Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Unravelling Phage-Host Attachment and Recognition Modalities (PHARM) in the dairy bacterium *Streptococcus thermophilus*

Thesis presented by

Zoe Kampff, MRes

for the degree of

Doctor of Philosophy

University College Cork

School of Microbiology.

Head of School/Department: Professor David Clarke

Supervisor(s): Professor Jennifer Mahony and Professor Douwe van Sinderen

2026

Table of Contents

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Signature:

Date: 12/06/2026

This thesis is dedicated to my family and Gill. I love you and thank you for everything.

List of Publications

Chapter I: Kampff Z, Van Sinderen D, Mahony J. Cell wall polysaccharides of streptococci: A genetic and structural perspective. Biotechnology Advances. 2023 Dec;69:108279. doi:10.1016/j.biotechadv.2023.108279

Chapter II: Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, Van Sinderen D, Mahony J. The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Elkins CA, editor. Appl Environ Microbiol. 2025 Oct 22;91(10):e01238-25. doi:10.1128/aem.01238-25

The following publications represent research contributions completed during the course of this PhD but are not included in the presented thesis:

Ruta S, Murray M, **Kampff Z**, McDonnell B, Lugli GA, Ventura M, et al. Microbial Ecology of Pecorino Siciliano PDO Cheese Production Systems. Fermentation. 2023 Jun 29;9(7):620. doi:10.3390/fermentation9070620

Kelleher P, Ortiz Charneco G, **Kampff Z**, Diaz-Garrido N, Bottacini F, McDonnell B, et al. Phage defence loci of *Streptococcus thermophilus*— tip of the anti-phage iceberg? Nucleic Acids Research. 2024 Oct 28;52(19):11853–69. doi:10.1093/nar/gkae814

Olupot CK, Sheehan O, **Kampff Z**, McDonnell B, Woods DF, Lugli GA, et al. Raw Milk Cheese Microbiomes: A Paradigm for Interactions of Lactic Acid Bacteria in Food Ecosystems. Foods. 2026 Mar 30;15(7):1160. doi:10.3390/foods15071160

Abstract

Streptococcus thermophilus is one of the most economically important bacterial starter cultures employed in the global dairy industry and is used extensively in the manufacture of yoghurt and cheese. Despite its technological importance, bacteriophage infection remains a major cause of fermentation delays and/or failure, resulting in reduced and inconsistent product quality and significant economic losses. Successful phage infection is dependent upon interactions between phage-encoded receptor-binding proteins and specific bacterial receptors. In *S. thermophilus*, cell wall polysaccharides including rhamnose-glucose polysaccharides (RGPs) and exopolysaccharides (EPSs) are central to phage recognition and binding processes. The overall aim of this thesis was to investigate the diversity, biosynthesis and biological functions of these polysaccharides and determine how variation in their structures influences phage-host interactions.

Comparative genomic analysis demonstrated that genes associated with RGP biosynthesis are widespread throughout the *Streptococcus* genus. Hierarchical clustering analyses revealed extensive diversity within *rgp* loci and facilitated the identification of multiple distinct *rgp* genotypes among members of the salivarius group of streptococci. Additionally, the elucidation of the chemical structures of previously uncharacterised RGPs established direct links between the *rgp* genotype and the associated RGP structure, providing new insights into the evolution and diversification of streptococcal cell wall polysaccharides.

Functional genomic analyses of bacteriophage-insensitive mutants (BIMs) possessing mutations within the *rgp* locus of the *S. thermophilus* strain UCCSt50 provided important insights into the genetic basis of RGP side chain biosynthesis.

Structural characterisation of the RGPs of specific BIMs, together with transcriptional analyses and heterologous complementation studies, enabled the construction of a preliminary model of RGP side chain assembly and provided experimental evidence supporting the functional assignment of several genes involved in RGP side chain biosynthesis. Furthermore, these studies demonstrated that structural variation within the RGP directly influences phage susceptibility. The RGP was identified as the receptor for phage P738 and distinct RGP-attracted phages were shown to possess distinct structural requirements for host binding.

Comparative genomic and structural analyses of the EPS biosynthetic loci further revealed extensive diversity among *eps* genotypes and enabled the assignment of experimentally resolved EPS structures to established genotype groups. Collectively, these findings establish direct relationships between polysaccharide genotype and biochemical structure for both RGP and EPS in *S. thermophilus* and provide a foundation for predicting polysaccharide structure from genome sequence information.

Overall, the work presented throughout this thesis demonstrates that cell wall polysaccharide diversity is a major determinant of phage-host interactions in *S. thermophilus*. More broadly, these findings position *S. thermophilus* as a valuable model through which fundamental aspects of streptococcal cell wall polysaccharide biology can be explored.

Chapter I

Introduction

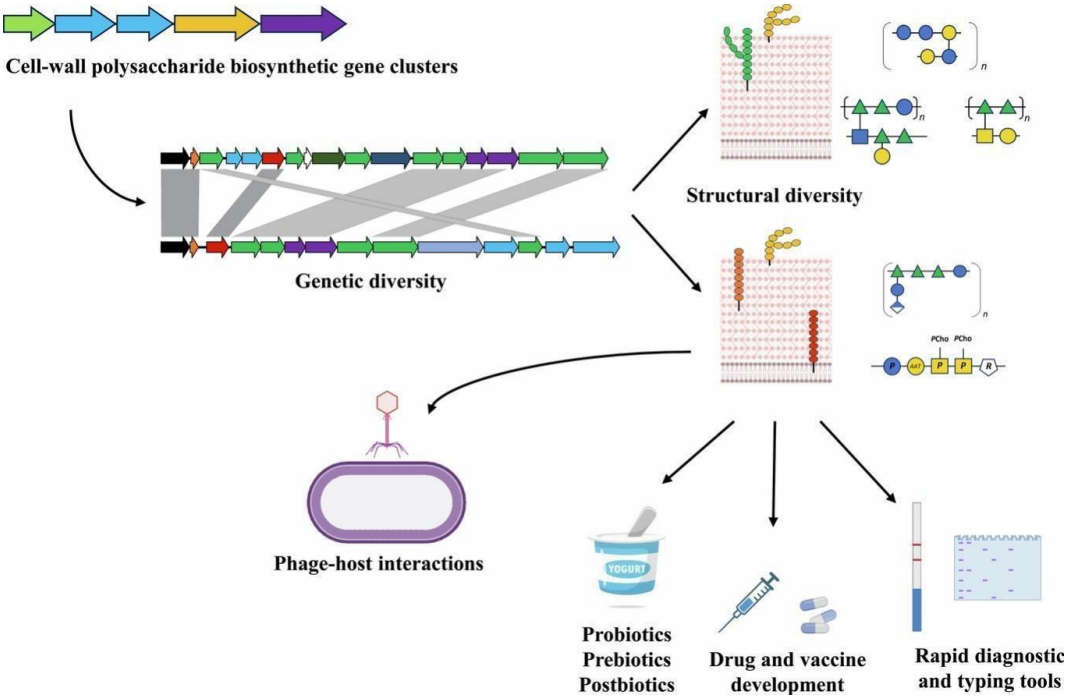
Section I

Cell wall polysaccharides of streptococci: A genetic and structural perspective

This chapter has been published as: Cell wall polysaccharides of streptococci: A genetic and structural perspective. Biotechnology Advances. 2023. Dec;69:108279. doi: 10.1016/j.biotechadv.2023.108279. Epub 2023 Oct 31. PMID: 37913948. **Kampff Z, van Sinderen D, Mahony J.**

Abstract

The *Streptococcus* genus comprises both commensal and pathogenic species. Additionally, *Streptococcus thermophilus* is exploited in fermented foods and in probiotic preparations. The ecological and metabolic diversity of members of this genus is matched by the complex range of cell wall polysaccharides that they present on their cell surfaces. These glycopolymers facilitate their interactions and environmental adaptation. Here, current knowledge on the genetic and compositional diversity of streptococcal cell wall polysaccharides including rhamnose-glucose polysaccharides, exopolysaccharides and teichoic acids is discussed. Furthermore, the species-specific cell wall polysaccharide combinations and specifically highlighting the presence of rhamnose-glucose polysaccharides in certain species, which are replaced by teichoic acids in other species. This review highlights model pathogenic and non-pathogenic species for which there is considerable information regarding cell wall polysaccharide composition, structure and genetic information. These serve as foundations to predict and focus research efforts in other streptococcal species for which such data currently does not exist.



Graphical abstract. Biotechnology Advances Volume 69, December 2023, 108279.

Introduction

The bacterial cell wall provides protection against the external environment as well as contributing to cell shape, integrity and strength. The Gram-positive cell wall contains a carbohydrate-rich layer of peptidoglycan that surrounds the cell membrane (1). The thick layer of peptidoglycan is a mesh-like structure composed of alternating β-1,4-linked *N*-acetylmuramic acid and *N*-acetylglucosamine strands covalently linked by penta-peptides (2). In addition to its primary functions in morphology and protection, it also acts as a scaffold for proteins and other glycopolymers including cell wall polysaccharides (CWPS) in Gram-positive bacteria (3). *Bacillus subtilis* and *Staphylococcus aureus* are among the best studied Gram-positive bacteria with respect to their cell walls and in which wall teichoic acids (WTAs) represent more than half of the total cell wall mass (4,5). Unlike these model organisms however, not all Gram-positive bacteria synthesise WTA to the same degree or even produce and incorporate WTA structures into their cell wall (5). Instead, some Gram-positive bacteria, including species of *Streptococcus* can produce peptidoglycan-anchored polysaccharides to the same extent e.g. more than 60% of the peptidoglycan dry weight of *Streptococcus agalactiae* (6). For streptococcal species, CWPSs can encompass lipoteichoic acids (LTAs), rhamnose-glucose polysaccharides (RGPs) and exopolysaccharides (EPS) which decorate the exposed surface of the cell (7,8). The RGP and/or EPS may form part of the capsular polysaccharides produced by many streptococci, which may inhibit complement activity and phagocytosis and in so doing, is an important virulence factor in pathogenic streptococci (9). Capsular and exopolysaccharides are also implicated in biofilm formation in pathogenic bacteria, therefore, these structures are essential to the virulence, establishment and maintenance of these organisms in their

respective ecological niches (10). *Streptococcus thermophilus* is a non-pathogenic streptococcal species that is associated with the dairy niche and many strains of this species are renowned for their ability to produce EPS although in many cases it presents more like a capsular polysaccharide being closely associated with the cell and with limited shedding into the surrounding medium. Therefore, it is timely to evaluate the state of the art of research pertaining to the cell wall polysaccharides of streptococcal species at the genetic and structural levels and to define their commonalities as well as species-specific cases that will guide future research in this field.

Historic classification of streptococci

Streptococcal diseases have affected human populations for centuries. Hippocrates described the disease "erysipelas" in the 4th century, writings of ancient scholars describe red tonsils and sore throats, and puerperal sepsis (or 'childbed fever'). The latter of these was associated with high maternal mortality rates following childbirth across Europe and North America in the 18th and 19th century (11). The first formal description of pathogenic streptococci was that of isolates from blood samples of women with puerperal fever by Louis Pasteur in 1879 (11). *Streptococcus pyogenes* was named following its isolation from suppurative lesions while *Streptococcus erysipelas* was isolated from a patient with erysipelas. Early differentiation and nomenclature strategies were based on the origin/disease associated with the streptococcal strain. Subsequently, distinct haemolytic phenotypes were observed on blood agar plates and supported a method of distinguishing streptococcal species i.e., those with a clear zone surrounding a colony were named *Streptococcus haemolyticus*; and those presenting a green zone surrounding a colony were named *Streptococcus viridans*. On this basis, pathogenic streptococci are largely divided

into these two phenotypic groups, i.e. the β-haemolytic pyogenic group or the α-haemolytic viridans group. Seminal work by Lancefield commencing in the 1920s facilitated the development of a new streptococcal classification system of β-haemolytic streptococci based on cell-surface antigen differences. The antigenic properties of group specific polysaccharides referred to as 'C-substances' initially enabled the serotype classification of Group A-E, these groupings expanded to include as many as 20 serotypes (Groups A - V) (5). The typing system encompasses pathogenic streptococcal species linked to mild and severe infections in human and animal populations as well as non-pathogenic species that primarily occupy the human oral cavity.

In contrast to the pathogenic streptococcal species, one species of the *Streptococcus* genus is a major starter bacterial species used in dairy fermentation processes i.e. *S. thermophilus*. The earliest evidence of cheese making is documented as far back as 6000 years before the common era (BCE) in northern Europe (12), while yoghurt production is also documented in Indian Ayurvedic scripts, dating from the same time (13). In addition to its technological application in the food fermentation industry, *S. thermophilus* has also gained interest as a 'probiotic' with several purported human health promoting functionalities (14).

While *S. thermophilus* is included in the heterogeneous group of lactic acid bacteria (LAB) and is related to *Lactococcus lactis,* it is phylogenetically closer to streptococcal species of the viridans group. Members of the viridans group include streptococcal species that are not beta-haemolytic, cannot tolerate high pH or salt growth conditions and cannot grow at 10°C. The salivarius group of viridans streptococci includes *S. thermophilus*, *Streptococcus salivarius* and *Streptococcus vestibularis* (15). The group are closely related based on 16S rRNA gene sequence

(16). The taxonomic status of *S. thermophilus* has evolved in recent years. Traditionally, it was grouped as part of the viridans group as a *S. salivarius* subspecies (17). Based on DNA-DNA reassociation experiments under stringent conditions, *S. thermophilus* gained full species status (15). It is also the only streptococcal species to be granted the 'Generally Recognized As Safe' (GRAS) status by the Food and drug administration (FDA) and the 'Qualified Presumption of Safety' (QPS) status by the European Food Safety Authority based on its long history of safe human consumption (EFSA) (18). Although *S. salivarius* and *S. vestibularis* are inhabitants of the human oral microbiome, several cases of meningitis, bacteraemia, and endocarditis have been associated with these organisms (19,20). The adaptation of *S. thermophilus* to the nutrient-rich milk substrate has been associated with genome decay and horizontal gene transfers (21,22). Consequently, a high proportion (approximately 10%) of the *S. thermophilus* genome is constituted by pseudogenes (21) with their original functions being non-essential for growth in milk. As a starter culture, *S. thermophilus* rapidly converts lactose into lactic acid, causing a decrease in pH and subsequent coagulation of milk proteins. Additionally, many *S. thermophilus* strains confer flavour and textural properties to the final product (23,24).

Role and significance of streptococcal cell wall glycans

As mentioned above, the three major glycans associated with the cell wall of streptococci are RGP, EPS and TAs. The RGP structures are particularly well characterised for many pathogenic and dairy streptococci due to their clinical and economic importance in their respective industries. RGPs of pathogenic streptococcal species have been associated with virulence, antimicrobial resistance, and immune evasion (5). These glycopolymers may serve as functional replacements for WTAs in certain Gram-positive species (Fig. 1) (5). Interestingly, RGPs are not universally present in all pathogenic streptococci. For example, *Streptococcus pneumoniae* instead incorporates WTAs and LTAs and the cell surface is covered with a capsule, the structure of which has high strain-to-strain variability prompting methods for pneumococcal capsular serotyping (2,25).

Due to their association with animal and human diseases, the CWPS structure and composition of pathogenic streptococci have been studied extensively and particularly their RGPs (26). Similarly, owing to their role in biofilm formation, EPS production in pathogenic streptococci has received significant research attention (27–29). Peptidoglycan anchored LTAs have been identified as virulence factors in *S. pneumoniae* (30) and more recently LTA synthesis of the closely related oral commensals *Streptococcus mitis* and *Streptococcus oralis* has been of interest due to their involvement in bacteremia and infective endocarditis (31,32). The recent resurgence of group A streptococcal disease in children in a number of European countries has focused research efforts towards the identification of virulence determinants (including CWPS) as well as vaccine development (33).

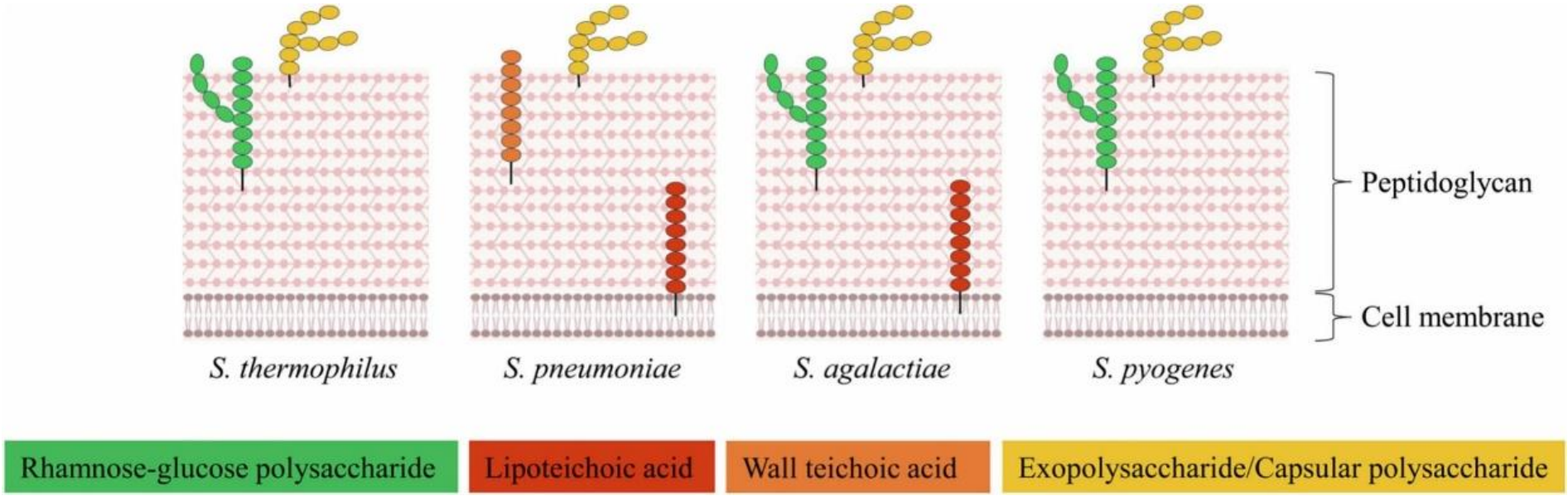


Fig. 1. Schematic representation of the major cell wall polysaccharides in the cell wall of important streptococcal species. Rhamnose-glucose polysaccharides and exopolysaccharide/capsular polysaccharides are depicted by branched chains of circles highlighting their more complex structures compared to lipoteichoic and wall-teichoic acid structures. Exopolysaccharides and capsular polysaccharides have been grouped as there is a lack of clarity regarding the genetic and/or structural relationship between exopolysaccharides and capsular polysaccharides produced by streptococcal species.

In the food industry, both RGP and EPS components of *S. thermophilus* have been implicated in (bacterio)phage-host interactions where they act as receptors for several phages (34). Phage infection of such starter cultures may lead to disrupted or failed dairy fermentations generating considerable food waste volumes and with economic consequences. However, from an applied perspective EPS produced by *S. thermophilus* is linked to desirable textures, mouthfeel and flavour of fermented dairy products and particularly so in the case of yoghurt-style products (35).

Comprehensive reviews of CWPS components of streptococci have predominantly focused on RGP analysis, highlighting their genetic diversity, biosynthesis, and physiological functions (5,36,37). Among streptococcal species, the most thoroughly characterised with respect to CWPS composition and biosynthesis include *S. pyogenes, S. agalactiae, Streptococcus mutans, Streptococcus dysgalactiae,* and *S. thermophilus*. This review will discuss current knowledge of the diversity, biosynthesis, implications and/or applications of these cell wall glycans in both pathogenic and non-pathogenic streptococci with respect to their RGP and EPS production as well the LTA synthesis for certain species. Among these saccharides, RGPs are undoubtedly the best studied cell wall polysaccharides and are, therefore, the primary focus of this review.

Rhamnose-glucose polysaccharides produced by streptococci

The polysaccharide components that are rich in L-rhamnose are described in literature as rhamnose-containing cell wall polysaccharides (5,37), rhamnose-rich cell wall polysaccharides (36), polyrhamnose polysaccharides, rhamnopolysaccharides (38), rhamnose polysaccharides (39) and rhamnose-glucose polysaccharides (RGPs) (40) among others. Despite the lack of a universal nomenclature, they are consistently described as complex heteropolysaccharides comprising a polyrhamnose chain (rhamnan backbone) accompanied by variations of oligosaccharide or polysaccharide chains (variable side chain) which can contain glucose, *N*-acetylglucosamine, galactose, *N*-acetyl galactosamine and phosphate (5). Similar to WTA, the rhamnan backbone is proposed to be covalently attached to the peptidoglycan *N*-acetyl muramic acid residues. Both the rhamnan backbone and side chain components of pathogenic and dairy streptococci display strain-to-strain variability. Among streptococcal species, RGP structures share common features most notably the relatively conserved rhamnan backbone. However, the number of L-rhamnose residues and the glycosidic linkages may differ between species and even within species. In certain *S. thermophilus* strains and in *S. mutans* serotype c, e, and f strains, the repeating rhamnose unit also contains glucose (accounting for the RGP nomenclature) (5,41) and group G streptococcal RGPs (*S. dysgalactiae* subsp. *equisimilis*) contain galactose in the rhamnan backbone (42). The side chain components appear to account for much of the structural diversity among and within streptococcal species. This structural diversity has been the basis for pathogenic and non-pathogenic streptococcal typing systems (43).

The pioneering work by Lancefield on group-specific polysaccharides facilitated a link between human and animal diseases to group specific bacteria within the genus

Streptococcus (groups A to V, excluding I and J) (15,17). The now 20 groups encompass key pathogenic species including *S. pyogenes*, responsible for a variety of diseases ranging from throat and skin infections to invasive infections including streptococcal toxic shock syndrome (44) and *S. agalactiae*, a causative agent of bovine mastitis and associated with skin and soft tissue infections in neonates, pregnant women, and immunocompromised individuals (26,45). Group A carbohydrate (GAC) (*S. pyogenes*), and group C carbohydrate (GCC) (certain *S. dysgalactiae*) as well as *S. mutans* share a common disaccharide rhamnan backbone (repeating α-1,2−/α-1,3-linked polyrhamnose backbone), with variation seen in their side chain components (36). The gene cluster harbouring the biosynthetic machinery to form the final RGP structure (herein referred to as the *rgp* gene cluster) has been identified for many streptococcal species.

The leftward end of the *rgp* gene cluster, of GAC, GCC and *S. mutans* species exhibit high sequence identity and is proposed to encode the biosynthetic functions relating to the rhamnan backbone structure (5). The decoration of the rhamnan backbone with specific side chain units produces structural diversity between these groups and yields the discriminating epitopes for serotyping. The group specific antigen of Group B carbohydrates (GBC) (*S. agalactiae*) is unique as it contains a complex multiantenna branching with a final structure containing rhamnose, *N*-acetylglucosamine, glucitol, galactose, and phosphate (46).

Despite the many merits of the Lancefield typing system, it is important to highlight that it cannot distinguish streptococci at the species level. For example, *S. dysgalactiae* subsp. *equisimilis* can express Group A, C and G antigens (5) and *S. mutans* strains can be classified into serotypes c, e, f or k based on their expressed polysaccharides (47). Interestingly, *S. pneumoniae* cannot be classified into a

Lancefield group. Instead of RGP structures described for other streptococcal species, the cell wall contains lipoteichoic acid (LTA) and wall teichoic acid (WTA) structures attached to the membrane or peptidoglycan, respectively (48). A predominant component of the pneumococcal teichoic acid is a ribitol teichoic acid termed the C-polysaccharide whose structure contains glucose, 2-acetamido-2,4,6-trideoxygalactose, galactosamine, ribitol phosphate, and phosphorylcholine covalently bound to the peptidoglycan layer. Each of these components have been identified as constituents of the capsular polysaccharide. Similar to RGP structures in the described pathogenic streptococci, the capsule structure is highly variable among pneumococcal species enabling the identification of at least 90 serotypes (49) and the successful development of highly effective pneumococcal conjugate vaccines (50).

Undoubtedly, the varied nomenclature and typing systems that have been invoked over the past decades have impeded a universal understanding of the unifying features of distinct species in terms of cell wall polysaccharide composition. Furthermore, since strains of a given species may produce different group antigens (e.g. GAC or GCC), it is more useful to define the genetic regions associated with their biosynthesis and classify strains on this basis. This would allow a more generalised understanding of the conserved and variable elements of these moieties and allow a rapid classification of emerging isolates. Therefore, it may be useful to refine or reduce the nomenclature to RGP, L/WTA and/or EPS. Since RGP is a well-established naming scheme in place for several well studied species, it seems a natural choice to use this for studies pertaining to peptidoglycan-anchored cell wall polysaccharides. EPS should be only used as the nomenclature of genuine EPS phenotypes where the polysaccharide is shed into the surrounding medium and

should be clearly distinguished from capsular polysaccharide-producing strains or species as there is a considerable lack of clarity on this point, particularly with respect to non-pathogenic streptococci.

Gene clusters associated with rhamnose-glucose polysaccharide (RGP) biosynthesis in streptococci

The genetic regions associated with RGP biosynthesis in streptococci were first identified in *S. mutans* through the characterisation of four genes (*rmlA*, *rmlB, rml*C and *rmlD*) whose products are indispensable for dTDP-L-rhamnose production (51). The *rml* genes in several streptococcal species including *S. mutans, S. pyogenes, S. agalactiae* and *S. thermophilus* are not all co-located. The *rmlA, B, and C* genes are typically clustered in the *rmlABC* locus and *rmlD* has been identified as part of a separate locus, termed the *rgp* locus. For *S. pneumoniae* the *rmlABCD* locus encodes the RmlA-D proteins that are required for dTDP-L-rhamnose synthesis (52,53). In addition to the *rmlD* gene, the *S. mutans rgp* loci also encodes six genes, *rgpA* through *rgpF*, essential for the synthesis of the rhamnan backbone (40) as well as several glycosyltransferase-encoding genes that were implicated in side chain formation and an ABC membrane transport system encoded by *rgpC* and *rgpD*. Two additional genes (*rgpI* and *rgpH)* which encode proteins involved in the formation of the glucose side chain (RgpH) and branching frequency of side chain linkages (RgpI) have extended the boundary of the *rgp* locus downstream of *rgpF* in *S. mutans* (54). Model pathways for the biosynthesis of the rhamnan backbone and subsequent addition of the variable side chain component for pathogenic streptococcal species have been proposed and will be outlined further below (5,36).

The biosynthesis of RGPs of several streptococcal species shares common processes including initiation, elongation, translocation and modifications following cell wall anchoring (5). The first enzymatic step in the synthesis of the rhamnan backbone involves the activation of the lipid carrier undecaprenyl phosphate by a UDP-GlcNAc:lipid phosphate transferase. In *S. mutans, rgpG,* a gene located outside of the *rgp* locus performs the UDP-GlcNAc:Und-P-GlcNAc-1-P-transferase activity that is functionally equivalent to the TarO/TagO/WecA integral proteins of *B. subtilis, S. aureus* and *Escherichia coli,* respectively (40,55) and is presumed to carry out the initiation step in rhamnan synthesis. Homologs of the phosphoglycosyltransferase RgpG, have also been identified (also located outside of the *rgp* locus) in *S. pyogenes* and *S. agalactiae* termed *gacO* and *gbcO,* respectively (6,56). Deficiency of RgpG has been shown to affect cell division and biofilm formation in *S. mutans* (57). Following lipid carrier activation, biosynthesis of the complete RGP structure likely occurs through the addition of monosaccharides by specific glycosyltransferases contained within the *rgp* cluster. It is proposed that the step-wise additions are initiated by the rhamnosyltransferase RgpA, which transfers the first rhamnose residue from the dTDP-L-rhamnose precursor onto a *N*-acetylglucosamine-PP-undecaprenyl (GlcNAc-PP-Und) a lipid intermediate acceptor (42). Subsequently, RgpB and RgpF are proposed to elongate the rhamnan chain through the addition of additional rhamnose moieties. RgpE is hypothesised to prevent further elongation of the rhamnan chain thus regulating rhamnan chain length in *L. lactis* (58). In *S. mutans,* RgpE contributes to the formation of glucose side chains and is responsible for glucose side chain linkage (59). An *rgpE*$^{-}$ mutant strain was capable of growing to the same maximal density (however with a reduced growth rate) as the parent strain and overall displayed little

overall fitness defects, suggesting RgpE functions after the backbone has been assembled and alludes to a protective role of the rhamnan backbone in *S. mutans* (59). Similarly, the inactivation of *gacF* in *S. pyogenes* (which shares moderate sequence identity with *rgpE* of *S. mutans*) had no effect on cell viability, and the GAC produced by the *gacF* mutant strain displayed a wild-type antigenic profile. The ABC transport system encoded by *rgpC* and *rgpD*, subsequently transports the complete rhamnose-containing glycan across the membrane (54). Although *S. mutans* can present four separate serotypes (based on the site-specific linkages of the glucose side chain) the rhamnan synthesis-associated genes *rgpA* through *rgpF* are conserved between strains reflective of their identical α-1,2/α −1,3-linked rhamnan backbone structures (54). The final step involving the rhamnan backbone is the covalent anchoring of the polymer to the peptidoglycan layer. This is most likely achieved by a transferase in the LytR-CpsA-Psr (LCP) family (60).

Decoration of the RGP backbone with specific side chain moieties is accomplished through the activity of multiple glycosyltransferases encoded within the *rgp* locus. These side chain decorations underpin the Lancefield streptococcal classification system and most notably can distinguish between GAC, GCC and *S. mutans* who share a common rhamnan backbone structure. For *S. mutans*, glycosyltransferases RgpE and RgpI add α1,2-linked glucose to rhamnan backbone (40), whereas *gacI* in *S. pyogenes* is proposed to transfer the characteristic β-linked GlcNAc side chain (56).

Detailed analysis of GAC synthesis and its assembly pathway in *S. pyogenes* has been established (56,61). With a similar architecture to that of *S. mutans* RGP synthesis, the gene cluster involved in GAC synthesis comprises twelve genes, *gacA-L,* with *gacO* typically located beyond the GAC gene cluster (56). As with *S.*

mutans, the components of the GAC cluster include a *rmlD* homolog, *gacA*, rhamnosyltransferase-encoding genes *gacB, gacC* and *gacG* and an ABC-type transport system encoded by *gacD* and *gacE* (61). Recently a common initiation step for RGP synthesis (Group A, B, C and G antigen synthesis) was discovered through heterologous expression and complementation studies whereby the substitution of *S. pyogenes gacB* with the homologous gene from *S. agalactiae*, *S. equi* subsp. *zooepidemicus*, *S. dysgalactiae* subsp. *equisimilis*, or *S. mutans* complemented the GAC biosynthesis pathway (42). Similarities in architecture and gene content between important pathogenic streptococci has immense importance for future studies in understanding pathogenesis and potential therapeutics.

While the early studies of CWPS biosynthesis were aimed at the development of clinical diagnostics and a basis to predict pathogenicity of streptococcal species, recent studies seek to apply the CWPS composition and structure information to improve drug targeting, vaccine design and phage therapy systems (62–64). Furthermore, there is a significant and growing volume of data pertaining to the diversity, composition, structure and biosynthesis of CWPSs of the dairy-associated species *S. thermophilus* (65–67). The *rgp* gene cluster of *S. thermophilus* shares synteny with those of their pathogenic counterparts (Fig. 2). However, in contrast to *S. mutans*, the *rgpA-F* cluster of *S. thermophilus* associated with rhamnan backbone biosynthesis is co-located with genes associated with variable side chain production (22,65). A model for the biosynthesis of the RGP in *S. thermophilus* has been recently proposed (65). The RGP component is universally present in all analysed *S. thermophilus* strains and is closely associated with the cell wall. RGPs of *S. thermophilus* are characterised as structures containing a polyrhamnose backbone with some strains also incorporating glucose in their backbone structure.

The final RGP structure of *S. thermophilus* strains can also contain galactose, *N*-acetylglucosamine or *N*-acetylgalactosamine as well as the previously noted rhamnose and/or glucose subunits primarily in the side chain component (Fig. 3) (65,68).

