the left. The length of some of these modules can be extended to include the *wzx* and/or the GT gene, as indicated by the dashed boxes.

more variable with sequence identity as low as 54%, and then the GT gene for addition of the side-branch sugar to the main chain.

Symmetry of the S and M gene modules provides an explanation for the observed patterns

For both S and M gene modules, the most common genes are at one end of the gene clusters, and those that are present in only one or few S or M modules are towards the centre of the gene cluster. The details are such that for almost any pairwise comparison, all of the genes with significant sequence differences are contiguous and form a segment in the middle of the gene cluster, which is bracketed by flanking genes that are very similar. Figure S2 (Supporting Information) gives an example alignment of all combinations for three gene clusters, which illustrates how having the variable genes in a central location can lead to genes being in function order.

We propose that the reason for the two trends documented in Fig. 8 is 2-fold:

i. During assembly of the seven O units, the structures diverge, so that for each successive additional sugar, there are more possible acceptors than previously, which, as they become part of the acceptor structure, may require a different GT for the next round. All modules start assembly with GlcNAc. In M2, M3 and M4, this is then converted to GalNAc. In the next round, the first GTs generate four different sugar linkages to GlcNAc or GalNAc, and in the next round there are again four GTs for different linkages. As yet we do not know how specific the GTs are for their cognate acceptor structure. However, the pattern that results has the effect of ensuring that in almost any pairwise alignment, the genes with significant sequence differences are in a single block at the bottom of the chart in Fig. 8, and thus adjacent to the gene blocks for the variable genes in the S modules.

Note that the seven *wzy* genes are very different even where they make the same polymerisation linkage, as for *wzy3* and *wzy5*, making a D-Gal*p*NAc-(1→2)-D-Man*p* linkage (see Fig. S1), and are always in the central block of main-chain specific genes.

ii. There is selection for the genes which differ in any pairwise alignment to be in one block, with shared genes on either side, because this facilitates homologous recombination events in which the genes that determine the structure in a donor strain replace the resident gene cluster for the other structure by recombination in the shared genes on either side of the divergent genes. Any other arrangement would also allow homologous recombination that generated novel gene clusters. Note that the two *manC* genes appear to not comply with this, but as they are only 61% identical recombination between them will be relatively rare. If this hypothesis is

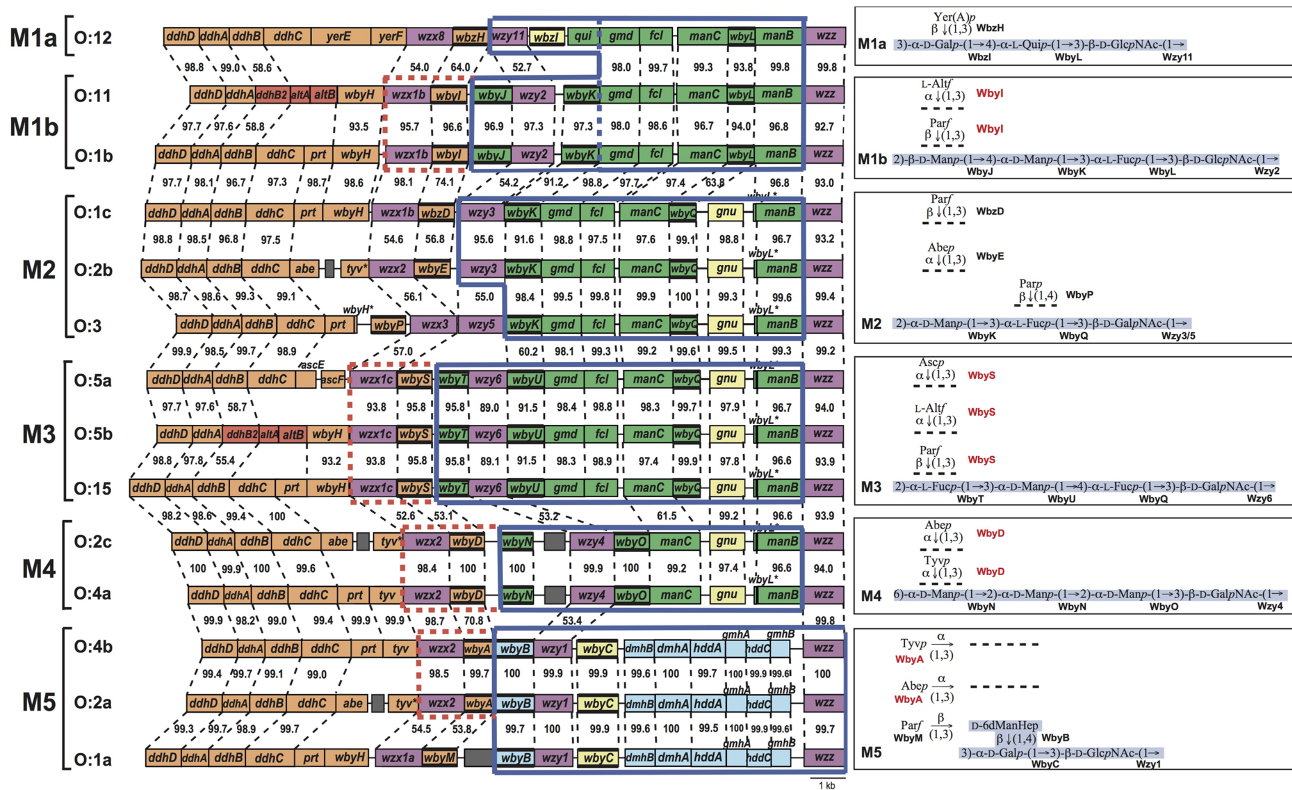


Figure 6. Gene modules for O-unit main-chain synthesis in *Y. pseudotuberculosis* O-antigen gene clusters. Gene cluster details as in legend for Fig. 3. Boxes highlight the gene modules required for the synthesis of main-chain sugars. Names of modules and serotypes are shown on the left. The length of some of these modules can be extended to include the *wzx* and/or the GT gene, as indicated by the dashed boxes.

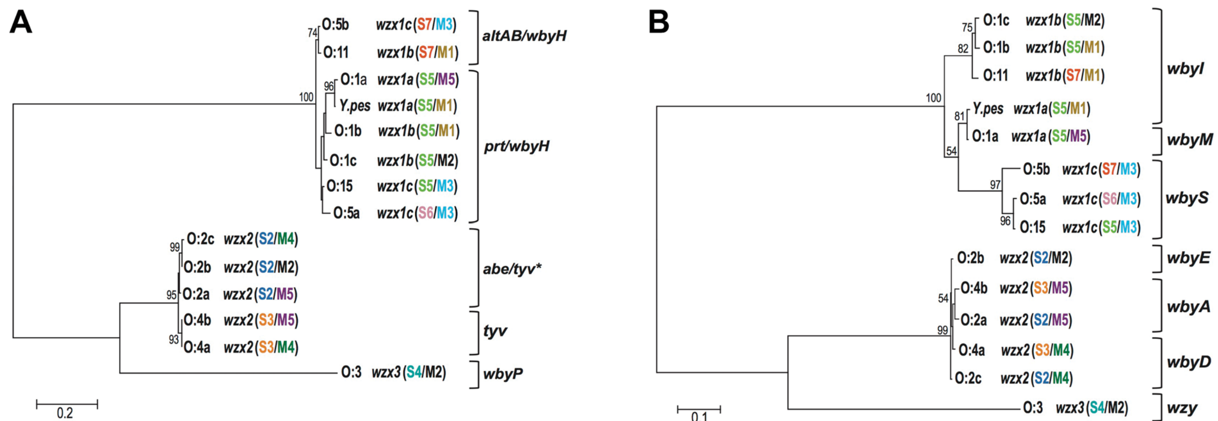


Figure 7. Phylogenies of *Y. pseudotuberculosis wzx* ends partitioned at base position 575. Neighbour-joining tree with and bootstrap values based on 100 replicate trees. Serotype names, module groups and *wzx* names are shown, along with scale bar. Groupings indicated on the right of each phylogeny denote gene immediately upstream or downstream of *wzx*.

correct, then selection to facilitate replacement of one gene cluster by another would indirectly apply selection for genes to be in function order.

