Title	The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected Lactococcus lactis strains
Authors	Mahony, Jennifer;Frantzen, Cyril;Vinogradov, Evgeny;Sadovskaya, Irina;Theodorou, Ilias;Kelleher, Philip;Chapot-Chartier, Marie-Pierre;Cambillau, Christian;Holo, Helge;van Sinderen, Douwe
Publication date	2020-06-09
Original Citation	Mahony, J., Frantzen, C., Vinogradov, E., Sadovskaya, I., Theodorou, I., Kelleher, P., Chapot-Chartier, M.-P., Cambillau, C., Holo, H. and van Sinderen, D. (2020) 'The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected Lactococcus lactis strains', Molecular Microbiology, 114(4), pp. 582-596. doi: 10.1111/mmi.14561
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1111/mmi.14561
Rights	© 2020, John Wiley & Sons Ltd. This is the peer reviewed version of the following item: Mahony, J., Frantzen, C., Vinogradov, E., Sadovskaya, I., Theodorou, I., Kelleher, P., Chapot-Chartier, M.-P., Cambillau, C., Holo, H. and van Sinderen, D. (2020) 'The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selectedLactococcus lactisstrains', Molecular Microbiology, 114(4), pp. 582-596, doi: 10.1111/mmi.14561, which has been published in final form at: https://doi.org/10.1111/mmi.14561 This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.
Download date	2024-05-06 18:22:02
Item downloaded from	https://hdl.handle.net/10468/12217

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

DR. JENNIFER MAHONY (Orcid ID : 0000-0001-5846-6303)

DR. MARIE-PIERRE CHAPOT-CHARTIER (Orcid ID : 0000-0002-4947-0519)

DR. CHRISTIAN CAMBILLAU (Orcid ID : 0000-0001-5502-4729)

Article type : Research Article

Title: The CWPS Rubik's cube: linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains.

Running title: Lactococcal cell wall polysaccharide biosynthesis

ΨJennifer Mahony*[1,2], ΨCyril Frantzen[3], ΨEvgeny Vinogradov[4], Irina Sadovskaya[5], Ilias Theodorou[1], Philip Kelleher[1,2], Marie-Pierre Chapot-Chartier[6], Christian Cambillau[1,7], Helge Holo[3] and Douwe van Sinderen*[1,2]

[1]School of Microbiology, University College Cork (UCC), Cork, Ireland; [2]APC Microbiome Ireland, University College Cork, Cork, Ireland; [3]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NULS), Ås, Norway; [4]National Research Council Canada, Institute for Biological Sciences, 100 Sussex Drive, Ottawa, ON K1A 1L1, Canada; [5]Equipe BPA, Université du Littoral-Côte d'Opale, Institut Charles Violette EA 7394 USC Anses, Bd Bassin Napoléon, BP 120, 62327 Boulogne-sur-mer, France; [6]Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France; [7]Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique (CNRS), Campus de Luminy, 13009 Marseille, France.

ΨThese authors have contributed equally

*Corresponding authors: School of Microbiology, University College Cork, Western Road, Cork T12 K8AF, Ireland; Telephone: +353 21 4901365; Fax: +353 21 4903101; E-mail: j.mahony@ucc.ie or d.vansinderen@ucc.ie

Keywords: Bacteriophage, *Lactococcus lactis*, dairy, phage receptor, polysaccharide, genomics, cell wall polysaccharide (CWPS)

Conflict of interest statement: The authors declare no competing interests.

Summary. The biosynthetic machinery for cell wall polysaccharide (CWPS) production in lactococci is encoded by a large gene cluster, designated *cwps*. This locus displays considerable variation among lactococcal genomes, previously prompting a classification into three distinct genotypes (A-C). In the present study, the *cwps* loci of 107 lactococcal strains were compared, revealing the presence of a fourth *cwps* genotype (type D). Lactococcal CWPSs are comprised of two saccharidic structures: a peptidoglycan-embedded rhamnan backbone polymer to which a surface-exposed, poly/oligosaccharidic side-chain is covalently linked. Chemical structures of the side-chain of seven lactococcal strains were elucidated, highlighting their diverse and strain-specific nature. Furthermore, a link between *cwps* genotype and chemical structure was derived based on the number of glycosyltransferase-encoding genes in the *cwps* cluster and the presence of conserved genes encoding the presumed priming glycosyltransferase. This facilitates predictions of several structural features of lactococcal CWPSs including (i) whether the CWPS possesses short oligo/polysaccharide side-chains, (ii) the number of component monosaccharides in a given CWPS structure, (iii) the order of monosaccharide incorporation into the repeating units of the side-chain (for C-type strains), (iv) the presence of Gal*f* and phosphodiester bonds in the side-chain, and (v) the presence of glycerol phosphate substituents in the side-chain.

Introduction

The initial interaction between members of several lactococcal phage groups (i.e. members of the 936, P335, 1358, 949 and P087 groups) and their host is mediated by a particular component of the cell wall polysaccharide (CWPS) layer that is presented on the cell surface, being termed the polysaccharide pellicle (PSP) (Ainsworth et al., 2014; Chapot-Chartier et al., 2010; Farenc et al., 2014; Mahony et al., 2013; Mahony, Randazzo, Neve, Settanni, & van Sinderen, 2015). This PSP layer is (in many cases) made of a phosphate-containing polysaccharide chain that is believed to be covalently attached to a second CWPS, a peptidoglycan-embedded rhamnan, which was shown to be a polymer of a neutral rhamnose-containing trisaccharide subunit in the model strain MG1363 (Sadovskaya et al., 2017). The PSP exhibits structural diversity between strains which appears as a major factor explaining the narrow specificity of lactococcal phages. Noteworthy, however, is that while this receptor moiety allows a given phage to bind to a host, it is not necessarily synonymous with an ability to infect and lyse a strain due to the presence of inhibitory elements such as phage-resistance systems or the absence of a DNA injection trigger, for example. The gene cluster responsible for the biosynthesis of these two CWPS elements (i.e. rhamnan and PSP) is designated *cwps*. This genetic locus varies in size (between 20 and 30 kb) and genetic composition among lactococcal strains and has permitted the identification of three distinct CWPS genotypes (CWPS types A, B and C) (Mahony et al., 2013), as well as subtypes of CWPS type C strains (Ainsworth et al., 2014). The distinction between these genotypes corresponds to the presence of distinct glycosyltransferase-encoding genes in the 3' region of this cluster.

To date, the chemical structure of the CWPS (oligo/polysaccharide side-chain and/or rhamnan component) of five lactococcal strains has been defined, i.e. that of the A-type strain UC509.9 (Vinogradov, Sadovskaya, Grard, et al., 2018); the B-type strain IL1403 (Vinogradov, Sadovskaya, Courtin, et al., 2018) and the C-type strains MG1363 (Chapot-Chartier et al., 2010), 3107 (Ainsworth et al., 2014) and SMQ-388 (Farenc et al., 2014). The polysaccharide side-chain of MG1363 and SMQ-388 is composed of repeating phosphohexasaccharide subunits (Chapot-Chartier et al., 2010; Farenc et al., 2014), while that of *L. lactis* 3107 comprises a repeating phosphopentasaccharide subunit (Ainsworth et al., 2014). In contrast to the linear rhamnan and the regular, repeating and polymeric side chain structures observed among the analyzed C-type strains, the CWPS structures of type A and B strains appear more heterogeneous and complex, comprising a rhamnose-rich backbone that is partially substituted with short oligosaccharide side

chains. The CWPS of *L. lactis* UC509.9 (type A) consists of a linear backbone of tetrasaccharide repeating subunits containing one Glc and three Rha, with approximately every third unit substituted with a branched phosphate-containing oligosaccharide having a common trisaccharide and three non-stoichiometric substitutions including a glycerophosphate group (Vinogradov, Sadovskaya, Grard, et al., 2018). Similarly, the CWPS of the *L. lactis* type B strain IL1403 exhibits a linear backbone of an α-L-Rha repeating disaccharide subunit which is irregularly substituted with a trisaccharide that occasionally carries a glycerophosphate group (Vinogradov, Sadovskaya, Courtin, et al., 2018).

A dual chain biosynthetic pathway for rhamnan and its attached PSP structure has recently been proposed in the lactococcal strain MG1363 (Theodorou et al., 2019). In this model, rhamnan and PSP chains are synthesized on two different lipid-sugar precursor and joined at the extracellular side of the cytoplasmic membrane. Rhamnan is synthesized by an ABC-transporter dependent pathway with the first step consisting in the transfer of GlcNAc-P from UDP-GlcNAc onto the lipid carrier undecaprenyl-phosphate (Und-P) achieved by TagO, located outside of the *cwps* gene cluster leading to Und-PP-GlcNAc (Sadovskaya et al., 2017). Regarding PSP, its biosynthesis commences with the transfer of a GlcNAc residue from UDP-GlcNAc to Und-P by a priming glycosyltransferase encoded by WpsA (Wps: wall polysaccharide) and assisted by a small membrane protein WpsB generating Und-P-GlcNAc (Chapot-Chartier et al., 2010; Theodorou et al., 2019). The PSP subunit is then further elongated in the cytoplasm by several glycosyltransferases (WpsC, D, E and F in MG1363), then the lipid-linked PSP subunit is flipped outside the cytoplasmic membrane by WpsG, is polymerized by the GT-C fold glycosyltransferase WpsI assisted by the membrane protein WpsH, and finally attached to rhamnan by another GT-C fold glycosyltransferase, WpsJ. Using the MG1363 PSP synthesis model, we hypothesize that it is possible to identify the priming glycosyltransferase of many lactococcal strains based on sequence homology. Furthermore, having identified the priming glycosyltransferase, it may be possible to establish the number and identity of the component monosaccharides incorporated into the growing oligosaccharide subunit in other lactococcal strains based on similarity and gene order in the *cwps* cluster. Ultimately, this will permit the prediction of the CWPS structure of uncharacterized, yet sequenced *L. lactis* strains, which in turn may provide information on phage sensitivity of such strains. Beyond the *cwps* loci, two distinct, so-called three component glycosylation systems (TGSs) have been shown to add a glucose moiety to either the rhamnan or

PSP component of lactococcal CWPS structures (Theodorou et al., 2020) and thus contribute to CWPS structural diversity. A TGS is composed of three proteins: an Und-P-sugar activating glycosyltransferase (Und-P GT), a flippase transferring the sugar moiety outside the cytoplasmic membrane, and a polytopic glycosyltransferase (PolM GT) dedicated to attaching sugar residues to a specific glycopolymer. The presence or absence of homologues of these TGSs in lactococcal strains will thus guide predictions of such glucose side chains on CWPS structures.