S. thermophilus has historically been considered to lack genetic diversity due to its widespread application as an industrial starter culture (16). However, the *rgp* gene cluster displays high variability between strains (34,43,67). Similar to Lancefield classification or pneumococcal serotyping techniques, the genetic diversity of *rgp* clusters has been harnessed to devise multiplex PCR systems for their differentiation and classification (65,73). Hierarchical clustering analysis of the *rgp* gene clusters *S. thermophilus* strains genomes revealed considerable strain level diversity, enabling the identification of seven distinct groups of *S. thermophilus rgp* clusters termed Rgp1–7 (34,43,65). Detailed analysis of the rhamnan (backbone) and side chain (variable)-associated regions of the *rgp* locus revealed three distinct backbone genotypes (Bt) and five variable genotypes (Vt) (65). Theoretically 15 unique combinations of backbone and side chain may exist for *S. thermophilus* strains. To date however, just seven of the potential 15 combinations have been identified and five complete RGP structures have been elucidated, each representing a unique combination of backbone and side chain components and associated *rgp* genotype (65).

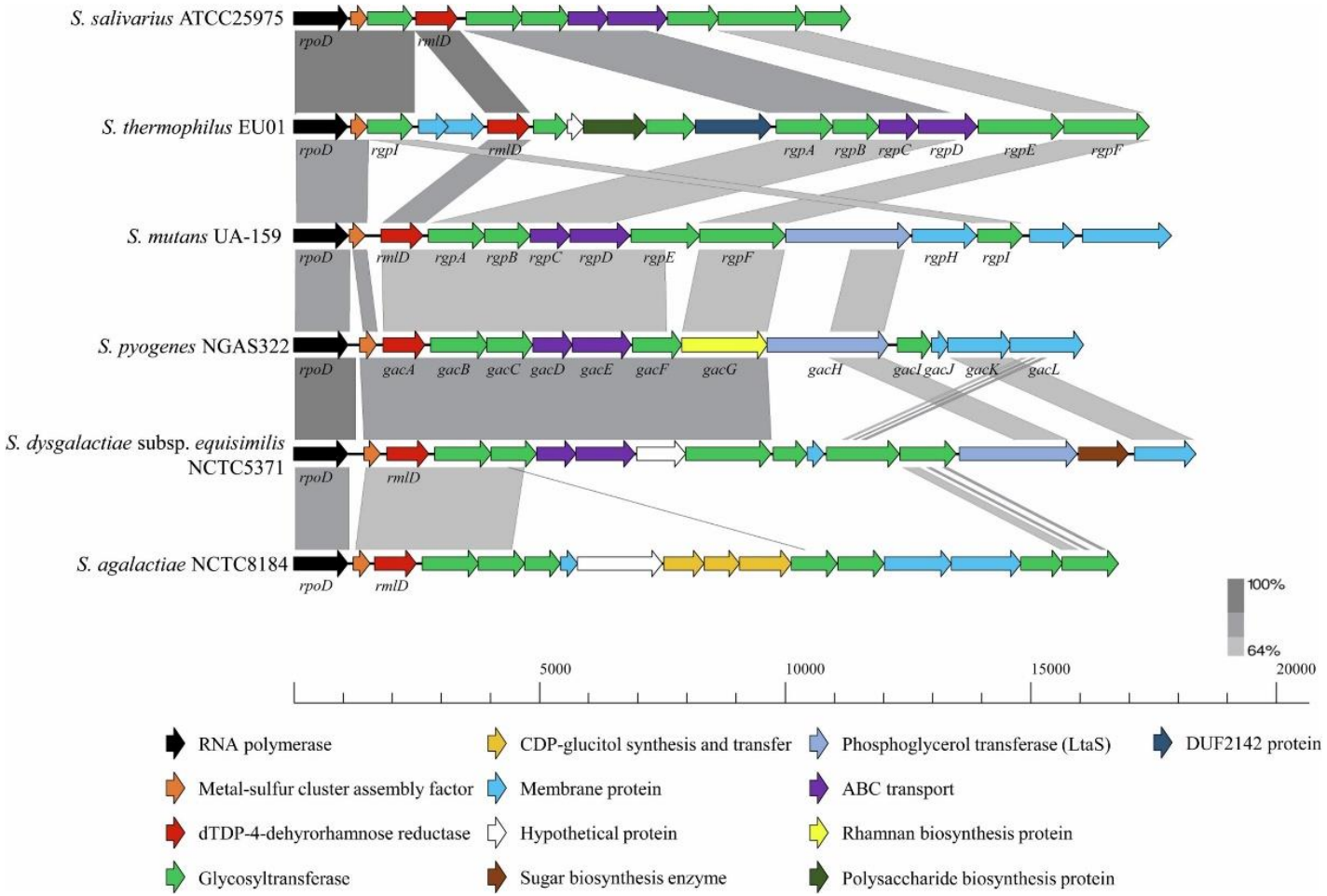


Fig. 2. Comparison of rhamnose-glucose polysaccharide (RGP) biosynthetic gene clusters of representative pathogenic and non-pathogenic streptococcal species. The 5′ end of the *rgp* gene cluster is flanked by *rpoD* in the analysed species. The proposed functions of the gene products are indicated by colour-coded arrows with functions annotated under the scale bar. Regions of homology are joined by blocks in grayscale and generated using Easyfig (69). The scale bar is measured in base pairs. The gene clusters of the following strains are shown: *S. salivarius* ATCC 25975, *S. thermophilus* EU01, *S. mutans* UA-159, *S. pyogenes* NGAS322, *S. dysagalataie* subsp. *equisimilis* NCTC5371, *S. agalactiae* NCTC5371.

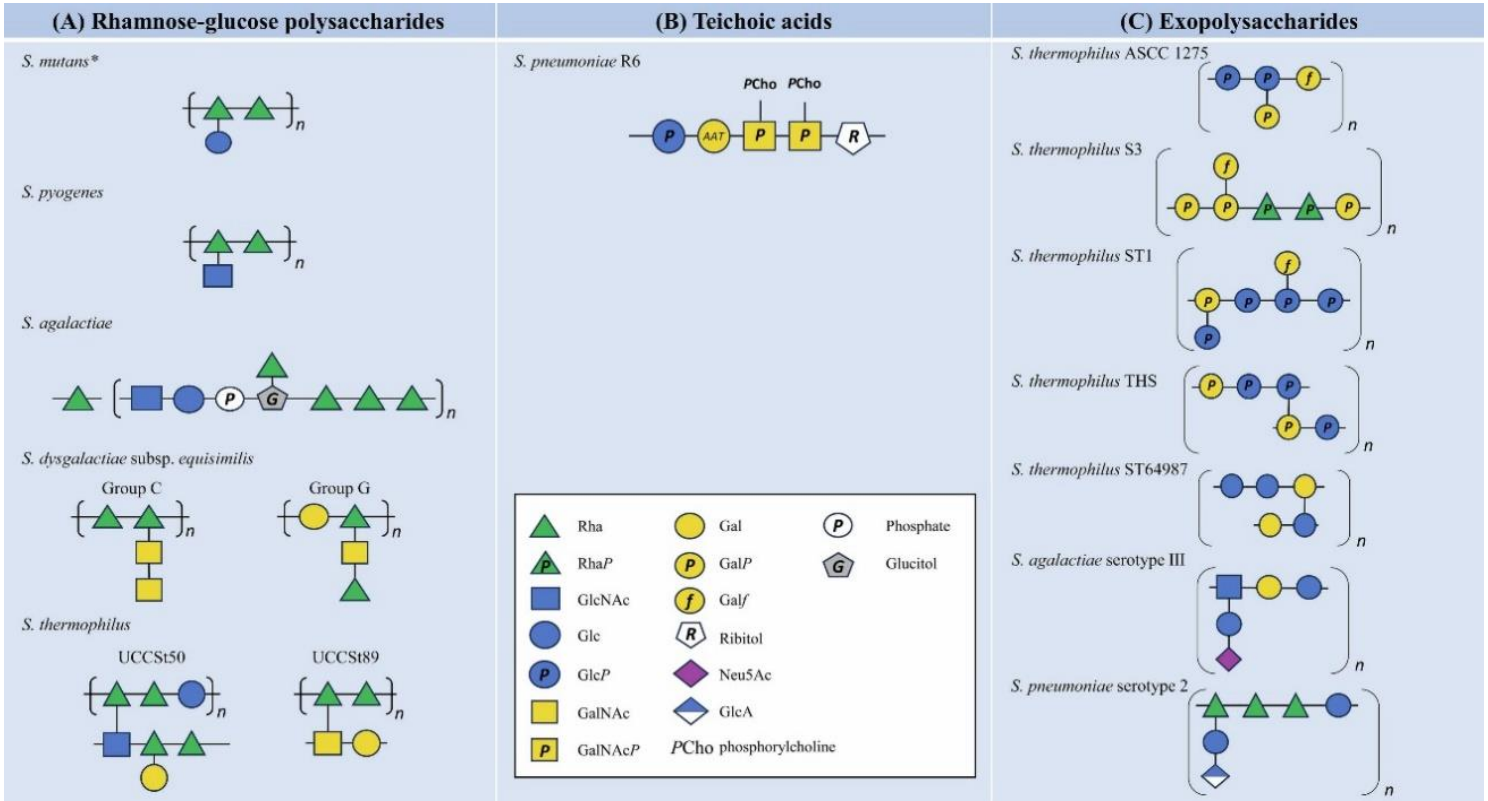


Fig. 3. Overview of the main cell-wall polysaccharide structures within different *Streptococcus* species. **(A)** Schematic representation of the rhamnose-glucose polysaccharide (RGPs) structures. *For *S. mutans* the glucose side chain decoration varies for each serotype (the glucose side chain be absent (serotype k) or be linked to the rhamnan backbone in a α-1,2 configuration (serotype c), β-1,2 configuration (serotype e), or a α-1,3 configuration (serotype f). In addition, GroP (glycerolphosphate) attachments have been identified on the serotype c carbohydrate (70). GroP modifications have also been identified on the GlcNAc substituents in *S. pyogenes* (39). **(B)** Schematic representation of the teichoic acid (TA) structure of *S. pneumoniae* R6 where ATTGal is 2-acetamido-4-amino-2,4,6-trideoxygalactose. The repeating units can be attached to the glycolipid glucose-diacylglycerol (Glc-DAG) to yield membrane-anchored lipoteichoic acid (LTA) or to peptidoglycan via Nacetylmuramic acid (MurNAc) to produce wall teichoic acid (WTA) (71). **(C)** Schematic representation of the exopolysaccharide (EPS) structures produced by *S. thermophilus, S. pneumoniae* and *S. agalactiae* (72).

Biosynthesis of rhamnose glucose polysaccharides

The *rgp* locus, encoding the biosynthetic machinery for both backbone and side chain synthesis, is a region of approximately 20–30 kb based on a recent study of 23 *S. thermophilus* strains (67,68). While the size and gene content of the *rgp* locus appears to be strain-specific, the general organisation of the gene cluster is conserved among strains. In *S. thermophilus*, the *rgp* locus comprises a conserved set of genes associated with rhamnan backbone synthesis, including rhamnosyltransferases RgpA, RgpB, and RgpF as well as an ABC transporter/transport system (RgpC and RgpD), involved in exporting the side chain component to the outer surface of the cell membrane (22,36). The *S. thermophilus rmlD* and *rgpA* through *rgpF* homologs share high identity with their *S. mutans* counterparts. Based on *rgp* gene cluster analysis of a total of 78 strains, the ABC transport system (RgpC and RgpD) is present in all *S. thermophilus* strains analysed (34,67). As with pathogenic streptococci, the attachment of the variable side chain is likely executed via a multicomponent glycosyltransferase system with a number of encoded glycosyltransferases present in *S. thermophilus* strains. The exact mechanisms by which the RGP backbone and side chain components are synthesised and their ultimate convergence to produce the final RGP structure remain to be experimentally proven for *S. thermophilus*.

dTDP-L-rhamnose is also essential for RGP synthesis in the non-pathogenic *S. thermophilus.* The biosyntheis of dTDP-L-rhamnose is reliant on the *rml* operon which contains *rmlA, B* and *C* and is located outside of the *rgp* locus of all *S. thermophilus* strains. *rmlD* however is located at the beginning of the *rgp* locus and shares high identity between strains and with pathogenic streptococci (Fig. 2). For *S. thermophilus,* the initiation of rhamnan backbone synthesis is mediated by an

RgpG homolog (encoded by *stu163*) (65). The rhamnosyltransferases RgpA, B and F are proposed to be functional analogues of those described in pathogenic streptococcal species. RgpA transfers the first rhamnose residue, RgpB and RgpF are proposed to elongate the rhamnan backbone through the addition of individual monosaccharide units while RgpE is hypothesised to regulate rhamnan chain length. The ABC-transport system (encoded by *rgpC* and *rgpD*) transports the backbone polymer across the membrane following its intracellular synthesis. The incorporation of the backbone into the peptidoglycan layer is most likely performed by LCP family proteins acknowledging the similarities between *S. thermophilus* and *L. lactis* in terms *rgp* gene content and architecture (LcpA being the major LCP protein involved in rhamnan backbone attachment to peptidoglycan in *L. lactis*) (58).

Biosynthesis of the side chain structure in the *S. thermophilus* is believed to be initiated by (ScbA) protein A, and its associated activator, ScbB (65). These proteins initiate side chain synthesis by the addition of a GlcNAc (or GalNAc in certain strains) moiety to the lipid carrier Und-P, representing functional homologs of *S. pyogenes* GacI/J (61). The elongation of the side chain is believed to be performed by glycosyltransferases in the *rgp* locus which are predicted to add monosaccharides sequentially to the lipid-linked GlcNAc foundation (65).

Translocation of the side chain across the cytoplasmic membrane typically requires proteins which 'flip' the saccharidic component following their intracellular synthesis often via a Wzy/Wzx-dependent pathway (74). The *S. thermophilus* ScbF of strain UCCSt50 shares structural similarly to an *E. coli* flippase and is hypothesised to be responsible for the translocation or 'flipping' of the saccharidic

side chain across the cytoplasmic membrane (65). Interestingly, a putative Wzx-like flippase encoding gene is present for a majority of analysed *S. thermophilus* strains. The attachment of the side chain component to the rhamnan backbone is believed to be performed by a homolog of WpsJ, a GT-C fold glycosyltransferase discovered in the *L. lactis* strain MG1363 (58). Regulation of the side chain branching is likely the function of *rgpI* in *S. thermophilus* as it shares significant sequence identity to *rgpI* of *S. mutans*. Adding to the unique architecture, *rgpI* is the first gene in the *S. thermophilus rgp* cluster and one of distal genes in pathogenic streptococcal species (Fig. 2).

Teichoic acids produced by streptococci

The term TA has historically been used to describe all bacterial cell wall, membrane and capsular polymers which contain glycerolphosphate (GroP) or ribitolphosphate (RboP) (75). Currently, the term encompasses two cell wall polymers in Gram-positive bacteria: wall teichoic acid (WTA) and lipoteichoic acid (LTA). LTAs are tethered to the plasma membrane and extend into the peptidoglycan layer, while WTAs are covalently attached to the *N*-acetyl muramic acid component of peptidoglycan (via a phosphodiester linkage) and extend through and beyond the cell wall (4,55). WTA glycopolymers are well studied in organisms such as *Bacillus* and *Staphylococcus* due to their crucial roles in bacterial physiology, cell division, antibiotic resistance and protection against host defence peptides and enzymes (39). Interestingly, many streptococci lack the classically defined WTA and instead produce rhamnose-containing polysaccharides. Therefore, the idea that these glycopolymers act as functional homologs of TAs is one of note in the current literature (5,6,39,59). Although early work reported the presence of glycerol and

phosphate residues in a number of streptococcal species glycopolymers there remains a knowledge gap pertaining to TA synthesis in streptococci.

Recently, the presence of GroP in *S. pyogenes* and serotype c carbohydrate *S. mutans* was determined to be mediated by the product of *gacH* (or a *gacH* homolog) (39). GacH, a membrane protein belonging to the alkaline phosphatase superfamily, catalyses GroP addition to the C6 hydroxyl group of GlcNAc substituents (39). The GroP attachment present on ~25% of the GlcNAc side chains, converts the saccharide to polyanionic cell wall glycopolymer decorations (39). Similarly, GroP decorates the Glc side chain of the RGP in a *S. mutans* serotype c strain (70). Based on phylogenetic analysis, GacH homologs are present in numerous streptococci, suggesting other species may also express glycopolymers with GroP additions on their side chain components (39). Interestingly, GroP-modified side chains may have been largely overlooked due to acid-based methodologies for RGP extraction that may lead to GroP removal (36).

The cell wall of *S. pneumoniae* is a complex structure that contains WTA and LTA surrounded by one of the 90 capsular polysaccharides produced by members of the species (48). *S. pneumoniae* is unique among streptococcal species in presenting identical WTA and LTA repeating units and only the anchoring location (peptidoglycan or membrane) defines their classification (Fig. 3B) (30,48,71). In most bacteria, these chain structures are different for the two polymers (76). The TA biosynthesis pathway has been proposed based on bioinformatic analysis of various pneumococcal strains (71). Genes associated with TA biosynthesis are located in three genomic loci, termed *lic1*, *lic2* and *lic3* (Fig. 4). These regions contain genes associated with choline uptake, CDP-choline and CDP-ribitol synthesis, teichoic acid flippase functions, and phosphotransferases (48). Of note

TacF, a putative teichoic flippase, binds specifically to the choline-containing TA precursor chains before transferring the structure across the cytoplasmic membrane, providing evidence that all mature TAs are decorated with choline residues, essential for the choline-dependent growth of *S. pneumoniae* (77). Subsequently, the mature TA chain is anchored to peptidoglycan or cell membrane to form WTA or LTA, respectively (48). Similar to RGP anchoring to peptidoglycan, members of the LCP family of proteins have been implicated in the cell wall maintenance and assembly of *S. pneumoniae* (7). For LTA structures, the membrane protein TacL, transfers the TA precursor chains onto the glycolipid anchor (30). Similar LTA structures to those described for *S. pneumoniae* are described in *S. mitis* and *S. oralis* (32).

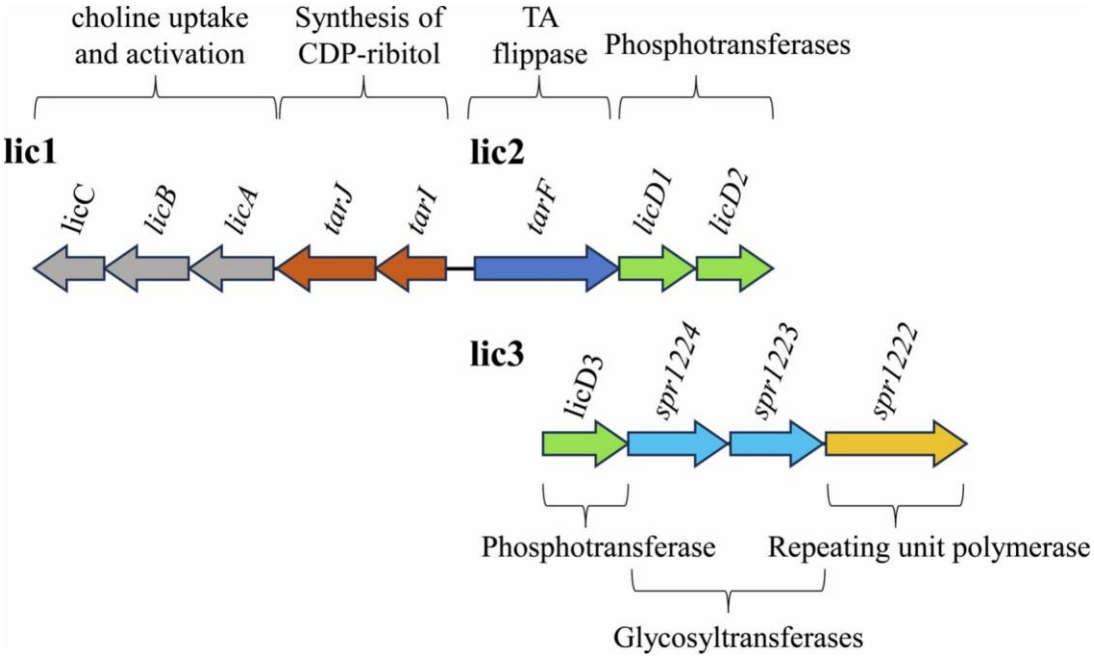


Fig. 4. Genetic organisation of the *lic* regions of *Streptococcus pneumoniae* R6 and associated functions of the gene products (71). The *lic1* operon contains genes associated with choline uptake, CDP-choline and CDP-ribitol synthesis. The adjacent *lic2* region encodes a putative teichoic acid (TA) flippase in addition to phosphotransferases. The *lic3* region is present in a disparate location to the *lic1* *lic2* gene clusters on the chromosome with proposed functions of gene products including a phosphotransferase, glycosyltransferase(s) and integral membrane protein(s).

Exopolysaccharides of *Streptococcus thermophilus*

While pathogenic streptococci such as *S. mutans* are renowned for their biofilm forming activity, the descriptions of the associated polysaccharides, their composition, structure and genetics is surprisingly limited. Biofilm formation may involve capsular polysaccharides or exopolysaccharides; however, this is not clearly delineated among pathogenic streptococci. It seems most likely that this property is invoked by capsular polysaccharides (CPS), which are described to be a widespread phenomenon among pathogenic streptococci (78). The size (~16–36 kb) and organisation of the *cps* loci appears similar to those of the *rgp* loci (78) but remains to be experimentally established. Reports of EPS among pathogenic streptococci are rare and may, in fact, be CPS. The importance of TA and CPS in pathogenic species has dominated and likely overshadowed research of EPS in these species. In contrast, the EPS of *S. thermophilus* are thoroughly investigated with extensive reports of the genetic mechanisms underpinning their biosynthesis as well as their composition and structure. Here, the review presents an overview of the current knowledge of the genetics, composition and structure of the EPS of *S. thermophilus*.

Similar to RGP, EPS produced by *S. thermophilus* strains are heteropolysaccharides that primarily contain glucose, galactose, and rhamnose, and in some strains can contain fucose, *N*-acetylglucosamine, *N*-acetylgalactosamine, ribose, and glucuronic acid as well as non-carbohydrate units such as glycerol, phosphate, pyruvyl, and acetyl groups (66,79–82). The gene cluster that encodes the biosynthetic machinery for EPS production in *S. thermophilus*, termed the *eps* cluster exhibits considerable strain-to-strain variations in terms of size and gene content (67). Furthermore, the locus termini appear to be highly conserved with

variation primarily observed in the central regions of the cluster in which genes encoding glycosyltransferases are located (67,83). Hierarchical clustering analysis established the presence of six distinct groups of *S. thermophilus eps* genotypes designated as EPS A to F (43). Recent analysis of dairy streptococcal isolates identified a further four EPS genotypes (A to J) (67). The genetic diversity of these clusters has yet to be linked to the compositional and structural diversity of these polysaccharides.

The *eps* gene cluster is approximately 15–30 kb in size and contains ∼18 genes on average (67,68). Despite the genetic diversity observed in the *eps* cluster of various *S. thermophilus* strains, the loci harbour genes that encode conserved functions associated with EPS chain length regulation, glycosyltransferase activity, polymerization and translocation (Fig. 5) (23,67). The *eps* gene cluster contains three functional modules i.e. *epsAB* are located at the leftward end of the cluster and are involved in regulation of EPS production as well as *epsCD* whose products regulate EPS chain length (84). Two pairs of *epsC-epsD* genes were identified in the model *S. thermophilus* strain ASCC 1275, which are hypothesised to facilitate the strain producing two types of EPSs - a capsular EPS attached to the cell surface and a ropy EPS secreted into the extracellular environment (Fig. 5) (85). The second functional module is associated with the synthesis of repeating units of EPS (*epsE, epsF, epsG, epsH* and *epsI* and additional genes in some cases) and the third module is responsible for the polymerization of the repeating units and the export of the intracellularly synthesised EPS to the extracellular environment (*epsJ, epsK, epsL* and *epsM*) (84,86). The membrane-associated glycosly-1-phosphate transferase (EpsE) catalyses the first step in the assembly of the EPS through the transfer of a sugar-1-phosphate to a lipid carrier located on the cytoplasmic surface of the

membrane (86,87). Similar to RGP synthesis the transfer of the first unit triggers the addition of subsequent repeating units to the EPS (84).

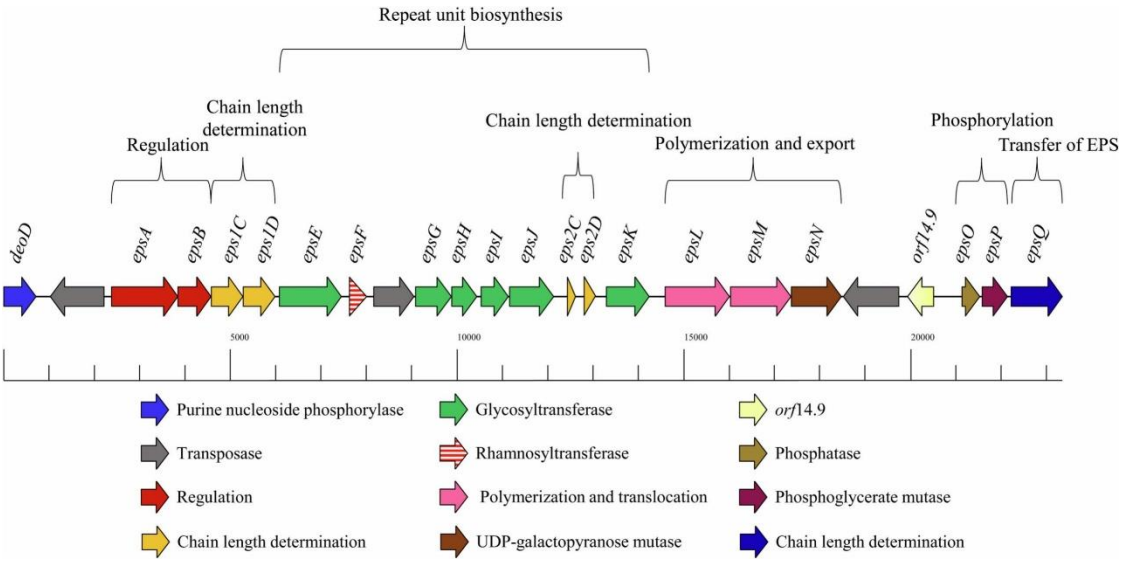


Fig. 5. Schematic depicting the architecture of the *eps* gene cluster in *S. thermophilus* ASCC 1275 as a representative of the species (85). Predicted functions are indicated by differently coloured arrows and highlight the main functions encoded by genes within the cluster including regulation, chain length determination, glycosyltransferases involved in repeat unit biosynthesis and polymerization and export.

The EPS produced by *S. thermophilus* strains is reported to be primarily composed of glucose, rhamnose and galactose (88) while strain-specific combinations of galactose and rhamnose; galactose, glucose and *N*-acetylgalactosamine; galactose, glucose and rhamnose; galactose, glucose, rhamnose and *N*-acetylgalactosamine have also been reported (Fig. 3C) (66,79–82) The compositional diversity of these structures can be linked to the highly polymorphic nature of *S. thermophilus eps* cluster-encoded glycosyltransferases which commonly include galactosyltransferases, rhamnosyltransferases, glucosyltransferases, and *N*-acetylglucosamine transferases and less commonly include galactofuranose transferases, mannosyltransferases and fructosyltransferases (84). Polymerization-associated genes as well as export-associated genes are located at the 3′ end of the *eps* cluster. The nomenclature associated with these functions is inconsistent and is entirely dependent on the number of genes preceding those associated with polymerization functions. For example, *epsJ, epsL* and *epsM* in one *S. thermophilus* strain are associated with polymerization and translocation of the EPS repeating units while *epsQ* is associated with the transfer of EPS between the membrane and peptidoglycan layer in another strain (85). Therefore, while there is an abundance of sequence and compositional data associated with EPS production in *S. thermophilus*, it is clear that there is significant diversity in both their genetic and chemical composition causing lack of clarity on some fundamental points including (i) the lack of a universal nomenclature for the *eps* associated genes and their encoded products; (ii) a lack of clarity regarding the nature of CPS and its genetic/structural relationship to either EPS or RGPs and; (iii) the extent of chemical diversity of EPS in this species. To date the chemical composition of the EPS of a large number of strains of *S. thermophilus* have been elucidated while the

chemical structure of only a handful have been completely resolved. This will likely represent an area of research growth in the coming years as it is imperative to understand the importance of these structures on microbiota- and immune-modulation in humans as well as the well-studied technological properties of fermented foods.

Implications of cell wall polysaccharide diversity in the food and healthcare industries

Starter cultures, probiotics, prebiotics and postbiotics

Bacteriophages, also known as phages, continue to be among the most significant and consistent threats to dairy fermentation processes (34). Disruption caused by phages in the dairy industry have significant technological and economic consequences while failed fermentations contribute to global food waste (34). Therefore, food fermentation processes depend on reliable, technologically favourable, and robust starter cultures to reduce waste and increase the sustainability of production practices. Recent studies have highlighted the role of both RGP and EPS in strain-specific phage-host interactions and the development of multiplex PCR systems facilitate the rapid classification of dairy streptococcal strains based on their *rgp* genotype (65). Such classification may be applied to identifying strains that present distinct RGP structures on the cell surface and that reduce the risk of phage infection and loss of the fermentate. Such strategies could also be developed for the classification of *eps* genotypes for this species and further enhance the robustness of starter cultures. Currently, five genetically distinct phage groups of dairy streptococcal phages are known to exist and among these the Brussowviruses are known to recognise RGP receptors while EPS is the implicated

receptor for 987 group phages (65). Limited information exists for the receptor of the remaining three phage groups although it is likely that one or both of these saccharides is required by these phages also. Owing to the pervasiveness of phages in the industrial setting, it is suggested that this pressure has underpinned the diversity of the *rgp* and *eps* loci. As the food industry adapts to growing demands for fermented products derived from plant-based alternatives to dairy, knowledge of *S. thermophilus* genetic composition and metabolic capabilities will underpin industry's ultimate success and ability to adapt to these, and other emerging, demands and challenges.

The metabolism of lactose by *S. thermophilus* has long been regarded a defining probiotic trait of strains of this species. However, the survivability of *S. thermophilus* strains in the human gastrointestinal tract is the subject of debate and is likely a strain-specific attribute. Despite this, there are numerous possible benefits associated with *S. thermophilus* EPS including such as the limitation of pathogen colonisation/biofilm formation, immune system activation, anti-inflammatory and prebiotic activities (89). Furthermore, even in the absence of viability, the EPS may impart such effects as a "post-biotic" further highlighting the potential of this species and its associated CWPS in contributing to human health. While the true benefits of *S. thermophilus* may emanate from its co-association with proto-cooperative organisms such lactobacilli, the importance of *S. thermophilus* in the food and probiotics industries is undeniable.

Classical phage typing of pathogenic bacterial species was based on the unique interaction profiles of clinical isolates with phage panels known to infect strains of given serotype(s). These serotypes, being associated to cell wall polysaccharides, highlight the long-standing implications of cell wall polysaccharides including

RGP, EPS and teichoic acids. As the exact composition and/or structure of the majority of pathogenic streptococcal species is incomplete, it is not currently possible to define if there is an explicit link between cell wall polysaccharides that are involved in virulence and phage interactions or if these are distinct moieties; however, this is certainly an aspect that will become clear as an increasing number of cell wall polysaccharide structures and genome sequences of emerging strains become available.

Drug development and therapeutics

RGPs of pathogenic streptococcal species present promising opportunities in human therapeutic vaccine design, drug targeting and phage therapy approaches (62,64,90). The loss of virulence and suboptimal cell growth and cell morphology exhibited by a number of pathogenic streptococcal species following depletion of rhamnose, has highlighted the streptococcal L-rhamnose biosynthesis as a useful interesting drug target (64). Moreover, as dTDP-L-rhamnose is not produced or used by humans (5), drugs targeting the L-rhamnose biosynthesis pathway would have limited undesirable side effects. A recent study identified compounds capable of inhibiting dTDP-L-rhamnose biosynthesis in streptococcal species (64). Further knowledge regarding this pathway in additional pathogenic species could optimize steps in drug design in the healthcare industry.