We are not aware of data to support recombinational replacement of O-antigen gene clusters in *Y. pseudotuberculosis* but there are good examples elsewhere. For *S. enterica* (Li and Reeves 2000) and *S. pneumoniae* (Jiang, Wang and Reeves 2001), there is gene cluster sequence evidence for recombination within shared genes of the O-antigen and capsule gene clusters, respectively, and for *S. pneumoniae* there is also data from genome sequences of strains from a clone shown by multilocus sequence typing (MLST) to have undergone antigenic shift (Croucher *et al.* 2011). In each case, homologous recombination was observed, which could be either in DNA flanking the gene cluster or in DNA within the gene cluster in shared regions that flank a central divergent block of genes.

The pattern of gene differences between serotypes being grouped in the middle of the gene cluster has also been observed in polysaccharide gene clusters of other species such as *Acinetobacter baumannii* (Hu *et al.* 2013; Kenyon and Hall 2013). In *S. pneumoniae* capsule gene clusters, the genes at the 5′ end are shared

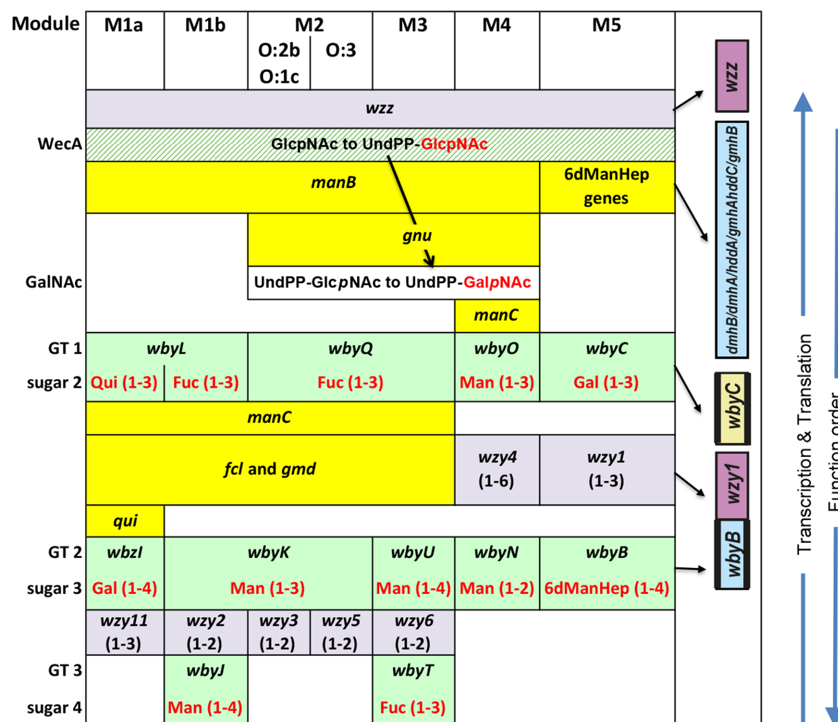


Figure 8. Pattern representation of the main-chain sugars and their transferases. Each column shows the genes for one of the seven main chains in map order from bottom to top, which puts most of the genes close to function order from top to bottom. The sugars and their linkages are shown in red, below the corresponding GT gene name. GlcpNAc, the first sugar, is shown attached to UndPP at the top, although the *wecA* gene is not in the gene cluster, followed by rows for the GT genes for the second, third and fourth sugars, which all occur in inverse map order. Intermingled in ascending map order are the other genes in each module. Cells for the initial sugar transferase gene and GT genes are coloured green and cells for other genes are coloured yellow. The significance of the gene order is discussed in the text.

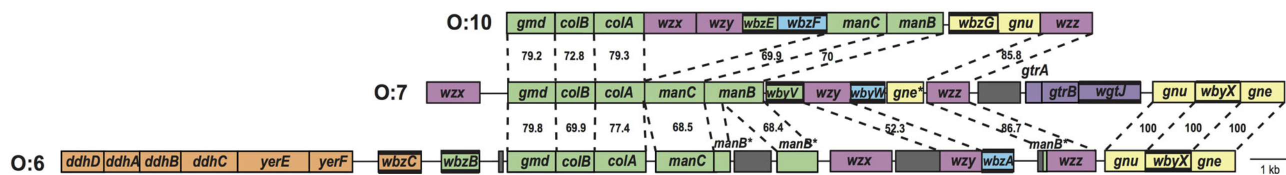


Figure 9. Comparison of the *Y. pseudotuberculosis* O:6, O:7 and O:10 gene clusters. Gene cluster details as in legend for Fig. 3. Nucleotide sequence identities are shown between clusters, and the serotype is indicated on the left.

by all serotypes (Aanensen *et al.* 2007), but this is not common at the 3' end, and it was found that serotype switching was more frequent between related serotypes with more shared sequence (Croucher *et al.* 2011). The function of many of the *S. pneumoniae* GT genes could be inferred and in about 80% of cases where all were predicted, GT function was in map order (Aanensen *et al.* 2007), in contrast to the inverse map order for *Y. pseudotuberculosis*. This is as expected, as in *S. pneumoniae* the initial transferase genes are transcribed before the GT genes. Despite the shortage of direct evidence, it appears that selection for ease of transfer of surface polysaccharide gene clusters is widespread.

EVOLUTION OF OPS GENE CLUSTERS IN THE COMPLEX

OPS genes move by homologous recombination

It is common for O-antigen gene clusters of a species to all be at the same locus, and it has long been recognised that this allows a gene cluster to be replaced by another by homologous recombination involving the shared sequence that flanks the gene clusters (Lawrence 1997). It appears that at least some of the observed substitutions are beneficial due to loss of the original gene cluster, as in the situation of *Vibrio cholerae* seventh pandemic strain when a 35-kb O139 gene cluster replaced the 22-kb O1 gene cluster in a recombination event (Faruque *et al.* 2003). This enabled the organism to colonise adults who were immune to the O1 form in areas with endemic cholera (Ramamurthy *et al.* 2003). Also in *S. pneumoniae,* the introduction of a vaccine based on capsule antigens led to 'serotype switching' away from antigens in the vaccine by a similar process (Croucher *et al.* 2011) We will refer to the existence of a shared locus as 'co-location' of the gene clusters.

Co-location of O-antigen gene clusters facilitates the simultaneous loss of the current OPS gene cluster and the gain of a different OPS gene cluster from another strain, because both are accomplished by a single homologous recombination event. We usually observe it as the gain of a new O antigen but, as in the examples above, it may commonly be selection for loss of the OPS gene cluster in the recipient that drives the replacement, with gain of the donor OPS gene cluster being essential to maintain the protection of an O antigen in the LPS (Cunneen and Reeves 2007). OPS are major antigens and the capability for one gene cluster to displace another is thought to have been the

evolutionary driver for co-location of the gene clusters (Cunneen and Reeves 2007). There are many reasons why an O unit could become deleterious, including an immune response in a host species during an outbreak, or the rise of a bacteriophage that targets the O-antigen structure as a receptor, or to avoid being eaten by a range of organisms (Atzinger, Butela and Lawrence 2016). It is also probable that some O-antigen structures are intrinsically more suitable than others in a given niche, but this is not so easily demonstrated. O antigens encoded by a gene cluster that is easily transferred in this way will persist over long periods of time, providing strong selection for all genes involved to be in a gene cluster at the common locus or at least close to it. The prevalence of such transfer in *Y. pseudotuberculosis* is shown by the presence of strains with different O antigens within an MLST or clonal cluster (Ch'ng *et al.* 2011; Laukkanen-Ninios *et al.* 2011).