In the current study, structural analysis of CWPS components of seven lactococcal strains combined with a comprehensive assessment of *cwps* gene clusters present in 107 lactococcal strains was undertaken in an effort to establish a link between the ever-increasing sequence data and the saccharidic structures that they correspond to.

Results

Diversity of the cwps gene cluster among lactococcal strains.

In the present study, we extracted the *cwps* cluster sequences of 107 lactococcal strains (68 sequenced as part of this study and 39 extracted from publicly available complete genome sequences) (Table S1). We evaluated the extent of genetic diversity of this cluster among the assessed strains, paying particular attention to the 3' variable region encoding the predicted glycosyltransferases presumed to be involved in the biosynthesis of the saccharidic structures of the rhamnan-linked side chains. A presence/absence matrix of protein families encoded by the assessed loci was derived using the Markov Clustering (MCL) algorithm followed by hierarchical clustering (HCL) analysis of the *cwps* gene clusters derived from 107 lactococcal strains, revealing 68 protein families and 37 distinct MCL profiles (Fig. 1).

Four distinct CWPS genotypes were identified: the three previously described genotypes (CWPS types A, B and C) plus one additional, genetically distinct group, designated here as CWPS type D, which incorporates the *cwps* clusters of strains 184 and KLDS4.0325. In addition to the four CWPS genotypes, eight subtypes of CWPS type C ($C_{1\text{-}8}$) strains were identified based on the presence/absence heatmap generated through HCL analysis. Subtypes were assigned where at least one protein family was identified to be present/absent in a subset of strains. Where it is based on the presence of a protein family, this protein family was required to be unique among strains of the same type (e.g. C_4 strains encode a single protein family in their *cwps* gene cluster that is not observed among those of any other C-subtype strains above a threshold of 50 % aa similarity). Interestingly, it was established through phylogenetic analysis of the 16S rRNA-encoding genes that CWPS genotypes are not linked to, or defined by, *L. lactis* subspecies (Table S1). However, it is noteworthy that a majority of *L. lactis* subsp. *lactis* strains are classified among the B-type and the majority of *L. lactis* subsp. *cremoris* strains belong to C-type strains, while strains of the A-type are a mixture of both subspecies.

A schematic representation of the diversity of glycosyltransferase-encoding genes within their associated gene clusters is presented in Figure 2. This comparative analysis highlights the high level of inter-CWPS type (A, B, C and D) diversity of glycosyltransferase-encoding genes, while the intra-type ($C_{1\text{-}8}$) diversity is much lower. For example, the *cwps* gene clusters of C_1 strains MG1363 and JM1 are identical with the exception of a transposase-encoding gene between the

fourth and fifth glycosyltransferase-encoding genes in the JM1 cluster, while the gene predicted to encode WpsH, the presumed co-polymerase of the side chain polysaccharide structures is also interrupted by the incorporation of a stop codon. Furthermore, C_4 strains W34 and 1196 possess identical *cwps* clusters with the exception of one gene within the glycosyltransferase-encoding region that is predicted to encode a nucleotidyltransferase (84 % aa identity), which does not appear to be intact. While the subtypes are somewhat arbitrarily assigned (based on the presence/absence analysis employed in this study), they may be useful in defining specific phage-host interactions.

Determination of novel CWPS structures from seven *L. lactis* strains

Based on the comparative analysis of the *cwps* clusters of the lactococcal strains described above, subtle differences were observed among individual CWPS C subtype strains. The apparent relative dominance of type C strains in the assessed collection and the genetic novelty associated with the D type strains invoked an analysis of the chemical diversity of PSP structures among members of these groups. Consequently, PSPs of *L. lactis* JM1 (C_1), SK11 (C_3), W34 (C_4), 1196 (C_4), IO-1 (C_5), A76 (C_6) and 184 (D) were extracted and purified, and their chemical structures were investigated by monosaccharide composition, methylation analysis and 2D NMR techniques (Supplementary Fig. 1-9 and Tables S2-9). Representative strains of C_7 and C_8 subtypes were not available for analysis in the context of this study.

L. lactis JM1 was selected for compositional analysis since the genetic composition of its *cwps* gene cluster is almost identical to that of MG1363, for which the CWPS structure and composition has previously been determined (Chapot-Chartier et al., 2010). While the CWPS of MG1363 has been defined as a repeating subunit of a phosphohexasaccharide, that of JM1 appears as repeating disaccharide subunits with what appears to be the initial GlcNAc and Glc substituent residues of the PSP of MG1363 (Table 1). The PSP of SK11 (C_3) was shown to be composed of repeating phosphohexasaccharide subunits containing Gal, Glc, Gal*f* and GlcNAc, while those of W34 and 1196 (both CWPS C_4 type strains) were determined to be made of pentasaccharide subunits devoid of phosphate (Table 1). The structure of the PSP of IO-1 is quite different from other C type strains, exhibiting repeating phosphopentasaccharide subunits that contain arabinitol (Ara-ol) residues unlike any previously studied lactococcal PSP, a finding that is consistent with its unique genetic content within the predicted glycosyltransferase-encoding region of the *cwps* gene cluster (Fig. 2). The C_6 representative strain A76 appears to exhibit a unique structure with repeating phosphohexasaccharide subunits containing glucose, galactose and O-acetylated rhamnose (Table 1). Similarly, strain 184 (CWPS type D) is entirely unique among lactococcal PSP structures to date, composed of repeating subunits of a ribitol-containing phosphopentasaccharide (Table 1). The genetic complexity associated with the central region of the *cwps* gene cluster that is predicted to encode the glycosyltransferase activities of these strains is mirrored by the structural diversity of their associated PSPs substantiating the role of the formerly proposed glycosyltransferases.

From genome sequence to structure predictions

With the availability of (i) a very substantial number of *cwps* cluster sequences provided through the current and previous studies, (ii) several CWPS structures, and (iii) recently developed models for rhamnan and PSP biosynthesis in MG1363, it is now possible to make predictions on a number of key characteristics of CWPS components. These include predictions of conserved groups of rhamnan structures, the number and order of component monosaccharides in CWPS structures and identifying strains that would produce monomeric or polymeric oligosaccharide side chains. Furthermore, the chemical structures of CWPS components elucidated for a number of representative strains of each type/subtype in the present study were used to substantiate the above-mentioned hypotheses. While it cannot be precluded that strain-specific factors including the impact of minor sequence variations on the activity of glycosyltransferases may impede accurate predictions of structures, bioinformatic interrogation allows us to envisage the potential and capacity of a given strain to encode structural features in a way that was thus far not possible.

The enzymes responsible for biosynthesis of the rhamnan component of lactococcal CWPSs are encoded by the proximal end of the *cwps* loci and this region of the cluster appears to be conserved across the majority of lactococcal strains, particularly among C-type strains (Fig. 2). This led to an early assumption that the rhamnan structure of the majority of lactococcal strains is highly conserved; however, it is now clear that the rhamnan subunits of lactococci may represent di-, tri- or tetrasaccharides (Fig. 2).

In contrast to the considerable degree of conservation of the rhamnan-specifying region of the *cwps* cluster of C-type strains, those of A- and B-type strains are quite divergent. Despite this, the rhamnan-encoding region and its associated functions in all evaluated lactococcal genomes display high levels of similarity and synteny with the following general organization: genes encoding the rhamnose precursor proteins (*rmlACBD*) followed by the rhamnosyltransferase-encoding genes *rgpA* and *rgpB*, the transport protein-encoding genes *rgpC* and *rgpD*; *wpsJ* whose product is predicted to transfer the saccharidic side chain on to the rhamnan component; *rgpE* proposed to encode the rhamnan chain termination glycosyltransferase and *rgpF*, the final rhamnosyltransferase of the rhamnan biosynthetic cluster.

The *rmlA-D* genes (whose encoded proteins produce nucleotide-linked rhamnose precursors namely dTDP-Rha, for rhamnan biosynthesis) are highly conserved among all lactococcal strains as may be expected given the conserved nature of the rhamnan backbone structures (Fig. 2). The rhamnan components of JM1, W34 and 1196 were defined as being composed of -2-α-Rha-3-α-Rha-2-α-Rha- repeating subunits, identical to those of the previously characterized strains, MG1363 and 3107 (Ainsworth et al., 2014; Sadovskaya et al., 2017). The rhamnan subunit of the C-type strain SK11 is similar to those of all other analyzed C-type strains, though it contains a stoichiometric Glc side chain decoration (Table 1). The transfer of a glucose side-chain to the rhamnan structure was recently shown to be mediated by a three-component glycosylation system (TGS) incorporating two glycosyltransferases (CsdAB, of which the corresponding genes are located outside of the *cwps* locus) and a flippase-encoding gene (Theodorou et al., 2020). The glucosylated rhamnan structure of SK11 is consistent with the presence of homologues of *csdAB* and the flippase-encoding gene in the genome of SK11 (Theodorou et al., 2020). Downstream of *rmlA-D* are the rhamnosyltransferase-encoding genes, which appear less conserved between the different CWPS type strains (A-D). This genetic diversity is also borne out in the unique biochemical structure of the CWPS (i.e. rhamnan and side chain) characterized from the A- and B-type strains UC509.9 and IL1403, respectively (Table S10). The rhamnose-rich backbone of IL1403 has previously been defined as 3-α-Rha-2-α-Rha polymeric chains, while that of UC509.9 is composed of -6-α-Glc-2-α-Rha-2-α-Rha-2-α-Rha- tetrasaccharide subunits (Vinogradov, Sadovskaya, Courtin, et al., 2018; Vinogradov, Sadovskaya, Grard, et al., 2018). The *rgpA* and *rgpB* genes are conserved between strains IL1403 and UC509.9, while *rgpC, D, E* and *F* are quite divergent, which is consistent with the compositional and structural differences of their encoded rhamnan moieties. Interestingly, the *rgpEF* genes appear fused into a single gene in UC509.9, which may imply a dual function for the encoded protein in the transfer of rhamnose moieties on to the growing rhamnan component (RgpF function) and chain length termination (RgpE function). The rhamnan component of lactococci appears to be transferred to the outer face of the cell membrane via a highly conserved ABC transport system (RgpCD, Fig. 2). Downstream of the RgpF-encoding gene of UC509.9 and IL1403 is a glycosyltransferase-encoding gene that is common to both strains, yet absent among the C-type strains (named *ycbA* in IL1403 (Dupont et al., 2004)). The function of this glycosyltransferase is presently unclear.