In *S. mutans,* RgpG plays a crucial role in cell division which impacts biofilm formation, revealing a new potential target to develop anticaries vaccines or therapeutics (57). Similar strategies have been exploited in *S. aureus*, whereby the inhibition of the UDP-GlcNAc:lipid phosphate transferase, TarO, renders the

bacterium avirulent (91) and in *S. epidermidis* TagO deficiency was reported to cause major defects in biofilm development and formation (92).

CPSs of *S. pneumoniae* are immunoreactive components of the pathogen that have formed the basis of carbohydrate-based vaccines (93). Similarly, research into capsular polysaccharide conjugate vaccine development for *S. agalactiae* is underway, however no licensed vaccine is currently available. A *gacI* mutant of *S. pyogenes* prevented GlcNAc side chain addition in *S. pyogenes* and was attenuated for virulence in two infection models (56). This represents additional opportunities to develop alternate vaccine antigens devoid of the presumptive infection-causing side chain components. As neither *S. pyogenes* nor *S. agalactiae* vaccines are currently available more research into understanding pathogenesis relating to CWPS is necessary for future vaccine designs.

There is growing interest in exploiting bacteriophages to combat bacterial infections due to the rise in nosocomial multidrug-resistant bacterial infections. Phages have evolved to recognise and adhere to bacterial cell surfaces with subsequent infection and replication within the cell. Targeted elimination of pathogenic streptococcal infections is gaining creedance once again given the current antibiotic crisis. Endolysins, which are phage-encoded peptidoglycan hydrolases responsible for the degradation of peptidoglycan and subsequent cell lysis (94) have become increasingly prevalent in the fight against streptococcal infections. An extensive review of the numerous streptococcal phages that target pathogenic species as well as the challenges that limit phages and endolysin application in the healthcare at present has recently been published (95).

Definition of the diversity and role of EPS and CPS in streptococcal species, especially in biofilm-forming species may lead to the development of novel antimicrobial agents by enhancing biofilm dispersion or by inhibiting the reproduction and metabolism of pathogens (96). While biofilm formation and pathogenesis are well studied in *S. mutans*, it is less well defined and understood in other streptococcal species associated with the oral cavity and may represent an area of emerging interest in the coming years (27).

Conclusions

Streptococci have evolved to diversify the cell surface glycans to adapt to a range of ecological niches. These glycans may be interchangeable depending on the species and the specific roles that they play in that host. Recent advances in the understanding of the genetic, compositional and structural diversity simultaneously increases awareness of the opportunities for tailored drug/vaccine development and sustainable food production systems. However, species- or niche-specific studies have also created a challenge as there is currently a lack of universal definition of the associated biosynthetic pathways and structures. While this may create a temporary impediment to progress in streptococcal glycobiology research in streptococci, the breadth of depth of activity that is currently underway gives cause for optimism in the context of therapeutics, diagnostics and the healthcare and food industries. To achieve this, it is imperative that the field unifies its efforts in linking the genetic diversity of the biosynthetic clusters to the composition and chemical structure of the final cell wall polysaccharide moieties. Secondly, it is essential that the field embarks on a systematic analysis of representatives of a breadth of streptococcal species to define the links between the various species and understand the biosynthetic processes that will have implications for the development of vaccines and therapeutics. Finally, it will be important to have a universal nomenclature to provide transparency in the field and a broader understanding of the common and unique aspects of the cell wall polysaccharides of streptococcal species that have profound impact on human health.

References

1. Silhavy TJ, Kahne D, Walker S. The Bacterial Cell Envelope. Cold Spring Harbor Perspectives in Biology. 2010 May 1;2(5):a000414–a000414. doi:10.1101/cshperspect.a000414

2. Vollmer W, Blanot D, De Pedro MA. Peptidoglycan structure and architecture. FEMS Microbiol Rev. 2008 Mar;32(2):149–67. doi:10.1111/j.1574-6976.2007.00094.x

3. Garde S, Chodisetti PK, Reddy M. Peptidoglycan: Structure, Synthesis, and Regulation. Slauch JM, editor. EcoSal Plus. 2021 Dec 15;9(2):ecosalplus.ESP-0010-2020. doi:10.1128/ecosalplus.ESP-0010-2020

4. Brown S, Santa Maria JP, Walker S. Wall Teichoic Acids of Gram-Positive Bacteria. Annu Rev Microbiol. 2013 Sep 8;67(1):313–36. doi:10.1146/annurev-micro-092412-155620

5. Mistou MY, Sutcliffe IC, Van Sorge NM. Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria. Bitter W, editor. FEMS Microbiology Reviews. 2016 Jul;40(4):464–79. doi:10.1093/femsre/fuw006

6. Caliot É, Dramsi S, Chapot-Chartier MP, Courtin P, Kulakauskas S, Péchoux C, et al. Role of the Group B Antigen of *Streptococcus agalactiae*: A Peptidoglycan-Anchored Polysaccharide Involved in Cell Wall Biogenesis. Wessels MR, editor. PLoS Pathog. 2012 Jun 14;8(6):e1002756. doi:10.1371/journal.ppat.1002756

7. Kawai Y, Marles-Wright J, Cleverley RM, Emmins R, Ishikawa S, Kuwano M, et al. A widespread family of bacterial cell wall assembly proteins. EMBO J. 2011 Dec 14;30(24):4931–41. doi:10.1038/emboj.2011.358

8. Weidenmaier C, Peschel A. Teichoic acids and related cell-wall glycopolymers in Gram-positive physiology and host interactions. Nat Rev Microbiol. 2008 Apr;6(4):276–87. doi:10.1038/nrmicro1861

9. Hyams C, Camberlein E, Cohen JM, Bax K, Brown JS. The *Streptococcus pneumoniae* Capsule Inhibits Complement Activity and Neutrophil Phagocytosis by Multiple Mechanisms. Infect Immun. 2010 Feb;78(2):704–15. doi:10.1128/IAI.00881-09

10. Balducci E, Papi F, Capialbi DE, Del Bino L. Polysaccharides' Structures and Functions in Biofilm Architecture of Antimicrobial-Resistant (AMR) Pathogens. IJMS. 2023 Feb 17;24(4):4030. doi:10.3390/ijms24044030

11. Ferretti J, Köhler W. History of Streptococcal Research. In: Ferretti JJ, Stevens DL, Fischetti VA, editors. Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet]. Oklahoma City (OK): University of Oklahoma Health Sciences Center; 2016 [cited 2026 Apr 9]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK333430/ PubMed PMID: 26866232.

12. Salque M, Bogucki PI, Pyzel J, Sobkowiak-Tabaka I, Grygiel R, Szmyt M, et al. Earliest evidence for cheese making in the sixth millennium bc in northern Europe. Nature. 2013 Jan;493(7433):522–5. doi:10.1038/nature11698

13. Fisberg M, Machado R. History of yogurt and current patterns of consumption. Nutr Rev. 2015 Aug;73(suppl 1):4–7. doi:10.1093/nutrit/nuv020

14. Uriot O, Denis S, Junjua M, Roussel Y, Dary-Mourot A, Blanquet-Diot S. *Streptococcus thermophilus*: From yogurt starter to a new promising probiotic candidate? Journal of Functional Foods. 2017 Oct;37:74–89. doi:10.1016/j.jff.2017.07.038

15. Facklam R. What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. Clin Microbiol Rev. 2002 Oct;15(4):613–30. doi:10.1128/CMR.15.4.613-630.2002

16. Delorme C, Abraham AL, Renault P, Guédon E. Genomics of *Streptococcus salivarius*, a major human commensal. Infection, Genetics and Evolution. 2015 Jul;33:381–92. doi:10.1016/j.meegid.2014.10.001

17. Kohler W. The present state of species within the genera *Streptococcus* and *Enterococcus*. International Journal of Medical Microbiology. 2007 Jun 11;297(3):133–50. doi:10.1016/j.ijmm.2006.11.008

18. Martinović A, Cocuzzi R, Arioli S, Mora D. *Streptococcus thermophilus*: To Survive, or Not to Survive the Gastrointestinal Tract, That Is the Question! Nutrients. 2020 Jul 22;12(8):2175. doi:10.3390/nu12082175

19. Simsek AD, Sezer S, Ozdemir NF, Mehmet H. *Streptococcus vestibularis* bacteremia following dental extraction in a patient on long-term hemodialysis: a case report. Clinical Kidney Journal. 2008 Aug 1;1(4):276–7. doi:10.1093/ndtplus/sfn071

20. Wilson M, Martin R, Walk ST, Young C, Grossman S, McKean EL, et al. Clinical and Laboratory Features of *Streptococcus salivarius* Meningitis: A Case Report and Literature Review. Clinical Medicine & Research. 2012 Feb 1;10(1):15–25. doi:10.3121/cmr.2011.1001

21. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, et al. Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. Nat Biotechnol. 2004 Dec;22(12):1554–8. doi:10.1038/nbt1034

22. Hols P, Hancy F, Fontaine L, Grossiord B, Prozzi D, Leblondbourget N, et al. New insights in the molecular biology and physiology of revealed by comparative genomics. FEMS Microbiology Reviews. 2005 Aug;29(3):435–63. doi:10.1016/j.femsre.2005.04.008

23. Cui Y, Xu T, Qu X, Hu T, Jiang X, Zhao C. New Insights into Various Production Characteristics of *Streptococcus thermophilus* Strains. IJMS. 2016 Oct 12;17(10):1701. doi:10.3390/ijms17101701

24. Iyer R, Tomar SK, Uma Maheswari T, Singh R. *Streptococcus thermophilus* strains: Multifunctional lactic acid bacteria. International Dairy Journal. 2010 Mar;20(3):133–41. doi:10.1016/j.idairyj.2009.10.005

25. Jauneikaite E, Tocheva AS, Jefferies JMC, Gladstone RA, Faust SN, Christodoulides M, et al. Current methods for capsular typing of *Streptococcus pneumoniae*. Journal of Microbiological Methods. 2015 Jun;113:41–9. doi:10.1016/j.mimet.2015.03.006

26. Sitkiewicz I, Hryniewicz W. Pyogenic Streptococci - Danger of Re-Emerging Pathogens. Polish Journal of Microbiology. 2010;59:219–26.

27. Cugini C, Shanmugam M, Landge N, Ramasubbu N. The Role of Exopolysaccharides in Oral Biofilms. J Dent Res. 2019 Jul;98(7):739–45. doi:10.1177/0022034519845001

28. Klein MI, Hwang G, Santos PHS, Campanella OH, Koo H. *Streptococcus mutans*-derived extracellular matrix in cariogenic oral biofilms. Front Cell Infect Microbiol. 2015 Feb 13;5. doi:10.3389/fcimb.2015.00010

29. Li X, Liu Y, Yang X, Li C, Song Z. The Oral Microbiota: Community Composition, Influencing Factors, Pathogenesis, and Interventions. Front Microbiol. 2022 Apr 29;13:895537. doi:10.3389/fmicb.2022.895537

30. Heß N, Waldow F, Kohler TP, Rohde M, Kreikemeyer B, Gómez-Mejia A, et al. Lipoteichoic acid deficiency permits normal growth but impairs virulence of *Streptococcus pneumoniae*. Nat Commun. 2017 Dec 12;8(1):2093. doi:10.1038/s41467-017-01720-z

31. Kim SL, Gordon SM, Shrestha NK. Distribution of streptococcal groups causing infective endocarditis: a descriptive study. Diagnostic Microbiology and Infectious Disease. 2018 Jul;91(3):269–72. doi:10.1016/j.diagmicrobio.2018.02.015

32. Wei Y, Joyce LR, Wall AM, Guan Z, Palmer KL. *Streptococcus pneumoniae*, *S. mitis*, and *S. oralis* Produce a Phosphatidylglycerol-Dependent, *ltaS* -Independent Glycerophosphate-Linked Glycolipid. Ellermeier CD, editor. mSphere. 2021 Feb 24;6(1):e01099-20. doi:10.1128/mSphere.01099-20

33. Castro SA, Dorfmueller HC. A brief review on Group A *Streptococcus* pathogenesis and vaccine development. R Soc open sci. 2021 Mar;8(3):rsos.201991, 201991. doi:10.1098/rsos.201991

34. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. Dairy lactococcal and streptococcal phage–host interactions: an industrial perspective in an evolving phage landscape. FEMS Microbiology Reviews. 2020 Nov 24;44(6):909–32. doi:10.1093/femsre/fuaa048

35. Broadbent JR, McMahon DJ, Welker DL, Oberg CJ, Moineau S. Biochemistry, Genetics, and Applications of Exopolysaccharide Production in *Streptococcus thermophilus*: A Review. Journal of Dairy Science. 2003 Feb;86(2):407–23. doi:10.3168/jds.S0022-0302(03)73619-4

36. Guérin H, Kulakauskas S, Chapot-Chartier MP. Structural variations and roles of rhamnose-rich cell wall polysaccharides in Gram-positive bacteria. Journal of Biological Chemistry. 2022 Oct;298(10):102488. doi:10.1016/j.jbc.2022.102488

37. Lavelle K, Sinderen DV, Mahony J. Cell wall polysaccharides of Gram positive ovococcoid bacteria and their role as bacteriophage receptors. Computational and Structural Biotechnology Journal. 2021;19:4018–31. doi:10.1016/j.csbj.2021.07.011

38. Rush JS, Parajuli P, Ruda A, Li J, Pohane AA, Zamakhaeva S, et al. PplD is a de-N-acetylase of the cell wall linkage unit of streptococcal rhamnopolysaccharides. Nat Commun. 2022 Feb 1;13(1):590. doi:10.1038/s41467-022-28257-0

39. Edgar RJ, Van Hensbergen VP, Ruda A, Turner AG, Deng P, Le Breton Y, et al. Discovery of glycerol phosphate modification on streptococcal rhamnose polysaccharides. Nat Chem Biol. 2019 May;15(5):463–71. doi:10.1038/s41589-019-0251-4

40. Yamashita Y, Tsukioka Y, Tomihisa K, Nakano Y, Koga T. Genes Involved in Cell Wall Localization and Side Chain Formation of Rhamnose-Glucose Polysaccharide in *Streptococcus mutans*. J Bacteriol. 1998 Nov;180(21):5803–7. doi:10.1128/JB.180.21.5803-5807.1998

41. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its Streptococcus thermophilus Host. Alexandre G, editor. Appl Environ Microbiol. 2022 Jan 11;88(1):e01723-21. doi:10.1128/AEM.01723-21

42. Zorzoli A, Meyer BH, Adair E, Torgov VI, Veselovsky VV, Danilov LL, et al. Group A, B, C, and G Streptococcus Lancefield antigen biosynthesis is initiated

by a conserved α-d-GlcNAc-β-1,4-l-rhamnosyltransferase. Journal of Biological Chemistry. 2019 Oct;294(42):15237–56. doi:10.1074/jbc.RA119.009894

43. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in *Streptococcus thermophilus* bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

44. Jespersen MG, Lacey JA, Tong SYC, Davies MR. Global genomic epidemiology of *Streptococcus pyogenes*. Infection, Genetics and Evolution. 2020 Dec;86:104609. doi:10.1016/j.meegid.2020.104609

45. Ya'qoub L, Rehan L, Parikh S, Enriquez J. *Streptococcus Agalactiae* Infective Endocarditis in a Healthy Middle-aged Man: Uncommon but Life-threatening. Cureus. 2018 May 16. doi:10.7759/cureus.2632

46. Wang Z, Enotarpi J, Buffi G, Pezzicoli A, Gstöttner CJ, Nicolardi S, et al. Chemical Synthesis and Immunological Evaluation of Fragments of the Multiantennary Group-Specific Polysaccharide of Group B *Streptococcus*. JACS Au. 2022 Jul 25;2(7):1724–35. doi:10.1021/jacsau.2c00302

47. Nakano K, Ooshima T. Serotype classification of *Streptococcus mutans* and its detection outside the oral cavity. Future Microbiol. 2009 Sep;4(7):891–902. doi:10.2217/fmb.09.64

48. Vollmer W, Massidda O, Tomasz A. The Cell Wall of *Streptococcus pneumoniae*. Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Braunstein M, Rood JI, editors. Microbiol Spectr. 2019 May 31;7(3):7.3.19. doi:10.1128/microbiolspec.GPP3-0018-2018

49. Eletu SD, Sheppard CL, Rose S, Smith K, Andrews N, Lim WS, et al. Re-validation and update of an extended-specificity multiplex assay for detection of

Streptococcus pneumoniae capsular serotype/serogroup-specific antigen and cell-wall polysaccharide in urine specimens. Access Microbiology. 2020 Mar 1;2(3). doi:10.1099/acmi.0.000094

50. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, et al. Pneumococcal Capsules and Their Types: Past, Present, and Future. Clin Microbiol Rev. 2015 Jul;28(3):871–99. doi:10.1128/CMR.00024-15

51. Tsukioka Y, Yamashita Y, Oho T, Nakano Y, Koga T. Biological function of the dTDP-rhamnose synthesis pathway in *Streptococcus mutans*. J Bacteriol. 1997 Feb;179(4):1126–34. doi:10.1128/jb.179.4.1126-1134.1997

52. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, et al. Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. Frasier CM, editor. PLoS Genet. 2006 Mar 10;2(3):e31. doi:10.1371/journal.pgen.0020031

53. Rockett RJ, Oftadeh S, Bachmann NL, Timms VJ, Kong F, Gilbert GL, et al. Genome-wide analysis of *Streptococcus pneumoniae* serogroup 19 in the decade after the introduction of pneumococcal conjugate vaccines in Australia. Sci Rep. 2018 Nov 16;8(1):16969. doi:10.1038/s41598-018-35270-1

54. Shibata Y, Ozaki K, Seki M, Kawato T, Tanaka H, Nakano Y, et al. Analysis of Loci Required for Determination of Serotype Antigenicity in *Streptococcus mutans* and Its Clinical Utilization. J Clin Microbiol. 2003 Sep;41(9):4107–12. doi:10.1128/JCM.41.9.4107-4112.2003

55. Swoboda JG, Campbell J, Meredith TC, Walker S. Wall Teichoic Acid Function, Biosynthesis, and Inhibition. ChemBioChem. 2010 Jan 4;11(1):35–45. doi:10.1002/cbic.200900557

56. van Sorge NM, Cole JN, Kuipers K, Henningham A, Aziz RK, Kasirer-Friede A, et al. The Classical Lancefield Antigen of Group A *Streptococcus* Is a Virulence Determinant with Implications for Vaccine Design. Cell Host & Microbe. 2014 Jun;15(6):729–40. doi:10.1016/j.chom.2014.05.009

57. De A, Liao S, Bitoun JP, Roth R, Beatty WL, Wu H, et al. Deficiency of RgpG Causes Major Defects in Cell Division and Biofilm Formation, and Deficiency of LytR-CpsA-Psr Family Proteins Leads to Accumulation of Cell Wall Antigens in Culture Medium by *Streptococcus mutans*. Nojiri H, editor. Appl Environ Microbiol. 2017 Sep;83(17):e00928-17. doi:10.1128/AEM.00928-17

58. Theodorou I, Courtin P, Palussière S, Kulakauskas S, Bidnenko E, Péchoux C, et al. A dual-chain assembly pathway generates the high structural diversity of cell-wall polysaccharides in *Lactococcus lactis*. Journal of Biological Chemistry. 2019 Nov;294(46):17612–25. doi:10.1074/jbc.RA119.009957

59. Kovacs CJ, Faustoferri RC, Bischer AP, Quivey RG. *Streptococcus mutans* requires mature rhamnose-glucose polysaccharides for proper pathophysiology, morphogenesis and cellular division. Molecular Microbiology. 2019 Sep;112(3):944–59. doi:10.1111/mmi.14330

60. Stefanović C, Hager FF, Schäffer C. LytR-CpsA-Psr Glycopolymer Transferases: Essential Bricks in Gram-Positive Bacterial Cell Wall Assembly. IJMS. 2021 Jan 18;22(2):908. doi:10.3390/ijms22020908

61. Rush JS, Edgar RJ, Deng P, Chen J, Zhu H, Van Sorge NM, et al. The molecular mechanism of N-acetylglucosamine side chain attachment to the Lancefield group A carbohydrate in *Streptococcus pyogenes*. Journal of Biological Chemistry. 2017 Nov;292(47):19441–57. doi:10.1074/jbc.M117.815910

62. Dale JB, Walker MJ. Update on group A streptococcal vaccine development. Current Opinion in Infectious Diseases. 2020 Jun;33(3):244–50. doi:10.1097/QCO.0000000000000644

63. Mangalea MR, Duerkop BA. Fitness Trade-Offs Resulting from Bacteriophage Resistance Potentiate Synergistic Antibacterial Strategies. Ottemann KM, editor. Infect Immun. 2020 Jun 22;88(7):e00926-19. doi:10.1128/IAI.00926-19

64. Van Der Beek SL, Zorzoli A, Çanak E, Chapman RN, Lucas K, Meyer BH, et al. Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification. Molecular Microbiology. 2019 Apr;111(4):951–64. doi:10.1111/mmi.14197

65. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. Bipartite *rgp* Locus Diversity in *Streptococcus thermophilus* Corresponds to Backbone and Side Chain Differences of Its Rhamnose-Containing Cell Wall Polysaccharide. Ercolini D, editor. Appl Environ Microbiol. 2022 Dec 13;88(23):e01504-22. doi:10.1128/aem.01504-22

66. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Molecular Microbiology. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494

67. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

68. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. Cell Wall Glycans Mediate Recognition of the Dairy Bacterium *Streptococcus*

thermophilus by Bacteriophages. Dudley EG, editor. Appl Environ Microbiol. 2018 Dec;84(23):e01847-18. doi:10.1128/AEM.01847-18

69. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011 Apr 1;27(7):1009–10. doi:10.1093/bioinformatics/btr039

70. Zamakhaeva S, Chaton CT, Rush JS, Ajay Castro S, Kenner CW, Yarawsky AE, et al. Modification of cell wall polysaccharide guides cell division in *Streptococcus mutans*. Nat Chem Biol. 2021 Aug;17(8):878–87. doi:10.1038/s41589-021-00803-9

71. Denapaite D, Brückner R, Hakenbeck R, Vollmer W. Biosynthesis of Teichoic Acids in *Streptococcus pneumoniae* and Closely Related Species: Lessons from Genomes. Microbial Drug Resistance. 2012 Jun 1;18(3):344–58. doi:10.1089/mdr.2012.0026

72. Whitfield C, Wear SS, Sande C. Assembly of Bacterial Capsular Polysaccharides and Exopolysaccharides. Annu Rev Microbiol. 2020 Sep 8;74(1):521–43. doi:10.1146/annurev-micro-011420-075607

73. Kouwen RHM, van Sinderen D, McDonnell B, Verloren van Themaat PE, Mahony J. *Streptococcus thermophilus* starter cultures. US11674117B2.

74. Islam ST, Lam JS. Synthesis of bacterial polysaccharides via the Wzx/Wzy-dependent pathway. Can J Microbiol. 2014 Nov;60(11):697–716. doi:10.1139/cjm-2014-0595

75. Percy MG, Gründling A. Lipoteichoic Acid Synthesis and Function in Gram-Positive Bacteria. Annu Rev Microbiol. 2014 Sep 8;68(1):81–100. doi:10.1146/annurev-micro-091213-112949

76. Bui NK, Eberhardt A, Vollmer D, Kern T, Bougault C, Tomasz A, et al. Isolation and analysis of cell wall components from *Streptococcus pneumoniae*. Analytical Biochemistry. 2012 Feb;421(2):657–66. doi:10.1016/j.ab.2011.11.026

77. Damjanovic M, Kharat AS, Eberhardt A, Tomasz A, Vollmer W. The Essential *tacF* Gene Is Responsible for the Choline-Dependent Growth Phenotype of *Streptococcus pneumoniae*. J Bacteriol. 2007 Oct;189(19):7105–11. doi:10.1128/JB.00681-07

78. Skov Sørensen UB, Yao K, Yang Y, Tettelin H, Kilian M. Capsular Polysaccharide Expression in Commensal *Streptococcus* Species: Genetic and Antigenic Similarities to *Streptococcus pneumoniae*. Klugman KP, editor. mBio. 2016 Dec 30;7(6):e01844-16. doi:10.1128/mBio.01844-16

79. Faber EJ, Van Den Haak MJ, Kamerling JP, Vliegenthart JFG. Structure of the exopolysaccharide produced by *Streptococcus thermophilus* S3. Carbohydrate Research. 2001 Mar;331(2):173–82. doi:10.1016/S0008-6215(01)00013-1

80. Nordmark EL, Yang Z, Huttunen E, Widmalm G. Structural Studies of an Exopolysaccharide Produced by *Streptococcus thermophilus* THS. Biomacromolecules. 2005 Jan 1;6(1):105–8. doi:10.1021/bm0496514

81. Padmanabhan A, Shah NP. Structural characterization of exopolysaccharide from *Streptococcus thermophilus* ASCC 1275. Journal of Dairy Science. 2020 Aug;103(8):6830–42. doi:10.3168/jds.2019-17439

82. Säwén E, Huttunen E, Zhang X, Yang Z, Widmalm G. Structural analysis of the exopolysaccharide produced by *Streptococcus thermophilus* ST1 solely by NMR spectroscopy. J Biomol NMR. 2010 Jun;47(2):125–34. doi:10.1007/s10858-010-9413-0

83. Alexandraki V, Kazou M, Blom J, Pot B, Papadimitriou K, Tsakalidou E. Comparative Genomics of *Streptococcus thermophilus* Support Important Traits Concerning the Evolution, Biology and Technological Properties of the Species. Front Microbiol. 2019 Dec 20;10:2916. doi:10.3389/fmicb.2019.02916

84. Cui Y, Jiang X, Hao M, Qu X, Hu T. New advances in exopolysaccharides production of *Streptococcus thermophilus*. Arch Microbiol. 2017 Aug;199(6):799–809. doi:10.1007/s00203-017-1366-1

85. Wu Q, Tun HM, Leung FCC, Shah NP. Genomic insights into high exopolysaccharide-producing dairy starter bacterium *Streptococcus thermophilus* ASCC 1275. Sci Rep. 2014 May 15;4(1):4974. doi:10.1038/srep04974

86. Kong L, Song X, Xia Y, Ai L, Xiong Z. Construction of a CRISPR/nCas9-assisted genome editing system for exopolysaccharide biosynthesis in *Streptococcus thermophilus*. Food Research International. 2022 Aug;158:111550. doi:10.1016/j.foodres.2022.111550

87. Jolly L, Stingele F. Molecular organization and functionality of exopolysaccharide gene clusters in lactic acid bacteria. International Dairy Journal. 2001 Jan;11(9):733–45. doi:10.1016/S0958-6946(01)00117-0

88. Pachekrepapol U, Lucey JA, Gong Y, Naran R, Azadi P. Characterization of the chemical structures and physical properties of exopolysaccharides produced by various *Streptococcus thermophilus* strains. Journal of Dairy Science. 2017 May;100(5):3424–35. doi:10.3168/jds.2016-12125

89. Angelin J, Kavitha M. Exopolysaccharides from probiotic bacteria and their health potential. International Journal of Biological Macromolecules. 2020 Nov;162:853–65. doi:10.1016/j.ijbiomac.2020.06.190

90. Szafrański SP, Winkel A, Stiesch M. The use of bacteriophages to biocontrol oral biofilms. Journal of Biotechnology. 2017 May;250:29–44. doi:10.1016/j.jbiotec.2017.01.002

91. Sewell EW, Brown ED. Taking aim at wall teichoic acid synthesis: new biology and new leads for antibiotics. J Antibiot. 2014 Jan;67(1):43–51. doi:10.1038/ja.2013.100

92. Holland LM, Conlon B, O'Gara JP. Mutation of tagO reveals an essential role for wall teichoic acids in *Staphylococcus epidermidis* biofilm development. Microbiology. 2011 Feb 1;157(2):408–18. doi:10.1099/mic.0.042234-0

93. Riu F, Ruda A, Ibba R, Sestito S, Lupinu I, Piras S, et al. Antibiotics and Carbohydrate-Containing Drugs Targeting Bacterial Cell Envelopes: An Overview. Pharmaceuticals. 2022 Jul 29;15(8):942. doi:10.3390/ph15080942

94. Roach DR, Donovan DM. Antimicrobial bacteriophage-derived proteins and therapeutic applications. Bacteriophage. 2015 Jul 3;5(3):e1062590. doi:10.1080/21597081.2015.1062590

95. Wong KY, Megat Mazhar Khair MH, Song AAL, Masarudin MJ, Chong CM, In LLA, et al. Endolysins against Streptococci as an antibiotic alternative. Front Microbiol. 2022 Aug 2;13:935145. doi:10.3389/fmicb.2022.935145

96. Zhang B, Zhao M, Tian J, Lei L, Huang R. Novel antimicrobial agents targeting the *Streptococcus mutans* biofilms discovery through computer technology. Front Cell Infect Microbiol. 2022 Dec 1;12:1065235. doi:10.3389/fcimb.2022.1065235

Section II

Bacteriophages of the dairy bacterium *Streptococcus thermophilus*

Abstract

Bacterio(phages) that infect *Streptococcus thermophilus* remain a major problem in the dairy fermentation industry. As a result, it is an extensively studied phage–host system among gram-positive bacteria/phages and has generated insights into phage diversity, evolution, and the phage-host interactome. Nevertheless, the continuous identification of novel groups of *S. thermophilus* phage, ongoing evolution of phage populations, and their sustained impact on industrial fermentations highlight the importance of continued evaluation of phage biodiversity in dairy fermentation environments and emphasise the need to further elucidate the mechanisms underpinning phage-host interactions. In this review we discuss current knowledge of the diversity, taxonomy and genomic organisation of *S. thermophilus* phages, the molecular mechanisms governing host recognition and infection, and the bacterial defence systems that protect against phage predation. We further examine phage counter-adaptation strategies, including mechanisms that enable evasion of host-encoded immunity, and consider how these interactions shape the co-evolution of phages and their bacterial hosts.

Introduction

The lactic acid bacterium *Streptococcus thermophilus* is one of the most extensively utilised starter culture-associated bacterial species in the global dairy industry. Owing to its rapid milk acidification abilities, efficient lactose metabolism and favourable technological properties, *Streptococcus thermophilus* is widely employed in the industrial production of fermented dairy products, including yoghurt and numerous cheese varieties, either as a single starter culture or in combination with other lactic acid bacteria such as *Lactobacillus delbrueckii* subsp. *bulgaricus* (1,2). The widespread and repeated industrial application of *S. thermophilus* strains often renders them susceptible to infection by bacteriophages (or phages) within dairy environments. Phages that infect *S. thermophilus* remain among the principal causes of fermentation failure/inconsistency (3–5).

Phages may be introduced into the processing environment through numerous routes, including in the substrate (milk), equipment and processing surfaces and personnel (6). Despite the implementation of extensive phage-monitoring procedures and starter culture rotation strategies, phage infections continue to occur frequently in industrial dairy environments (3,5,7). The continued prevalence of phage-related fermentation failures is largely attributed to the rapid adaptation and evolution of phage populations under the highly selective conditions present within industrial dairy fermentations. Additionally, large-scale production processes and the continual availability of susceptible bacterial hosts facilitate the emergence of mechanisms capable of overcoming host-associated defence systems (5,8,9).

Comparative genomic studies performed in recent decades have revealed extensive diversity among *S. thermophilus* infecting phages and have demonstrated the

ongoing emergence of novel phage groups within dairy fermentations (10–16). At the same time, considerable research efforts have investigated and defined the broad repertoire of antiphage defence systems in *S. thermophilus* strains that contribute to phage resistance and shape phage-host co-evolution within these environments (17–19).

S. thermophilus and its associated phages represent one of the most extensively studied phage-host systems within industrial microbiology. The following review discusses the current understanding of phage-host interaction in *S. thermophilus* including phage diversity, host recognition mechanisms and defence systems.