The pattern of 'divergence-level symmetry', described above for variation in the *Y. pseudotuberculosis* OPS gene clusters, is important for gene cluster replacement given the number of genes present in more than one gene cluster, as otherwise recombination involving shared genes in the two gene clusters would not always give a simple replacement of one gene cluster by another, and this would reduce the long-term prospects of that gene cluster. As it is, replacement of one OPS with another requires only transfer of the central region that distinguishes the donor and recipient gene clusters, and recombination within the homologous terminal segments of the gene clusters achieves this. Genome studies on the *Y. pseudotuberculosis* complex have shown that OPS serotypes often vary within an MLST, likely due to recombination (Ch'ng *et al.* 2011; Laukkanen-Ninios *et al.* 2011). We propose that the capacity for O-antigen replacement by a single recombination event gives very strong selection for gene clusters to have the genes that determine the differences between related structures in a single block. This allows genes that have the potential to determine the OPS structural difference to spread easily within a species or group of species. Newly successful clones that are being limited by host immune responses can simply change their surface antigen. Importantly, this has evolutionary benefit for the gene cluster as well as for the recipient strain.

Relationships of S modules

The S2 Abe module is clearly derived from a S3 Tyv module, as it includes a substantial remnant of a *tyv* gene. A plausible origin for the *Y. pseudotuberculosis* Abe module is that the *abe* gene was transferred from another species, which already had an Abe module, and replaced the *par* gene in a recombination event involving homologous recombination at the 5′ end of *ddhC*, and the other end within the *tyv* gene as now marked in the Abe module by presence of an IS element (Fig. S3, Supporting Information). A similar situation was proposed for the transfer of a *wzy* gene to generate the *S. enterica* group D2 gene cluster, for which it was hypothesised that a mobile element at one end of the inserted fragment initiated the process by forming a cointegrate that was resolved by homologous recombination to generate the final product (Reeves *et al.* 2013), and a similar mechanism can be postulated here.

The relationships of S3, S4 and S5 are interesting. The S3 and S5 modules look mature with start codons all very close to the stop codon of the previous gene, as is common in bacteria as it allows translational coupling for expression of all genes from the one Shine-Dalgarno sequence (Tian and Salis 2015). However, the S4 module has a large intergenic region between *prt* gene and the Par*p* transferase gene *wbyP*. We conclude that the S3 and S5 modules are both long standing whereas the S4 module could be derived from either an S3 or S5 module by replacement of the respective *wbyH* or *tyv* gene and GT gene, by the *wbyP* gene for the Par*p*-Man linkage. We suggest that this is the likely origin of the S4 module, but it is not possible to say if the ancestral strain was S3 or S5. This is the reverse of the situation in *S. enterica*, where paratose occurs only as Par*p*, and the Par*p* module is clearly derived from a Tyv module by mutation in the *tyv* gene, which leaves CDP-Par*p* as the end product, which is then used as the side-branch sugar of the Group A O unit (Reeves *et al.* 2013). A relatively recent replacement of *wbyH* by *wbyP* would account for *wbyP* being unique as a side-branch transferase in being before and not after the *wzx* gene.

The S6 Asc module appears to have the *ascE* and *ascF* genes inserted into an S5 module to replace the *prt* and *wbyH* genes. Recombination junctions can be identified, but there is no evidence of homologous recombination or mobile elements. There are intergenic gaps between *ascE* and *ascF*, and between *ascF* and *wzx*, suggesting that the Asc module is the derived form and S5 the parent form.

The S1 Yer module appears to be long standing as it has a very short gap between the *ddh* and *yer* genes, as in the S3 Tyv and S4/S5 Par modules. However, the O:6 and O:12 gene clusters do not contain the same *yerF* gene, as they are only 63% identical. This could be due to independent acquisition of *yerF* genes from different sources upon the introduction of new main-chain genes in one of the clusters, given that the regions immediately downstream are significantly different (De Castro *et al.* 2012). However, as YerF determines that the sugar is Yer(A)*p* and not Yer(B)*p*, it is possible that there is an error and one of the structures has Yer(B)*p*, most likely in O:6 as the structure was determined in the 1980s.

The S7 module was found to have sharp boundaries flanking *ddhB-altA-altB*, suggesting that this short gene module was imported into S5 in a single event, replacing *ddhB-ddhC-prt* (Cunneen *et al.* 2011). Interestingly, S7 includes *ddhD* although the DdhD product does not contribute to L-Alt*f* synthesis (Cunneen *et al.* 2009). However, as *ddhD* is always the first gene in most *Y. pseudotuberculosis* gene clusters, recombinational changes of O-antigen genes downstream would mean that if DdhD function is lost, that copy of the S module would have no future outside of S7, and this presumably is sufficient to maintain it.

RELATIONSHIPS OF *YERSINIA PSEUDOTUBERCULOSIS* GENE CLUSTERS TO THOSE OF OTHER SPECIES

Yersinia spp.

The gene cluster from *Y. enterocolitica* O:8 (GenBank accession number U46859) has S and M modules that resemble those of *Y. pseudotuberculosis* (Fig. 1). The O:8 structure includes a 6-deoxy-gulose (Gul) and a L-Fuc*p* side branch to a main chain that contains D-Man*p*, D-Gal*p* and D-Gal*p*NAc (Zhang *et al.* 1997). As for the *Y. pseudotuberculosis* S modules, the genes for Gul synthesis are found next to *hemH* at the beginning of the gene cluster. Unusually, there are two GT gene candidates for Gul. The *wbcC* gene is adjacent to the Gul synthetase genes and its product shares identity with the CDP-Abe/Tyv transferase from *Y. pseudotuberculosis* O:2c/O:4a, whereas *wbcD*, which is in the usual position for the side-branch GT gene on the other side of *wzx*, has a

product sharing identity with the WbyP CDP-Par*f* transferase from *Y. pseudotuberculosis* O:3. Genes for L-Fuc*p*, D-Man*p*, D-Gal*p* and D-Gal*p*NAc are found on the other side of the gene cluster, and resemble an M module.

The OPS gene cluster of *Y. mollaretii* ATCC 43969 (GenBank accession number AALD00000000.2) has an overall organisation similar to those in the *Y. pseudotuberculosis* complex (Fig. 1), but unfortunately the structure is not known. However, it has an S3 gene module suggesting a Tyv side branch. The *Y. pseudotuberculosis ddhDABC* genes share an average of 90% nucleotide sequence identity with those of *Y. mollaretii* ATCC 43969. These genes are followed by *prt*, *tyv*, a GT gene and then *wzx*. The *prt* and *tyv* genes share 74%–78% identity to *prt* and *tyv* from *Y. pseudotuberculosis* O:4a, while the GT is unrelated. Interestingly, the *wzx* gene in the *Y. mollaretii* ATCC 43969 gene cluster is not located immediately downstream of *tyv* as in *Y. pseudotuberculosis* and *S. enterica* OPS gene clusters, indicating that gene rearrangements may have occurred. A potential M gene module is also present, with the *gmd*, *fcl*, *manC*, *wbyL* and *manB* genes in the same order as in the *Y. pseudotuberculosis* M1b module and these share more than 90% identity to the O:1b genes. This strain at least has the modular structure of *Y. pseudotuberculosis*.