The saccharidic or PSP side chains of the lactococcal CWPS are encoded by the distal end of the *cwps* cluster and represent highly variable structures (Fig. 2; see below). Despite this overall variability, the genes encoding the priming glycosyltransferase (WpsA) and its partner stimulatory protein (WpsB) were found to be highly conserved among the *cwps* clusters of most C-type strains (3107, JM1, SK11, W34, 1196 and IO-1; Fig. 2), while they are also identifiable in the A- and B-type strains UC509.9 and IL1403 (downstream of *ycbA* in IL1403 and named *ycbB* and *C*). Based on the recently proposed oligosaccharide PSP biosynthetic model in *L. lactis* MG1363, we propose that the GlcNAc residue is the first sugar to be transferred to the undecaprenyl phosphate lipid carrier in the subunit biosynthesis process of the oligosaccharide side chain of these C-type strains similar to the case of MG1363. For strains W34, 1196 and 3107, which each possess two GlcNAc residues in their oligosaccharide subunits, the 3-β-linked GlcNAc residue was selected to mirror that of the primary GlcNAc of MG1363. Using the MG1363 model, the order of monosaccharide incorporation during biosynthesis of the repeating oligosaccharide subunit was derived for five C-type strains analyzed in this study as well as those of the previously described oligosaccharides of the lactococcal strains SMQ-388 and 3107 (Table 1 & Fig. 2). Downstream of the WpsA-encoding gene is the WpsA-stimulating partner protein-encoding gene, *wpsB*, followed by the strain-specific glycosyltransferase-encoding genes. Each glycosyltransferase is believed to sequentially add a monosaccharide in an ordered manner implying that it is now possible to predict the specific glycosyltransferase activities of each of the glycosyltransferase-encoding genes within each strain's *cwps* cluster (i.e. of those that have been biochemically analyzed in this study). The Und-P-oligosaccharide subunit is then transferred to the outside face of the cell membrane via a flippase (WpsG of NZ9000 and equivalents in other strains) and this approach is used for all oligosaccharide side-chains, being in contrast to the ABC transport mechanism employed for the externalization of the rhamnan. To evaluate if the predicted order of the monosaccharides in the oligosaccharide chains was accurate, the presence and genomic position of phosphosugar transferases was assessed. This was based on the recent finding that *wpsE* of MG1363 encodes an enzyme with a Stealth_CR2 (PF11380) domain at the N-terminus of the protein, which is present in D-hexose-1-phosphate transferases and that can generate interglycosidic phosphate diester linkages (Theodorou et al., 2019). This domain was identified using Pfam analysis in the fourth glycosyltransferase encoded by the *cwps* cluster of *L. lactis* 3107 (the same position as that in MG1363) and in the third glycosyltransferase encoded by the *cwps* cluster of *L. lactis* SK11. The genetic location of these predicted sugar-1-phosphate transferase-

encoding genes is consistent with the location of the phosphorylated sugars in the oligosaccharide chains (Figure 2 and Table 1). In SK11/MG1363 and 3107, a Glc-P (in third position) and a GlcNAc-P (in fourth position) is predicted to be transferred to the growing undecaprenyl-associated subunit precursor, respectively, through the activity of the sugar-phosphate transferases. This finding is therefore in line with the predicted order of the saccharide incorporation in the oligosaccharidic subunit. Furthermore, where such domains are identified in newly sequenced strains, predictions of those possessing a phosphate linkage in their PSP subunit may now be established.

In addition to the primary saccharidic side-chain biosynthesis, it may also be possible to predict if these structures are attached as a monomeric or polymeric oligosaccharide as based on the presence or absence of the genes encoding the presumed oligosaccharide polymerase and co-polymerase proteins, WpsH and WpsI (Theodorou et al., 2019). *In silico* analysis revealed that homologues of WpsH and WpsI are encoded by the *cwps* loci of all assessed C-type strains and that of the D-type strain 184. It is noteworthy that the WpsI homologues in W34 and 1196 are significantly shorter than those present in other strains, which do not appear to be sequence errors (Fig. 2). However, the PSPs of these strains are polymeric indicating that a functional homologue of WpsI may be present on the genomes of W34 and 1196 that complement this apparent deficiency in the *cwps* cluster. This prediction is confirmed by the resolved polymeric side-chain structures of the C- and D-type strains that were chemically analyzed in this study. Interestingly, the previously studied A- and B-type strains do not possess homologues of WpsH and WpsI, and thus are expected and in some cases have been shown to contain a monomeric side chain (Fig. 2). In contrast to these, the *cwps* locus of the B-type strain UC11 was observed to encode WpsH and WpsI homologues with similar characteristics to those of encoded by NZ9000 in terms of membrane topology suggesting that this strain produces a polymeric side-chain structure as discussed earlier ($WpsH_{UC11}$ possesses two transmembrane domains with a large external facing loop and $WpsI_{UC11}$ is an integral membrane protein with 11 transmembrane spanning domains) (Theodorou et al., 2019).

Connecting the rhamnan & its oligosaccharide side chains

As mentioned above, the polytopic membrane GT-C fold glycosyltransferase WpsJ is proposed to be involved in the transfer of the (polymerized) oligosaccharide side-chains on to the rhamnan backbone structure. Indeed, with the exception of JM1, the rhamnan of all strains evaluated in the present study appears to be covalently linked to the polysaccharidic side-chain component of the CWPS of C-type strains. WpsJ is not conserved at the sequence level in all strains (e.g. in A, B and D-type strains) and this sequence disparity is consistent with the unique characteristics of the structures being transferred to the rhamnan components of type A, B and C strains. This hypothesis is further corroborated by the identification of a unique WpsJ encoded by the D-type strain 184 (encoded by *ll184_0287*) with 12 predicted transmembrane domains. The unique WpsJ of strain 184 suggests that this strain may have a distinct rhamnan structure from those of the characterized A, B and C-type strains (Fig. 3A).

Strain 184 produces a distinct polysaccharidic side-chain structure relative to the other types that have been characterized to date (Fig. 2 and Table 1). Therefore, despite a lack of significant sequence similarity in many cases, it is now possible to identify WpsJ analogs and to predict many of the functions of the proteins encoded within the *cwps* cluster based on synteny and characteristics such as membrane topology. Furthermore, phylogenetic analysis of the WpsJ-encoding regions of lactococci identified that the phylogeny of this protein correlates to the *cwps* genotype in the vast majority of cases suggesting a specificity of this protein for the polysaccharide or oligosaccharide side chains to be attached to the rhamnan. However, there are some exceptions to this general trend. For example, the WpsJ sequences of the B-type strains UC08 and UC11 cluster with C-type strains suggesting that the saccharidic side chain of these strains resembles the C-type strains rather than B-type strains, which produce an oligosaccharide side chain. To evaluate this, the *cwps* clusters of UC08 and UC11 (whose *cwps* clusters are identical) and those of the B-type model strain IL1403 and the C-type strain SK11 (which possesses the most similar WpsJ; Fig. 3A) were compared using BlastP analysis. This analysis revealed that the rhamnan encoding region of UC08 and UC11 bears more similarity to that of SK11 than IL1403 (Fig. 3B). A striking feature of the *cwps* clusters of UC08 and UC11 is the presence of low identity (approx. 28 -30%) homologues of WpsH and I (polymerase and co-polymerase), which suggest that these strains produce polymeric side chain structures akin to the described C-type strains. The reduced identity of WpsJ encoded by IL1403 and UC509.9 may indicate that WpsJ sequences may identify strains that transfer an oligosaccharide (as is the case

for IL1403 and UC509.9) or a polysaccharide structure (as for all described C-type strains) and should thus correlate with the presence/absence of the polymerase/co-polymerase functions.

The rhamnan of UC509.9 and IL1403 is linked to the primary GlcNAc residue of the oligosaccharide side chain. In the other strains with known CWPS structures, the position of the linkage between the rhamnan and saccharidic side chain is not known at present. However, due to the high degree of conservation of a GlcNAc residue as the first sugar in the side chain structures in a majority of strains analysed to date, it seems likely that the PSP of C-type strains is attached to the rhamnan through this GlcNAc residue (where it is present).

Predicting the oligosaccharide side-chain subunit size

Based on the identification of glycosyltransferase-encoding genes using BlastP, HHPred and hydrophobicity analysis of components of the *cwps* loci of selected lactococcal strains, it is expected that *L. lactis* strain 3107, JM1, MG1363, SMQ-388, SK11, W34 and 1196 should produce pentasaccharidic oligosaccharide side-chains, whereas strains IO-1 and A76 are deduced to produce tetrasaccharidic and hexasaccharidic subunits, respectively. However, considering the role of CsdCD in glucosylation of these oligosaccharide structures, it was predicted that the oligosaccharides of SK11 and A76 possess an additional glucose side chain similar to that found in the PSP structure of MG1363 and SMQ-388. Furthermore, the genome of JM1 encodes the CsdC function and a predicted non-functional CsdD (Theodorou et al., 2020), thus the oligosaccharide of this strain was not predicted to possess the glucose substitution unless the function of CsdD is complemented by a functional homolog in the genome of this strain. Conversely, no *csdCD* homologues appear to be present in the genomes of strains 3107, W34, 1196 and IO-1, which are thus not expected to possess such glucose decorations on their PSP subunits. These predictions were matched by the established chemical structures in most cases (Table 1, Figure 2) (the exception, JM1 will be discussed in further detail below).

While the majority of the predictions of (a) the number of component monosaccharides in a given oligosaccharide; (b) the presence of glucose side-chains and; (c) the mono/polymeric nature of the oligosaccharide side-chain agree with the biochemical structures elucidated in this, and previous studies, exceptions were observed. For example, the *cwps* cluster of the C_1 type strain JM1 is

almost identical to that of MG1363 with the exception of the presence of a transposase located downstream of the fourth glycosyltransferase-encoding gene (Fig. 2). However, while the oligosaccharide side chain of MG1363 is composed of repeating subunits of a phospho-hexasaccharide, the oligosaccharide side-chain subunits of JM1 were composed of disaccharides corresponding to the primary GlcNAc and the Glc decoration of the PSP from MG1363. Therefore, it appears that the presence of the transposase-encoding gene within the cluster of glycosyltransferase-encoding region disrupts the biosynthetic process possibly through polar effects on the downstream genes or through saturation of the undecaprenyl phosphate. Furthermore, the apparently defective WpsH function in this strain may also prevent the polymerization of the oligosaccharide side-chain structures. In addition to the genetic factors, it is also possible that the harsh physical treatments applied in the extraction of these polysaccharides may have broken the phosphodiester- and/or Gal*f*-associated linkages in the structures as observed previously (Theodorou et al., 2019).