Current taxonomy of dairy streptococcal phages

Early classification of dairy streptococcal phages identified two distinct groups based on their DNA packaging mechanisms i.e., *cos* (cohesive ends) and *pac* (headful packaging) as well as their structural protein composition/profiles observed in SDS-PAGE profiles (20–23). Following recent taxonomic revision, these historically recognised groups were reclassified as members of the genera *Moineauvirus* and *Brussowvirus*, respectively. Advances in comparative genomics and whole genome sequencing have since revealed that the diversity of dairy streptococcal phages extends beyond these two classical lineages. Over the past two decades, three additional and genetically distinct genera of *S. thermophilus* phages have been identified, designated the *Vansinderenvirus* (formerly 5093) (14), *Piorkowskivirus* (formerly 987) (13,15) and more recently the P738 genera (16). Comparative genomic analyses have revealed that members of the *Vansinderenvirus* and P738 phage genera share sequence similarity with phages infecting pathogenic and non-dairy streptococcal species, while *Piorkowskivirus* group phages appear to be hybrid phages comprising modules associated with both dairy streptococcal and lactococcal phages (5,10,16).

All currently characterised *S. thermophilus* phages possess long non-contractile tails with isometric capsids (4,5). While many dairy streptococcal phages exhibit broadly similar overall morphology, several groups possess distinctive structural features that may aid in their identification (Fig. 1). Members of the *Moineauvirus*, *Brussowvirus* and *Vansinderenvirus* genera generally possess long tails that exceed 200 nm in length, in comparison, representatives of the *Piorkowskivius* and P738 groups exhibit shorter tails ranging from approximately 120–140 nm (11,14–16). Furthermore, the representative phages of the *Piorkowskivirus* genus possess

prominent appendage-like structures at the tail tip resembling those observed among certain lactococcal P335 phages (Fig. 1) while *Vansindenrenvirus* phages exhibit distinctive globular tail-tip structures (Fig. 1) (11,12,14,15).

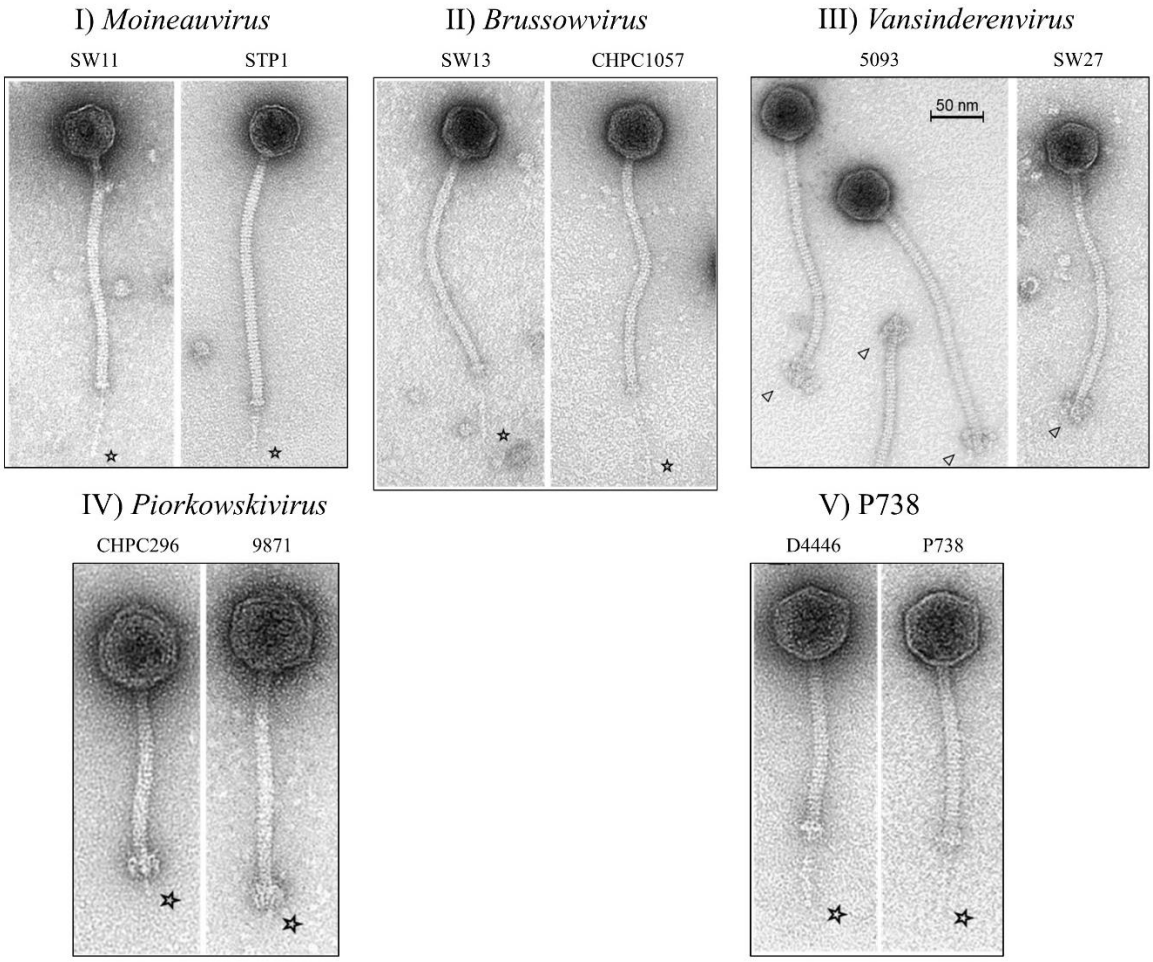


Fig. 1. Adapted from Hanemaaijer et al., 2021 (5). Transmission electron micrographs of negatively stained *S. thermophilus* phages representing the five distinct phage genera (**I**) *Moineauvirus* (phages SW11 (11) and STP1 (7)); (**II**) *Brussowvirus* (phages SW13 (11) and CHPC1057 (12)); (**III**) *Vansinderenvirus* (phages 5093 (14) and SW27 (11)); (**IV**) *Piorkowskivirus* (phages CHPC926 (12) and 9871 (15)); and (**V**) P738 (phages P738 and D4446 (16)). The star symbol indicates the distal ends of the central tail fibre structures (of various lengths). The triangles show the globular appendages at the distal end of the *Vansinderenvirus* phage tails. The (50 nm) scale bar is presented in the III) panel.

To assess the diversity and phylogenetic relationships of currently available dairy streptococcal phages, a phylogenetic analysis was performed using all publicly available *S. thermophilus*-infecting phage genomes available within the NCBI database at the time of analysis (accessed on 11 May 2026). Comparative analysis of 244 dairy streptococcal phage genomes and one *Streptococcus pyogenes* infecting phage genome (T12) was performed by all-against-all, bidirectional BLAST alignment using default parameters (24). Multiple nucleotide sequence alignment was performed using ClustalW v2.1 with default parameters (25). A phylogenetic tree was generated using the neighbour-joining method and bootstrapped with 1,000 replicates. The unrooted phylogenetic tree was visualised using iTOL v7 software (https://itol.embl.de/).

Phylogenetic analysis supported the current classification framework of dairy streptococcal phages and resolved the analysed phages into five major phylogenetic groupings corresponding to the currently recognised genera (Fig. 2). This is consistent with recent analyses of 100 (16) and 182 streptococcal phage genomes (5). Also consistent with previous genomic surveys of *S. thermophilus* phages, the majority of sequenced isolates belong to the *Moineauvirus* (n=157) and *Brussowvirus* (n=51) genera, highlighting the continued dominance of these phage groups within industrial dairy environments that has been repeatedly reported in the literature (7,10,11,26). In contrast, members of the *Piorkowskivirus* (n=14), *Vansinderenvirus* (n=20) and P738 (n=2) groups formed smaller but clearly distinct phylogenetic clusters, reflecting their greater genetic divergence from the historically dominant dairy streptococcal phage lineages. Interestingly, inclusion of the *Streptococcus pyogenes* infecting phage T12 within the comparative analysis demonstrated its close phylogenetic association with the P738 genus, further supporting previous observations that members of this phage group possess

genomic regions sharing similarity with phages infecting non-dairy streptococcal species (16). This observation reinforces the hypothesis that interspecies recombination and module exchange may contribute to the emergence and evolution of novel dairy streptococcal phage lineages.

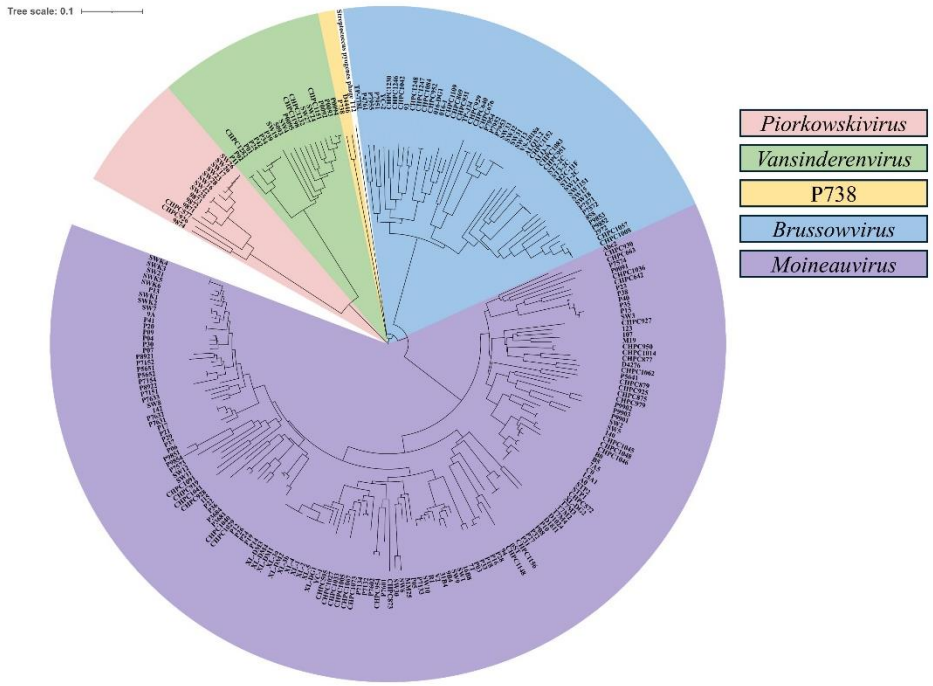


Fig. 2. Phylogenetic tree based on the complete nucleotide sequences of 244 publicly available streptococcal phage genomes alongside one *Streptococcus pyogenes* infecting phage genome retrieved from the NCBI database. Members of the *Moineauvirus* (purple branches), *Brussowvirus* (blue), P738 (yellow), *Vansinderenvirus* (green) and *Piorkowskivirus* (pink) are represented in this analysis.

The extensive diversity exhibited by dairy streptococcal phages is considered to be driven, at least in part, by the modular nature of their genomes and the frequent exchange of genetic material between phage populations (7,11,13,23,27–29). Despite the reported genetic diversity, dairy streptococcal phages generally possess genomes ranging from approximately 30–40 kb that exhibit a modular organisation. These genomes typically comprise discrete functional modules containing genes associated with DNA replication, morphogenesis and lysogeny (Fig. 3) (5).

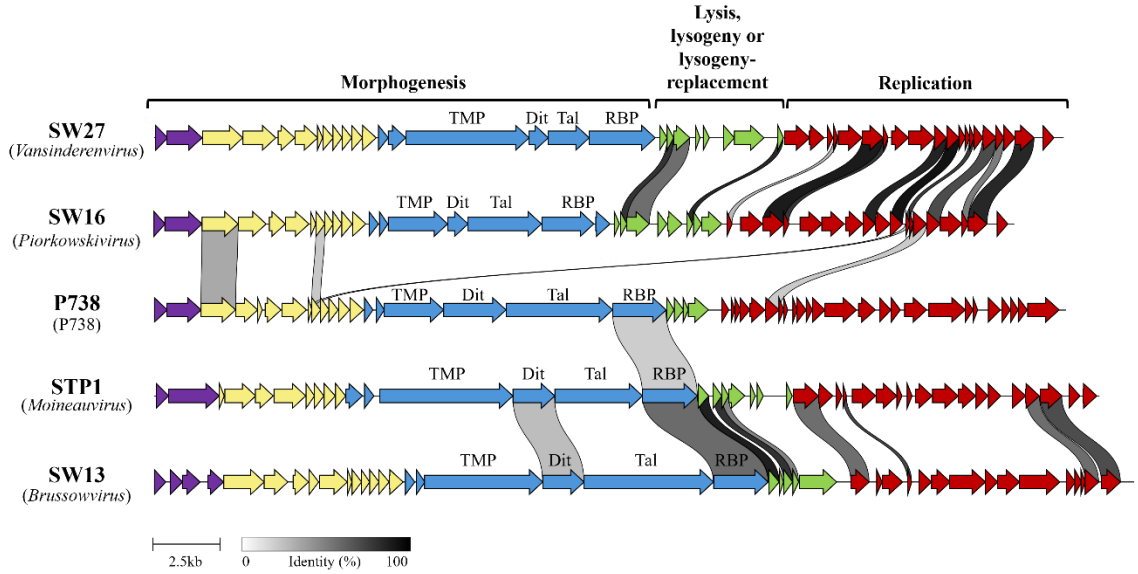


Fig. 3. Adapted from Hanemaaijer et al., 2021 (5). Schematic illustration of the five dairy streptococcal phage genera represented by phage members of each genus i.e., SW27 (*Vansinderenvirus*); SW16 (*Piorkowskivirus*); P738 (P738); STP1 (*Moineauvirus*) and SW13 (*Brussowvirus*). Each arrow represents a gene and are colour-coded to identify the functional module to which they belong, based on the analysis performed by Hanemaaijer et al., 2021 (5). Regions of homology are joined by blocks of different shades of grey to black and generated using CAGECAT (30).

While comparative genomic analyses have provided important insights into the diversity and evolution of dairy streptococcal phages, successful propagation ultimately depends upon the specific outcome of interactions between the phage and its bacterial host (19). Productive infection requires recognition of a susceptible host, adsorption to the bacterial cell surface and delivery of the phage genome into the host cytoplasm. *S. thermophilus* has evolved a diverse repertoire of defence mechanisms that act to prevent phage infection at multiple stages of the phage infection cycle while phages have evolved to evade said mechanisms. Phage-host interactions are therefore shaped by a continual evolutionary arms race that drives adaptation and diversification in both phage and host populations (19).

Phage-host interactions

Successful infection of *S. thermophilus* is dependent on the ability of the phage to recognise and adsorb/bind to specific host-encoded receptor(s) on the bacterial cell surface. This interaction is mediated by the phage-encoded adhesion device (or baseplate complex), a multi-protein complex located at the distal end of the phage tail that is responsible for host recognition and initiation of infection. Structural and bioinformatic analyses of dairy streptococcal phages have revealed that the phage adhesion device typically comprises (part of) the tail tape measure protein (TMP), distal tail protein (Dit), tail-associated lysin (Tal), receptor-binding protein (RBP) and, in certain phages, additional accessory proteins (Fig. 4) (31–34). A feature of several dairy streptococcal phages is the presence of so-called “evolved” Dit and Tal proteins. This term refers to these proteins encompassing additional carbohydrate-binding domain (CBD) extensions. While the RBP is generally regarded as the primary determinant of receptor specificity, the additional CBDs in these additional adhesion device-associated proteins are believed to enhance phage

attachment process by orienting the phage in proximity to its specific receptor, thereby facilitating specific RBP–receptor interactions (34).

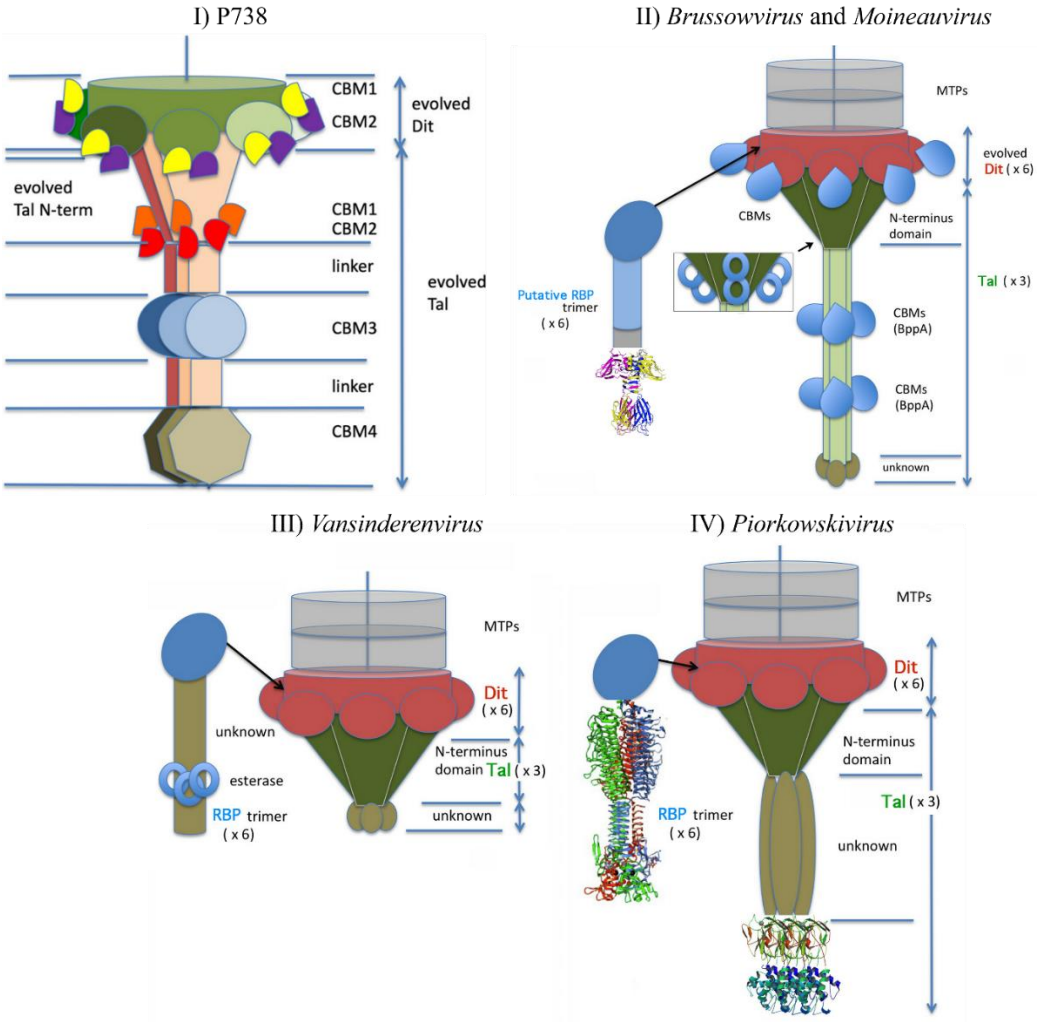


Fig. 4. Adapted from Philippe et al., 2020 (16) and Lavelle et al., 2020 (33). Schematic depictions of the baseplate structures for the five *S. thermophilus* phage genera. I) The identical architecture of the distal tail and Tal regions of phages D4446 and P738 (16). II) *Brussowvirus* and *Moineauvirus*. III) *Vansinderenvirus*. IV) *Piorkowskivirus*.

The current understanding of the *S. thermophilus* infection process suggests that phage adsorption occurs through a coordinated, multistep process involving several components of the adhesion device (19,33,34). The phage infection process commences with the initial contact of the phage to a host-encoded receptor on the bacterial cell surface. Binding of the RBP represents a critical step in host recognition and typically determines whether productive infection can proceed (12,33,35). Following successful attachment, structural rearrangements within the phage baseplate are believed to occur, positioning the virion for genome delivery; in support of this, the STP1 baseplate contains six elongated electron densities in the core of the baseplate that harbour the p2-like (a well characterised *Lactococcus lactis* phage) RBPs at their tip (33). The Tal protein is thought to play a particularly important role during this stage as many contain enzymatic domains that are characteristic of predicted peptidoglycan-degrading activities, suggesting that localised hydrolysis of the bacterial cell wall facilitates penetration of the cell envelope (33,34). Although many aspects of DNA injection remain poorly understood from a mechanistic viewpoint, recent advances in structural biology and in particular, the application of bioinformatic tools such as HHPred (36) and AlphaFold (37) has enabled the identification and structural prediction of adhesion device proteins across numerous phage groups, providing valuable insights into the organisation and function of these complexes (34). Such studies have highlighted the remarkable diversity of phage adhesion devices while simultaneously revealing a conserved overall strategy whereby phages employ specialised tail-tip structures to recognise, adsorb to and ultimately infect their bacterial hosts (33–35). Once injected into the host cytoplasm, phage genomes encounter a wide range of bacterial defence systems that act to prevent phage replication and dissemination. These

primarily intracellular defence mechanisms add an additional barrier to successful infection and play a central role in the evolutionary arms race between *S. thermophilus* and its associated phages.

As mentioned, the interactions between dairy streptococcal phages and their hosts are highly specific, with most phages exhibiting relatively narrow host ranges. This specificity is largely attributed to the diversity observed in the cell surface-associated polysaccharides produced by *S. thermophilus* strains, which have been shown to serve as receptors during the initial stages of phage adsorption (38–41). Experimental studies employing bacteriophage-insensitive mutants (BIMs), whole genome sequencing and genetic complementation have demonstrated that distinct phage groups utilise different saccharide structures during host recognition. Members of the *Brussowvirus* genus have been shown to recognise rhamnose-glucose polysaccharides (RGPs), while phages belonging to the *Piorkowskivirus* genus utilise exopolysaccharides (EPSs) as receptors (38–41). Additionally, correlations between *Moineauvirus* RBP phylogeny and host *eps* genotypes suggests that members of this genus may utilise the EPS as a possible receptor (12). Similarly, the EPS may also serve as the host-encoded receptor for members of the *Vansinderenvirus* genus although this has not been experimentally validated (10,33,42). As the most recently identified phage genus, little is known about the potential receptor for the two members of the P738 genus (16). Despite many advances in investigating the phage-host interactome, the receptor specificities of several dairy streptococcal phage groups remain unresolved and the precise carbohydrate motifs recognised by phage RBPs are unknown or remain to be experimentally validated. Elucidating these interactions will be essential for

understanding the molecular basis of host specificity and predicting the host range of newly emerging dairy streptococcal phages.

Host defence mechanisms

While adsorption and receptor recognition represent critical determinants of phage host range, successful infection ultimately depends on the ability of the phage to overcome the extensive arsenal of antiphage defence systems encoded by the bacterial host. In turn, phages acquire counter-defence strategies that restore their infectivity (19). *S. thermophilus* strains have evolved multiple defence mechanisms to counter phage infection at various stages of the phage infection cycle (Fig. 5). Among these, clustered regularly interspaced short palindromic repeated (CRISPR) and CRISPR-associated (Cas) gene systems as well as restriction-modification (R/M) systems represent the most extensively characterised forms of antiphage immunity and have historically dominated studies of phage resistance in this species (9,43,44). CRISPR-Cas systems provide adaptive immunity by incorporating short DNA fragments from the invading phage genomes and into their CRISPR spacer array, thereby enabling sequence-specific recognition and degradation of the same or closely related genetic elements (phage DNA) during subsequent infections (44). R/M systems, distinguish self from non-self DNA (invading phage DNA) through host-specific methylation patterns and typically target unmethylated foreign DNA for cleavage by restriction endonucleases (45). The widespread occurrence and effectiveness of these systems have made them central to our current understanding of bacterial resistance to phage infection, while also providing naturally occurring mechanisms that can be exploited to improve phage resistance in industrial starter cultures. In particular, exposure of phage-sensitive strains to problematic phages, at an industrial scale, can select for BIMs that have acquired additional CRISPR

spacers and exhibit enhanced phage resistance. As these derivatives typically retain the desirable technological properties of the parental strain and arise through natural adaptation rather than genetic engineering, they have become valuable components of starter culture blends and strain-rotation strategies employed throughout the dairy industry (18,46,47).

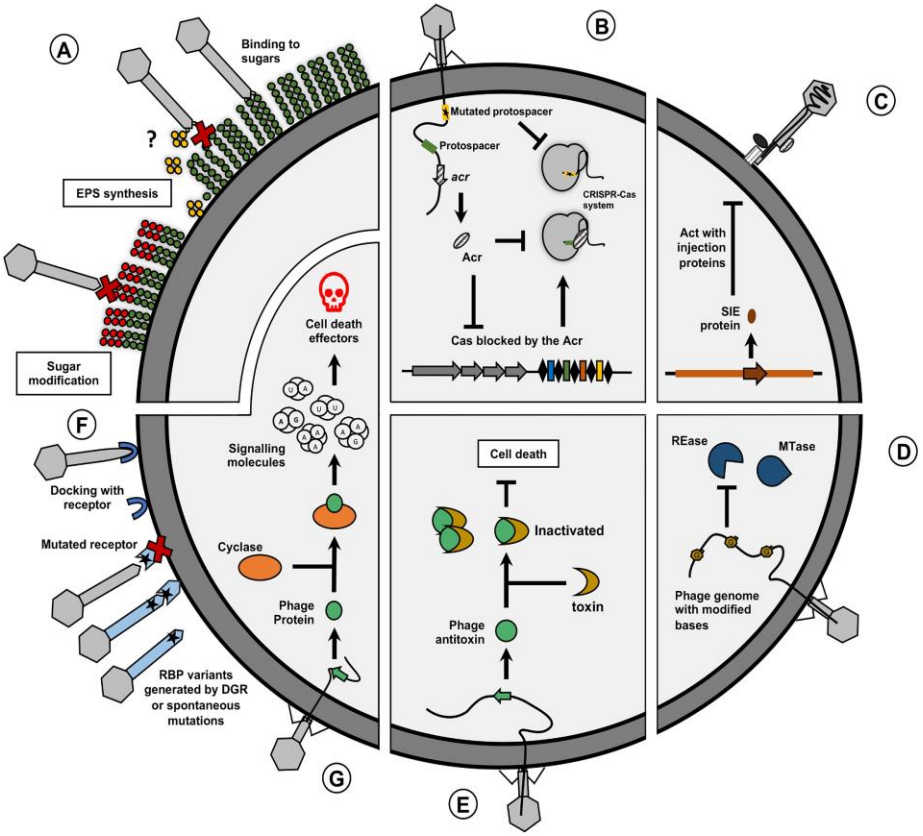


Fig. 5. Sourced from Philippe et al., 2023 (19) - Examples of phage counter defence strategies identified in lactic acid bacteria. **(A)** Modification of cell wall polysaccharides can lead to different host range (38–41). **(B)** Phage Acr proteins block the CRISPR-Cas systems of their host to avoid phage genome targeting (48). **(C)** Some prophages encode sie (superinfection exclusion) proteins that block genome injection into the host cytoplasm (49). **(D)** Bacterial RM systems restrict invading DNA with specific DNA sequences. *S. thermophilus* phages can contain specific methylases in their genome, which allow them to bypass the RM system (45). **(E)** Some phages encode antitoxin genes that interact with the bacterial toxin and impede cell death, which allows for the completion of their cycle (this has not been described in *S. thermophilus*). **(F)** Modified phage receptors prevent phage adsorption. Some phages can still adsorb by acquiring new receptor-binding proteins (RBP) that allow recognition of mutated or new receptors (9). **(G)** In the signal-based CBASS defence mechanism, the cyclase is activated by a phage protein during infection, which generates cyclic-oligonucleotide signalling molecules. Binding of signalling molecule to CBASS effector activates the effector proteins which trigger the death of the phage-infected cell (18,50).

In recent years, the diversity of antiphage systems have been extensively explored in many bacterial species (17,18,51–54). Recent advances in comparative genomics, bioinformatic tools and defence-system mining (through online databases) have revealed that the *S. thermophilus* antiphage repertoire is substantially more diverse than previously appreciated, with 21 accessory defence systems (non-CRISPR and non-R/M defence systems) identified (18). Additionally, a recent analysis of industrial *S. thermophilus* strains identified numerous additional defence systems beyond CRISPR-Cas and R/M, including homologues of AbiD, AbiE, Gao19, Gabija, Hachiman, Kiwa and SoFic, several of which were shown to confer measurable resistance against dairy streptococcal phages (17). This analysis further revealed that defence systems frequently occur within genomic 'Defence Islands', where multiple mechanisms are clustered together within discrete chromosomal regions (17). The co-localisation of these systems is thought to provide enhanced protection against phage attack by creating multiple barriers to infection that operate through distinct molecular mechanisms. Furthermore, it has also been demonstrated that antiphage defence systems can be introduced into industrial *S. thermophilus* strains through natural transformation and stably integrated into the chromosome. The integration of individual defence systems or small defence islands conferred enhanced phage resistance while maintaining growth and milk acidification performance, highlighting the potential of defence-system engineering as a strategy to improve the robustness of industrial starter cultures (18).

The effectiveness of *S. thermophilus* antiphage systems is continually challenged by the ability of phages to evolve mechanisms that can evade host immunity. As a consequence of this long-standing evolutionary arms race, phages have acquired a

diverse repertoire of counter-defence strategies that enable successful infection despite the presence of multiple host-encoded barriers (19). Phages can evade CRISPR-Cas mediated immunity through point mutation, target deletions and the evolution of anti-CRISPR (Acr) proteins (55). Over 100 Acr proteins have been identified in many bacterial species and across diverse CRISPR-Cas system types (56,57). There are many known *S. thermophilus* phage Acrs. These include AcrII3, AcrII5, AcrII6, AcrII24, AcrII25, with different Acr proteins displaying distinct specificities towards the CRISPR1 and CRISPR3 systems and that directly bind and inactivate the various nucleases of the CRISPR-Cas systems (8,48,58,59). Additionally, genomic surveys indicate that Acr proteins are widespread among dairy streptococcal phages, highlighting their importance in phage persistence within industrial fermentation environments (18). Furthermore, many phages infecting *S. thermophilus* encode one or more DNA methyltransferases, which are thought to protect phage genomes from cleavage by host R/M systems through DNA methylation (4,60). The prevalence of both Acr proteins and methyltransferases among dairy streptococcal phages highlights the importance of counter-defence mechanisms in facilitating phage persistence within industrial fermentation environments and suggests that additional, as yet uncharacterised, anti-defence systems remain to be discovered. As additional *S. thermophilus* phage genomes become available for analysis and annotation tools improve, it is likely that additional counter-defence systems will also be discovered, suggesting that this evolutionary arms race is far from over.

Conclusions

Phages infecting *S. thermophilus* remain a persistent challenge within industrial dairy fermentations and continue to represent one of the most extensively studied phage-host systems among lactic acid bacteria. Advances in comparative genomics, structural biology and molecular microbiology, in particular the continued improvement of web-based tools, have substantially improved our understanding of the diversity, evolution and biology of these phages, leading to the recognition of five distinct phage genera and revealing the remarkable genomic plasticity that underpins their adaptation within dairy environments.

Successful infection is dependent upon a series of highly specific interactions between phages and their bacterial hosts. In particular, receptor recognition mediated by the phage adhesion devices has emerged as a major determinant of host range and specificity in *S. thermophilus*. At the same time, *S. thermophilus* possesses a diverse arsenal of antiphage defence systems that can impede the phage infection cycle at multiple stages. While CRISPR-Cas and restriction-modification systems remain the most well characterised forms of immunity, recent studies have highlighted the presence of numerous additional defence mechanisms that collectively contribute to phage resistance. In response, phages have evolved an array of counter-defence strategies, including Acr proteins that enable continued infection despite host immunity.

Despite considerable progress, many aspects of the *S. thermophilus* phage-host interactome remain poorly understood. The receptors utilised by several phage groups have yet to be experimentally validated, and the precise carbohydrate motifs recognised by many receptor-binding proteins remain unresolved. Continued

investigation of these interactions will therefore be essential for understanding phage evolution, predicting host range and developing improved strategies for phage control in industrial fermentations. The following chapters build upon this foundation by investigating the molecular basis of phage-host interactions in *S. thermophilus*, with particular emphasis on the role of cell surface-associated polysaccharides in phage recognition and resistance.