Salmonella *enterica* side-branch and main-chain (S and M) modular structures

There is a set of eight related structures and gene clusters in *S. enterica* (Reeves *et al.* 2013) that resemble the set in the *Y. pseudotuberculosis* complex. All but one have essentially the same three-sugar main chain, encoded by genes at the 3′ end of the cluster equivalent to the M gene modules in *Y. pseudotuberculosis*, with variation in one linkage, which is either α-(1→4) or β-(1→4). *S. enterica* group C2 has a different main chain, and there is also variation in the Wzy polymerase.

All but group E have a DDH side branch, and there are also S gene modules that are very much like those in *Y. pseudotuberculosis*, with *S. enterica* group D1 having the same genes as the *Y. pseudotuberculosis* S2 module for the Tyv side branch. The *S. enterica* group D1 *wzx* and DDH GT genes also share sequence identity (Fig. S4, Supporting Information). However in *S. enterica*, there are only three alternative side-branch sugars, the other two being Par and Abe, and thus only three S gene modules. The *abe* gene is present in *S. enterica* groups B and C2, but unlike in the *Y. pseudotuberculosis* S2 gene module, the *abe* genes are not located next to remnants of a *tyv* gene or an IS (Reeves *et al.* 2013). However, group A strains with Par have a group D gene cluster with a mutation in the *tyv* gene, which prevents the conversion of CDP-Par*p* to CDP-Tyv*p* resulting in the addition of Par*p* as the side-branch sugar by mutation in *tyv* (Reeves *et al.* 2013). Thus, in contrast to *Y. pseudotuberculosis*, it appears that Tyv and Abe are the long-standing side-branch sugars in *S. enterica*, and that Par was very recently added as a side branch.

The *ddhDABC* genes from *Y. pseudotuberculosis* and *S. enterica* share an average of 66% pairwise identity. The presence of comparable gene clusters in different species with substantial sequence divergence suggests that this modular structure has been in existence for a long time. An exception is the *abe* gene, since the *S. enterica abe* genes are not significantly related to the *abe* gene in *Y. pseudotuberculosis* S2, suggesting either that these *abe* genes have different evolutionary origins or that the gene module has been present for an enormous length of time.

As for *Y. pseudotuberculosis*, the pairwise alignments for the *S. enterica* groups always have the divergent genes in the centre. However, without the number of related forms found in *Y. pseudotuberculosis*, the connection between selection for the more diverse genes to be in the centre and the GT genes being in inverse function order was not apparent.

THE *YERSINIA PSEUDOTUBERCULOSIS* O:6, O:7, AND O:10 GENE CLUSTERS

The *Y. pseudotuberculosis* O:6, O:7 and O:10 gene clusters include the *gmd*, *colA*, *colB*, *manB* and *manC* genes required for synthesis of GDP-L-Col*p* (Fig. 9). However, although colitose is a DDH and forms a side branch, the genes do not fit into the S modular pattern, but fit better with the M gene modules that commonly have the Man pathway genes that lead into GDP-colitose synthesis. The O:6 and O:7 gene clusters are also the only clusters in the set in which *wzz* is not the most distal gene. These gene clusters share an identical three-gene module, which contains *gnu*, *wbyX* and *gne* that together code for synthesis of UndPP-GalNAc and UDP-GalNAc, and the GT for addition of GalNAc to UndPP-GalNAc. The module lies between the bulk of the gene cluster and *wzz*, and is likely to have been recently acquired by either the O:6 or O:7 gene cluster, and transferred to the other, with *wzz* having previously been the terminal gene (Cunneen *et al.* 2011). The gain of this three-gene module would have the effect of replacing the original D-Glc*p*NAc first sugar with the α-D-Gal*p*NAc-(1→3)-D-Gal*p*NAc pair of sugars, the implication being that the parent form did have D-Glc*p*NAc as first sugar.

FUTURE RESEARCH

Very little is known of O antigens in other species of *Yersinia*, but two of the three *Y. enterocolitica* gene clusters studied have the ABC-transporter pathway, and the few structures reported suggest that this may be common in this species. *Yersinia enterocolitica* O:8 has a gene cluster and structure that fits into the modular pattern found in *Y. pseudotuberculosis*, as does the other reported gene cluster that we are aware of, for *Y. mollarettii*. The *Y. pseudotuberculosis* set that we have studied may turn out to be part of a larger set, and inspection of genomic data would provide further insights.

For many Gram-negative bacteria, OPS serotypes exhibit an enormous amount of structural and genetic diversity, although the evolution and formation of OPS can be difficult to trace. Studies in *E. coli* have reported that OPS diversification within strains is likely the result of the recombinational replacement of large chromosomal segments that include the OPS locus (Milkman, Jaeger and McBride 2003; Zhou *et al.* 2010). However, while long recombinant segments have been reported in *E. coli*, there is good support for recombination having occurred within the gene clusters in the closely related species *S. enterica* (Li and Reeves), and in *S. pneumoniae* it is common for serotype switching to involve recombination junctions within the gene clusters (Croucher *et al.* 2011). For *Y. pseudotuberculosis*, there have been two MLST studies showing that the species has a clonal population structure (Ch'ng *et al.* 2011; Laukkanen-Ninios *et al.* 2011). Both of these studies show that there is serotype variation within sequence types and clonal clusters, almost certainly due to recombination, as was proposed. However, we are not aware of any data for *Y. pseudotuberculosis* on the length of the recombinant fragments, or how often recombination involves the shared segments of the gene clusters.

The genomic locus that harbours different OPS biosynthesis gene clusters has also been previously described as a

hypermutable hotspot (Moxon, Bayliss and Hood 2006), for which natural selection is thought to drive genetic diversification of OPS within a species (Milkman, Jaeger and McBride 2003). Recent studies have shown that variation of polysaccharides, including capsules, is occurring within clonal lineages of many species (Hu *et al.* 2013; Kenyon and Hall 2013; Alqasim *et al.* 2014; Iguchi *et al.* 2015; Wyres *et al.* 2015; Holt *et al.* 2016) indicating strong selective pressures on genetic regions harbouring biosynthesis genes. At this stage, it is not clear what factors apply selective pressure. The variation at the O-antigen gene cluster is thought to be maintained in part at least by frequency-dependent selection (Atzinger, Butela and Lawrence 2016), in which phenotypic variation enhances fitness through the presentation of different antigenic epitopes that may confer protection against predators, environmental stresses or the host immune response (Wildschutte *et al.* 2004). It is also probable that some surface polysaccharides are intrinsically beneficial in particular niches occupied by a species. Future studies are expected to shed light on details of these phenomena.

SUMMARY

The complete set of OPS gene clusters for the 18 validated serotypes in the *Y. pseudotuberculosis* complex is reviewed here for the first time. Fifteen of them include related modules and these comprise one of two sets of related gene clusters for related structures known to us, the other being a set of eight *S. enterica* gene clusters that were reviewed recently (Reeves *et al.* 2013). It is interesting to compare the two sets. The *S. enterica* set is for OPS that have D-Gal as first sugar, rather than D-Glc*p*NAc or D-Gal*p*NAc as in *Y. pseudotuberculosis*, and many other species of the *Enterobacteriaceae* family. However, both sets generally have a DDH or related sugar as a side branch, and both have been extremely successful within their species. The *S. enterica* set comprises only 8 of 54 *S. enterica* serotypes, but they dominate in the strains isolated. In *Y. pseudotuberculosis*, they are so dominant that only three serotypes (O:7, O:9 and O:10) are outside of the set of related serotypes.