Predicting novel/strain-specific features

It is now possible to identify strains that may possess unique structures or unique elements/modifications within their CWPS structures. For example, the predicted nucleotidyltransferases (such as those encoded by strains W34, 1196, AI06 and 184; Fig. 2) may produce the substrate for the glycosyltransferases and may highlight strain-specific monosaccharide components that are incorporated into the oligo/polysaccharide structure. Among the C-type strains, the presence UDP-galactopyranose mutase encoding genes (*wpsF* in MG1363) directly correlates with the presence of Gal*f* residues in their associated PSP structures (e.g. 3107, MG1363, SK11, W34, 1196 and IO-1). Furthermore, several strains incorporate epimerase functions to isomerize component saccharides (e.g. strains 184, A76, IO-1, IL1403 encode such activities), while others again encode unique functions such as alcohol dehydrogenase (which may be involved in CDP-ribitol precursor formation) (e.g. strain 184) and glycerophosphodiester phosphodiesterases (e.g. strain UC509.9) and provide insights into the incorporation of unique attributes and modifications of the CWPS structures.

Biosynthetic pathway for the construction of the oligosaccharide side chains of IL1403

As mentioned previously, a model for the predicted biosynthetic pathway of the rhamnan and PSP of the model C-type strain MG1363 has been put forward (Sadovskaya et al., 2017; Theodorou et al., 2019). Given the significant disparities between the genetic composition and CWPS structures of C/D- and A/B-type strains, we interrogated the *cwps* gene cluster of the B-type strain IL1403 to predict the biosynthetic pathway of its oligosaccharide side chains and their incorporation into the rhamnan backbone as a model for the A- and B-type strains. Furthermore, IL1403 has served as a prototypical lactococcal strain of the *lactis* subspecies and is host to a number of phages, thus it was deemed a relevant strain on which to perform such a focused analysis. Based on sequence similarity to WpsA and WpsB of MG1363, it is predicted that the biosynthesis of the oligosaccharide side chains in IL1403 commences with the addition of the primary GlcNAc residue on to Und-P by YcbB and its stimulating partner YcbC. Downstream of *ycbC* is the predicted epimerase-encoding gene *ycbD*, which we propose is involved in the conversion of UDP-GlcNAc to UDP-GalNAc, which could be incorporated into the oligosaccharide side chain by the glycosyltransferase YcbF. HHpred analysis of YcbF identified a 100 % probability identity with a group 1 glycosyl transferase protein and members of this glycosyltransferase group are known to catalyze transfer of a glycosyl moiety from UDP-sugars to an acceptor moiety corroborating the likely role of this protein.

YcbG possesses a LicD domain, which is associated with phosphocholine and phosphosugar transferase activities in a number of species (Kuchta, Knizewski, Wyrwicz, Rychlewski, & Ginalski, 2009). Therefore, YcbG is a likely candidate for the incorporation of the Gro-P moiety into the growing oligosaccharide. The production of CDP-Gro precursor is proposed to be effected through the combined activities of YcbJ and TagD1, encoded at the rightmost end of the cluster (Fig. 4). YcbJ is annotated as a glycerophosphodiester phosphodiesterase (GDPD) and is thus likely to be involved in the formation of sn-glycerol-3-phosphate, which is then converted to CDP-glycerol through the activity of TagD1, which is annotated as a glycerol-3-cytidyltransferase. A glucose residue is incorporated on the oligosaccharide chain via YcbH and/or YcbI and the final oligosaccharide structure is transferred to the outer face of the cytoplasmic membrane by the flippase, YcbK, which possesses 12 predicted transmembrane domains. The transfer of the oligosaccharide side on to the rhamnan backbone is likely mediated by YcaG, potentially in collaboration with YcaF. Certain conserved functions are present among various A- and B-type

strains, such as the aforementioned GDPD and LicD domain-containing proteins, which may allow the identification of strains that possess Gro-P in their CWPS structures.

Discussion

The functional role of CWPSs and particularly rhamnose-containing CWPSs (termed RhaCWPS) in bacterial cell wall biogenesis and architecture, cell division/septation and pathogenesis in some cases (e.g. *Streptococcus pyogenes* and *Streptococcus agalactiae*) have been demonstrated across several Gram-positive bacterial species (Chapot-Chartier et al., 2010; Mistou, Sutcliffe, & van Sorge, 2016; Sadovskaya et al., 2017). Furthermore, RhaCWPS act as receptors for phages in several Gram-positive bacteria including *L. lactis* and pathogenic streptococci (Chapot-Chartier et al., 2010; Mistou et al., 2016). While the known RhaCWPS of pathogenic streptococcal species are repeating rhamnose-rich subunits with mono or disaccharide substituents, their recognized equivalents in lactococci are often much more structurally diverse as shown in this study. The genetic regions that encode these structures in Gram-positive bacteria most often surround the *rmlA-D* genes that specify the rhamnose precursor biosynthesis pathway as is the case with the lactococcal *cwps* gene cluster.

L. lactis strains have received considerable research attention in recent decades owing to their industrial significance. In the current study, the *cwps* gene clusters of 107 lactococcal strains were assessed for their genomic diversity so as to ascertain if additional CWPS genotypes are present in addition to the three previously described types (A, B & C) (Mahony et al., 2013). These three CWPS genotypes were assigned based on the comparative genomic analysis of *cwps* loci of six lactococcal strains that were available in 2013. A fourth *cwps* genotype was identified in the present study (and corresponding strains were designated as CWPS type D) substantiating previous suggestions of additional genetic diversity (Mahony et al., 2013). In addition to the four major CWPS types (A-D), subtypes of the C (C_{1-8}) type strains are distinguished based on hierarchical clustering analysis (Fig. 1). The structural flexibility of lactococcal CWPSs (among others) is likely a response to external forces such as phage predation. The co-evolution of lactococcal phages and their hosts renders it probable that the host CWPS genotypes / structures will continue to evolve to maintain competitiveness in an environment rich in parasitic viruses.

Structural analysis of the CWPSs of the C and D type strains assessed in the present study suggest that all C (and D) type strains possess two (covalently linked) CWPS elements- the rhamnan and their polymeric side chains. This is in contrast to CWPS type A and B strains UC509.9 and IL1403, whose CWPS structures appear more like the RhaCWPS of pathogenic streptococci (although seemingly more complex) with decorated rhamnan structures. This may support the proposed shared ancestry of lactococci and streptococci, while also highlighting the role and diversification of these structures in order to evade phagocytosis, phage infection and likely numerous other external influences on these bacterial species.

The recently proposed models for the biosynthesis of lactococcal rhamnan and PSP structures have formed the foundation for evaluating the construction and diversity of these essential structures among lactococci. Based on the PSP biosynthesis model for MG1363, it is now possible to identify the priming glycosyltransferase-encoding gene, *wpsA* most particularly among C-type strains for which this gene is highly conserved. This finding vastly improves our ability to predict the order of component monosaccharides in a given lactococcal CWPS structure. Furthermore, recent and growing structural analyses of glycosyltransferases (among other proteins) has significantly improved our ability to assign glycosyltransferase functions thereby allowing predictions of the number of component monosaccharides within the CWPS subunits of any given lactococcal strain. Additionally, it is now known that the glycosylation of CWPS component structures of lactococci is due to the activity of gene pairs located outside the *cwps* locus and the identification of homologues of these genes in lactococcal genomes permits predictions of the presence of glucosylated rhamnan and associated side chain structures. The culmination of these data now renders it possible to predict and appreciate the astonishing diversity of lactococcal CWPS structures. With representative structures for each of the major types now available, the predictive power has increased dramatically. Ultimately, this infers that as additional genome sequences become available, comparative analysis with the previously sequenced (and biochemically characterized) strains will provide sufficient information to establish the CWPS type or subtype that the strain will produce and by inference, the likelihood of phage-relatedness to other strains of *L. lactis*, which is highly informative to the dairy industry, yet also on a fundamental level to predict a strains' likely interactions with other organisms co-existing in its environmental niche.

Conclusions

The current study represents the largest comparative genomic, biological and structural analysis of lactococcal strains for the presence and diversity of cell wall-associated polysaccharidic structures with implications for phage sensitivity. Genomic analysis of 107 *cwps* loci defined 37 possible MCL profiles and thus potential CWPS structures, highlighting the extent of genetic and potential structural diversity of these glycans. As the loci encoding the biosynthetic machinery for the CWPS structure are observed to be diverse and flexible, it appears as though lactococci "mix and match" glycosyltransferase-encoding genes culminating in a vast number of possible combinations of structural arrangements of sugars to ensure their continued success in the industrial context and beyond. The wealth of information generated in this study has established a clear link between the *cwps* loci and their corresponding CWPS structures, and makes it possible to predict the number of component monosaccharides in the constituent subunits of the CWPS, the order in which they may be synthesized, the presence of phosphodiester-linkages and whether the rhamnan side chains are oligosaccharides or polysaccharides.

Materials & Methods

Bacterial growth conditions. Lactococcal strains used in this study are listed in Table S1 and were grown without agitation at 30 °C in M17 broth (Oxoid Ltd., Hampshire, U.K.) supplemented with 0.5 % glucose (GM17). 4-8 L cultures of *L. lactis* JM1, SK11, W34, 1196, IO-1, A76 and 184 were produced as described above and the resulting cells harvested by centrifugation at 6000 × *g* in a Sorvall RC6+ centrifuge (Thermo Scientific, UK). Lactococcal strains JM1, SK11, W34, 1196, IO-1, A76 and 184 were assessed for their CWPS composition and structure, each with unique aspects to their extraction as described in detail further below.

DNA extraction & genome sequencing, assembly and annotation. For strains derived from the NULS collection, lactococcal genomic DNA was isolated with the DNeasy tissue and culture DNA isolation kit (Qiagen, Germany) according to manufacturer's instructions. Sequencing libraries were prepared using the Nextera XT DNA library preparation kit (Illumina, San Diego, USA), in combination with Agencourt AMpure XP (Beckman Coulter, Oslo, Norway) to a target insert size of 700 bp and were sequenced using Illumina MiSeq technology (Illumina, San Diego, USA). Strains from the UCC collection were sequenced using PacBio SMRT sequencing technology (GATC Biotech Ltd., Germany) as described previously (Kelleher, Bottacini, Mahony, Kilcawley, & van Sinderen, 2017). The genomes of the NULS strains were assembled using SPAdes (Version 3.5.1; (Bankevich et al., 2012)) and putative open reading frames annotated using the Prokka pipeline software (Version 1.11;(Seemann, 2014)). Strains derived from the UCC collection were assembled via the Pacific Biosciences SMRTPortal analysis platform (version 2.3.1) using the RS_HGAP_Assembly.2 protocol. Putative open reading frames were predicted using the Prodigal v2.5 software (http://prodigal.ornl.gov) and all annotations were checked manually in Artemis v16 (http://www.sanger.ac.uk/science/tools/artemis). Multiple alignment of amino acid sequences of the WpsJ sequences of the strains applied in this study was performed using ClustalW software. The alignment was employed to generate an unrooted phylogenetic tree using the "itol" software (http://itol.embl.de/), applying the neighbor-joining method.