References

1. Iyer R, Tomar SK, Uma Maheswari T, Singh R. *Streptococcus thermophilus* strains: Multifunctional lactic acid bacteria. International Dairy Journal. 2010 Mar;20(3):133–41. doi:10.1016/j.idairyj.2009.10.005

2. Ge Y, Yu X, Zhao X, Liu C, Li T, Mu S, et al. Fermentation characteristics and postacidification of yogurt by *Streptococcus thermophilus* CICC 6038 and Lactobacillus delbrueckii ssp. bulgaricus CICC 6047 at optimal inoculum ratio. Journal of Dairy Science. 2024 Jan;107(1):123–40. doi:10.3168/jds.2023-23817

3. Garneau JE, Moineau S. Bacteriophages of lactic acid bacteria and their impact on milk fermentations. Microb Cell Fact. 2011;10(Suppl 1):S20. doi:10.1186/1475-2859-10-S1-S20

4. Lavelle K, McDonnell B, Fitzgerald G, van Sinderen D, Mahony J. Bacteriophage-host interactions in *Streptococcus thermophilus* and their impact on co-evolutionary processes. FEMS Microbiology Reviews. 2023 Jul 5;47(4):fuad032. doi:10.1093/femsre/fuad032

5. Hanemaaijer L, Kelleher P, Neve H, Franz C, De Waal P, Van Peij N, et al. Biodiversity of Phages Infecting the Dairy Bacterium *Streptococcus thermophilus*. Microorganisms. 2021 Aug 27;9(9):1822. doi:10.3390/microorganisms9091822

6. Marcó MB, Moineau S, Quiberoni A. Bacteriophages and dairy fermentations. Bacteriophage. 2012 Jul;2(3):149–58. doi:10.4161/bact.21868

7. Lavelle K, Murphy J, Fitzgerald B, Lugli GA, Zomer A, Neve H, et al. A Decade of *Streptococcus thermophilus* Phage Evolution in an Irish Dairy Plant.

Vieille C, editor. Appl Environ Microbiol. 2018 May 15;84(10):e02855-17. doi:10.1128/AEM.02855-17

8. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018 Jul 25;9(1):2919. doi:10.1038/s41467-018-05092-w

9. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, et al. Phage Response to CRISPR-Encoded Resistance in *Streptococcus thermophilus*. J Bacteriol. 2008 Feb 15;190(4):1390–400. doi:10.1128/JB.01412-07

10. McDonnell B, Mahony J, Hanemaaijer L, Neve H, Noben JP, Lugli GA, et al. Global Survey and Genome Exploration of Bacteriophages Infecting the Lactic Acid Bacterium *Streptococcus thermophilus*. Front Microbiol. 2017 Sep 12;8:1754. doi:10.3389/fmicb.2017.01754

11. Lavelle K, Martinez I, Neve H, Lugli GA, Franz CMAP, Ventura M, et al. Biodiversity of *Streptococcus thermophilus* Phages in Global Dairy Fermentations. Viruses. 2018 Oct 22;10(10):577. doi:10.3390/v10100577

12. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in *Streptococcus thermophilus* bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

13. Szymczak P, Janzen T, Neves AR, Kot W, Hansen LH, Lametsch R, et al. Novel Variants of *Streptococcus thermophilus* Bacteriophages Are Indicative of Genetic Recombination among Phages from Different Bacterial Species. Dudley

EG, editor. Appl Environ Microbiol. 2017 Mar;83(5):e02748-16. doi:10.1128/AEM.02748-16

14. Mills S, Griffin C, O'Sullivan O, Coffey A, McAuliffe OE, Meijer WC, et al. A new phage on the 'Mozzarella' block: Bacteriophage 5093 shares a low level of homology with other *Streptococcus thermophilus* phages. International Dairy Journal. 2011 Dec;21(12):963–9. doi:10.1016/j.idairyj.2011.06.003

15. McDonnell B, Mahony J, Neve H, Hanemaaijer L, Noben JP, Kouwen T, et al. Identification and Analysis of a Novel Group of Bacteriophages Infecting the Lactic Acid Bacterium *Streptococcus thermophilus*. Pettinari MJ, editor. Appl Environ Microbiol. 2016 Sep;82(17):5153–65. doi:10.1128/AEM.00835-16

16. Philippe C, Levesque S, Dion MB, Tremblay DM, Horvath P, Lüth N, et al. Novel Genus of Phages Infecting *Streptococcus thermophilus*: Genomic and Morphological Characterization. Ercolini D, editor. Appl Environ Microbiol. 2020 Jun 17;86(13):e00227-20. doi:10.1128/AEM.00227-20

17. Kelleher P, Ortiz Charneco G, Kampff Z, Diaz-Garrido N, Bottacini F, McDonnell B, et al. Phage defence loci of *Streptococcus thermophilus*— tip of the anti-phage iceberg? Nucleic Acids Research. 2024 Oct 28;52(19):11853–69. doi:10.1093/nar/gkae814

18. Leprince A, Lefrançois J, Millen AM, Magill D, Horvath P, Romero DA, et al. Strengthening phage resistance of *Streptococcus thermophilus* by leveraging complementary defense systems. Nat Commun. 2025 Aug 4;16(1):7142. doi:10.1038/s41467-025-62408-3

19. Philippe C, Cornuault JK, de Melo AG, Morin-Pelchat R, Jolicoeur AP, Moineau S. The never-ending battle between lactic acid bacteria and their phages.

FEMS Microbiology Reviews. 2023 Jul 5;47(4):fuad035. doi:10.1093/femsre/fuad035

20. Le Marrec C, Van Sinderen D, Walsh L, Stanley E, Vlegels E, Moineau S, et al. Two groups of bacteriophages infecting *Streptococcus thermophilus* can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. Appl Environ Microbiol. 1997 Aug;63(8):3246–53. doi:10.1128/aem.63.8.3246-3253.1997

21. Desiere F, Lucchini S, Brüssow H. Comparative Sequence Analysis of the DNA Packaging, Head, and Tail Morphogenesis Modules in the Temperate cos-Site *Streptococcus thermophilus* Bacteriophage Sfi21. Virology. 1999 Aug;260(2):244–53. doi:10.1006/viro.1999.9830

22. Lucchini S, Desiere F, Brüssow H. The Genetic Relationship between Virulent and Temperate *Streptococcus thermophilus* Bacteriophages: Whole Genome Comparison of cos-Site Phages Sfi19 and Sfi21. Virology. 1999 Aug;260(2):232–43. doi:10.1006/viro.1999.9814

23. Lucchini S, Desiere F, Brüssow H. Comparative Genomics of *Streptococcus thermophilus* Phage Species Supports a Modular Evolution Theory. J Virol. 1999 Oct;73(10):8647–56. doi:10.1128/JVI.73.10.8647-8656.1999

24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2

25. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 1994;22(22):4673–80. doi:10.1093/nar/22.22.4673

26. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. Dairy lactococcal and streptococcal phage–host interactions: an industrial perspective in an evolving phage landscape. FEMS Microbiology Reviews. 2020 Nov 24;44(6):909–32. doi:10.1093/femsre/fuaa048

27. Neve H, Zenz KI, Desiere F, Koch A, Heller KJ, Brüssow H. Comparison of the Lysogeny Modules from the Temperate *Streptococcus thermophilus* Bacteriophages TP-J34 and Sfi21: Implications for the Modular Theory of Phage Evolution. Virology. 1998 Feb;241(1):61–72. doi:10.1006/viro.1997.8960

28. Brüssow H, Canchaya C, Hardt WD. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. Microbiol Mol Biol Rev. 2004 Sep;68(3):560–602. doi:10.1128/MMBR.68.3.560-602.2004

29. Brüssow H, Desiere F. Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. Molecular Microbiology. 2001 Jan;39(2):213–23. doi:10.1046/j.1365-2958.2001.02228.x

30. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Robinson P, editor. Bioinformatics. 2021 Aug 25;37(16):2473–5. doi:10.1093/bioinformatics/btab007

31. Duplessis M, Moineau S. Identification of a genetic determinant responsible for host specificity in *Streptococcus thermophilus* bacteriophages. Molecular Microbiology. 2001 Jul;41(2):325–36. doi:10.1046/j.1365-2958.2001.02521.x

32. Duplessis M, Lévesque CM, Moineau S. Characterization of *Streptococcus thermophilus* Host Range Phage Mutants. Appl Environ Microbiol. 2006 Apr;72(4):3036–41. doi:10.1128/AEM.72.4.3036-3041.2006

33. Lavelle K, Goulet A, McDonnell B, Spinelli S, Van Sinderen D, Mahony J, et al. Revisiting the host adhesion determinants of *Streptococcus thermophilus* siphophages. Microbial Biotechnology. 2020 Nov;13(6):1765–79. doi:10.1111/1751-7915.13593

34. Goulet A, Joos R, Lavelle K, Van Sinderen D, Mahony J, Cambillau C. A structural discovery journey of streptococcal phages adhesion devices by AlphaFold2. Front Mol Biosci. 2022 Aug 19;9:960325. doi:10.3389/fmolb.2022.960325

35. Leprince A, Mahillon J. Phage Adsorption to Gram-Positive Bacteria. Viruses. 2023 Jan 10;15(1):196. doi:10.3390/v15010196

36. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology. 2018 Jul;430(15):2237–43. doi:10.1016/j.jmb.2017.12.007

37. Broendum SS, Williams DE, Hayes BK, Kraus F, Fodor J, Clifton BE, et al. High avidity drives the interaction between the streptococcal C1 phage endolysin, PlyC, with the cell surface carbohydrates of Group A *Streptococcus*. Molecular Microbiology. 2021 Aug;116(2):397–415. doi:10.1111/mmi.14719

38. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. Cell Wall Glycans Mediate Recognition of the Dairy Bacterium *Streptococcus*

thermophilus by Bacteriophages. Dudley EG, editor. Appl Environ Microbiol. 2018 Dec;84(23):e01847-18. doi:10.1128/AEM.01847-18

39. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Molecular Microbiology. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494

40. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Alexandre G, editor. Appl Environ Microbiol. 2022 Jan 11;88(1):e01723-21. doi:10.1128/AEM.01723-21

41. Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, Van Sinderen D, Mahony J. The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Elkins CA, editor. Appl Environ Microbiol. 2025 Oct 22;91(10):e01238-25. doi:10.1128/aem.01238-25

42. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

43. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007 Mar 23;315(5819):1709–12. doi:10.1126/science.1138140

44. Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in

Streptococcus thermophilus. J Bacteriol. 2008 Feb 15;190(4):1401–12. doi:10.1128/JB.01415-07

45. Dupuis MÈ, Villion M, Magadán AH, Moineau S. CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. Nat Commun. 2013 Jul 2;4(1):2087. doi:10.1038/ncomms3087

46. Chirico D, Gorla A, Verga V, Pedersen PD, Polgatti E, Cava A, et al. Bacteriophage-insensitive mutants for high quality Crescenza manufacture. Front Microbiol. 2014 May 6;5. doi:10.3389/fmicb.2014.00201

47. Achigar R, Magadán AH, Tremblay DM, Julia Pianzzola M, Moineau S. Phage-host interactions in *Streptococcus thermophilus*: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Sci Rep. 2017 Mar 6;7(1):43438. doi:10.1038/srep43438

48. Pastuszka A, Rousseau GM, Somerville V, Levesque S, Fiset JP, Goulet A, et al. Dairy phages escape CRISPR defence of *Streptococcus thermophilus* via the anti-CRISPR AcrIIA3. International Journal of Food Microbiology. 2023 Dec;407:110414. doi:10.1016/j.ijfoodmicro.2023.110414

49. Ali Y, Koberg S, HeÃŸner S, Sun X, Rabe B, Back A, et al. Temperate *Streptococcus thermophilus* phages expressing superinfection exclusion proteins of the Ltp type. Front Microbiol. 2014 Mar 13;5. doi:10.3389/fmicb.2014.00098

50. Huiting E, Cao X, Ren J, Athukoralage JS, Luo Z, Silas S, et al. Bacteriophages inhibit and evade cGAS-like immune function in bacteria. Cell. 2023 Feb;186(4):864-876.e21. doi:10.1016/j.cell.2022.12.041

51. Millman A, Melamed S, Leavitt A, Doron S, Bernheim A, Hör J, et al. An expanded arsenal of immune systems that protect bacteria from phages. Cell Host & Microbe. 2022 Nov;30(11):1556-1569.e5. doi:10.1016/j.chom.2022.09.017

52. Grafakou A, Mosterd C, Beck MH, Kelleher P, McDonnell B, de Waal PP, et al. Discovery of antiphage systems in the lactococcal plasmidome. Nucleic Acids Research. 2024 Sep 9;52(16):9760–76. doi:10.1093/nar/gkae671

53. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018 Mar 2;359(6379):eaar4120. doi:10.1126/science.aar4120

54. Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. Nat Rev Microbiol. 2023 Oct;21(10):686–700. doi:10.1038/s41579-023-00934-x

55. Stanley SY, Maxwell KL. Phage-Encoded Anti-CRISPR Defenses. Annu Rev Genet. 2018 Nov 23;52(1):445–64. doi:10.1146/annurev-genet-120417-031321

56. Davidson AR, Lu WT, Stanley SY, Wang J, Mejdani M, Trost CN, et al. Anti-CRISPRs: Protein Inhibitors of CRISPR-Cas Systems. Annual Review of Biochemistry. 2020 Jun 20;89(1):309–32. doi:10.1146/annurev-biochem-011420-111224

57. Bondy-Denomy J, Davidson AR, Doudna JA, Fineran PC, Maxwell KL, Moineau S, et al. A Unified Resource for Tracking Anti-CRISPR Names. The CRISPR Journal. 2018 Oct 1;1(5):304–5. doi:10.1089/crispr.2018.0043

58. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits *Streptococcus pyogenes* Cas9. Nat Microbiol. 2017 Aug 7;2(10):1374–80. doi:10.1038/s41564-017-0004-7

59. Song G, Zhang F, Tian C, Gao X, Zhu X, Fan D, et al. Discovery of potent and versatile CRISPR–Cas9 inhibitors engineered for chemically controllable genome editing. Nucleic Acids Research. 2022 Mar 21;50(5):2836–53. doi:10.1093/nar/gkac099

60. Takahashi M, Hiraoka S, Matsumoto Y, Shibagaki R, Ujihara T, Maeda H, et al. Host-encoded DNA methyltransferases modify the epigenome and host tropism of invading phages. iScience. 2025 Apr;28(4):112264. doi:10.1016/j.isci.2025.112264

Chapter II

The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738

This chapter has been published as: The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Applied and Environmental Microbiology. 2025. 91:e01238-25. **Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, van Sinderen D, Mahony J.**

Contributions: Irina Sadovskaya and Evgeny Vinogradov: Biochemical analysis and NMR spectroscopy.

Abstract

Streptococcus thermophilus is a key lactic acid bacterial species extensively used in dairy fermentations, where bacteriophage (phage) infections pose a significant threat to production efficiency. While the cell surface receptors for four of the five currently known streptococcal phage genera have previously been identified, the specific host-encoded receptor for the P738 phage group has not been elucidated to date. In this study, P738 bacteriophage-insensitive mutants of *S. thermophilus* UCCSt50 were isolated, and whole genome sequencing revealed mutations in a glycosyltransferase-encoding gene within the *rgp* locus, responsible for the biosynthesis of the cell wall-associated rhamnose-glucose polysaccharide (RGP). Compositional and methylation analysis demonstrated that loss of a terminal rhamnose residue from the RGP side chain underpinned the observed resistance to P738 binding and infection. Fluorescent binding assays validated the biological functionality of the predicted receptor binding protein of P738, confirming the critical role of the RGP side chain in phage adsorption. These findings reveal a strict requirement for a complete tetrasaccharide side chain structure for P738 binding, contrasting with the less stringent requirements observed for the *Brussowvirus* SW13. This work offers valuable insights to guide the development of phage-robust starter cultures through knowledge-based strain selection and rotation strategies, thereby supporting more sustainable and reliable dairy fermentation processes.

Importance

Streptococcus thermophilus is one of the most extensively applied members of the lactic acid bacteria, being extensively used as a bacterial starter culture in the production of fermented foods, including cheeses and yoghurt. Bacteriophage infections present a significant threat to industrial dairy fermentations, which can lead to substandard production processes and product quality. Understanding phage-host interactions, which commence with the recognition and binding of a given bacteriophage to a particular host-encoded, cell surface-associated receptor, is essential for developing strategies to minimise phage-related risks in dairy fermentations. This study identifies the specific cell surface polysaccharide structure produced by *Streptococcus thermophilus* as the receptor required for infection by phage P738, bridging a key knowledge gap in phage-host interactions in this important industrial bacterial species. By elucidating how phage P738 recognises and binds to its host, this work offers valuable information to aid strain selection and the development of phage-robust starter cultures.

Introduction

Streptococcus thermophilus is a member of the lactic acid bacteria (LAB) that has long been associated with the production of fermented dairy products (1). Based on its long history of safe use in food fermentations and human consumption, *S. thermophilus* is the only streptococcal species to be granted "generally recognized as safe" status by the Food and Drug Administration and "qualified presumption of safety" status by the European Food Safety Authority (2). *S. thermophilus* is commercially one of the most valuable LAB globally, as strains of this species are used extensively in bacterial starter cultures in the production of fermented foods. From an applied perspective, certain metabolic products of *S. thermophilus,* particularly exopolysaccharides (EPSs), contribute to desirable textures, mouthfeel, and flavour of yoghurts and hard cheeses (3,4).

One of the most significant and consistent threats to these fermentation processes is infection of starter bacteria by bacterial viruses, commonly called bacteriophages (or phages). Phage infection of strains within starter cultures may disrupt growth and milk acidification rates, leading to slower or inconsistent production processes or, in severe cases, failed fermentations (5). The persistence of phages in the dairy industry negatively impacts production efficiency and product quality with serious economic consequences. Historically, two groups of dairy streptococcal phages were identified based on their mode of DNA packaging, as well as the number of major structural proteins (6). These groups are now termed the Moineauviruses (originally termed the *cos* phages) and Brussowviruses (originally termed the *pac* phages) (7). Subsequently, three additional and genetically distinct phage groups were identified, termed the *Vansinderenvirus* (8), *Piorkowskivirus* (formerly termed the 987 group) (9,10), and P738 (11) groups. Phages belonging to the

Vansinderenvirus and P738 genera exhibit sequence similarity to phages of pathogenic and non-dairy streptococci, while genomes representative of *Piorkowskivirus* group phages appear to be hybrids of dairy streptococcal and lactococcal phage genome modules (12,13).

Considerable research efforts have been invested in defining the mechanisms by which *S. thermophilus* responds to the challenge of phage infection. Most notable among these are studies pertaining to clustered regularly interspaced short palindromic repeats (CRISPR) systems in *S. thermophilus* (14–17). The CRISPR-Cas systems in *S. thermophilus* provide adaptive immunity against bacteriophages by incorporating short DNA segments from invading phages into the CRISPR spacer arrays, enabling the bacterium to recognise and defend against subsequent infections (16). Beyond these systems, receptor modification has been shown to play a crucial role in phage resistance, in addition to other strategies (18,19). The host-encoded receptors for a handful of dairy streptococcal phages have been identified as saccharidic moieties, which are presented on the surface of the host cell as tightly or loosely cell-associated structures and which correspond to cell wall rhamnose-glucose polysaccharides (RGPs) and exopolysaccharides, respectively (20–23). Biosynthesis of these complex glycans is performed by an enzymatic machinery encoded by two distinct loci termed the *rgp* and *eps* gene clusters, respectively, which are located in the chromosomes of *S. thermophilus* strains. Members of the *Piorkowskivirus* phage group recognise an EPS receptor (20,21), while the RGP is essential for *Brussowvirus* infection (23). Furthermore, a potential link has been proposed between *Moineauvirus* phylogeny and host *eps* genotypes, implicating EPS as a possible receptor for members of this group, though this has not been experimentally validated to date (22). Similarly, the EPS may serve as the

receptor for members of the *Vansinderenvirus* phage group, based on host range data and receptor binding protein (RBP) analyses (12,24,25). Collectively, these findings emphasize the essential role of *S. thermophilus* cell wall polysaccharides (CWPSs) in mediating phage interactions. Isolation of non-CRISPR-mediated bacteriophage-insensitive mutants (BIMs) has been instrumental in identifying host-encoded phage receptors, further informing the selection and development of robust, phage-resistant *S. thermophilus* strains for industrial applications (26).

As mentioned above, the *rgp* locus encodes the biosynthetic machinery for RGP production. The RGP molecule of a given *S. thermophilus* strain is composed of two distinct saccharidic components: a conserved rhamnan-rich backbone and a variable side chain. The backbone structure is believed to be embedded in and covalently attached to the peptidoglycan layer, while the side chain structure is covalently linked to the backbone structure and exposed at the cell surface, where it may become the specific "docking" point for phages with a cognate receptor binding protein (27). The variable side chain may differ in subunit composition, length, and it is this unique composition and structure that dictate the highly specific interactions of certain dairy streptococcal phages and their hosts. This has been proven experimentally, where the side chain structure is required for host binding by a member of the *Brussowvirus* phage group (23). Synthesis of the rhamnan backbone structure commences with the transfer of a GlcNAc moiety from UDP-GlcNAc to Und-P to form the precursor Und-P-P-GlcNAc, which is mediated by the RgpG/TagO equivalent encoded by a glycosyltransferase encoded beyond the *rgp* gene cluster (28). Subsequently, RgpA transfers the first rhamnose residue to the Und-P-P-GlcNAc foundation, and the extension of the backbone structure is a function of the rhamnosyltransferases/glycosyltransferases encoded within the 3′

end of the *rgp* gene cluster. The rhamnan subunits are transferred to the outer face of the cell membrane through the activity of the ABC-transport proteins RgpC/D. The side chain structure is synthesized independently and initiated by the transfer of GlcNAc/GalNAc moiety to the lipid carrier Und-P by the priming glycosyltransferase. The glycosyltransferases encoded within the 5′ region of the *rgp* gene cluster contribute to the extension of the side chain subunit structure, and the subunits are transferred to the outer face of the membrane by a flippase also encoded in this region of the cluster. Hierarchical clustering analysis of the *rgp* gene clusters of 78 *S. thermophilus* strains revealed considerable genetic diversity of *rgp* loci at the strain level, enabling the identification of seven distinct *rgp* genotypes (termed *rgp*1–7). Additionally, *S. thermophilus* strains can be classified based on the gene content associated with the biosynthesis of the rhamnan backbone (Bt) and variable side chain (Vt) (28). A detailed analysis of the rhamnan backbone and side chain-associated regions of *rgp* loci has so far revealed three distinct backbone genotypes (Bt1-3) and five variable side chain genotypes (Vt1-5), each corresponding to distinct RGP chemical structures (28).

To date, the receptors recognised by members of four of the five genera of dairy streptococcal phages have been identified. Currently, only two members of the recently described P738 genus have been isolated and characterised, namely P738 and D4446, which were isolated in Germany and Senegal, respectively (11). Despite the relatively recent identification of these phages, their isolation from two distant geographical locations and high level of sequence relatedness suggest that these phages are likely to be present in other factories and may remain undetected due to the current lack of tools to identify them. It is also likely that these phages will increase in abundance if suitable host strains continue to be applied in industrial

fermentations. Therefore, in the current study, we sought to identify the specific host-encoded receptor of the P738 group through analysis of bacteriophage-insensitive mutants of the host strain of phage P738. Whole genome sequencing and single nucleotide polymorphism (SNP) analysis of two such BIMs were applied to identify the genes responsible for the observed resistance. Additionally, the structure of the RGP moiety of a representative mutant was determined and compared to the parent strain to establish the impact of the mutation on the host cell wall RGP structure. Finally, the biological functionality of the predicted RBP of P738 was validated through fluorescent binding assays, while these assays also confirmed that the adsorption step of the phage cycle was impaired in the generated BIMs. This study provides the first molecular-level insights into P738 phage-host interactions, characterising the genetic and structural basis of phage receptor binding.

Results

P738-insensitive derivatives of *S. thermophilus* UCCSt50 are non-CRISPR mediated

The host-encoded receptors of members of four of the five known genera of dairy streptococcal phages have been defined to date (20–23). In the present study, we aimed to identify the receptor for the fifth genus of dairy streptococcal phages, i.e., the P738 genus (11). Following exposure of the host strain, *S. thermophilus* UCCSt50, to phage P738 (10^8 PFU/mL, multiplicity of infection ~0.5–1), a panel of 20 individual survivors that were presumed to be BIMs was colony-purified for further characterisation. Presumptive P738-resistant BIMs were generated at a frequency of ~10^{-5}. All isolates selected for analysis retained an incomplete or complete phage-resistance phenotype after colony purification and overnight incubation in LM17 broth. Full resistance of a BIM was defined as the complete absence of plaques following exposure to phage P738 (at a level of 10^8 PFU/mL, obtained for parent strain UCCSt50), while incomplete resistance was characterised by a significant reduction in phage susceptibility, where BIMs exhibited more than a five-log reduction in phage titre compared to the wild-type strain (Supplementary Table S1). To validate the integrity of the BIMs as derivatives of *S. thermophilus* UCCSt50, *rgp* genotyping using the recently described dual multiplex PCR system was performed on each isolate as well as the parent strain. Furthermore, the CRISPR1 and CRISPR3 loci and spacer arrays of the isolates were also amplified and compared to those of the parent strain. In all cases, the PCR-generated DNA fragments were shown to be identical in size to those of the parent strain. Among the panel of 20 strains, two BIMS named F1B and F2A were selected for whole

genome sequencing to determine the underlying genetic alterations that mediate the observed P738 phage resistance.

To investigate potential CRISPR-mediated resistance mechanisms, the genomes of UCCSt50 and the selected BIMs were sequenced to completion with whole genome sequencing performed using long read sequencing technology paired with Illumina short-read sequencing (a summary of the general genome characteristics is presented in Supplementary Table S2). To establish if CRISPR spacer acquisitions underpinned the observed resistance in one or both strains, the genome sequences of BIMs F1B and F2A were analysed using CRISPRCasFinder (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index) to compare the composition of their spacer arrays against that of the parent strain and were observed to be identical (Supplementary Table S3).

P738 phage resistance is mediated by a mutation in a gene within the *rgp* cluster

To establish which genes were responsible for the observed P738-resistant phenotype, SNP analysis was performed. This revealed distinct single point mutations in a gene within the *rgp* gene cluster for both BIMs, while F2A also exhibited an additional mutation (Table 1).

Table 1. Summary of SNPs identified in derived BIMs of *S. thermophilus* UCCSt50

BIM	SNP position	Base modification	Amino acid changes	*orf* no.	Predicted function
F1B	1,344,528	g → a	Trp^{269} → stop	*orf*07020	Glycosyltransferase
	1,371,556	c → a	Ser^{28} → stop	*orf*07170	DMT family transporter
F2A	1,230,230	c → t	Pro^{52} → Ser^{52}	*orf*06475	UDP-glucose-4-epimerase
	1,345,056	a → c	Asp^{93} → Ala^{93}	*orf*07020	Glycosyltransferase

Both BIMs were found to harbour a distinct mutation within *orf*07020, a gene encoding a predicted glycosyltransferase within the *rgp* cluster of *S. thermophilus* UCCSt50, which is presumed to be involved in RGP side chain biosynthesis (28). This gene corresponds to *orf*06960 in a previous genome annotation and is located immediately upstream of *orf*06955, which has been shown to be essential for a *Brussowvirus*-type phage infection (23). In the case of F1B, the incorporated mutation introduces a stop codon in place of a tryptophan (Trp269) residue, resulting in the truncation of the encoded protein (269 aa instead of 322 aa). For the derivative F2A, the mutation within *orf*07020 culminates in an amino acid substitution (aspartic acid to alanine). Additionally, F1B harbours a mutation in *orf*07170, a gene predicted to encode a DMT family transporter. F2A also contains a mutation in *orf*06475, which is annotated as *galE*, and predicted to encode a UDP-glucose-4-epimerase, which is involved in the interconversion of UDP-galactose to UDP-glucose. Given that *orf*07020 was mutated in both BIMs and since low concentrations of UDP-Glucose and UDP-Galactose are associated with exopolysaccharide production in *Lactococcus* (29), the functional role of this gene in mediating P738 phage resistance was also evaluated.

To establish if the identified mutations within *orf*07020$_{UCCSt50}$ were responsible for the acquired P738 phage resistance phenotype observed in the BIMs, the native *orf*07020$_{UCCSt50}$ was cloned into the high copy expression vector pNZ44 (30) and introduced into BIM F1B and F2A (designated F1B::pNZ44-07020$_{UCCSt50}$ and F2A::pNZ44-07020$_{UCCSt50}$, respectively; Table 2). The introduction of the native gene resulted in the restoration of phage sensitivity to the same order of magnitude as that of UCCSt50 for F2A (efficiency of plaquing [EOP] = 0.9 ± 0.1) and F1B (EOP = 0.35 ± 0.33) (Table 3), confirming the role of *orf*07020$_{UCCSt50}$ in the P738 infection process. Furthermore, F1B harbouring the control plasmid (F1B::pNZ44) and a plasmid-cured derivative (F1B::pNZ44-07020C$_{UCCSt50}$) exhibited complete phage resistance (Table 3). To evaluate the possible role of *galE* in the observed phage resistance of BIM F2A, the native *orf*06475$_{UCCSt50}$ was cloned and transformed into BIM F2A, but its presence was not associated with restoration of phage sensitivity, indicating that it does not contribute to P738 phage resistance. These findings reinforce the involvement of *orf*07020$_{UCCSt50}$ in P738 phage resistance in both BIMs.

Table 2. Bacterial strains, bacteriophages, plasmids, and primers used in this study

Strains		
	Description	**Source**
S. thermophilus strains		
UCCSt50	Wild-type strain sensitive to phage P738/host strain for P738	(23)
UCCSt50::pNZ44	Wild-type strain harbouring the control/cloning vector pNZ44	This study
F1B	Phage insensitive derivative of UCCSt50/BIM of UCCSt50 generated against phage P738	This study
F2A	Phage-insensitive derivative of UCCSt50/BIM of UCCSt50 generated against phage P738	This study
F1B::pNZ44	F1B harbouring the control plasmid, pNZ44	This study
F1B_pNZ44-07020$_{UCCSt50}$	F1B harbouring the complementing plasmid	This study
F1B_pNZ44-07020C$_{UCCSt50}$	F1B derivative cured of the complementing plasmid	This study
F2A::pNZ44	F2A harbouring the control plasmid, pNZ44	This study
F2A::pNZ44-07020$_{UCCSt50}$	F2A harbouring the complementing plasmid	This study
F2A::pNZ44-galE$_{UCCSt50}$	F2A harbouring the *galE* complementing plasmid	This study
Brie28	Non-host control strain used in binding studies	(24)
Phage		
P738 (accession number MK911750)	Lytic P738 group phage infecting UCCSt50 used for BIM generation and subsequent assays	(11)
SW13 (accession number MH892362)	*Brussowvirus* of *S. thermophilus* UCCSt50	(31)
Plasmids		
pNZ44	High copy expression vector	(30)
pNZ44-07020	pNZ44 harbouring the native *orf*07020 from UCCSt50	This study
pNZ44-galE	pNZ44 harbouring the native *orf*06475 from UCCSt50	This study

pHTP9	Green fluorescent protein (GFP) fusion vector used to produce the recombinant GFP-RBP-P738 protein	NZYTech, Portugal

Primers

Primer	Sequence (5′–3′)	Amplicon size (bp)	Source
RGP			
MSControl_F	gctggtcgtaattacctcg	2,724	(28)
MSControl_R	caacatcttccaaggtacg		
Var1_F	gtatataatgcacaagaggg	895	
Var1_R	gaaactaatcttaagcgttcc		
Fg1_F	gtcatgtgctctaccaattg	1,481	
Fg1_R	gatggagcttataacgttc		
CRISPR loci			
CRISPR1_F	tgctgagacaacctagtctctc	Template dependent	(16)
CRISPR1_R	taaacagagcctccctatcc		
CRISPR2_F	ttagcccctaccatagtgctg	Template dependent	
CRISPR2_R	tagtctaacactttctggaagc		
CRISPR3_F	ctgagattaatagtgcgattacg	Template dependent	
CRISPR3_R	gctggatattcgtataacatgtc		
Plasmid constructs			
pNZ44_F	ctaatgtcactaacctgccccg	Template dependent	(20)
pNZ44_R	gctttatcaactgctgct		
07020_F	aaaaaaccatggaaggaggcactcacatgtctaatccaaccgtttcag	969	This study
07020_R	aaaaaactgcagtacttctcctttcctttacagtaa		
galE_F	aaaaaaccatggaaggaggcactcacatggtggaccgtttggttgaa	972	This study
galE_R	aaaaaactgcagttaatctctatcgtcataacctt		
RBP			
P738-RBPopt_F	tcagcaagggctgagggcggtgaccttctggagcaacagcgat	2001	This study
P738-RBPopt_R	tcagcggaagctgaggttaaatcatgatggtaatctggccacggt		
pHTP9_F	gaatgaaaaacgcgaccacatggtg	RBP amplicon size + 294 bp	NZYTech, Portugal
pHTP9_R	ggttatgctagttattgctcagcg		

To evaluate the impact of the *orf*07020$_{UCCSt50}$ mutations on P738 phage infection, adsorption assays of P738 with UCCSt50, BIMs F1B and F2A, and the complemented F1B derivative were undertaken. Indeed, P738 adsorption was reduced by more than ~90% in F1B and F2A compared to the wild-type strain. Furthermore, phage adsorption was restored in the presence of the complementing plasmid for F1B (Table 3), while the plasmid-cured derivative was adsorption deficient. Therefore, it was inferred that the assumed glycosyltransferase activity encoded by *orf*07020$_{UCCSt50}$ is essential to produce the (part of) RGP saccharidic receptor for P738 phage adsorption.