One of the most striking features of the *Y. pseudotuberculosis* set is that many of the gene clusters contain shared gene modules responsible for the synthesis of defined structural components of the O unit (side branch, main chain or other features). The relationships between modules make it possible to correlate the genetics with O-unit structures, and assign GT genes to the formation of specific sugar linkages (Table 2). These assignments revealed the connection between two previously described characteristics of gene clusters for OPS and similar structures. First, that when some genes are shared among serotypes, the serotype-specific genes are located in the middle of the gene cluster flanked by the shared genes, which allows replacement of one OPS by another by homologous recombination in the shared gene cluster DNA, as shown for example in *S. enterica* (Li and Reeves 2000) and *S. pneumoniae* (Jiang, Wang and Reeves 2001). Second, the GT genes are generally in inverse order to their location in O-antigen gene clusters, an observation that has been a mystery discussed at meetings but rarely in print. Also puzzling was that in *St. pneumoniae* capsule gene clusters the GT genes had a similar relationship, but the first GT was the first and not the last gene in the cluster (Aanensen *et al.* 2007).

A convincing explanation is revealed by the larger number of related forms in *Y. pseudotuberculosis*. There is more variation in later steps of repeat unit synthesis than in early steps, with generally very few options for the first sugar, but commonly several sugars that can be added to it, often via alternative linkages. The same applies to each of these disaccharides when the third sugar is added, and so on for each step. If the variable genes are to be in a central block, given the pattern of gene variation described above, then this is best achieved by having genes in function order at the 5′ end of the cluster and in inverse function order at the 3′ end as illustrated in Fig. 8, and which we refer to as 'divergence-level symmetry'. This will apply particularly to GT genes as the encoded GTs act directly on the structure. There is flexibility in location of sugar pathway genes as they maintain a pool of their enzyme product, but these genes are usually present only in strains with GTs for that sugar, and they tend to follow the same pattern. We propose that it is selection for keeping the strain-specific genes as a single block in all pairwise combinations, which maintains divergence-level symmetry, and in doing so keeps many of the genes in function order. It does not matter if the first acting or last acting gene is at the 5′ end, and in *Y. pseudotuberculosis*, having the main-chain modules and side-branch modules at the two ends appears to have allowed re-assortment of the modules by occasional recombination between them in the *wzx* gene. A similar situation can now be seen to apply to the much smaller *S. enterica* Gal-initiated set of O antigens. It should be noted that the divergence-level symmetry does not require comparable numbers of genes at each end, and in *St. pneumoniae* the serotype-specific genes are all at the 3′ end (Aanensen *et al.* 2007). The selection arises because where the 'rule' is broken, it is less likely that the appropriate set of genes will be transferred by recombination and this is critical over the long term. These genetic observations suggest that the organisation of the genetic content of the many OPS gene clusters in the *Y. pseudotuberculosis* complex has been refined over a very long time period frame to achieve such a good fit to the ideal pattern, as selection pressure would be reduced by the fact that recombination in genes on either side of the gene cluster would be effective even if the genes were randomly distributed in the gene cluster.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSRE* online.

ACKNOWLEDGEMENTS

The genetics of O-antigen synthesis is only truly meaningful in the context of the structures they code for, and we acknowledge the importance to our story of the structural work carried out by collaborators Cristina De Castro, Otto Holst, Yuriy Knirel and Antonio Molinaro. We also acknowledge the major contributions of collaborator Mikael Skurnik to work on Yersinia, and coordinating some of the above collaborations. We are grateful for very helpful suggestions by the assessors.

FUNDING

Work at the University of Sydney has been supported by Australian Research Council grants.

Conflict of interest. None declared.

REFERENCES

Aanensen DM, Mavroidi A, Bentley SD *et al.* Predicted functions and linkage specificities of the products of the

Streptococcus pneumoniae capsular biosynthetic loci. *J Bacteriol* 2007;**189**:7856–76.

Achtman M, Zurth K, Morelli G *et al*. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *P Natl Acad Sci USA* 1999;**96**:14043–8.

Alam J, Beyer N, Liu HW. Biosynthesis of colitose: expression, purification, and mechanistic characterization of GDP-4-keto-6-deoxymannose-3-dehydrase (ColD) and GDP-L-coltiose synthase (ColC). *Biochemistry* 2004;**43**:16450–60.

Aleksic S, Suchan G, Bockemühl J *et al*. An extended antigenic scheme for *Yersinia pseudotuberculosis*. In: Une T, Maruyama T, Tsubokura M (eds). *Contributions to Microbiology and Immunology; Current Investigations of the Microbiology of Yersiniae*. Basel, Switzerland: Karger, 1991, 235–8.

Alqasim A, Scheutz F, Zong Z *et al*. Comparative genome analysis identifies few traits unique to the *Escherichia coli* ST131 H30Rx clade and extensive mosaicism at the capsule locus. *BMC Genomics* 2014;**15**:830.

Atzinger A, Butela K, Lawrence JG. The O antigen mediates differential survival of *Salmonella* against communities of natural predators. *Microbiology* 2016;**162**:610–21.

Bailey MJA, Hughes C, Koronakis V. RfaH and the *ops* element, components of a novel system controlling bacterial transcription elongation. *Mol Microbiol* 1997;**26**:845–51.

Beczala A, Ovchinnikova O, Datta N *et al*. Structure and genetic basis of *Yersinia similis* serotype O:9 O-specific polysaccharide. *Innate Immun* 2013;**21**:3–16.

Bengoechea JA, Najdenski H, Skurnik M. Lipopolysaccharide O antigen status of *Yersinia enterocolitica* O:8 is essential for virulence and absence of O antigen affects the expression of other *Yersinia* virulence factors. *Mol Microbiol* 2004;**52**:451–69.

Bengoechea JA, Pinta E, Salminen T *et al*. Functional characterization of Gne (UDP-*N*-acetylglucosamine-4-epimerase), Wzz (chain length determinant), and Wzy (O-antigen polymerase) of *Yersinia enterocolitica* serotype O:8. *J Bacteriol* 2002;**184**: 4277–87.

Bercovier H, Mollaret HH, Alonso JM *et al*. Intra- and interspecies relatedness of *Yersinia pestis* by DNA hybridization and its relationship to *Yersina pseudotuberculosis*. *Curr Microbiol* 1980;**4**:225–9.

Bogdanovich T, Carniel E, Fukushima H *et al*. Use of O-antigen gene cluster-specific PCRs for the identification and O-genotyping of *Yersinia pseudotuberculosis* and *Yersinia pestis*. *J Clin Microbiol* 2003;**41**:5103–12.

Bruneteau M, Minka S. Lipopolysaccharides of bacterial pathogens from the genus *Yersinia*: a mini-review. *Biochimie* 2003;**85**:145–52.

Butty FD, Aucoin M, Morrison L *et al*. Elucidating the formation of 6-Deoxyheptose: biochemical characterization of the GDP-D-*glycero*-D-*manno*-heptose C6 dehydratase, DmhA, and its associated C4 reductase, DmhB. *Biochemistry* 2009;**48**: 7764–75.

Carniel E, Autenrieth I, Cornelis G *et al*. *Y. enterocolitica* and *Y. pseudotuberculosis*. In: Dworkin M, Falkow S, Rosenberg E (eds). *The Prokaryotes: A Handbook on the Biology of Bacteria*. New York: Springer, 2006, 270–398.