Genbank accession numbers. The genome sequences of the CWPS-encoding operons sequenced in the context of this study were deposited in the Genbank database and their associated Genbank accession numbers are presented in Table S1.

Comparative genomic analysis of *cwps* gene clusters. Comparison of *cwps* gene clusters was performed by all-against-all, bi-directional BLAST alignment (Altschul, Gish, Miller, Myers, & Lipman, 1990) with an alignment (or E-value) cut-off value of 0.0001 and greater than 50 % identity across at least 50 % of deduced amino acid sequences. To identify the closest relatives of the sequenced CWPS-specifying genomic regions and to establish their phylogeny based on their overall deduced proteomic content (based on the presence/absence of identified protein families with at least 50 % identity over 50 % of the protein), the Markov Clustering (MCL) algorithm was executed via the mclblastline pipeline v12-0678 as described previously (Enright, Van Dongen, & Ouzounis, 2002) followed by hierarchical clustering (HCL) analysis performed and viewed in the multi experiment viewer (MeV) (Saeed et al., 2003). CWPS genotype groupings were assigned as type A, B, C and D based on similarity to previously defined CWPS type strains (Mahony et al., 2013) and/or through the identification of distinct groups observed in the heatmap generated through the HCL analysis. Subspecies identification of the lactococcal strains was based on 16S rRNA sequence analysis as previously described (Kelleher et al., 2017).

CWPS extraction. *L. lactis* JM1 cells from 4 L of an overnight culture were suspended in water and stirred at room temperature overnight. The suspension was centrifuged at 6000 × *g* to separate cells from the clear supernatant (Water extract 1). Cells were re-suspended in water, stirred in a water bath at 100 °C for 2 h, cooled and centrifuged to separate cells (and cell debris) from Water extract 2. The collected cells were then separated in two equal parts. One of which was suspended in water and autoclaved (115 °C, 20 min), cooled and centrifuged to obtain Extract 3, while the other was extracted by trichloroacetic acid (TCA) (5 %, 48 h, 6°C). DNA and proteins were removed from the extracts 1, 2 and 3 through the addition of 50 % TCA to a final concentration of 5 %. After centrifugation, the extracts were dialyzed against water and freeze-dried. A similar extraction procedure was employed for *L. lactis* W34 and 1196 with the exception that the washing step with hot water was omitted. For *L. lactis* SK11 autoclaving was not applied as it caused degradation of the polysaccharide, thus the cells were washed with water and extracted with TCA. For strains 184 and IO-1, cells were washed with water and extracted directly with TCA (5 %, 48 hr). Cell debris was further extracted with hot HCl (0.01 N followed by 0.1 N), where relevant (for strain JM1) as described previously (Sadovskaya et al., 2017). All extracts were further fractionated by gel-filtration chromatography on a Sephadex G-50 column (2.6 x 80 cm, eluted with 0.1 % AcOH). For strains W34 and 1196, high molecular weight (HMW; >7-10

kDa) carbohydrate material eluted from the column as a single broad peak, close to the void volume (unlike strain JM1, where the eluted CWPS- and rhamnan-containing fractions were well separated by molecular weight (MW)). All HMW preparations were collected and screened by ^{1}H-NMR. For each strain, samples contained a mixture of rhamnan, other polysaccharide and peptidoglycan. The fractions were combined and separated on a Hitrap Q anion-exchange column. Neutral fractions were shown to contain less rhamnan and were used for 2D NMR structural analysis. For SK11, a HMW fraction was not observed and the preparation eluted as a broad peak which was further purified on a Hitrap Q anion-exchange column.

For strain 184, PS preparations from TCA extract were treated with hydrofluoric acid (HF) to afford an oligosaccharide (OS) fraction. Additional portions of the OS were prepared by extraction of the cell debris with the hot 0.01 HCl (Sadovskaya et al., 2017). The deproteinated HCl extract (100 mg) was treated with HF (48 %, 24 h, 4°C), evaporated, and fractionated on a Sephadex G-50 column (1 x 40 cm). The low MW fraction (<2 kDa) was further purified on a Biogel P2 column (2.6 x 60 cm). The OS preparations were further purified on a cation exchange Hitrap SP and a Sephadex G-15 column.

Chromatographic methods. Gel-permeation chromatography was performed on Sephadex G-50 (GE Healthcare, 2.6 x 80 cm and 1 x 40 cm), BioGel P-2 (Biorad, 2.6 x 60 cm), and Sephadex G-15 columns, eluted with 0.1 % AcOH. Fractions were screened for neutral sugars (Dubois, Gilles, Hamilton, Rebers, & Smith, 1956). Anion exchange chromatography was performed on Hitrap Q column (5 mL size, GE-Healthcare). Cation exchange chromatography was performed on Hitrap SP (5 mL size, GE-Healthcare). Ion-exchange chromatography was monitored with UV at 220 nm in a linear gradient of NaCl (10 min water, then gradient to 1 M NaCl over 1 h, 3 mL/min). Fractions of 1 min were collected and tested for the presence of eluted compounds by spotting on SiO_2 TLC plate, dipping in 5 % H_2SO_4 in ethanol and heating with a heat-gun. All fractions of interest were dried in a Savant drying centrifuge and the ^{1}H NMR spectra were recorded for each fraction without desalting. For 2D NMR desalting was performed on Sephadex G15 column.

Monosaccharide and methylation analysis. Monosaccharides were detected as reduced and acetylated derivatives (alditol acetates), prepared by conventional methods (Lindberg & Lonngren, 1978). Polysaccharide samples (0.2–1 mg) were hydrolyzed with 4 MTFA (110 °C, 3 h), dried, and reduced with $NaBH_4$. The reagent was destroyed with 1 mL of 10 % AcOH in MeOH, the

solution was dried under a stream of nitrogen, dried twice with addition 10 % AcOH in MeOH (1 mL), and then MeOH (2x). Alditols were acetylated with 0.4 mL Ac_2O–0.4 mL pyridine mixture for 1 h at 100 °C, dried, and analyzed by GC–MS.

Methylation analysis was performed as described previously (Ciucanu & Kerek, 1984) with modifications (Read, Currie, & Bacic, 1996). Briefly, a carbohydrate sample (0.5–1 mg) was dissolved in 1 mL of dry DMSO. Powdered NaOH (about 50 mg) was added and the mixture was stirred for 15 min, then 0.2 mL of MeI was added, the mixture was stirred for 1 or 2 h; then another 0.2 ml of MeI was added and stirred for 1 h. The reaction was stopped by addition of 5 mL of 10 % aqueous $Na_2S_2O_3$. The permethylated product was extracted into $CHCl_3$ (2x, 2 mL). The organic phase was washed with water (5 × 4 mL) and evaporated. The product was hydrolyzed with 4 M trifluoroacetic acid (TFA) (110 °C, 3 h), dried, reduced with $NaBD_4$, converted into alditol acetates as described above, and analyzed by GC–MS (Lindberg & Lonngren, 1978).

GC–MS was performed on a Trace GC ULTRA system (Thermo Scientific) equipped with a capillary column NMTR-5MS (30 m × 0.25 mm) using a temperature gradient of 170 °C (3 min) → 250 °C at 5 °C/min and with a DSQ II MS detector. Methylated derivatives were identified using the Complex Carbohydrate Research Center partially methylated alditol acetates (PMAA) database (www.ccrc.uga.edu/specdb/ms/pmaa/pframe.html), and by comparison with the authentic standards of methylation analysis of polysaccharides obtained from various (previously analyzed) *L. lactis* and *Lactobacillus* strains (Chapot-Chartier et al., 2010; Vinogradov, Sadovskaya, Cornelissen, & van Sinderen, 2015; Vinogradov, Sadovskaya, Grard, & Chapot-Chartier, 2016; Vinogradov et al., 2013).

NMR spectroscopy. NMR experiments were carried out on a Varian INOVA 500 MHz (1H) spectrometer with 3 mm Z-gradient probe at 25 °C with an acetone internal reference (2.225 ppm for 1H and 31.45 ppm for 13C) using standard pulse sequences gCOSY, TOCSY (mixing time 120 ms), NOESY (for polysaccharide) and ROESY (for oligosaccharide) (mixing time 500 ms), HSQC and HMBC (100 ms long range transfer delay). Acquisition time (AQ) was kept at 0.8 s for H-H correlations and 0.25 s for HSQC, 256 increments was acquired for t1. The spectra were processed and analyzed using the Bruker Topspin 2.1 program. Assignment of spectra was performed using Topspin 2 (Bruker Biospin) program for spectra visualization and overlap.

Acknowledgements

JM is in receipt of a Starting Investigator Research Grant (SIRG) (Ref. No. 15/SIRG/3430) funded by Science Foundation Ireland (SFI). DvS is supported by a Principal Investigator award (Ref. No. 13/IA/1953) through SFI and by APC Microbiome Ireland (under Science Foundation Ireland (SFI) grant number: SFI/12/RC/2273-P1 and SFI/12/RC/2273-P2). HH & CF are funded by the Norwegian Research Council grant number 225246 and TINE SA.

Data availability statement

All sequences generated in this study are available in the NCBI database (https://www.ncbi.nlm.nih.gov/nuccore/) through the Genbank accession numbers presented in Table S1.

Abbreviated summary

The chemical structures of cell wall polysaccharides (CWPS) of seven distinct lactococcal strains were elucidated. Comparative genome analysis of the gene clusters that encode these structures of 107 lactococcal strains has led to the identification of four distinct genotypes that correlates to distinct chemical structures. The combined genome and structural data allow predictions of characteristics of lactococcal CWPS based on sequence information, while a model for the biosynthetic pathway of the CWPS of the prototypical B-type strain IL1403 is proposed.