Table 3. Relative efficiencies of plaquing and adsorption levels of phage P738 on its host UCCSt50 and derivatives

Strain	**EOP of P738**	**Adsorption level of phage P738 (%)**
UCCSt50	1	83.33 ± 0.37
UCCSt50::pNZ44	0.77 ± 0.12	80.39 ± 0.17
F1B	$\leq 3.6 \times 10^{-8}$	9.08 ± 0.11
F1B::pNZ44	$\leq 3.6 \times 10^{-8}$	14.71 ± 0.26
F1B::pNZ44-07020$_{UCCSt50}$	0.35 ± 0.33	79.41 ± 0.65
F1B::pNZ44-07020C$_{UCCSt50}$	$\leq 3.6 \times 10^{-8}$	14.71 ± 0.07
F2A	$\leq 3.6 \times 10^{-8}$	5.88 ± 0.14
F2A::pNZ44	$\leq 3.6 \times 10^{-8}$	
F2A::pNZ44-07020$_{UCCSt50}$	0.9 ± 0.1	
F2A::pNZ44-galE$_{UCCSt50}$	$\leq 3.6 \times 10^{-8}$	

*orf*07020 is predicted to encode a glycosyltransferase family 2 protein

BLASTP, Pfam, and HHpred analysis of the protein product of *orf*07020$_{UCCSt50}$ revealed significant similarity to family 2 glycosyltransferases involved in cell wall biosynthesis with specific roles in transferring sugar moieties to acceptor molecules. The HHpred analysis returned a 99.94% probability match to GalT1, a glycosyltransferase of *Streptococcus parasanguinis* (32), while BLASTP analysis revealed several significant alignments (E-value 0.0) to *S. thermophilus* glycosyltransferase family 2 proteins. Additionally, Pfam analysis identified a glycosyltransferase domain (PF00535). Interestingly, the mutation in this gene in F2A leads to an amino acid substitution in Asp93 (to an alanine residue), which is one of five conserved residues that are predicted to form the active site of the encoded protein (cd00761; residues 11, 13, 39, 91, and 93 are conserved in GT2 proteins).

BLASTN analysis of *orf*07020$_{UCCSt50}$ revealed its widespread presence in the genomes of *S. thermophilus* strains with high levels of nucleotide sequence identity (>99% sequence identity across 100% of the query sequence), while it is also found in the genomes of other *Streptococcus* species, including *Streptococcus salivarius, S. parasanguinis,* and *Streptococcus lactarius,* albeit with sequence identity below 95% across the full gene sequence and with lower representation in these species. Specifically, this gene was present in 98 of 517 *S. salivarius* genomes available in the NCBI database (at all assembly levels, i.e., contig, scaffold, chromosome, and complete), 50 of 258 *S. parasanguinis* genomes (at the same assembly level), and interestingly, it was also present in the only published *S. lactarius* genome. Additionally, in each of these species, the gene was located within the *rgp* locus, identified by the presence of several glycosyltransferase-encoding genes and

containing the *rmlD* gene typically found in the gene cluster associated with RGP biosynthesis in streptococci (33). This conservation among species that share overlapping ecological niches, such as the oral cavity (*S. parasanguinis*), food-associated environments (*S. salivarius*), dairy products (*S. thermophilus*), and human milk (*S. lactarius* (34)), suggests a strong evolutionary pressure to maintain the functional role of this gene in RGP biosynthesis, which may contribute to niche adaptation, interspecies interactions, and susceptibility to phages. In contrast, this gene was absent in human-associated or niche-divergent streptococci, including *S. pyogenes*, *S. mutans,* and *S. pneumoniae,* inferring this gene is not required for virulence or infection mechanisms typical of pathogenic streptococci. This gene is unique to *S. thermophilus* strains with a Vt1 *rgp* genotype (Fig. 1). As proposed previously (28), the biosynthesis of the variable side chain structure in UCCSt50 is initiated by the priming glycosyltransferase and its associated activator. *Orf*07020$_{\text{UCCSt50}}$ (previously termed *scbD*$_{\text{UCCSt50}}$) is believed to be involved in the subsequent elongation (addition of monosaccharides) of the side chain structure. However, the exact nature of the substrate and order of assembly are unknown (28). Based on these observations, we speculate that mutations in this gene impact the sequential addition of monosaccharide residues to the lipid-linked GlcNAc intermediate during UCCSt50 RGP side chain biosynthesis, potentially altering the structure or assembly of the final side chain structure, thereby affecting phage sensitivity.

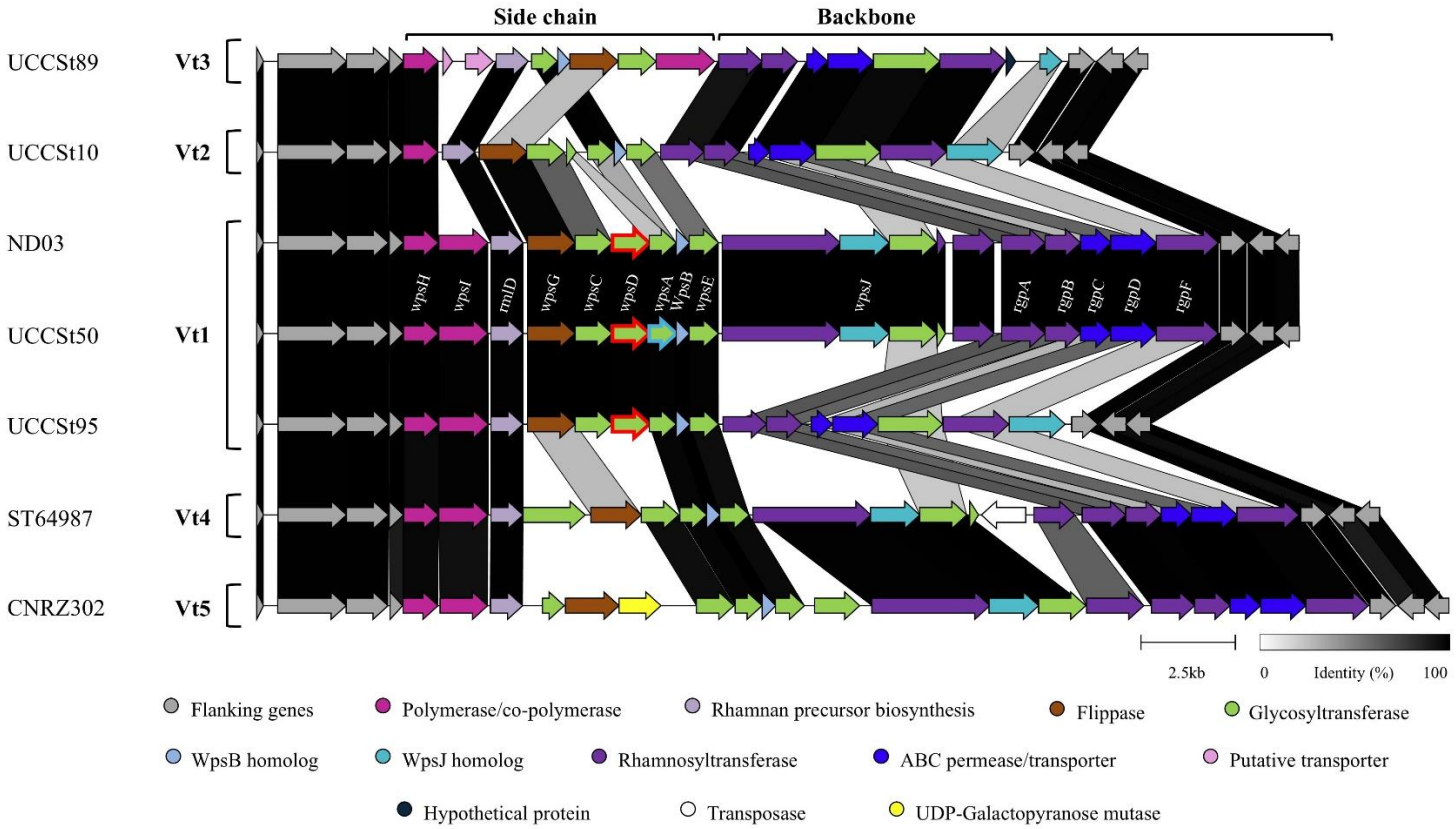


Fig. 1. Comparative analysis of *rgp* gene clusters from reference strains. The predicted functions of each of the protein-encoding open reading frames are colour coded and indicated at the base of the figure. The *rgp* locus can be split into two distinct regions: the variable side chain (5′) and the more conserved backbone (28) indicated on the figure. The mutated gene (shared between Vt1-type strains) is outlined in red. The mutated gene in BIM B1 (generated in a previous study (23)) is outlined in blue directly downstream of *wpsD*. Regions of homology are joined by blocks of different shades of grey to black and generated using CAGECAT (https://cagecat.bioinformatics.nl/tools/clinker). Gene names (written above the UCCSt50 cluster) were proposed based on comparisons to *Lactococcus lactis* CWPS clusters (35). GenBank accession numbers for strains—UCCSt89 (JANFMW000000000); UCCSt10 (CP065483); ND03 (CP002340); UCCSt50 (CP065477); ST64987 (CP049053); and CNRZ302 (CP065489).

It has previously been demonstrated that P738, as well as the other identified member of the P738 genus, D4446, possesses broad host ranges (11). We speculate that the P738-sensitive strains applied in this study possess a Vt1 genotype. A high prevalence of Vt1-type strains has also been reported, with Vt1-type *S. thermophilus* strains in an industrial strain collection being reported at a level of 50% (RGPA and RGPB possess a Vt1 genotype) (22). Similarly, in a recent study, approximately 17% of all isolates from a range of fermented dairy products possessed a Vt1 genotype (24), while 85% of *S. thermophilus* isolates from Siciliano PDO cheeses were shown to display the Vt1 genotype (36). These studies highlight the prevalence of strains possessing this genotype in both industrial and artisanal contexts. Furthermore, there are currently 156 *S. thermophilus* genome sequences available in the NCBI database (at the chromosome or complete genome assembly level); BLASTN analysis of the unique Vt1 gene against these genomes returned 61 sequence hits (all with E values of 0.0), representing 39.1% of the total publicly available genomes. These results reinforce the prevalence of Vt1-type strains in *S. thermophilus* populations and underscore the industrial importance of our findings, as they reveal a P738 phage infection profile dependent on the unique Vt1 gene, highlighting a potential vulnerability in industrial strains.

Preparation and structural analysis of the RGP from the P738 phage-resistant derivative of UCCSt50, F1B

To verify if mutations within *orf*07020$_{UCCSt50}$ affect the RGP structure (as was previously observed for the priming glycosyltransferase-encoding gene (23)), we prepared and analysed the RGPs of the parent strain, UCCSt50 (23), and the BIM F1B. Compositional analysis revealed the presence of rhamnose (Rha), glucose (Glc), N-acetylglucosamine (GlcNAc), and galactose (Gal) for both the wild-type and the phage-resistant derivative. However, the amount of Rha in the RGP of the mutant was lower compared to the wild type (with the ratio ~3:4 BIM:WT).

Detailed 2D NMR analysis of the RGP of the BIM F1B (Fig. 2; Supplementary Table S4) revealed that the backbone structure is composed of trisaccharide-repeating units (–2-α-Rha-6-α-Glc-2-α-Rha–), with linear tri- or disaccharide side chains on the α-Rha residue A (Fig. 2). Compared to the RGP of the wild-type strain UCCSt50, the terminal α-Rha residue is absent in all RGP side chains, and the β-Gal residues are absent in 50% of the RGP side chains (Fig. 3). The results of methylation analysis agreed with the proposed structure (Fig. 3). We propose that the absence of the terminal Rha (-α-Rha-2 moiety in WT structure, Fig. 3) is associated with a reduced efficiency of incorporating the Gal side chain.

a)

D F,F' G
β-Gal-4-α-Rha-3-β-GlcNAc
|3
-2-α-Rha-6-α-Glc-2-α-Rha-
C B A

β-Gal D 50%, F - no D, F' with D.

b)

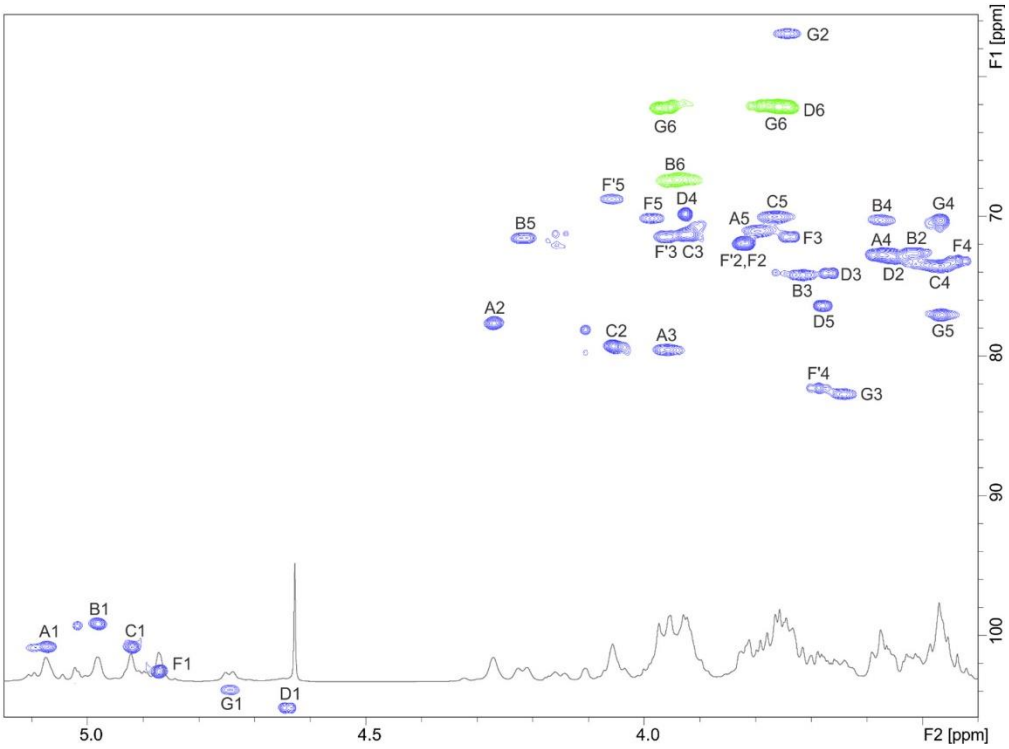


Fig. 2. Structure of the repeating unit (**a**) and fragments (**b**) of ^{1}H-^{13}C HSQC and ^{1}H NMR spectra of the RGP from the BIM F1B. In the HSQC spectra (**b**), different colours (green and blue) distinguish between positive and negative NMR signals.

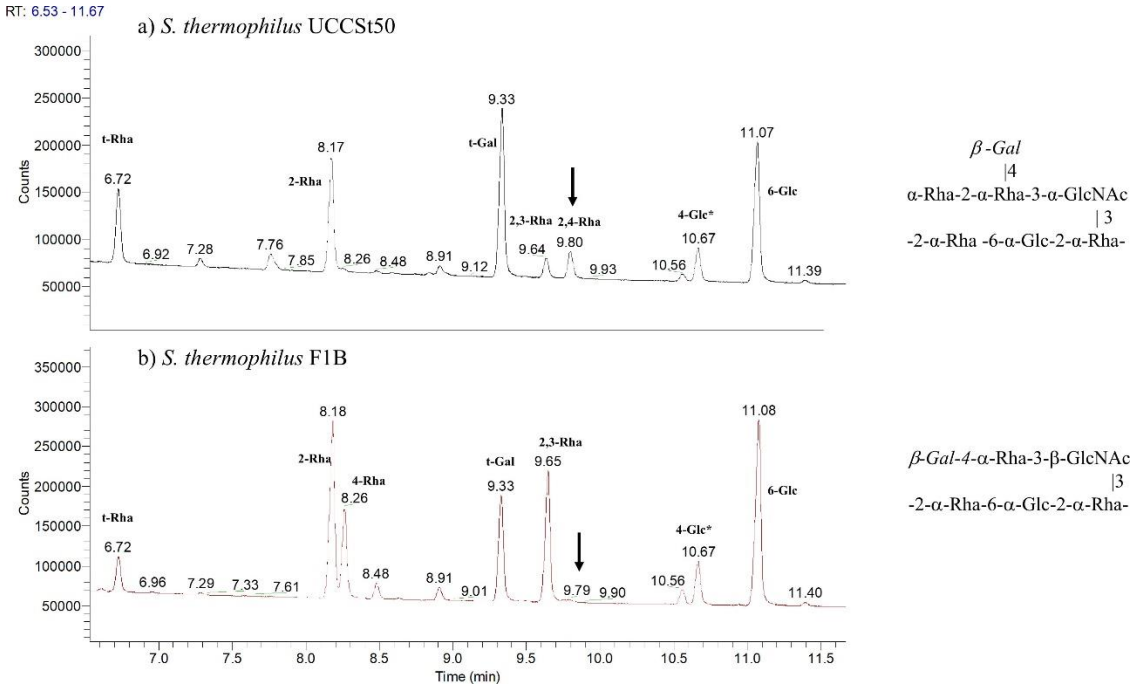


Fig. 3. Methylation analysis profiles of the RGP preparations of the wild-type *S. thermophilus* UCCSt50 (**a**) and its P738-resistant derivative F1B (**b**) and the associated RGP structures. In the absence of the terminal Rha residue, the branched 2,4-Rha becomes 4-Rha (corresponding peaks are marked with arrows). Approximately 50% of the terminal β-Gal residue is absent in F1B RGP. Peaks labelled with an asterisk represent polysaccharide(s) other than RGPs.

Therefore, we conclude that *orf*07020$_{UCCSt50}$ is a glycosyltransferase-encoding gene responsible for the addition of the terminal rhamnose residue to the side chain structure. It may be inferred that this alteration to the RGP structure results in resistance to phage P738.

The RGP side chain structure is required for P738 phage infection

Alterations to the CWPS structures of *Lactococcus lactis* and *S. thermophilus* have been implicated in phage resistance, primarily by disrupting phage adsorption at the cell surface (20,23,37,38). Recent work has demonstrated that the RGP side chain structure of *S. thermophilus* UCCSt50 is also essential for host adsorption by the *Brussowvirus* SW13 (23). To establish if the impact of the *rgp* locus mutations on phage adsorption was similar or distinct in these two members of distinct phage genera, the sensitivity of UCCSt50 and selected BIMs to SW13 (*Brussowvirus*) (31) and P738 was assessed (Table 4; all reported results are the average of at least triplicate biological assays). The wild-type UCCSt50 strain was, as expected, confirmed to be susceptible to either of these two phages, while BIM B1 (generated in a previous study and which lacks the entire RGP side chain structure (23)) was resistant to both phages, confirming the necessity of the RGP tetrasaccharidic side chain for infection. Notably, F1B, which has a structurally altered side chain, and F2A are susceptible to SW13 (EOP = 0.46 ± 0.08 and 0.69 ± 0.16, respectively) and resistant to P738. These results highlight the distinct structural requirements for SW13 and P738 (partial vs complete side chain structure), i.e., while SW13 can infect strains with a tri- or disaccharidic side chain structure, P738 infection is dependent on the presence of the intact RGP structure (comprising a backbone composed of trisaccharide-repeating units (–2-α-Rha-6-α-Glc-2-α-Rha–) and a tetrasaccharide side chain containing GlcNAc, Rha, Glc, and Gal residues (23). This underscores the role of specific host-encoded glycan features in mediating phage–host interactions and establishes that complete structural integrity or a specific part of the RGP side chain is required for P738 infection.

Table 4. Relative efficiencies of plaquing of phages SW13 (*Brussowvirus*) and P738 on their primary host UCCSt50 and derived BIMs

Strain	EOP of SW13	EOP of P738
UCCSt50	1	1
B1	$\leq 1.6 \times 10^{-6}$	$\leq 1.17 \times 10^{-7}$
F1B	0.46 ± 0.08	$\leq 1.17 \times 10^{-7}$
F2A	0.69 ± 0.16	$\leq 1.17 \times 10^{-7}$

RGP mutations cause growth aberrations

It has previously been shown that alterations to cell wall polysaccharide structures in LAB can have an impact on bacterial growth kinetics, including delayed growth and reduced overall cell density (37,39,40). Similar findings have been observed in pathogenic streptococcal species. For example, deficiency of RgpG affects cell division and biofilm formation in *Streptococcus mutans* (41), while mutants of *rgpE* elicit a reduced growth rate compared to the wild-type strain in the same species (42). Interestingly, modifications to the RGP structure in *Streptococcus pyogenes* and *Enterococcus faecalis* (the RGP structures are termed GAC and EPA, respectively) were not observed to affect bacterial growth yet significantly impact virulence (43–45). To determine the impact of RGP-associated mutations on planktonic growth and possible sedimentation in this study, growth characteristics of UCCSt50 and its phage-resistant derivatives over an 8-hour period were monitored using OD readings at a fixed point on the culture-containing tube (Fig. 4). All tested BIMs exhibited a slower growth rate during the early phases compared to the wild-type strain. Notably, B1 demonstrated the longest lag phase (defined in

this study as the period between hours 1 and 4), growing consistently slower than the other tested BIMs and failing to reach the peak OD observed in the wild type (or BIMs lacking the terminal rhamnose) (Fig. 4). Strains F1B and F2A showed an initial lag in growth but ultimately reached optical densities comparable to that observed for UCCSt50. All strains, including the wild type, displayed a rise-and-fall growth profile (Fig. 4), which is expected given the visual observation of the sedimenting phenotype commonly associated with *S. thermophilus* BIMs that possess altered CWPS structures (20,23).

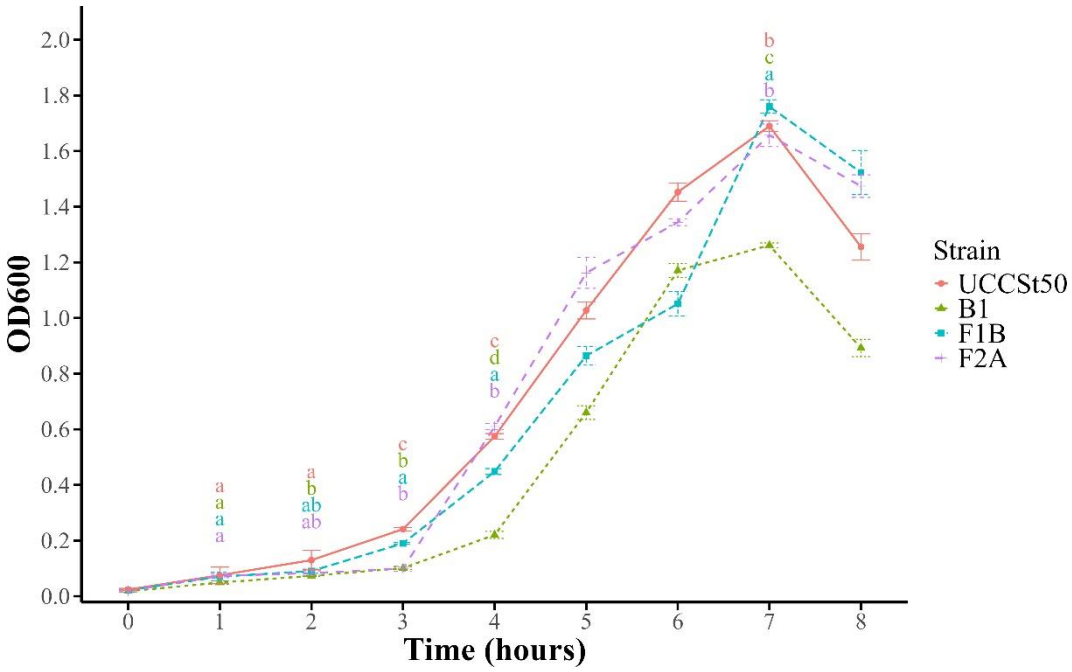


Fig. 4. Growth in liquid LM17 medium of *S. thermophilus* UCCSt50 and its phage-resistant derivatives. Each point represents the mean value at an optical density of 600 nm (OD_{600}) of three replicates. The standard deviations for each value are represented by error bars. One-way ANOVA followed by Tukey's honestly significant difference test was performed at hours 1, 2, 3, 4, and 7 using the multcompLetters function in R to assign group letters based on the adjusted *P*-values from Tukey's test. Compact letter displays indicate significant differences between strains ($P < 0.05$); letters are coloured by strain; strains sharing a letter are not significantly different.

Functional assignment of the receptor binding protein of phage P738

In order to further assess the molecular interactions involved in host recognition by P738, we wanted to validate the functionality of the proposed RBP of this phage. As a member of the most recently recognised genus of *S. thermophilus* phages, P738 has not been extensively studied in terms of host recognition or protein functions. However, several open reading frames (ORFs) have been identified with predicted roles in phage morphogenesis and host interaction (11). $Orf16_{P738}$ has been annotated as encoding the tail tape measure protein, $Orf17_{P738}$ as the distal tail protein (Dit), and $Orf18_{P738}$ as the Tal protein. HHpred analysis of a previously predicted structural protein (protein id = QDP43720.1) encoded by a gene located downstream of the Tal-encoding gene (11) revealed a phage RBP-like domain, identifying it as the putative RBP of phage P738. HHPred analysis revealed the presence of several high probability hits to receptor binding proteins or tail fibre proteins of phages, including T4 (99.7% PDB_5HX 2_I gp10 baseplate wedge protein), TP901-1 (96.0% probability PDB_4IOS_B receptor binding protein), and *Staphylococcus* phage phi812 baseplate protein (98.6% probability PDB_9EUF_q). These high-probability hits are all located within the C-terminal region, which is predicted to represent the so-called “head” or host binding domain of the receptor binding protein. Alignment of the protein sequences of the RBP sequences of TP901-1 and P738 also highlighted that the most significant region of sequence conservation was located within the C-terminal region of the respective proteins, with 22 conserved amino acids in this region likely accounting, in part, for the structural conservation of the RBP head domains of these proteins.

To establish if the proposed RBP is indeed responsible for host recognition, a green fluorescent protein (GFP)-tagged derivative of the predicted RBP was constructed

and applied in host binding studies using the wild-type strain, its P738-resistant derivatives, and a non-host strain (Fig. 5). Binding assays using 30 µg of the heterologously expressed GFP-RBP_{P738} fusion protein exhibited complete cell surface binding of the cognate host UCCSt50, consistent with the predicted functional activity of the P738 RBP in host recognition. In contrast, a significantly reduced fluorescence signal (and thus binding) was observed for the BIMs, indicating reduced host binding consistent with the adsorption assay findings. RBP binding was restored in the complementing strain (F1B::pNZ44-$07020_{UCCSt50}$), with observed fluorescence comparable to the parent strain UCCSt50. Binding was not observed (no fluorescence) for the non-host control strain, *S. thermophilus* Brie28, which possesses a distinct *rgp* genotype, i.e., V2B2 (24). These results demonstrate that the host-encoded RGP structure in UCCSt50 acts as the specific receptor for P738 phage and confirms the role of this gene product as the primary RBP of phage P738.

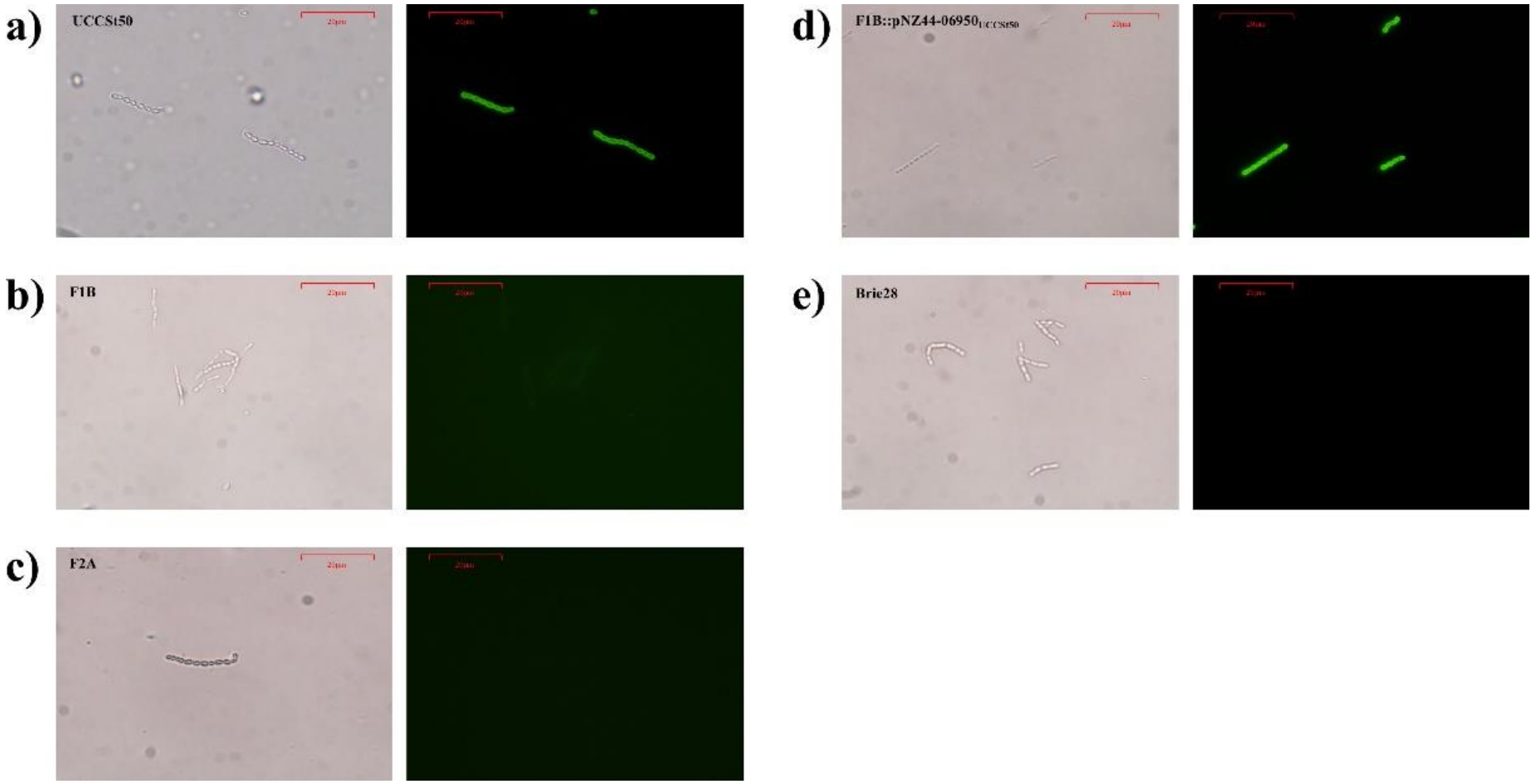


Fig. 5. Brightfield (left panels) and fluorescent labelling images (right panels) of the (**a**) parent strain UCCSt50, its phage-resistant derivatives (**b**) F1B and (**c**) F2A, (**d**) the genetically complemented strain F1B::pNZ44-07020$_{UCCSt50}$, and (**e**) a non-host strain (Brie28) using 30 µg protein at 100× magnification. The scale bars are indicated in the individual images.