Carniel E. Plasmids and pathogenicity islands of *Yersinia*. In: Hacker J, Kaper JB (eds). *Pathogenicity Islands and the Evolution of Pathogenic Microbes*. Berlin: Springer, 2002, 89–108.

Ch'ng SL, Octavia S, Xia Q *et al*. Population structure and evolution of pathogenicity of *Yersinia pseudotuberculosis*. *Appl Environ Microb* 2011;**77**:768–75.

Chen H, Guo Z, Liu H. Biosynthesis of Yersiniose: attachment of the two-carbon branched-chain is catalyzed by a thiamine pyrophosphate-dependent flavoprotein. *J Am Chem Soc* 1998;**120**:11796–7.

Conde-Álvarez R, Arce-Gorvel V, Iriarte M *et al*. The lipopolysaccharide core of *Brucella abortus* acts as a shield against innate immunity recognition. *PLoS Pathog* 2012;**8**:e1002675.

Croucher NJ, Harris SR, Fraser C *et al*. Rapid pneumococcal evolution in response to clinical interventions. *Science* 2011;**331**:430–4.

Cunneen MM, De Castro C, Kenyon JJ *et al*. The O-specific polysaccharide structure and biosynthetic gene cluster of *Yersinia pseudotuberculosis* serotype O:11. *Carbohydr Res* 2009;**344**:1533–40.

Cunneen MM, Liu B, Wang L *et al*. Biosynthesis of UDP-GlcNAc, UndPP-GlcNAc and UDP-GlcNAcA involves three easily distinguished 4-epimerase enzymes, Gne, Gnu and GnaB. *PLoS One* 2013;**8**:e67646.

Cunneen MM, Pacinelli E, Song WC *et al*. Genetic analysis of the O-antigen gene clusters of *Yersinia pseudotuberculosis* O:6 and O:7. *Glycobiology* 2011;**21**:1140–6.

Cunneen MM, Reeves PR. The *Yersinia kristensenii* O11 O-antigen gene cluster was acquired by lateral gene transfer and incorporated at a novel chromosomal locus. *Mol Biol Evol* 2007;**24**:1355–65.

De Castro C, Kenyon JJ, Cunneen MM *et al*. Genetic characterisation and structural analysis of the O-specific polysaccharide of *Yersinia pseudotuberculosis* serotype O:1c. *Innate Immun* 2011;**17**:183–90.

De Castro C, Kenyon JJ, Cunneen MM *et al*. The O-specific polysaccharide structure and gene cluster of serotype O:12 of the *Yersinia pseudotuberculosis* complex, and the identification of a novel L-quinovose biosynthesis gene. *Glycobiology* 2012;**23**:346–53.

De Castro C, Skurnik M, Molinaro A *et al*. Characterization of the O-polysaccharide structure and biosynthetic gene cluster of *Yersinia pseudotuberculosis serotype* O:15. *Innate Immun* 2009;**15**:351–9.

DebRoy C, Roberts E, Fratamico PM. Detection of O antigens in *Escherichia coli*. *Anim Health Res Rev* 2011;**12**:169–85.

Duan R., Liang J, Shi G *et al*. Homology analysis of pathogenic *Yersinia*: *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, and *Yersinia pestis* based on multilocus sequence typing. *J Clin Microbiol* 2014;**52**:20–9.

Erickson DL, Lew CS, Kartchner B *et al*. Lipopolysaccharide biosynthesis genes of *Yersinia pseudotuberculosis* promote resistance to antimicrobial chemokines. *PLoS One* 2016;**11**:e0157092.

Erridge C, Stewart J, Bennett-Guerrero E *et al*. The biological activity of a liposomal complete core lipopolysaccharide vaccine. *Innate Immun* 2002;**8**:39–46.

Faruque SM, Sack DA, Sack RB *et al*. Emergence and evolution of *Vibrio cholerae* O139. *P Natl Acad Sci USA* 2003;**100**:1304–9.

Finlay B, McFadden G. Anti-Immunology: Evasion of the host immune system by bacterial and viral pathogens. *Cell* 2006;**124**:767–82.

Frirdich E, Whitfield C. Review: Lipopolysaccharide inner core oligosaccharide structure and outer membrane stability in human pathogens belonging to the *Enterobacteriaceae*. *Innate Immun* 2005;**11**:133–44.

Fukushima H, Matsuda Y, Seki R *et al*. Geographical heterogeneity between Far Eastern and Western countries in prevalence of the virulence plasmid, the superantigen *Yersinia pseudotuberculosis*-derived mitogen, and the high-pathogenicity island among *Yersinia pseudotuberculosis* strains. *J Clin Microbiol* 2001;**39**:3541–7.

Garzetti D, Susen R, Fruth A *et al.* A molecular scheme for *Yersinia enterocolitica* patho-serotyping derived from genome-wise analysis. *Int J Med Microbiol* 2014;**304**:275–83.

Gibson B, Melaugh W, Phillips N *et al.* Investigation of the structural heterogeneity of lipooligosaccharides from pathogenic *Haemophilus* and *Neisseria* species and of R-type lipopolysaccharides from *Salmonella typhimurium* by electrospray mass spectrometry. *J Bacteriol* 1993;**175**:2702–12.

Gorshkova RP, Korchahina NI, Ovodov YS. Structural studies on the O-specific side-chain polysaccharide of lipopolysaccharide from the *Yersinia pseudotuberculosis* VA serovar. *Eur J Biochem* 1983;**131**:345–7.

Gorshkova RP, Zubkov VA, Isakov VV *et al.* Structural features of O-specific polysaccharide from lipopolysaccharide of *Yersinia pseudotuberculosis* VI serovar. *Bioorg Khim* 1983;**9**:1068–73.

Hata H, Natori T, Mizuno T *et al.* Phylogenetics of family *Enterobacteriaceae* and proposal to reclassify *Escherichia hermannii* and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter subterranea* gen. nov., comb. nov. *Microbiol Immunol* 2016;**60**:303–11.

Heinrichs D, Yethon J, Whitfield C. Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*. *Mol Micro* 2002;**30**:221–32.

Ho N, Kondakova AN, Knirel YA *et al.* The biosynthesis and biological role of 6-deoxyheptose in the lipopolysaccharide O-antigen of *Yersinia pseudotuberculosis*. *Mol Microbiol* 2008;**68**:424–47.

Hobbs M, Reeves PR. The JUMPstart sequence: a 39 bp element common to several polysaccharide gene clusters. *Mol Microbiol* 1994;**12**:855–6.

Holt KE, Kenyon JJ, Hamidian M *et al.* Five decades of genome evolution in the globally distributed, extensively antibiotic resistant *Acinetobacter baumannii* global clone 1. *Microbial Genomics* 2016;**2**:1–16.

Hu D, Liu B, Dijkshoorn L *et al.* Diversity in the major polysaccharide antigen of *Acinetobacter baumannii* assessed by DNA sequencing, and development of a molecular serotyping scheme. *PLoS One* 2013;**8**:e70329.

Iguchi A, Iyoda S, Kikuchi T *et al.* A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. *DNA Res* 2015;**22**:101–7.

Jalava K, Hallanvuo S, Nakari UM *et al.* Multiple outbreaks of *Yersinia pseudotuberculosis* infections in Finland. *J Clin Microbiol* 2004;**42**:2789–91.

Jiang SM, Wang L, Reeves PR. Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8 and 18C capsular polysaccharide gene clusters. *Infect Immun* 2001;**69**:1244–55.