References

Ainsworth, S., Sadovskaya, I., Vinogradov, E., Courtin, P., Guerardel, Y., Mahony, J., . . . van Sinderen, D. (2014). Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. *MBio, 5*(3), e00880-00814. doi: 10.1128/mBio.00880-14

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol, 19*(5), 455-477. doi: 10.1089/cmb.2012.0021

Bolotin, A., Aucouturier, A., Sorokin, A., & Bidnenko, E. (2019). Genomic sequence of the prophage-free *Lactococcus lactis* strain IL6288. *Microbiol Resour Announc, 8*(1). doi: 10.1128/mra.01545-18

Bolotin, A., Quinquis, B., Ehrlich, S. D., & Sorokin, A. (2012). Complete genome sequence of *Lactococcus lactis* subsp. *cremoris* A76. *J Bacteriol, 194*(5), 1241-1242. doi: 10.1128/JB.06629-11

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., . . . Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res, 11*(5), 731-753. doi: 10.1101/gr.169701

Chapot-Chartier, M. P., Vinogradov, E., Sadovskaya, I., Andre, G., Mistou, M. Y., Trieu-Cuot, P., . . . Kulakauskas, S. (2010). Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. *J Biol Chem, 285*(14), 10464-10471. doi: 10.1074/jbc.M109.082958

Ciucanu, I., & Kerek, F. (1984). A simple and rapid method for the permethylation of carbohydrates. *Carbohydr Res, 131*(2), 209-217. doi: Doi 10.1016/0008-6215(84)85242-8

Dubois, M., Gilles, K., Hamilton, J. K., Rebers, P. A., & Smith, F. (1956). Colorimetric method for determination of sugars and related substances. *Analytical Chemistry, 28*, 350-356.

Dupont, K., Janzen, T., Vogensen, F.K., Josephsen, J. & Stuer-Lauridsen, B. (2004). Identification of *Lactococcus lactis* genes required for bacteriophage adsorption. *Appl Environ Microbiol, 70*(10), 5825-32.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res, 30*(7), 1575-1584.

Farenc, C., Spinelli, S., Vinogradov, E., Tremblay, D., Blangy, S., Sadovskaya, I., . . . Cambillau, C. (2014). Molecular insights on the recognition of a *Lactococcus lactis* cell wall pellicle by the phage 1358 receptor binding protein. *J Virol, 88*(12), 7005-7015. doi: 10.1128/JVI.00739-14

Gao, Y., Lu, Y., Teng, K. L., Chen, M. L., Zheng, H. J., Zhu, Y. Q., & Zhong, J. (2011). Complete genome sequence of *Lactococcus lactis* subsp. *lactis* CV56, a probiotic strain isolated from the vaginas of healthy women. *J Bacteriol, 193*(11), 2886-2887. doi: 10.1128/JB.00358-11

Guellerin, M., Passerini, D., Fontagne-Faucher, C., Robert, H., Gabriel, V., Loux, V., . . . Le Bourgeois, P. (2016). Complete genome sequence of *Lactococcus lactis* subsp. *lactis* A12, a strain isolated from wheat sourdough. *Genome Announc, 4*(5). doi: 10.1128/genomeA.00692-16

Kato, H., Shiwa, Y., Oshima, K., Machii, M., Araya-Kojima, T., Zendo, T., . . . Yoshikawa, H. (2012). Complete genome sequence of *Lactococcus lactis* IO-1, a lactic acid bacterium that utilizes xylose and produces high levels of L-lactic acid. *J Bacteriol, 194*(8), 2102-2103. doi: 10.1128/JB.00074-12

Kelleher, P., Bottacini, F., Mahony, J., Kilcawley, K. N., & van Sinderen, D. (2017). Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation. *BMC Genomics, 18*(1), 267. doi: 10.1186/s12864-017-3650-5

Kelly, W. J., Altermann, E., Lambie, S. C., & Leahy, S. C. (2013). Interaction between the genomes of *Lactococcus lactis* and phages of the P335 species. *Front Microbiol, 4*, 257. doi: 10.3389/fmicb.2013.00257

Kuchta, K., Knizewski, L., Wyrwicz, L. S., Rychlewski, L., & Ginalski, K. (2009). Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. *Nucleic Acids Res, 37*(22), 7701-7714. doi: 10.1093/nar/gkp854

Linares, D. M., Kok, J., & Poolman, B. (2010). Genome sequences of *Lactococcus lactis* MG1363 (revised) and NZ9000 and comparative physiological studies. *J Bacteriol, 192*(21), 5806-5812. doi: 10.1128/JB.00533-10

Lindberg, B., & Lonngren, J. (1978). Methylation analysis of complex carbohydrates: General procedure and application for sequence analysis. *Methods in Enzymology, 50*, 3-33.

Mahony, J., Kot, W., Murphy, J., Ainsworth, S., Neve, H., Hansen, L. H., . . . van Sinderen, D. (2013). Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol, 79*(14), 4385-4392. doi: 10.1128/AEM.00653-13

Mahony, J., Randazzo, W., Neve, H., Settanni, L., & van Sinderen, D. (2015). Lactococcal 949 group phages recognize a carbohydrate receptor on the host cell surface. *Appl Environ Microbiol, 81*(10), 3299-3305. doi: 10.1128/AEM.00143-15

McCulloch, J. A., de Oliveira, V. M., de Almeida Pina, A. V., Perez-Chaparro, P. J., de Almeida, L. M., de Vasconcelos, J. M., . . . Nunes, M. R. (2014). Complete genome sequence of *Lactococcus lactis* strain AI06, an endophyte of the amazonian Acai palm. *Genome Announc, 2*(6). doi: 10.1128/genomeA.01225-14

Mistou, M. Y., Sutcliffe, I. C., & van Sorge, N. M. (2016). Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria. *FEMS Microbiol Rev, 40*(4), 464-479. doi: 10.1093/femsre/fuw006

Nakano, K., Minami, M., Shinzato, M., Shimoji, M., Ashimine, N., Shiroma, A., . . . Hirano, T. (2018). Complete genome sequence of *Lactococcus lactis* subsp. *lactis* G50 with immunostimulating activity,isolated from Napier grass. *Genome Announc, 6*(8). doi: 10.1128/genomeA.00069-18

Oliveira, L. C., Saraiva, T. D., Soares, S. C., Ramos, R. T., Sa, P. H., Carneiro, A. R., . . . Azevedo, V. (2014). Genome sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-producing strain. *Genome Announc, 2*(5). doi: 10.1128/genomeA.00980-14

Ostergaard Breum, S., Neve, H., Heller, K. J., & Vogensen, F. K. (2007). Temperate phages TP901-1 and phiLC3, belonging to the P335 species, apparently use different pathways for DNA injection in *Lactococcus lactis* subsp. *cremoris* 3107. *FEMS Microbiol Lett, 276*(2), 156-164. doi: 10.1111/j.1574-6968.2007.00928.x

Radziwill-Bienkowska, J. M., Le, D. T., Szczesny, P., Duviau, M. P., Aleksandrzak-Piekarczyk, T., Loubiere, P., . . . Kowalczyk, M. (2016). Adhesion of the genome-sequenced *Lactococcus lactis* subsp. *cremoris* IBB477 strain is mediated by specific molecular determinants. *Appl Microbiol Biotechnol, 100*(22), 9605-9617. doi: 10.1007/s00253-016-7813-0

Read, S. M., Currie, G., & Bacic, A. (1996). Analysis of the structural heterogeneity of laminarin by electrospray-ionisation-mass spectrometry. *Carbohydr Res, 281*(2), 187-201.

Sadovskaya, I., Vinogradov, E., Courtin, P., Armalyte, J., Meyrand, M., Giaouris, E., . . . Chapot-Chartier, M. P. (2017). Another brick in the wall: a rhamnan polysaccharide trapped inside peptidoglycan of *Lactococcus lactis*. *MBio, 8*(5). doi: 10.1128/mBio.01303-17

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., . . . Quackenbush, J. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques, 34*(2), 374-378.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069. doi: 10.1093/bioinformatics/btu153

Siezen, R. J., Bayjanov, J., Renckens, B., Wels, M., van Hijum, S. A., Molenaar, D., & van Hylckama Vlieg, J. E. (2010). Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium. *J Bacteriol, 192*(10), 2649-2650. doi: 10.1128/JB.00276-10

Theodorou, I., Courtin, P., Palussiere, S., Kulakauskas, S., Bidnenko, E., Pechoux, C., . . . Chapot-Chartier, M. P. (2019). A dual-chain assembly pathway generates the high structural diversity of cell-wall polysaccharides in *Lactococcus lactis*. *J Biol Chem*. doi: 10.1074/jbc.RA119.009957

Theodorou, I., Courtin, P., Sadovskaya, I., Palussiere, S., Fenaille, F., Mahony, J., . . . van Sinderen, D. (2020). Three distinct glycosylation pathways are involved in the decoration of *Lactococcus lactis* cell wall glycopolymers. *J Biol Chem*. doi: 10.1074/jbc.RA119.010844

Tian, K., Li, Y., Wang, B., Wu, H., Caiyin, Q., Zhang, Z., & Qiao, J. (2019). The genome and transcriptome of *Lactococcus lactis* ssp. *lactis* F44 and G423: Insights into adaptation to the acidic environment. *Journal of Dairy Science, 102*(2), 1044-1058. doi: https://doi.org/10.3168/jds.2018-14882

Tran, T. D., Huynh, S., Parker, C. T., Han, R., Hnasko, R., Gorski, L., & McGarvey, J. A. (2018). Complete genome sequence of *Lactococcus lactis* subsp. *lactis* strain 14B4, which inhibits the growth of *Salmonella enterica* serotype Poona *in vitro*. *Microbiol Resour Announc, 7*(19). doi: 10.1128/mra.01364-18

van Mastrigt, O., Abee, T., & Smid, E. J. (2017). Complete genome sequences of *Lactococcus lactis* subsp. *lactis* bv. diacetylactis FM03 and *Leuconostoc mesenteroides* FM06 isolated from cheese. *Genome Announc, 5*(28). doi: 10.1128/genomeA.00633-17

Vinogradov, E., Sadovskaya, I., Cornelissen, A., & van Sinderen, D. (2015). Structural investigation of cell wall polysaccharides of *Lactobacillus delbrueckii* subsp. *bulgaricus* 17. *Carbohydr Res, 413*, 93-99. doi: 10.1016/j.carres.2015.06.001

Vinogradov, E., Sadovskaya, I., Courtin, P., Kulakauskas, S., Grard, T., Mahony, J., . . . Chapot-Chartier, M. P. (2018). Determination of the cell wall polysaccharide and teichoic acid structures from *Lactococcus lactis* IL1403. *Carbohydr Res, 462*, 39-44. doi: 10.1016/j.carres.2018.04.002

Vinogradov, E., Sadovskaya, I., Grard, T., & Chapot-Chartier, M. P. (2016). Structural studies of the rhamnose-rich cell wall polysaccharide of *Lactobacillus casei* BL23. *Carbohydr Res, 435*, 156-161. doi: 10.1016/j.carres.2016.10.002

Vinogradov, E., Sadovskaya, I., Grard, T., Murphy, J., Mahony, J., Chapot-Chartier, M. P., & van Sinderen, D. (2018). Structural studies of the cell wall polysaccharide from *Lactococcus lactis* UC509.9. *Carbohydr Res, 461*, 25-31. doi: 10.1016/j.carres.2018.03.011

Vinogradov, E., Valence, F., Maes, E., Jebava, I., Chuat, V., Lortal, S., . . . Sadovskaya, I. (2013). Structural studies of the cell wall polysaccharides from three strains of *Lactobacillus helveticus* with different autolytic properties: DPC4571, BROI, and LH1. *Carbohydr Res, 379*, 7-12. doi: 10.1016/j.carres.2013.05.020

Vos, P., van Asseldonk, M., van Jeveren, F., Siezen, R., Simons, G., & de Vos, W. M. (1989). A maturation protein is essential for production of active forms of *Lactococcus lactis* SK11 serine proteinase located in or secreted from the cell envelope. *J Bacteriol, 171*(5), 2795-2802.

Yang, X., Wang, Y., & Huo, G. (2013). Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KLDS4.0325. *Genome Announc, 1*(6). doi: 10.1128/genomeA.00962-13

Zhao, F., Ma, H., Lu, Y., Teng, K., Kang, X., Wang, F., . . . Zhong, J. (2015). Complete genome sequence of *Lactococcus lactis* S0, an efficient producer of nisin. *J Biotechnol, 198*, 15-16. doi: 10.1016/j.jbiotec.2015.01.024

Table 1. Chemical structures and characteristics of the CWPS components of selected lactococcal strains

Strain	*CWPS $type_{subtype}$	PSP subunit structure	Rhamnan subunit structure	Covalently linked rhamnan-PSP	Reference
MG1363	C_{1a}	-2-β-Gal*f* -α-6-Glc-*P*-6-β-GlcNAc-3-α-Rha-3-β-GlcNAc- |6 α-Glc	-2-α-Rha-2-α-Rha-3-α-Rha-	Yes	(Chapot-Chartier et al., 2010)
SMQ-388	C_{1b}	-2-β-Gal*f* -6-α-GlcNAc-*P*-6-α-GlcNAc-3-β-Gal*f*-3-β-GlcNAc- |6 α-Glc	N.D.	Unknown	(Farenc et al., 2014)
♣JM1	C_{1c}	-3-β-GlcNAc- |6 α-Glc	-2-α-Rha-2-α-Rha-3-α-Rha-	No	This study
3107	C_2	-2-β-Gal*f* -6-α-GlcNAc-*P*-6-α-Glc-3-β-Gal*f*-3-β-GlcNAc-	-2-α-Rha-2-α-Rha-3-α-Rha-	Yes	(Ainsworth et al., 2014)

♣SK11	**C_3**	-6-β-Gal*f* -6-α-Gal-6-α-Glc-*P*-6-α-Gal-3-β-GlcNAc- |6 α-Glc	-2-α-Rha-2-α-Rha-3-α-Rha- |2 α-Glc	Yes	This study
♣W34	**C_{4a}**	-6-β-Gal*f*-6-α-Glc-6-β-GlcNAc-3-β-Gal*f*-3-β-GlcNAc-	-2-α-Rha-2-α-Rha-3-α-Rha-	Yes	This study
♣1196	**C_{4b}**	6-β-Gal*f*-6-α-Gal-6-β-GlcNAc-3-β-Gal*f*-3-β-GlcNAc-	-2-α-Rha-2-α-Rha-3-α-Rha-	Yes	This study
♣IO-1	**C_5**	-2-β-Gal*f*-5-Ara-ol-1-*P*-2-β-Gal*f*-3-α-GlcNAc |3 α-GlcNAc	N.D.	Yes	This study
♣A76	**C_6**	-6-β-Gal-2-α-Gal-3-β-RhaOAc-4-β-Glc6*P*-6-β-Gal*f*-6-β-Glc-	N.D.	Yes	This study
♣184	**D**	β-Glc |3 -6-α-GalNAc-3-β-GalNAc-5-Ribitol-1-*P*- |4 β-Glc*6OAla*	N.D.	Yes	This study

* Where the CWPS elements of a CWPS genotype of more than one strain were assessed, the different resulting biochemical structures were assigned unique identifiers e.g. $C_{1a, b, c}$ and $C_{4a, b}$. N.D. = not determined.

♣ The strains for which the structure of CWPS components were determined in the present study.

Figure legends

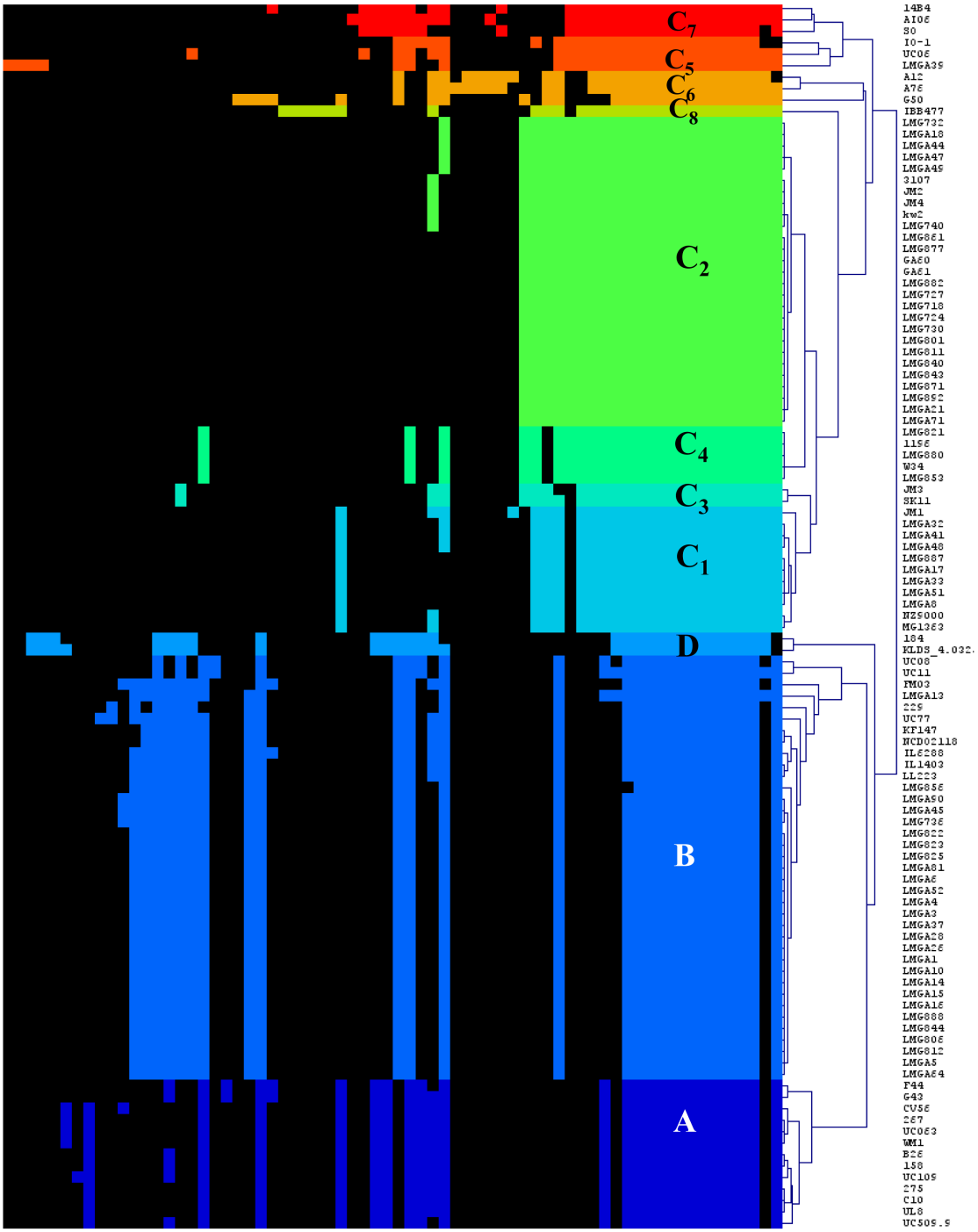
Fig. 1. Heatmap displaying the presence (Colour) or absence (black) of individual protein families within the CWPS biosynthesis operon of 107 lactococcal strains applied in this study. The assigned CWPS genotypes (A-D) are indicated on the right side of the figure and by colour within the heatmap, where CWPS A, B, C and D strains protein families are presented in red, yellow, blue and green, respectively. Subtypes are indicated within the heatmap, where relevant.

Fig. 2. Schematic overview of the organisation and sequence relatedness of the CWPS-encoding operons of lactococcal strains representing each of the CWPS types (A-D) for which chemical structures have been resolved. Regions of homology are joined by blocks of different shades of grey to black where black indicates 97-100 % aa identity, mid-grey indicates 80-96 % aa identity and light grey indicates 50-79 % aa identity. The proposed functions of the individual protein-encoding regions are colour coded and indicated within the figure. To the right is a schematic depicting the biochemical structures of the rhamnan (where known) and oligosaccharide side chain. For UC509.9 and IL1403 the exact site and the nature of the linkage between the rhamnan and the oligosaccharide side chains are defined while those of the CWPS C and D type strains have not been defined and thus their rhamnan (upper left) and side chain (right) structures are separated in this schematic. The symbols used to represent individual monosaccharides is based on the standard Symbol Nomenclature for Glycans (SNFG) while the linkages are provided in Table1 and Table S10.