Kenyon JJ, De Castro C, Cunneen MM *et al.* The genetics and structure of the O-specific polysaccharide of *Yersinia pseudotuberculosis* serotype O:10 and its relationship with *Escherichia coli* O111 and *Salmonella enterica* O35. *Glycobiology* 2011;**21**:1131–9.

Kenyon JJ, Duda K, De Felice A *et al.* Serotype O:8 isolates in the *Yersinia pseudotuberculosis* complex have different O-antigen gene cluster and produce various forms of rough LPS. *Innate Immun* 2016;**22**:205–17.

Kenyon JJ, Hall RM. Variation in the complex carbohydrate biosynthesis loci of *Acinetobacter baumannii* genomes. *PLoS One* 2013;**8**:e62160.

Kenyon JJ, Reeves PR. The Wzy O-antigen polymerase of *Yersinia pseudotuberculosis* O:2a has a dependence on the Wzz chain-length determinant for efficient polymerization. *FEMS Microbiol Lett* 2013;**349**:163–70.

Knirel YA. Structure of O-antigens. In: Knirel YA, Valvano M (eds). *Bacterial Lipopolysaccharides*. Vienna: Springer, 2011. 42–108.

Kondakova A, Sevillano A, Shaikhutdinova R *et al.* Revision of the O-polysaccharide structure of *Yersinia pseudotuberculosis* O:1a; confirmation of the function of WbyM as paratosyltransferase. *Carbohydr Res* 2012;**350**:98–102.

Kondakova AN, Bystrova OV, Shaikhutdinova RZ *et al.* Reinvestigation of the O-antigens of *Yersinia pseudotuberculosis*: revision of the O2c and confirmation of the O3 antigen structures. *Carbohydr Res* 2008a;**343**:2486–8.

Kondakova AN, Bystrova OV, Shaikhutdinova RZ *et al.* Structure of the O-polysaccharide of *Yersinia pseudotuberculosis* O:2b. *Carbohydr Res* 2009a;**344**:405–7.

Kondakova AN, Bystrova OV, Shaikhutdinova RZ *et al.* Structure of the O-antigen of *Yersinia pseudotuberculosis* O:4a revised. *Carbohydr Res* 2009b;**344**:531–4.

Kondakova AN, Bystrova OV, Shaikhutdinova RZ *et al.* Structure of the O-antigen of *Yersinia pseudotuberculosis* O:4b. *Carbohydr Res* 2009c;**344**:152–4.

Kondakova AN, Ho N, Bystrova OV *et al.* Structural studies of the O-antigens of *Yersinia pseudotuberculosis* O:2a and mutants thereof with impaired 6-deoxy-d-*manno*-heptose biosynthesis pathway. *Carbohydr Res* 2008b;**343**:1383–9.

Kondakova AN, Shaikhutdinova RZ, Ivanov SA *et al.* Revision of the O-polysaccharide structure of *Yersinia pseudotuberculosis* O:1b. *Carbohydr Res* 2009d;**344**:2421–3.

Korchagina NI, Gorshkova RP, Ovodov YS. Studies on O-specific polysaccharide from *Yersinia pseudotuberculosis* VB serovar. *Bioorg Khim* 1982;**8**:1666–9.

Kotandrova NA, Gorshkova RP, Zubkov VA *et al.* The structure of the O-specific polysaccharide chain of the lipopolysaccharide of *Yersinia pseudotuberculosis* serovar VII. *Bioorg Khim* 1989;**15**:104–10.

Laukkanen-Ninios R, Didelot X, Jolley KA *et al.* Population structure of the *Yersinia pseudotuberculosis* complex according to multilocus sequence typing. *Environ Microbiol* 2011;**13**:3114–27.

Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 1997;**44**:383–97.

Lehrer J, Vigeant KA, Tatar LD *et al.* Functional characterization and membrane topology of *Escherichia coli* WecA, a sugar-phosphate transferase initiating the biosynthesis of enterobacterial common antigen and O-antigen lipopolysaccharide. *J Bacteriol* 2007;**189**:2618–28.

Li Q, Reeves PR. Genetic variation of dTDP-L-rhamnose pathway genes in *Salmonella enterica*. *Microbiology* 2000;**146**:2291–307.

Liu B, Knirel YA, Feng L *et al.* Structure and genetics of *Shigella* O antigens. *FEMS Microbiol Rev* 2008;**32**:627–53.

Liu B, Knirel YA, Feng L *et al.* Structural diversity in *Salmonella* O antigens and its genetic basis. *FEMS Microbiol Rev* 2014;**38**:56–89.

McNally A, Thomson NR, Reuter S *et al.* 'Add, stir and reduce': *Yersinia spp.* as model bacteria for pathogen evolution. *Nat Rev Microbiol* 2016;**14**:177–90.

March C, Cano V, Moranta D *et al.* Role of bacterial surface structures on the interaction of *Klebsiella pneumoniae* with phagocytes. *PLoS One* 2013;**8**:e56847.

Matsuhashi S, Matsuhashi M, Brown JG *et al.* Enzymatic synthesis of cytidine diphosphate 3,6-dideoxyhexoses. *J Biol Chem* 1966;**241**:4283–7.

Mecsas J, Bilis I, Falkow S. Identification of attenuated *Yersinia pseudotuberculosis* strains and characterization of an orogastric infection in BALB/c mice on day 5 postinfection by signature-tagged mutagenesis. *Infect Immun* 2001;**69**:2779–87.

Meier-Dieter U, Barr K, Starman R *et al.* Nucleotide sequence of the *Escherichia coli rfe* gene involved in the synthesis of enterobacterial common antigen. *J Biol Chem* 1992;**267**:746–53.

Milkman R, Jaeger E, McBride RD. Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. *Genetics* 2003;**163**:475–83.

Moxon R, Bayliss C, Hood D. Bacterial contigency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* 2006;**40**:307–33.

Nagano T, Kiyohara T, Suzuki K *et al.* Identification of pathogenic strains within serogroups of *Yersinia pseudotuberculosis* and the presence of non-pathogenic strains isolated from animals and the environment. *J Vet Med Sci* 1997;**59**:153–8.

Nakano T, Kawaguchi H, Nakao K *et al.* Two outbreaks of *Yersinia pseudotuberculosis* 5a infection in Japan. *Scand J Infect Dis* 1989;**21**:175–9.

Nuorti JP, Niskanen T, Hallanvuo S *et al.* A widespread outbreak of *Yersinia pseudotuberculosis* O:3 infection from iceberg lettuce. *J Infect Dis* 2004;**189**:766–74.

Pacinelli E, Wang L, Reeves PR. Relationship of *Yersinia pseudotuberculosis* O antigens IA, IIA and IVB: The IIA gene cluster was derived from that of IVB. *Infect Immun* 2002;**70**:3271–6.

Paradis S, Boissinot M, Paquette N *et al.* Phylogeny of the *Enterobacteriaceae* based on genes encoding elongation factor Tu and F-ATPase β-subunit. *Int J Syst Evol* **Micr** 2005;**55**: 2013–25.

Pärn T, Hallanvuo S, Salmenlinna S *et al.* Outbreak of *Yersinia pseudotuberculosis* O:1 infection associated with raw milk consumption, Finland, spring 2014. *Euro Surveill* 2015;**20**:30033.

Pier GB. *Pseudomonas aeruginosa* lipopolysaccharide: a major virulence factor, initiator of inflammation and target for effective immunity. *Int J Med Micrbiol* 2007;**297**:277–95.

Plainvert C, Bidet P, Peigne C *et al.* A new O-antigen gene cluster has a key role in the virulence of the *Escherichia coli* meningitis clone O45:K1:H7. *J Bacteriol* 2007;**189**:8528–36.