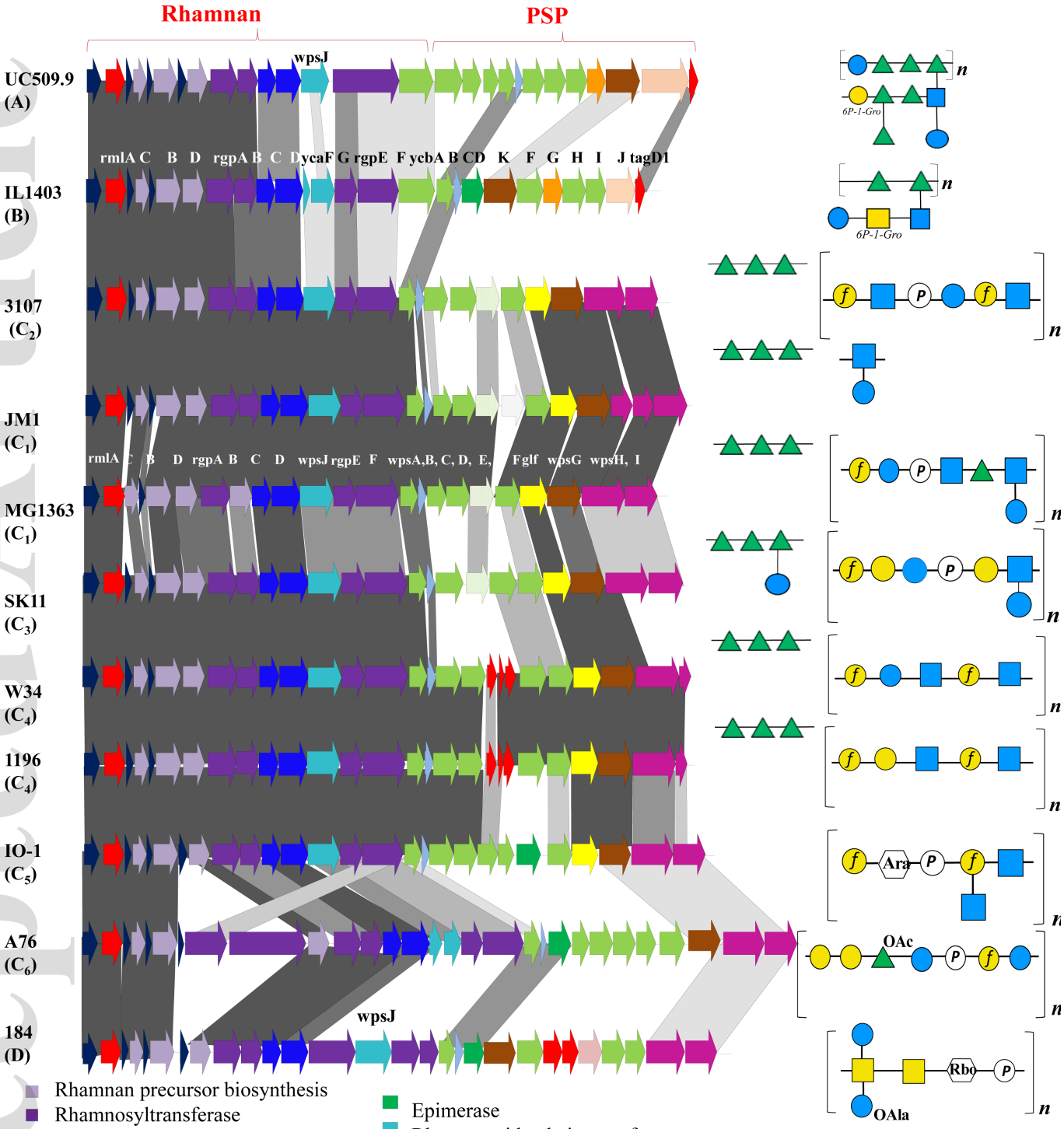
Fig. 3. Evaluation of the presence of oligomeric or monomeric side-chain decorations on rhamnan based on WpsJ phylogenetic analysis and the presence of the polymerisation functions, WpsH and WpsI. (A) Unrooted phylogenetic tree of the amino acid sequences of WpsJ homologues encoded by lactococci. The most closely related groups are highlighted by colour (purple, dark/light blue, dark/light green, orange, bright red, maroon). The purple and blue coloured clades are largely dominated by C-type strains, with the exception of the C_7 strains (light green) whose WpsJ equivalents align between those of the A-type (dark green) and D-type strains (orange). Within the dark blue coloured clade are two B-type strains UC08 and UC11. The maroon coloured clade represents the B-type strains. (B) A schematic representation of the *cwps* loci of UC08/11 in comparison to those of the B-type strain IL1403 (their closest relative based on the complete *cwps*

locus) and the C-type strain SK11 (their closest relative based on WpsJ phylogeny). It is notable that UC08/11 encodes a WpsJ homologue with over 90 % aa identity with that of the C-type strain SK11, while these strains also possess WpsH and WpsI homologues (pink) indicating the likely presence of polymerisation functions while these functions are absent in the model B-type strain IL1403. Two glycosyltransferase-encoding genes are observed in the variable 3' end of the *cwps* cluster of IL1403 relative to UC08/11 suggesting unique elements in their PSP structures.

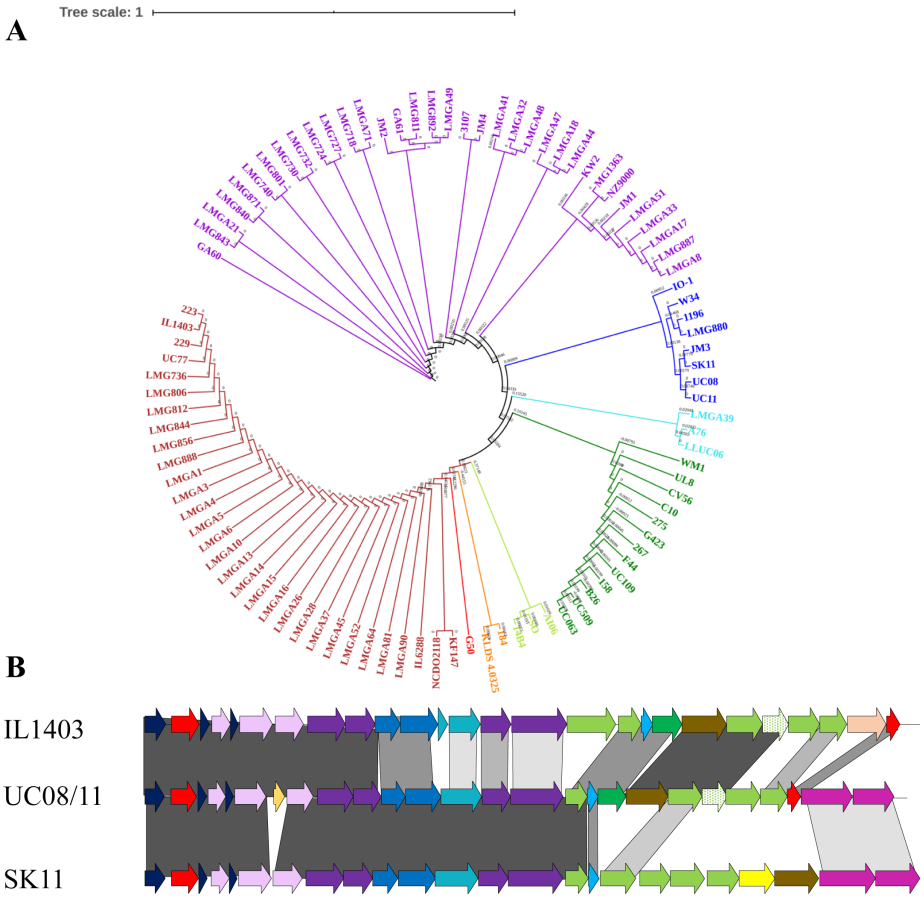
Fig. 4. Proposed model for the biosynthesis of CWPS of *L. lactis* IL1403. At the top is a representation of the *cwps* cluster in IL1403 with the associated gene names below each arrow. The lower schematic is a representation of the proposed pathway towards the construction of the rhamnan backbone and the oligosaccharide side chain and its subsequent attachment to the rhamnan component. The pathway is described in more details in the text.



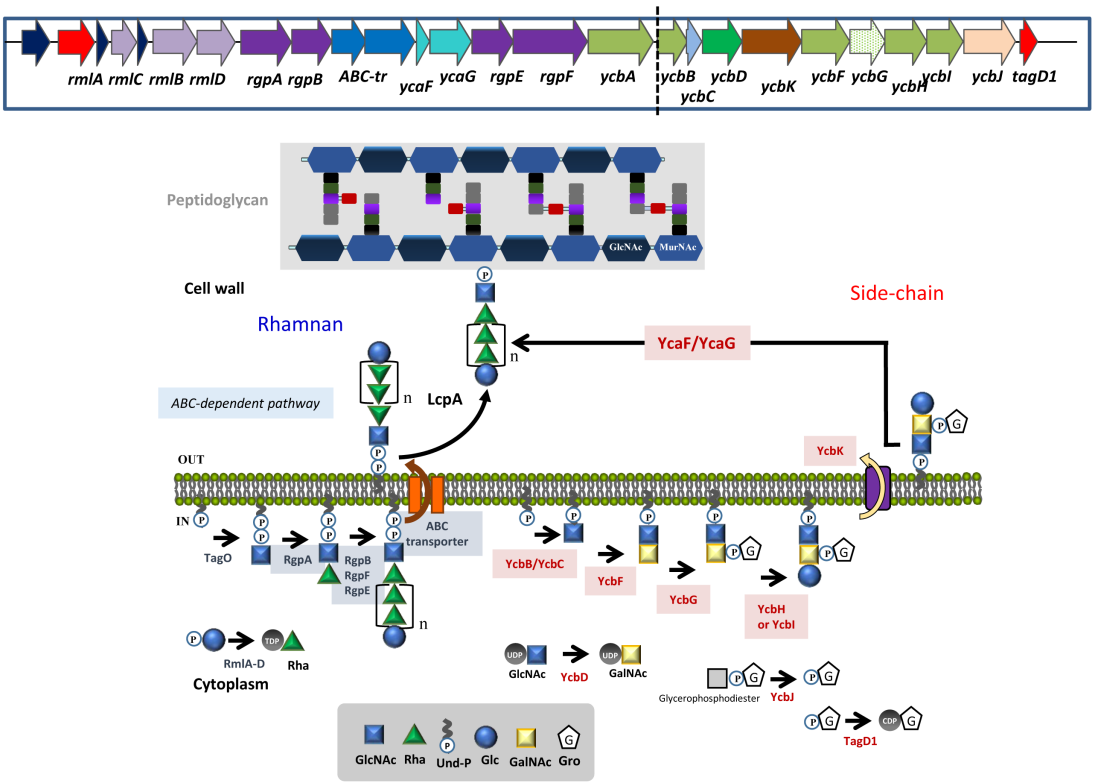
mmi_14561_f1.tiff



mmi_14561_f2.tif



mmi_14561_f3.tiff



mmi_14561_f4.tiff