Porat R, McCabe WR, Brubaker RR. Lipopolysaccharide-associated resistance to killing of *Yersiniae* by complement. *J Endotox Res* 1995;**2**:91–7.

Prior JL, Hitchen PG, Williamson DE. Characterisation of the lipopolysaccharide of *Yersinia pestis*. *Microb Pathog* 2001;**30**:49–57.

Ramamurthy T, Yamasaki S, Takeda Y *et al.* *Vibrio cholerae* O139 Bengal: odyssey of a fortuitous variant. *Microb Infect* 2003;**5**:329–44.

Reuter S, Connor TR, Barquist L *et al.* Parallel independent evolution of pathogenicity within the genus *Yersinia*. *PNAS* 2014;**111**:6768–73.

Reeves P, Cunneen MM, Liu B *et al.* Genetics and evolution of the *Salmonella* galactose-initiated set of O antigens. *PLoS One* 2013;**8**:e69306.

Reeves PR, Cunneen MM. Biosynthesis of O-antigen chains and assembly. In: Moran AP, Holst O, Brennan PJ *et al.* (eds). *Microbial Glycobiology: Structures, Relevance and Applications*. London: Academic Press, 2009, 319–35.

Reeves PR, Cunneen MM. Evolution of lipopolysaccharide biosynthesis genes. In: Valvano MA, Knirel YA (eds). *Bacterial Lipopolysaccharides: Structure, Chemical Synthesis, Biogenesis, and Interaction with Host Cells*. London, UK: Springer, 2011, 283–307.

Reeves PR, Pacinelli E, Wang L. O-antigen gene clusters of *Yersinia pseudotuberculosis*. *Adv Exp Med Biol* 2003;**529**:199–206.

Reeves PR, Wang L. Genomic organization of LPS-specific loci. In: Hacker J, Kaper JB, Compans W *et al.* (eds). *Pathogenicity Islands and the Evolution of Microbes (Vol. 2)*; Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer, 2002, 109–35.

Samuel G, Reeves PR. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr Res* 2003;**338**: 2503–19.

Savin C, Martin L, Bouchier C *et al.* The *Yersinia pseudotuberculosis* complex: Characterisation and delineation of a new species, *Yersinia wautersii*. *Int J Med Microbiol* 2014;**304**:452–63.

Silhavy TJ, Kahne D, Walker S. The bacterial cell envelope. *Cold Spring Harb Perspect Biol* 2010;**2**:a000414.

Skurnik M, Bengoechea JA. The biosynthesis and biological role of lipopolysaccharide O-antigens of pathogenic *Yersiniae*. *Carbohydr Res* 2003;**338**:2521–9.

Skurnik M, Biedzka-Sarek M, Lübeck PS *et al.* Characterization and biological role of the O-polysaccharide gene cluster of *Yersinia enterocolitica* serotype O:9. *J Bacteriol* 2007;**189**:7244–53.

Skurnik M, Peippo A, Ervela E. 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 Y. *pseudotuberculosis* serotype O:1b. *Mol Microbiol* 2000;**37**:316–30.

Skurnik M, Venho R, Bengoechea JA *et al.* The lipopolysaccharide outer core of *Yersinia enterocolitica* serotype O:3 is required for virulence and plays a role in outer membrane integrity. *Mol Microbiol* 1999;**31**:1443–62.

Skurnik M, Venho R, Toivanen P *et al.* A novel locus of *Yersinia enterocolitica* O:3 involved in lipopolysaccharide outer core biosynthesis. *Mol Microbiol* 1995;**17**:575–94.

Sprague LD, Scholz HC, Amann S *et al.* *Yersinia similis sp*. nov. *Int J Syst Evol Micr* 2008;**58**:952–8.

Tian T, Salis HM. A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons. *Nucleic Acids Res* 2015;**43**:7137–51.

Thal E, Knapp W. A revised antigenic scheme of *Yersinia pseudotuberculosis*. *Symp Series Immunobiol Standard* 1971;**13**: 219–22.

Thorson JS, Lo SF, Ploux O *et al.* Studies of the biosynthesis of 3,6-dideoxyhexoses: molecular cloning and characterization of the *asc* (ascarylose) region from *Yersinia pseudotuberculosis* serogroup VA. *J Bacteriol* 1994;**176**:5483–93.

Timchenko NF, Adgamov RR, Popov AF *et al.* Far East Scarlet-like fever caused by a few related genotypes of *Yersinia pseudotuberculosis*, Russia. *Emerg Infect Dis* 2016;**22**:503–6.

Trent MS, Stead CM, Tran AX *et al.* Diversity of endotoxin and its impact on pathogenesis. *J Endotoxin Res* 2006;**12**:205–23.

Tsubokura M, Aleksic S. A simplified antigenic scheme for serotyping of *Yersinia pseudotuberculosis*: phenotypic characterization of reference strains and preparation of O and H factor sera. *Contrib Microbiol Immunol* 1995;**13**:99–105.

Tsubokura M, Aleksic S, Fukishima H *et al.* Characterization of *Yersinia pseudotuberculosis* serogroups O9, O10 and O11; subdivision of O1 serogroup into O1a, O1b, and O1c subgroups. *Zentralbl Bakteriol* 1993;**278**:500–9.

Tsubokura M, Otsuki K, Kawaoka Y *et al.* Addition of new serogroups and improvement of the antigenic designs of *Yersinia pseudotuberculosis*. *Curr Microbiol* 1984;**11**: 89–92.

Vinogradov E, Nossovaa L, Radziejewska-Lebrechtb J. The structure of the O-specific polysaccharide from *Salmonella cerro* (serogroup K, O:6,14,18). *Carbohydr Res* 2004;**339**:2441–3.

West NP, Sansonetti P, Mounier J *et al.* Optimization of virulence functions through glucosylation of *Shigella* LPS. *Science* 2005;**307**:1313–7.

Whitfield C, Kaniuk N, Frirdich E. Molecular insights into the assembly and diversity of the outer core oligosaccharide in lipopolysaccharides from *Escherichia coli* and *Salmonella*. *J Endotoxin Res* 2003;**9**:244–9.

Wildschutte H, Wolfe DM, Tamewitz A *et al*. Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonella*. *P Natl Acad Sci USA* 2004;**101**:10644–9.

Woodward R, Yi W, Li L *et al*. *In vitro* bacterial polysaccharide biosynthesis: defining the functions of Wzy and Wzz. *Nat Chem Biol* 2010;**6**:418–23.

Wyres KL, Gorrie C, Edwards DJ *et al*. Extensive capsule locus variation and large-scale genomic recombination within the *Klebsiella pneumoniae* clonal group 258. *Genome Biol Evol* 2015;**7**:1267–79.

Zhang G, Meredith T, Kahne D. On the essentiality of lipopolysaccharide to Gram-negative bacteria. *Curr Microbiol* 2013;**16**:779–85.

Zhang L, al-Hendy A, Toivanen P *et al*. Genetic organization and sequence of the rfb gene cluster of *Yersinia enterocolitica* serotype O:3: similarities to the dTDP-L-rhamnose biosynthesis pathway of *Salmonella* and to the bacterial polysaccharide transport systems. *Mol Microbiol* 1993;**9**:309–21.

Zhang L, Radziejewska-Lebrecht J, Krajewska-Pietrasik D *et al*. Molecular and chemical characterization of the lipopolysaccharide O-antigen and its role in the virulence of *Yersinia enterocolitica* serotype O:8. *Mol Microbiol* 1997;**23**:63–76.

Zhou Z, Li X, Liu B *et al*. Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. *PLoS One* 2010;**5**:e8700.