Discussion

Deciphering the host-encoded receptors of dairy phages has long been recognised as a cornerstone of generating starter cultures to include strains that are resistant to phages, thereby increasing the resilience of dairy fermentations (46–48). While previous research has highlighted the role of EPS in mediating phage adsorption for members of the *Piorkowskivirus*, *Moineauvirus,* and *Vansinderenvirus* genera (12,20–22,24), the present study provides evidence supporting the involvement of the RGP as the host-encoded receptor for P738 phages in addition to the *Brussowvirus* SW13 (23). Our findings highlight that host recognition and binding by P738 is underpinned by *orf*07020$_{UCCSt50}$, which is located within the *rgp* locus and encodes a glycosyltransferase. This gene is unique to dairy streptococcal strains possessing a Vt1 genotype and the associated RGP structure that is now confirmed as the receptor for both P738 and the *Brussowvirus* SW13. In the present study, two BIMs (among a panel of 20 BIMs) were selected for detailed characterisation; however, it is likely that genome sequencing of additional BIM genomes may reveal additional insights into the genes underpinning host binding processes and represents an area of future research.

It has previously been demonstrated that the two currently known P738 phages (P738 and D4446) have broad host ranges, which is not typical among dairy phages (11). In this study, a BLASTN search against 156 publicly available *S. thermophilus* genomes revealed the presence of the unique Vt1 gene in 61 strains, highlighting the possible prevalence of Vt1-type strains within the species. While it is not possible to infer that all *S. thermophilus* strains exhibiting a Vt1 genotype will be bound and subsequently infected by P738-type phages without experimental validation, the widespread presence of Vt1 genotype strains represents a risk factor

for such scenarios. Binding studies exploring the full binding potential of P738 phage receptor binding proteins are currently under investigation (and which are beyond the scope of this manuscript) and will serve to address this knowledge gap. CWPS biosynthesis elicited through the functions of genes within the *cwps* locus of *L. lactis* is well characterised. The lactococcal CWPS is organised into distinct functional regions, including a conserved region responsible for the synthesis of a rhamnan backbone component and a highly variable region encoding the surface-exposed oligosaccharide or polysaccharide pellicle (PSP) (46,49). A comprehensive CWPS biosynthesis scheme has been proposed where the rhamnan and PSP side chains are synthesized independently from separate lipid-sugar precursors, then joined extracellularly by a GT-C fold glycosyltransferase (35). The functional assignment of genes was based on structural similarity prediction by HHpred analysis combined with transmembrane helix prediction (35). In contrast, the *rgp* cluster of *S. thermophilus* remains less well defined. Although a proposed model for RGP biosynthesis has recently been developed (28), functional annotation of genes within the cluster is still emerging. In this study, we assigned gene function guided by approaches used for CWPS classification in *L. lactis*. This includes the analysis of predicted transmembrane domains to infer roles in membrane-associated steps of the pathway, and the identification of glycosyltransferases based on conserved GT2 and GT4 family domains. The resulting framework aligns with the dual-pathway model observed in *L. lactis* and adds to the proposed pathway previously developed for *S. thermophilus rgp* loci.

Genetic complementation of the two phage-resistant derivatives with the native *orf*07020$_{UCCSt50}$ restored P738 phage sensitivity, suggesting a direct link between this gene and phage susceptibility. This is consistent with previous studies in *S.*

thermophilus and other lactic acid bacterial species (*L. lactis*), where glycosyltransferases involved in CWPS biosynthesis have been implicated in host recognition and adsorption (21–23,38,49). Adsorption assay results demonstrate that mutations in *orf*07020$_{UCCSt50}$ significantly reduced phage adsorption by P738, a phenotype corroborated by fluorescence-based RBP labelling assays. Similar observations have been made in *S. mutans*, for which the sero-specific region of the RGP cluster plays a crucial role in phage adsorption (50), and in *E. faecalis*, where the enterococcal polysaccharide antigen cluster determines phage susceptibility (51).

Compositional analysis of the RGP in the wild-type and phage-resistant derivative F1B revealed the complete loss of the terminal rhamnose saccharide, as well as the absence of approximately 50% of the β-Gal residues in the RGP side chain structure, which is surface-exposed. Previous studies have demonstrated that mutations in the predicted priming glycosyltransferase of the *rgp* gene cluster are associated with an absence of the side chain structure and complete resistance to phage SW13 (*Brussowvirus*) (23). In the current study, we confirm that this mutant (B1) is also resistant to P738, which is consistent with our finding that P738 requires (part of) the RGP side chain for adsorption. Interestingly, we demonstrated in the current study that SW13 can infect BIMs that lack the terminal rhamnose monosaccharide of the RGP side chain with almost equal efficiency to the parent strain, while P738 cannot. This result provides the first insight into the minimal RGP structural requirements for phage infection of the P738 and *Brussowvirus* genera and suggests that P738 has a strict requirement for the complete or distinct/specific part of the tetrasaccharide side chain for binding, while SW13 requires at most three of the four component monosaccharides for successful

adsorption. While the saccharidic receptor moiety for representatives of each of the five dairy streptococcal phage genera is now defined, future studies may focus on deriving the exact and minimal receptor that these phages bind to, which in turn will improve the predictions of phage sensitivity of strains that are applied in dairy fermentations. These findings could play a crucial role in developing phage-robust cultures and strategies to enhance the stability of industrial dairy fermentations.

It has recently been shown that the reduction in the production of cell wall polysaccharides may allow bacterial starter cultures to evade phage infection (52). Here, amino acid substitutions in a cell wall-associated glycosyltransferase may contribute to the fine-tuning and reduction in the production of RGP side chain with an associated (and possibly reversible) cessation in phage recognition. Although attempts were made to isolate phage escape mutants of P738 capable of infecting the BIMs, none were recovered in this study. Nevertheless, it is likely that repeated exposure of BIMs with non-synonymous missense mutations to phages at an industrial scale would select for such phage mutants. Alternatively, where phage pressure is removed or significantly reduced, the BIMs may revert to the wild type as observed in *Lactococcus cremoris* (52). One of the most significant challenges with dairy streptococcal phages is the inability to reproducibly achieve high titre phage lysates as is possible within other bacterial host species. The inability to achieve high titres (beyond 10^7–10^8 PFU/mL) hampered our efforts to challenge the BIMs with a sufficiently high number of phages to generate phage escape mutants. It is also possible that several mutations are required for P738 to overcome the modified cell surface, and this may not be possible in a single round of exposure to the phage. Therefore, future studies may focus on repeated longitudinal studies to achieve such mutations notwithstanding the issue of the low titre of these phages.

Future studies pertaining to the isolation of such phage escape mutants will facilitate a deeper understanding of the amino acids within the receptor-binding or other adhesion proteins that are involved in the saccharide-binding process.

Here, we show that RGP-associated mutations in *S. thermophilus* UCCSt50 negatively affect growth, with mutant B1 showing the longest lag phase and mutants F1B and F2A exhibiting milder growth defects. These findings suggest that the extent of structural disruption to the RGP correlates with the severity of the growth defect. The pronounced growth impairment observed in B1, which completely lacks the side chain, contrasts with the milder growth impact observed for F1B, which has a structurally altered RGP side chain and F2A. This supports the notion that even subtle modifications to the CWPS structure can impact bacterial fitness as well as phage sensitivity.

Recently, we have shown that P738 is affected by several anti-phage systems encoded on *S. thermophilus* genomes, including GAO19, Hachiman, SoFic, AbiD, and AbiE (53). Therefore, while the presence of a suitable receptor is critical to the initial interactions of P738 phages with their host, it is imperative that the internal anti-phage landscape is also compatible with phage replication. Since P738 phages have only been identified recently, it is unclear how much exposure these phages may have had to the anti-phage repertoire of *S. thermophilus* or if their genomes encode counter defences to CRISPR-Cas, restriction/modification, and the plethora of other antiphage systems. Anti-CRISPR and methyltransferase-encoding genes have been identified on the genomes of other genera of dairy streptococcal phages (9,54,55). Therefore, future studies may focus on the identification of such counter-defence systems in P738 phages to enhance current understanding of the co-evolution of these phages with their hosts.

In conclusion, the current study has identified the specific RGP component required by P738 for host recognition and binding, while the predicted receptor binding protein of phage P738 has been functionally proven. These insights enhance our understanding of the mechanisms by which dairy streptococcal phages interact with their hosts and the specificity of streptococcal phages. This information will be pivotal in informing the development of phage-resistant bacterial strains and the selection of strains to enhance the robustness of starter cultures, thereby improving the sustainability of dairy fermentation practices. Further investigations into the biochemical structures of additional *S. thermophilus* strains and the role of additional genes within the *rgp* cluster may facilitate more detailed correlations between the *rgp* genotype and the associated RGP chemical composition and structures and predictable outcomes in dairy fermentations.

Materials and Methods

Bacterial strains and bacteriophages

Bacterial strains and bacteriophages used in this study are listed in Table 2. *S. thermophilus* strains were routinely grown overnight at 42°C from single colonies or bacterial stocks stored at −20°C (25% [wt/vol] glycerol) in M17 (Sigma, USA) supplemented with 0.5% lactose (Sigma, USA). For *S. thermophilus* strains containing plasmids (UCCSt50::pNZ44, F1B::pNZ44, F1B::pNZ44-07020$_{UCCSt50}$, and F2A::pNZ44-galE$_{UCCSt50}$), chloramphenicol (Sigma, USA) was added to LM17 at a final concentration of 3 µg/mL in agar and 5 µg/mL in broth.

Bacteriophage assays

Phage P738 was propagated in fresh LM17 broth supplemented with 10 mM $CaCl_2$. The LM17 broth was inoculated with 1% of a fresh overnight culture of *S. thermophilus* UCCSt50, and 10 µL P738 phage lysate was added and incubated until visible lysis was observed (approximately 6 hours) or overnight at 42°C. Phage lysates were filtered (0.45 µm, Sarstedt, Germany) and stored at 4°C.

Bacteriophages were enumerated using the standard overlay method (56), in which LM17 medium was supplemented with 10 mM $CaCl_2$ and 10 g/L (solid agar) or 4 g/L (semi-solid agar) technical agar (Neogen, USA). Briefly, 400 µL of a fresh UCCSt50 culture grown overnight and 10 µL of the P738 phage lysate ($\geq 10^7$ PFU/mL) were mixed in 4 mL semi-solid LM17 agar and poured onto LM17 solid agar before incubation at 42°C overnight.

Bacteriophage-insensitive mutant generation

BIMs of *S. thermophilus* UCCSt50 were generated using a standard plaque assay protocol (56). Briefly, 400 µL of a fresh UCCSt50 overnight culture and 10 µL of phage P738 lysate (at a titre of $\geq 10^7$ PFU/mL) were mixed in 4 mL semi-solid LM17 agar (4 g/L agar) supplemented with 10 mM $CaCl_2$ and poured onto LM17 agar plates and incubated anaerobically (Anaerocult A, Millipore, Germany) at 42°C for 24 hours. Surviving colonies on the semi-solid agar were picked and passaged three times in fresh LM17.

All generated BIMs were screened for insensitivity to phage P738 using the spot and plaque assay methods (as described above). BIMs exhibiting complete or incomplete P738 resistance were confirmed as UCCSt50 derivatives by *rgp* PCR (28) (Table 2) and by visually comparing the amplicons of the CRISPR 1, 2, and 3 repeat spacer loci (16) (Table 2) on a 1% agarose gel.

DNA extraction, genome sequencing, and SNP analysis

Genomic DNA of strain UCCSt50 and two BIMs (F1B and F2A) was extracted using a modified protocol of the Nucleobond AXG 100 with Buffer set III (Macherey-Nagel, Germany). This choice of BIMs was based on a previous observation by our group that SW13-resistant derivatives of UCCSt50 produced mutations strictly in the priming glycosyltransferase-encoding gene, which exhibited a complete phage-resistance phenotype. Therefore, to maximize the potential to identify distinct mutations, the two BIMs with incomplete phage resistance were selected for further characterisation. DNA extraction was conducted using a modified protocol of the Nucleobond AXG100 kit with Buffer Set III (Macherey-Nagel, Germany). Initially, 40 mL cells grown to an OD_{600} of ~1.0 were

pelleted by centrifugation (5,000 × *g* for 20 minutes), followed by pre-treatment of the cell pellets with lysozyme (at a final concentration of 0.8 mg/mL; Sigma, USA) and mutanolysin (at a final concentration of 50 units/mL; Sigma, USA) at 37°C for 1 hour. Proteinase K (supplied with Buffer set III) was subsequently added (at a final concentration of 100 µg/mL), and samples were incubated at 50°C for an additional hour. The extraction protocol was completed according to the manufacturer's protocol using AXG 100 columns. DNA was resuspended in 10 mM Tris buffer (Fischer Scientific, USA) prior to whole genome sequencing.

Bacterial genome sequencing was performed by Plasmidsaurus (USA) using Oxford Nanopore Technology (UK) long-read paired with Illumina short-read sequencing. Long-read sequencing was conducted on a PromethION P24 platform with R10.4.1 flow cells, and libraries were prepared using the Rapid Barcoding Kit 96 V14. Basecalling was performed in "super-accurate mode" using the ont-doradod-for-promethion algorithm version 7.4.12, with ONT reads meeting a minimum *Q*-score of 10 and adapters trimmed via MinKnow (57). Illumina (USA) short-read sequencing was performed on an Illumina NextSeq2000 platform using libraries prepared with the SeqWell ExpressPlex 96 kit, using paired-end reads 2 × 150 bp chemistry. Illumina-derived forward and reverse reads were aligned to the ONT assembly using BWA-MEM (58), which clips low-quality sequences from the ends of reads. Polypolish version 6.0 (59) was used to generate the final, polished genome assembly, which was subsequently annotated using Bakta version 1.6.1 (60) with default parameters.

The quality or completeness of genome assemblies was evaluated with the microbial genomes atlas (MiGA) webserver (61). To verify that no modifications had been made to the BIM CRISPR (1, 2, and 3) loci spacer content or number, all

sequenced genomes were uploaded to CRISPRCasFinder (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index) (62), and CRISPR spacer arrays were analysed. A single nucleotide polymorphism analysis of BIM F1B and F2A genome sequences against the genome of the parent strain UCCSt50 was performed using the Bowtie2 alignment tool (63) and SAMtools (64).

Molecular cloning and complementation studies

The native *orf*07020$_{UCCSt50}$ and *orf*06475$_{UCCSt50}$ were amplified with Phusion Hot Start II High-Fidelity PCR Master Mix (Thermo Fisher Scientific, USA) using primer combinations 07020_F and 07020_R and galE_F and galE_R (Table 2). The resulting amplicons were purified using the GenElute PCR Clean-Up Kit (Sigma, USA). The amplicons and pNZ44 cloning vector were digested with NcoI and PstI fast digest enzymes (Thermo Scientific, USA) at 37°C for 30 minutes. All digests were purified (GenElute PCR Clean-Up Kit; Sigma, USA) before ligation was performed overnight at room temperature using T4 DNA ligase (Promega, UK). The ligation mixture was dialyzed against sterile, distilled water for 20 minutes before being introduced into competent *Escherichia coli* EC101 cells. Briefly, 5 µL of the ligation mixture was added to 45 µL of competent EC101 cells in a pre-chilled electroporation cuvette held on ice. The cells were subjected to a 2.5 kV single pulse before 950 µL pre-warmed LB was added to the cuvette. Recovery was performed at 37°C with agitation (150 rpm) for at least 1 hour before plating on LB agar supplemented with 10 µg/mL chloramphenicol and incubated overnight at 37°C. The resulting presumptive transformants were screened by colony PCR using pNZ44_F and pNZ44_R (Table 2), and "positive" clones were grown overnight at 37°C in LB supplemented with 10 µg/mL chloramphenicol. Plasmid DNA was subsequently extracted from selected transformants using the GeneJet plasmid

miniprep kit (Thermo Scientific, USA), and the integrity of the construct was validated by Sanger sequencing (Eurofins, Germany) before being introduced into BIM F1B or F2A.

Fresh competent cells of BIM F1B and F2A were prepared as previously described (23) with the following modifications: cells were grown in LM17 supplemented with 0.3% glycine and a sucrose concentration of 0.25 M in place of 0.5 M. Briefly, 100 µL of cells and 10 µL of plasmid DNA were mixed and left on ice for 30 minutes before a heat shock was applied at 42°C for 45 seconds followed by holding on ice for 2 minutes (65). The mixture was transferred to a pre-chilled electroporation cuvette before applying a single 2.0 kV pulse followed by the immediate addition of 900 µL HJL recovery media (3% tryptone, 1% yeast extract, 0.5% KH_2PO_4, 0.2% beef extract, 0.5% lactose, 20 mM $CaCl_2$, and 200 mM $MgCl_2$) and incubating at 42°C for a minimum of 4 hours. Cells were plated on LM17 agar supplemented with 3 µg/mL chloramphenicol and incubated at 42°C for 24–72 hours. Positive transformants were verified to contain the recombinant plasmids by colony-based PCR using pNZ44_F and pNZ44_R primers (Table 2) and confirmed by Sanger sequencing (Eurofins, Germany).

The complementing plasmid was cured from F1B::pNZ44-07020$_{UCCSt50}$ by novobiocin treatment and serial passaging in the absence of chloramphenicol. Fresh LM17 broth was inoculated with 1% of an overnight culture of the complemented strain F1B::pNZ44-07020$_{UCCSt50}$ and allowed to grow at 42°C for 2.5 hours. Novobiocin (Sigma, USA) was added to the culture at a final concentration of 10 µg/mL followed by an additional 5-hour incubation. The culture was then plated in the presence or absence of chloramphenicol for several passages. The cured derivative, F1B::pNZ44-07020C$_{UCCSt50}$, was tested for P738 phage sensitivity. All

assays were performed in at least triplicate, and the average and standard deviation were calculated using the standard Student's *t*-test in MS Excel.

Efficiency of plaquing and adsorption assays

The efficiency of plaquing of phage P738 on all tested strains listed in Table 2 was determined using the standard overlay method described above to assess the role of *orf*07020$_{UCCSt50}$ in the P738 phage infection process. The relative EOP of phages on the listed strains was determined by dividing the observed titer of the phage on a given BIM/complement/cured host by that on the parent strain. All assays were performed in triplicate, and averages and standard deviations were calculated in MS Excel using a Student's *t*-test.

Adsorption assays were performed based on a previously adapted protocol (9). A volume of 10 mL LM17 broth was inoculated with the appropriate *S. thermophilus* strain from a fresh overnight culture and grown at 42°C until an OD_{600} of 0.6–0.7 was achieved. A 500 µL volume of the growing culture was transferred to a microcentrifuge tube and centrifuged at 5,000 × *g* for 10 minutes to harvest the cells. The cell pellet was resuspended in 500 µL of SM buffer (100 mM NaCl, 10 mM Tris-HCl [pH 7.5], 20 mM $CaCl_2$, and 10 mM $MgSO_4$) before an equal volume of P738 phage lysate ($\leq 10^5$ PFU/mL) was added to the cells or 500 µL of SM buffer (negative control). The mixture was incubated at 42°C for 15 minutes and centrifuged at 15,000 × *g* for 5 minutes to remove bacterial cells before removing the residual phage-containing supernatant for enumeration as described above. The adsorption levels (as a percentage of the total number of phages present) were determined using the following formula: ([control phage titre − free phage titre in

supernatant]/control phage titre) × 100. All assays were performed in triplicate, and averages and standard deviations were calculated as described above.

Comparison of bacterial growth behaviour

Bacterial strains were inoculated (2%, vol/vol) into 5 mL of LM17 broth in a clear, sterile round-bottom tube (Corning Science, Mexico). Growth was recorded by placing the tubes in a Jenway 7200 visual spectrophotometer (Bibby Scientific, UK) and measuring at an optical density of 600 nm (OD_{600}). Tubes were incubated at 42°C, and OD_{600} readings were taken hourly over an 8-hour period. Growth of each strain was measured in triplicate. OD_{600} readings were plotted as mean ± standard deviation using R (version 4.4.1) (R Core Team, 2024) and the ggplot2 package (66).

To compare growth differences between strains at each time point, one-way ANOVA was performed, followed by Tukey's Honestly Significant Difference *post hoc* test. Where statistically significant differences ($P < 0.05$) were found, compact letter displays were generated using the multcompLetters() function from the R package *multcompView* to assign groupings (67). Each strain was assigned a letter at each time point; strains sharing a letter are not significantly different from one another.

RBP production and purification/recombinant protein production and purification

All predicted ORFs from P738 were subject to HHpred analysis. The putative receptor binding protein of P738 (protein id = QDP43720.1) was codon optimised for *E. coli* to improve protein expression using the "codon optimization" tool on VectorBuilder (https://en.vectorbuilder.com/tool/codon-optimization.html). The

codon-optimised protein sequence was synthesized by Eurofins Genomics (Germany). The optimised RBP was amplified from the supplied vector using RBP primers listed in Table 2. The resulting amplicon was purified using a GenElute PCR cleanup kit (Sigma, USA) and cloned into the green fluorescent protein fusion vector pHTP9 (NZYTech, Portugal) following the manufacturer's protocol. The ligation mixture was dialyzed against sterile distilled water before being introduced to *E. coli* BL21 (DE3) via the heat shock method (30 minutes on ice; 42°C for 45 seconds; 2 minutes on ice (65)). After recovery at 37°C for 1 hour with agitation (150 rpm), the cells were plated on LB agar containing 50 µg/mL kanamycin (Sigma, USA) and incubated overnight. Recombinant colonies were screened using pHTP9-specific primers (listed in Table 2) and verified by Sanger sequencing (Eurofins, Germany).

For protein production and purification, 100 mL of LB broth supplemented with 50 µg/mL kanamycin was inoculated with 1 mL of an overnight culture of *E. coli* BL21(DE3) harbouring the GFP-P738 RBP fusion protein. Cells were grown to an OD_{600} of 0.5–0.6 before isopropyl-β-D-thiogalactoside (Melford Laboratories, UK) induction at a final concentration of 0.5 mM. The culture was incubated at 24°C for 24 hours with shaking (300 rpm). Cells were harvested by centrifugation at 4,700 × *g* for 30 minutes at 10°C. The resulting cell pellets were resuspended in lysis buffer (10 mM Tris-HCl [pH 7.5], 300 mM NaCl, 10 mM imidazole, and 25 mg/mL lysozyme) and stored at −70°C for a minimum of 24 hours. After thawing, the cells were subjected to five cycles of sonication (MSE Soniprep, Sanyo, Japan) at maximum amplitude (30 seconds on, 30 seconds off with a 2-minute rest on ice after the third cycle). Cellular debris was removed by centrifugation at 14,945 × *g*

for 30 minutes at 4°C, followed by an additional centrifugation at 4,696 × *g* at room temperature to eliminate aggregates before protein purification.

The recombinant protein was purified using a standard Ni-nitrilotriacetic acid agarose column (Qiagen, UK), according to the manufacturer's instructions, with the desired protein eluting in the 100–250 mM imidazole fractions. Protein concentrations were measured using the Qubit protein assay (Thermo Scientific, USA) on a Qubit 2.0 Fluorometer. Protein samples were dialyzed against protein buffer (10 mM Tris-HCl [pH 7.5] and 300 mM NaCl) before visualization via sodium dodecyl sulphate-polyacrylamide gel electrophoresis analysis. Protein fractions were stored at 4°C until use.

Labelling/binding assays

Fluorescent binding assays were performed as previously described, with minor modifications (23). Briefly, 300 µL of relevant cells (Fig. 5) was harvested (6,000 × *g* for 5 minutes) at an OD_{600} between 0.4 and 0.6 and resuspended in 300 µL SM buffer before 30 µg of protein was added. The mixture was incubated at 42°C for 12.5 minutes. Cells were washed three times in 120 µL SM buffer before fluorescent binding was visualized by fluorescence microscopy. Fluorescence labelling of cells was visualized using an Olympus BX53 microscope equipped with an LED illuminator and fluorescence cube (U-FBNA), and images were processed using cellSens software (Olympus, Japan). An equal volume of SM buffer (representative of 30 µg) was added to cells acting as a negative control. *S. thermophilus* strain Brie28 (24) was selected as a non-host control.

Preparation and structural analysis of RGP of *S. thermophilus* F1B

Approximately 8 L of *S. thermophilus* culture (UCCSt50 and F1B) was grown overnight and harvested by centrifugation at 4,400 × *g* at 4°C for 30 minutes. The cell pellets were washed twice with ice-cold sterile distilled H_2O (300 mL, followed by pooling of cells before washing in 40 mL) and stored at −20°C.

CWPS (RGP) was extracted with cold 5% TCA and hot diluted HCl and purified on a Sephadex G-50 column, as described previously (20), without the final treatment with HF. HCl preparations had identical monosaccharide composition; the preparation of 0.01 M HCl extraction was used for linkage analysis. It was further purified by reverse-phase HPLC using an Agilent Zorbax C18 column (250 × 9 mm) in solvent A (0.1% TFA, vol/vol) and B (80% MeCN, vol/vol) at a flow rate of 3 mL/minute, starting from 3% B for 10 minutes then gradient to 100% B over 30 minutes. The UV detector was set at 220 nm, and fractions were collected every 1 minute. The CWPS (RGP) purified by HPLC was used for NMR analysis.

NMR experiments were carried out on a Bruker AVANCE III 600 MHz (^{1}H) spectrometer with a 5 mm Z-gradient probe with acetone internal reference (2.225 ppm for ^{1}H and 31.45 ppm for ^{13}C) using standard pulse sequences cosygpprqf (gCOSY), mlevphpr (TOCSY, mixing time 120 ms), roesyphpr (ROESY, mixing time 500 ms), hsqcedetgp (HSQC), hsqcetgpml (HSQC-TOCSY, 100 ms TOCSY delay), and hmbcgplpndqf (HMBC, 70 ms long range transfer delay). Resolution was kept <3 Hz/pt in F2 in proton-proton correlations and <5 Hz/pt in F2 of H-C correlations. The spectra were processed and analysed using the Bruker Topspin program. Monosaccharides were identified by COSY, TOCSY, and NOESY cross-peak patterns and ^{13}C NMR chemical shifts. Amino group location was concluded

from the high field signal position of aminated carbons (CH at 45–60 ppm). Connections between monosaccharides were determined from transglycosidic NOE and HMBC correlations.

Monosaccharide and methylation analyses were carried out as described previously (46).

Data Availability

The wild-type strain UCCSt50 was resequenced as part of this study (previously deposited under the strain name *S. thermophilus* 4078, accession number CP065477 (23)). The genome sequences of *S. thermophilus* UCCSt50, F1B, and F2A have been deposited in GenBank under BioProject accession number PRJNA1267047, and their associated accession numbers are as follows: CP194174, CP194006, and CP194005.

Supplementary Material

Supplementary Table S1: Efficiencies of plaquing (EOP) of phage P738 on its *S. thermophilus* host UCCSt50 and panel of presumed BIMs.

Strain/BIM	**EOP of P738**
UCCSt50 WT	1
F1A	$\leq 3 \times 10^{-8}$
F1B	6.6×10^{-7}
F1C	$\leq 3 \times 10^{-8}$
F1D	3.3×10^{-7}
F1E	$\leq 3 \times 10^{-8}$
F2A	6.6×10^{-7}
F2B	$\leq 3 \times 10^{-8}$
F2C	$\leq 3 \times 10^{-8}$
F2D	$\leq 3 \times 10^{-8}$
F2E	$\leq 3 \times 10^{-8}$
F21A	$\leq 3 \times 10^{-8}$
F21B	$\leq 3 \times 10^{-8}$
F21C	$\leq 3 \times 10^{-8}$
F21D	6.6×10^{-7}
F21E	2.3×10^{-5}
F22A	$\leq 3 \times 10^{-8}$
F22B	$\leq 3 \times 10^{-8}$
F22C	$\leq 3 \times 10^{-8}$
F22D	$\leq 3 \times 10^{-8}$
F22E	3.3×10^{-7}

Supplementary Table S2: General genome characteristics of UCCSt50 and selected BIMs, F1B and F2A. The quality or completeness of genome assemblies was evaluated with the microbial genomes atlas (MiGA) (61).

Strain	**Genome size (bp)**	**Genome Completeness**	**Average GC content**	**No. of predicted proteins**	**Accession number**
UCCSt50	1841797	100 %	39.1	1951	CP194174
F1B	1839155	100 %	39.1	1949	CP194006
F2A	1839161	100 %	39.1	1947	CP194005

Supplementary Table S3: CRISPR loci and spacer array profiles of *S. thermophilus* UCCSt50 and BIMs F1B and F2A

Strain	**C1 (bp)**	**C1 (spacers)**	**C2 (bp)**	**C2 (spacers)**	**C3 (bp)**	**C3 (spacers)**
UCCSt50	2,409	36	101	1	1,358	20
F1B	2,409	36	101	1	1,358	20
F2A	2,409	36	101	1	1,358	20

Supplementary Table S4: NMR data for F1B polysaccharide (600 MHz, 40 °C, ppm).

Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
α-Rha A	H	5.07	4.27	3.96	3.57	3.80	1.32
	C	100.8	77.6	79.5	72.8	71.1	17.9
α-Glc B	H	4.98	3.52	3.72	3.57	4.22	3.94; 3.94
	C	99.2	72.7	74.2	70.3	71.5	67.4
α-Rha C	H	4.92	4.05	3.92	3.47	3.76	1.34
	C	100.8	79.3	71.3	73.6	70.0	18.1
β-Gal D	H	4.64	3.55	3.67	3.93	3.68	3.75; 3.75
	C	105.2	72.9	74.1	69.8	76.4	62.2
α-Rha F	H	4.87	3.81	3.74	3.43	3.99	1.24
	C	102.5	71.9	71.5	73.2	70.1	17.7
α-Rha F'	H	4.87	3.82	3.96	3.69	4.06	1.32
	C	102.5	71.9	71.5	82.3	68.7	17.9
β-GlcNAc G	H	4.74	3.74	3.64	3.47	3.47	3.75; 3.96
	C	103.9	56.9	82.7	70.3	77.1	62.2

References

1. Iyer R, Tomar SK, Uma Maheswari T, Singh R. *Streptococcus thermophilus* strains: Multifunctional lactic acid bacteria. International Dairy Journal. 2010 Mar;20(3):133–41. doi:10.1016/j.idairyj.2009.10.005

2. Martinović A, Cocuzzi R, Arioli S, Mora D. *Streptococcus thermophilus*: To Survive, or Not to Survive the Gastrointestinal Tract, That Is the Question! Nutrients. 2020 Jul 22;12(8):2175. doi:10.3390/nu12082175

3. Broadbent JR, McMahon DJ, Welker DL, Oberg CJ, Moineau S. Biochemistry, Genetics, and Applications of Exopolysaccharide Production in *Streptococcus thermophilus*: A Review. Journal of Dairy Science. 2003 Feb;86(2):407–23. doi:10.3168/jds.S0022-0302(03)73619-4

4. Lluis-Arroyo D, Flores-Nájera A, Cruz-Guerrero A, Gallardo-Escamilla F, Lobato-Calleros C, Jiménez-Guzmán J, et al. Effect of an Exopolysaccharide-Producing Strain of *Streptococcus thermophilus* on the Yield and Texture of Mexican Manchego-Type Cheese. International Journal of Food Properties. 2014 Sep 14;17(8):1680–93. doi:10.1080/10942912.2011.599091

5. Garneau JE, Moineau S. Bacteriophages of lactic acid bacteria and their impact on milk fermentations. Microb Cell Fact. 2011;10(Suppl 1):S20. doi:10.1186/1475-2859-10-S1-S20

6. Le Marrec C, Van Sinderen D, Walsh L, Stanley E, Vlegels E, Moineau S, et al. Two groups of bacteriophages infecting *Streptococcus thermophilus* can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. Appl Environ Microbiol. 1997 Aug;63(8):3246–53. doi:10.1128/aem.63.8.3246-3253.1997

7. Adriaenssens EM, Wittmann J, Kuhn JH, Turner D, Sullivan MB, Dutilh BE, et al. Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch Virol. 2018 Apr;163(4):1125–9. doi:10.1007/s00705-018-3723-z

8. Mills S, Griffin C, O'Sullivan O, Coffey A, McAuliffe OE, Meijer WC, et al. A new phage on the 'Mozzarella' block: Bacteriophage 5093 shares a low level of homology with other *Streptococcus thermophilus* phages. International Dairy Journal. 2011 Dec;21(12):963–9. doi:10.1016/j.idairyj.2011.06.003

9. McDonnell B, Mahony J, Neve H, Hanemaaijer L, Noben JP, Kouwen T, et al. Identification and Analysis of a Novel Group of Bacteriophages Infecting the Lactic Acid Bacterium *Streptococcus thermophilus*. Pettinari MJ, editor. Appl Environ Microbiol. 2016 Sep;82(17):5153–65. doi:10.1128/AEM.00835-16

10. Szymczak P, Janzen T, Neves AR, Kot W, Hansen LH, Lametsch R, et al. Novel Variants of *Streptococcus thermophilus* Bacteriophages Are Indicative of Genetic Recombination among Phages from Different Bacterial Species. Dudley EG, editor. Appl Environ Microbiol. 2017 Mar;83(5):e02748-16. doi:10.1128/AEM.02748-16

11. Philippe C, Levesque S, Dion MB, Tremblay DM, Horvath P, Lüth N, et al. Novel Genus of Phages Infecting *Streptococcus thermophilus*: Genomic and Morphological Characterization. Ercolini D, editor. Appl Environ Microbiol. 2020 Jun 17;86(13):e00227-20. doi:10.1128/AEM.00227-20

12. McDonnell B, Mahony J, Hanemaaijer L, Neve H, Noben JP, Lugli GA, et al. Global Survey and Genome Exploration of Bacteriophages Infecting the Lactic Acid Bacterium *Streptococcus thermophilus*. Front Microbiol. 2017 Sep 12;8:1754. doi:10.3389/fmicb.2017.01754

13. Hanemaaijer L, Kelleher P, Neve H, Franz C, De Waal P, Van Peij N, et al. Biodiversity of Phages Infecting the Dairy Bacterium *Streptococcus thermophilus*. Microorganisms. 2021 Aug 27;9(9):1822. doi:10.3390/microorganisms9091822

14. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007 Mar 23;315(5819):1709–12. doi:10.1126/science.1138140

15. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, et al. Phage Response to CRISPR-Encoded Resistance in *Streptococcus thermophilus*. J Bacteriol. 2008 Feb 15;190(4):1390–400. doi:10.1128/JB.01412-07

16. Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in *Streptococcus thermophilus*. J Bacteriol. 2008 Feb 15;190(4):1401–12. doi:10.1128/JB.01415-07

17. Philippe C, Moineau S. The endless battle between phages and CRISPR–Cas systems in *Streptococcus thermophilus*. Biochem Cell Biol. 2021 Aug;99(4):397–402. doi:10.1139/bcb-2020-0593

18. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010 May;8(5):317–27. doi:10.1038/nrmicro2315

19. Lavelle K, McDonnell B, Fitzgerald G, van Sinderen D, Mahony J. Bacteriophage-host interactions in *Streptococcus thermophilus* and their impact on co-evolutionary processes. FEMS Microbiology Reviews. 2023 Jul 5;47(4):fuad032. doi:10.1093/femsre/fuad032

20. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for

bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Molecular Microbiology. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494

21. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. Cell Wall Glycans Mediate Recognition of the Dairy Bacterium *Streptococcus thermophilus* by Bacteriophages. Dudley EG, editor. Appl Environ Microbiol. 2018 Dec;84(23):e01847-18. doi:10.1128/AEM.01847-18

22. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in *Streptococcus thermophilus* bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

23. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Alexandre G, editor. Appl Environ Microbiol. 2022 Jan 11;88(1):e01723-21. doi:10.1128/AEM.01723-21

24. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

25. Lavelle K, Goulet A, McDonnell B, Spinelli S, Van Sinderen D, Mahony J, et al. Revisiting the host adhesion determinants of *Streptococcus thermophilus* siphophages. Microbial Biotechnology. 2020 Nov;13(6):1765–79. doi:10.1111/1751-7915.13593

26. McDonnell B, Mahony J, Hanemaaijer L, Kouwen TRHM, Van Sinderen D. Generation of Bacteriophage-Insensitive Mutants of *Streptococcus thermophilus* via an Antisense RNA CRISPR-Cas Silencing Approach. Elkins CA, editor. Appl Environ Microbiol. 2018 Feb 15;84(4):e01733-17. doi:10.1128/AEM.01733-17

27. Guérin H, Kulakauskas S, Chapot-Chartier MP. Structural variations and roles of rhamnose-rich cell wall polysaccharides in Gram-positive bacteria. Journal of Biological Chemistry. 2022 Oct;298(10):102488. doi:10.1016/j.jbc.2022.102488

28. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. Bipartite *rgp* Locus Diversity in *Streptococcus thermophilus* Corresponds to Backbone and Side Chain Differences of Its Rhamnose-Containing Cell Wall Polysaccharide. Ercolini D, editor. Appl Environ Microbiol. 2022 Dec 13;88(23):e01504-22. doi:10.1128/aem.01504-22

29. Ramos A, Boels IC, De Vos WM, Santos H. Relationship between Glycolysis and Exopolysaccharide Biosynthesis in *Lactococcus lactis*. Appl Environ Microbiol. 2001 Jan;67(1):33–41. doi:10.1128/AEM.67.1.33-41.2001

30. McGrath S, Fitzgerald GF, Van Sinderen D. Improvement and Optimization of Two Engineered Phage Resistance Mechanisms in *Lactococcus lactis*. Appl Environ Microbiol. 2001 Feb;67(2):608–16. doi:10.1128/AEM.67.2.608-616.2001

31. Lavelle K, Martinez I, Neve H, Lugli GA, Franz CMAP, Ventura M, et al. Biodiversity of *Streptococcus thermophilus* Phages in Global Dairy Fermentations. Viruses. 2018 Oct 22;10(10):577. doi:10.3390/v10100577

32. Zhang H, Zhu F, Ding L, Zhou M, Wu R, Wu H. Preliminary X-ray crystallographic studies of an N-terminal domain of unknown function from a putative glycosyltransferase from *Streptococcus parasanguinis*. Acta Crystallogr F Struct Biol Cryst Commun. 2013 May 1;69(5):520–3. doi:10.1107/S1744309113007161

33. Kampff Z, Van Sinderen D, Mahony J. Cell wall polysaccharides of streptococci: A genetic and structural perspective. Biotechnology Advances. 2023 Dec;69:108279. doi:10.1016/j.biotechadv.2023.108279

34. Martín V, Mañes-Lázaro R, Rodríguez JM, Maldonado-Barragán A. *Streptococcus lactarius* sp. nov., isolated from breast milk of healthy women. International Journal of Systematic and Evolutionary Microbiology. 2011 May 1;61(5):1048–52. doi:10.1099/ijs.0.021642-0

35. Theodorou I, Courtin P, Palussière S, Kulakauskas S, Bidnenko E, Péchoux C, et al. A dual-chain assembly pathway generates the high structural diversity of cell-wall polysaccharides in *Lactococcus lactis*. Journal of Biological Chemistry. 2019 Nov;294(46):17612–25. doi:10.1074/jbc.RA119.009957

36. Ruta S, Murray M, Kampff Z, McDonnell B, Lugli GA, Ventura M, et al. Microbial Ecology of Pecorino Siciliano PDO Cheese Production Systems. Fermentation. 2023 Jun 29;9(7):620. doi:10.3390/fermentation9070620

37. Chapot-Chartier MP, Vinogradov E, Sadovskaya I, Andre G, Mistou MY, Trieu-Cuot P, et al. Cell Surface of *Lactococcus lactis* Is Covered by a Protective Polysaccharide Pellicle. Journal of Biological Chemistry. 2010 Apr;285(14):10464–71. doi:10.1074/jbc.M109.082958

38. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, et al. Differences in Lactococcal Cell Wall Polysaccharide Structure Are Major Determining Factors in Bacteriophage Sensitivity. Hendrix R, editor. mBio. 2014 Jul;5(3):e00880-14. doi:10.1128/mBio.00880-14

39. Sadovskaya I, Vinogradov E, Courtin P, Armalyte J, Meyrand M, Giaouris E, et al. Another Brick in the Wall: a Rhamnan Polysaccharide Trapped inside Peptidoglycan of *Lactococcus lactis*. Turner MS, Biswas I, editors. mBio. 2017 Nov 8;8(5):e01303-17. doi:10.1128/mBio.01303-17

40. Guérin H, Quénée P, Palussière S, Courtin P, André G, Péchoux C, et al. PBP2b Mutations Improve the Growth of Phage-Resistant *Lactococcus cremoris*

Lacking Polysaccharide Pellicle. Dozois CM, editor. Appl Environ Microbiol. 2023 Jun 28;89(6):e02103-22. doi:10.1128/aem.02103-22

41. De A, Liao S, Bitoun JP, Roth R, Beatty WL, Wu H, et al. Deficiency of RgpG Causes Major Defects in Cell Division and Biofilm Formation, and Deficiency of LytR-CpsA-Psr Family Proteins Leads to Accumulation of Cell Wall Antigens in Culture Medium by *Streptococcus mutans*. Nojiri H, editor. Appl Environ Microbiol. 2017 Sep;83(17):e00928-17. doi:10.1128/AEM.00928-17

42. Kovacs CJ, Faustoferri RC, Bischer AP, Quivey RG. *Streptococcus mutans* requires mature rhamnose-glucose polysaccharides for proper pathophysiology, morphogenesis and cellular division. Molecular Microbiology. 2019 Sep;112(3):944–59. doi:10.1111/mmi.14330

43. van Sorge NM, Cole JN, Kuipers K, Henningham A, Aziz RK, Kasirer-Friede A, et al. The Classical Lancefield Antigen of Group A *Streptococcus* Is a Virulence Determinant with Implications for Vaccine Design. Cell Host & Microbe. 2014 Jun;15(6):729–40. doi:10.1016/j.chom.2014.05.009

44. Teng JLL, Ma Y, Chen JHK, Luo R, Foo CH, Li TT, et al. Streptococcus *oriscaviae* sp. nov. Infection Associated with Guinea Pigs. Kovac J, editor. Microbiol Spectr. 2022 Jun 29;10(3):e00014-22. doi:10.1128/spectrum.00014-22

45. Xu Y, Singh KV, Qin X, Murray BE, Weinstock GM. Analysis of a Gene Cluster of *Enterococcus faecalis* Involved in Polysaccharide Biosynthesis. Tuomanen EI, editor. Infect Immun. 2000 Feb;68(2):815–23. doi:10.1128/IAI.68.2.815-823.2000

46. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, et al. The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide

structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. Molecular Microbiology. 2020 Oct;114(4):582–96. doi:10.1111/mmi.14561

47. Dupont K, Janzen T, Vogensen FK, Josephsen J, Stuer-Lauridsen B. Identification of *Lactococcus lactis* Genes Required for Bacteriophage Adsorption. Appl Environ Microbiol. 2004 Oct;70(10):5825–32. doi:10.1128/AEM.70.10.5825-5832.2004

48. Guérin H, Courtin P, Guillot A, Péchoux C, Mahony J, Van Sinderen D, et al. Molecular mechanisms underlying the structural diversity of rhamnose-rich cell wall polysaccharides in lactococci. Journal of Biological Chemistry. 2024 Jan;300(1):105578. doi:10.1016/j.jbc.2023.105578

49. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, et al. Investigation of the Relationship between Lactococcal Host Cell Wall Polysaccharide Genotype and 936 Phage Receptor Binding Protein Phylogeny. Appl Environ Microbiol. 2013 Jul 15;79(14):4385–92. doi:10.1128/AEM.00653-13

50. Shibata Y, Yamashita Y, Van Der Ploeg JR. The serotype-specific glucose side chain of rhamnose–glucose polysaccharides is essential for adsorption of bacteriophage M102 to *Streptococcus mutans*. FEMS Microbiology Letters. 2009 May;294(1):68–73. doi:10.1111/j.1574-6968.2009.01546.x

51. Ho K, Huo W, Pas S, Dao R, Palmer KL. Loss-of-Function Mutations in *epaR* Confer Resistance to ϕNPV1 Infection in *Enterococcus faecalis* OG1RF. Antimicrob Agents Chemother. 2018 Oct;62(10):e00758-18. doi:10.1128/AAC.00758-18

52. Giesbers CAP, Fagan J, Parlindungan E, Palussière S, Courtin P, Lugli GA, et al. Reduced synthesis of phospho-polysaccharide in *Lactococcus* as a strategy to

evade phage infection. International Journal of Food Microbiology. 2023 Dec;407:110415. doi:10.1016/j.ijfoodmicro.2023.110415

53. Kelleher P, Ortiz Charneco G, Kampff Z, Diaz-Garrido N, Bottacini F, McDonnell B, et al. Phage defence loci of *Streptococcus thermophilus*— tip of the anti-phage iceberg? Nucleic Acids Research. 2024 Oct 28;52(19):11853–69. doi:10.1093/nar/gkae814

54. Pastuszka A, Rousseau GM, Somerville V, Levesque S, Fiset JP, Goulet A, et al. Dairy phages escape CRISPR defence of *Streptococcus thermophilus* via the anti-CRISPR AcrIIA3. International Journal of Food Microbiology. 2023 Dec;407:110414. doi:10.1016/j.ijfoodmicro.2023.110414

55. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits *Streptococcus pyogenes* Cas9. Nat Microbiol. 2017 Aug 7;2(10):1374–80. doi:10.1038/s41564-017-0004-7

56. Lillehaug D. An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. J Appl Microbiol. 1997 Jun;83(1):85–90. doi:10.1046/j.1365-2672.1997.00193.x

57. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Birol I, editor. Bioinformatics. 2019 Jul 1;35(13):2193–8. doi:10.1093/bioinformatics/bty841

58. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60. doi:10.1093/bioinformatics/btp324

59. Wick RR, Holt KE. Polypolish: Short-read polishing of long-read bacterial genome assemblies. Schneidman-Duhovny D, editor. PLoS Comput Biol. 2022 Jan 24;18(1):e1009802. doi:10.1371/journal.pcbi.1009802

60. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microbial Genomics. 2021 Nov 30;7(11). doi:10.1099/mgen.0.000685

61. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Research. 2018 Jul 2;46(W1):W282–8. doi:10.1093/nar/gky467

62. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research. 2018 Jul 2;46(W1):W246–51. doi:10.1093/nar/gky425

63. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9. doi:10.1038/nmeth.1923

64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9. doi:10.1093/bioinformatics/btp352

65. Froger A, Hall JE. Transformation of Plasmid DNA into E. coli Using the Heat Shock Method. JoVE. 2007 Aug 1;(6):253. doi:10.3791/253

66. Wickham H. ggplot2. WIREs Computational Stats. 2011 Mar;3(2):180–5. doi:10.1002/wics.147

67. Graves S, Piepho H, Selzer L, Dorai-Raj S. multcompView: visualizations of paired comparisons. 2019.

Chapter III

Cell wall polysaccharides are a diverse, yet core component across salivarius group streptococci

Contributions: Irina Sadovskaya and Evgeny Vinogradov: Biochemical analysis and NMR spectroscopy. Hilary Browne, Francesco Vacca and Kelsey White: Phylogenetic and bioinformatic analyses. Anne de Jong: RNA extraction, RNA-Seq library preparation and sequencing.

Abstract:

Streptococcus is one of the earliest identified bacterial genera, and over its ~150 year history, has been extensively studied in the context of human health, pathogenesis and food fermentations. Several phylogenetic groups of streptococcal species have been identified in recent decades. In this study, we performed the most extensive analysis to date, incorporating almost 5,000 streptococcal genomes and metagenome-assembled genomes employed. This allowed the identification of eleven distinct phylogenetic groups. Among these groups, members of the salivarius group are among the earliest human colonisers. Cell wall-associated rhamnose-glucose polysaccharides (RGPs) are defined as major structural components of the cell envelope of certain streptococcal species. In this study, we combined comparative genomics, structural analysis, and transcriptional profiling to characterise RGP biosynthesis across the genus, using the salivarius group as a proxy for the broader genus. Comparative genomic analysis identified genes associated with RGP precursor synthesis and biosynthesis across the majority of analysed streptococcal species, indicating that RGPs likely represent a highly conserved feature among streptococci. Further comparative genomic analysis of 223 *rgp* loci from *Streptococcus thermophilus*, *Streptococcus salivarius*, *Streptococcus vestibularis* and *Streptococcus raffinosi* identified 14 distinct *rgp* genotypes together with multiple novel backbone-associated (Bt) and variable side chain-associated (Vt) sub-genotypes. Despite diversity in gene content, all *rgp* loci displayed a conserved modular organisation and genomic position. Structural elucidation of five previously uncharacterised RGPs significantly expanded the known structural repertoire of salivarius group RGPs and established clear relationships between *rgp* gene content and polysaccharide structure.

Transcriptional mapping demonstrates that *rgp* loci are organised into discrete transcriptional units that underpin their modular nature and unique organisation. Collectively, these findings demonstrate that RGP biosynthesis in streptococci is both conserved and highly adaptable, shaped by species-specific and niche-specific pressures. This work provides a foundation for future exploration into streptococcal cell envelope-associated polysaccharide diversity and the influence on host interactions, ecological adaptation and bacterial cell physiology across the genus.

Introduction

Streptococcus species are among the earliest colonisers of the human oral cavity and play a pivotal role in establishing the oral microbiota. *Streptococcus* is a genus that encompasses over 140 validated species that have traditionally been classified based on their serological and biochemical properties or their haemolytic capability, i.e. α-haemolytic, β-haemolytic or non-haemolytic (1). Among the α-haemolytic streptococci are the viridans streptococci, which are considered to be of low pathogenicity (2). The viridans streptococci are described to incorporate six biochemically distinct groups, i.e. the mitis, mutans, salivarius, anginosus, sanguinis and bovis groups of streptococci. However, streptococcal taxonomy remains a dynamic issue, with multiple analytical approaches and increasingly large genome datasets contributing to ongoing reclassification (1,3–6). While the traditional group names anginosus, bovis, mitis, mutans, pyogenic and salivarius have been widely recognised (7), more recent studies have reported between six and nine taxonomic groups depending on dataset composition, genome retrieval strategies, and the use of different pangenome pipelines. For example, based on phylogenetic analysis of core genes of >60 genomes of streptococci, nine taxonomic groups have been proposed, namely the bovis, gordonii, mitis, mutans, pluranimalium, pyogenic, salivarius, sobrinus, and suis groups (8). Additionally, analysis of core genes from 3,052 streptococcal genomes has identified seven taxonomic groups, with high order branches according to the traditional taxonomic group names and the addition of the suis group (6). Members of the salivarius group are predominantly commensal organisms in contrast to the majority of streptococcal species belonging to other biochemical groups, which are more commonly associated with pathogenic lifestyles. The salivarius group includes *Streptococcus*

salivarius, Streptococcus thermophilus, *Streptococcus vestibularis* (9) and the recently identified species, *Streptococcus raffinosi* (10).

The cell envelope of streptococci contains, next to its cytoplasmic membrane and peptidoglycan layer, a complex matrix of proteins, teichoic acids and polysaccharides. The latter include capsular or exopolysaccharides (EPSs) and cell wall-associated rhamnose-glucose polysaccharides (RGPs). Rhamnose-rich polysaccharides represent a major cell envelope component across several ovococcoid species, including streptococci, enterococci and lactococci (11–13). The most thoroughly characterised streptococcal species with respect to their cell envelope-associated polysaccharide composition and biosynthesis include *Streptococcus mutans* (14–22)*, Streptococcus pyogenes* (23–30), *Streptococcus agalactiae* (31–37) and *Streptococcus thermophilus* (38–43). In these species, RGPs constitute a major cell wall-linked glycopolymer that may serve as functional replacements for wall teichoic acids, being implicated in cell morphology, cell division, and biofilm formation processes in pathogenic species (44). The complete RGP structures of these well characterised species are composed of a polyrhamnose backbone decorated with species- and serotype-specific, glucose-containing side chains, which contribute to antigenic diversity and host immune recognition (26). Functional studies in *S. mutans* have demonstrated that RGPs are essential for maintaining cell envelope integrity, coordinating septum placement during cell division, and mediating tolerance to environmental stresses relevant to pathogenesis (15). Moreover, RGP structural features, particularly side chain modifications, can influence the highly specific interactions with between *S. thermophilus* and its infecting bacteriophages (39,40,45–47), while RGPs also serve as receptors for bacteriophage-encoded endolysins in *Streptococcus pyogenes* (48). Additionally, in

S. mutans, the α-1,2-linked glucose side chain of serotype *c* is recognised by specific phages, being essential for their adsorption (17).

S. thermophilus has become a useful paradigm for the study of cell envelope-associated polysaccharides among streptococci and, indeed, among the broader ovococcoid Gram-positive bacteria. The *rgp* locus encodes the biosynthetic machinery for RGP production (16,38). Seven distinct *rgp* genotypes (termed *rgp*1-7) are currently known to exist among *S. thermophilus* strains based on the gene content associated with the biosynthesis of the backbone and side chain units (38). Detailed analysis of the rhamnan and side chain (biosynthesis)-associated regions of the *rgp* locus has revealed three distinct backbone genotypes (Bt1-3) and five variable genotypes (Vt1-5) (38). To date, the RGP structures of five *S. thermophilus* strains have been elucidated, representing distinct *rgp* genotypes, *rgp*1, 2, 3, 4, and 6 (38,39,47). Comparative analysis of these structures demonstrated that highly conserved genetic regions (≥ 95 %) of the *rgp* locus are matched by identical structures while distinct or variable regions are associated with the unique components of the complete RGP structure (38). However, as the availability of *S. thermophilus* genome sequences has increased, it has become apparent that the true extent of *rgp* locus diversity extends beyond the previously established framework. Beyond *S. thermophilus*, it is currently not defined if all species belonging to the salivarius group possess *rgp* loci and if so, how these resemble or differ from those of *S. thermophilus*. Furthermore, it is not defined if the *rgp* locus is a conserved component across all streptococcal species (notwithstanding likely genetic diversity).

The genomes of several thousand streptococcal isolates and metagenome assembled genomes are currently available (49–52). Therefore, to establish if the phylogeny

of these groups remains valid, broad scale comparisons of all available high quality sequence information was undertaken. Multiple streptococcal species are known to form part of the human oral microbiota including *S. salivarius, Streptococcus mitis, S. mutans* and *Streptococcus gordonii*, among others. Extending analysis of the *rgp* loci encoded by all members of the salivarius group streptococci (among others) is essential to improve our understanding of the broader significance of these polysaccharides across streptococci and Gram-positive bacteria.

In the current study, we performed phylogenetic analysis of ~5,000 high quality streptococcal genomes to establish the validity of the currently proposed taxonomic groups of streptococci. We also evaluated the presence, diversity and inter-species relationships of the *rgp* loci of members of the salivarius group streptococci. The chemical structures of five strains representing distinct *rgp* genotypes were elucidated. Finally, using computational promoter and terminator predictions and RNA-Seq analysis, we present the first insights into the transcriptional units of *rgp* loci in salivarius group streptococci. Collectively, this work provides the most extensive genetic, transcriptional and structural analysis of streptococcal RGPs to date, and establishes a foundation for predictive, genotype-based understanding of cell envelope-associated polysaccharide diversity with important implications for core biological functions of streptococci.

Results

Streptococcus species can be classified into eleven phylogenetic groups

The *Streptococcus* genus has been reported to be comprised of between six and nine major phylogenetic groups (1,4,6,7). With an ever-growing repository of genome sequences, it was possible to conduct the most comprehensive phylogenetic analysis of the streptococcal genus through the pangenome analysis of 4,966 high quality streptococcal genomes and with an added lactococcal outgroup (Supplementary Table S1). This dataset consists of both high-quality MAGs (2,737) and genomes of isolated streptococcal strains (2,229) retrieved from a diverse set of databases including SPIRE (a Searchable Planetary-scale mIcrobiome Resource) (50), curated FoodMetagenomicData (cFMD) (52), NCBI Genome database, and MGnify Genome catalogues (Unified Human Gastrointestinal Genome (UHGG) collection and Skin biome collection) (49,51). A phylogenetic analysis of the relaxed core (shared by >90 % of all genomes) genome consisting of 558 single-copy genes allowed for the classification of the *Streptococcus* genus into eleven phylogenetic groups i.e. salivarius, bovis, pyogenic, mitis, mutans, suis, sobrinus, orisratti, pluranimalium, parauberis and agalactiae (Fig. 1). A recent study of 3,052 representative streptococcal genomes retrieved from RefSeq determined that the genus had a relaxed core genome of 292 single-copy genes (shared across > 95 % of all genomes) and classified the genus into seven phylogenetic groups (6). Taxonomic classification using GTDB-TK (53) assigned the 4,966 genomes to 315 species-level taxa, with 48 genomes classified only to the genus level (Supplementary Table S1). These 315 species included draft species names that have not been validly published under the International Code of Nomenclature of Prokaryotes (ICNP). When restricted to only validly published ICNP nomenclature

(limited to the species level), the dataset comprised 135 recognised *Streptococcus* species, with *S. thermophilus* being the most abundant species (n=993), followed by *S. pyogenes* (n= 430) and *Streptococcus pneumoniae* (n=376). This is a significant expansion from a recent study spanning ~ 80 species of *Streptococcus* (6). Additionally, 229 genomes belong to currently potentially novel or, as yet, unvalidated species demonstrating the diversity of the *Streptococcus* genus.

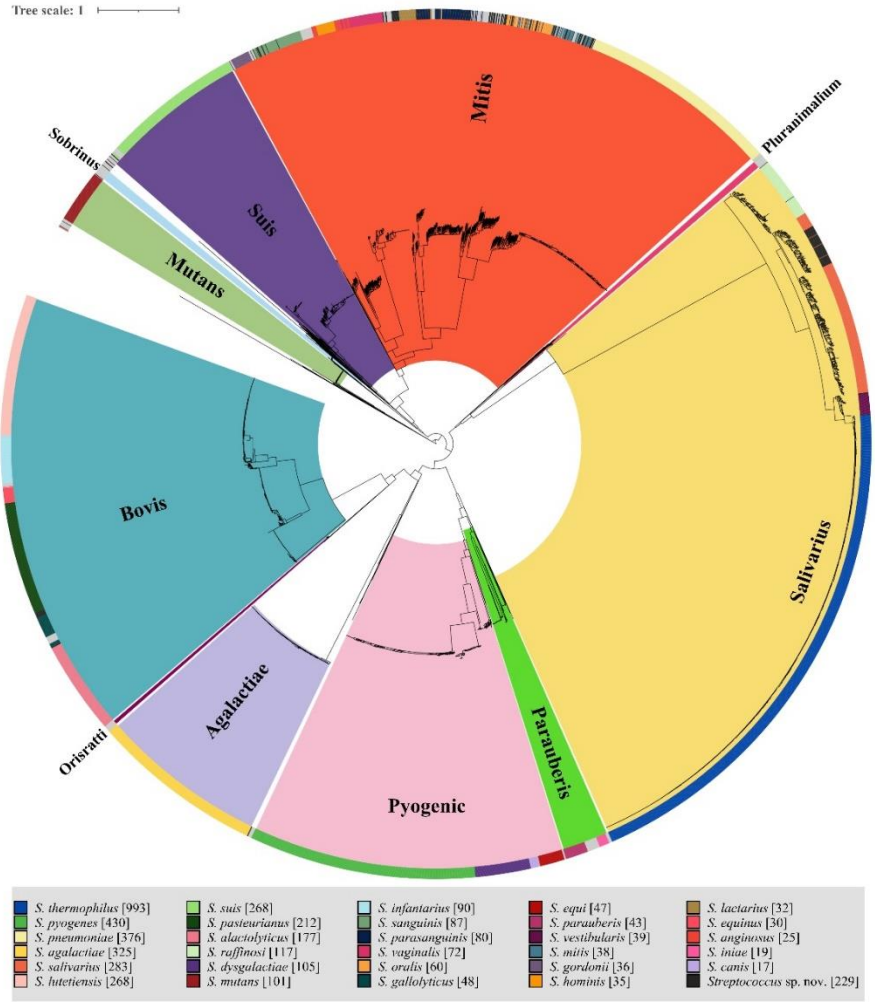


Fig. 1. Phylogenetic tree of the *Streptococcus* genus constructed using pangenome analysis of 4,966 streptococcal genomes. The inner ring indicates the phylogenetic groups which are colour-coded and labelled. The outer ring indicates representative streptococcal species within the dataset. The analysis supports the division of the genus into eleven phylogenetic groups.

Gene clusters associated with RGP synthesis are a prevalent feature among streptococci

Cell surface-exposed molecules, and in particular saccharidic moieties such as RGPs, have historically been a differentiating attribute of streptococci, as they underpin the Lancefield classification of streptococci. The gene clusters that encode the biosynthetic functions associated with RGP structures are described to be diverse in well characterised species such as *S. mutans, S. agalactiae* and *S. thermophilus* (16,38). These gene clusters are not believed to be universally present across all streptococci, however. For example, they are believed to be absent in *S. pneumoniae* whose cell envelope instead incorporates wall teichoic acids and lipoteichoic acids, and a capsular cover (28, 29). This capsular structure is highly variable among pneumococcal strains and has enabled the identification of at least 90 serotypes (56). The *rgp* clusters are comprised of two functional modules; one which encodes the biosynthetic functions of the peptidoglycan-embedded rhamnan backbone component and the second which is associated with the biosynthesis of a surface-exposed side chain structure. The organisation of the rhamnan- and side chain associated modules within the *rgp* gene clusters differ between *S. thermophilus* and *S. mutans rgp* loci (38,43); however, those of other species are not well investigated or defined.

To establish the possible presence and prevalence of *rgp* loci among species within the genus *Streptococcus*, we applied a consistent annotation to reference genomes of 150 species from the NCBI Genome database and performed an automated search for genes associated with RGP precursor synthesis (*rmlA, rmlB, rmlC, rmlD*), rhamnan backbone biosynthesis (*rgpA*) and side chain biosynthesis

(DUF2304/*wpsB*). Gene identification was based on an annotation-based search using a custom Python script that screened each Bakta-annotated genome for predefined gene names and annotated gene products associated with RGP biosynthesis (see Methods section). For each genome, the script reported the presence or absence of each target gene based on these annotations. This analysis established that 137 reference genomes of the 150 analysed were positive for the precursor and/or RGP structure biosynthetic genes (Fig. 2). Therefore, it is highly likely that RGPs are, in fact, a typical feature among most, or all, streptococci. It should be noted that 13 of the 150 reference genomes did not possess any of the targeted RGP-associated genes identified in this analysis. This should be interpreted with caution, as the annotation-based search strategy relied on exact matches to predefined gene names and annotated products (see Methods section). Consequently, although consistent annotation (using Bakta) was applied for all analysed genomes, RGP-associated genes may not have been identified if they were assigned alternative annotations or remained unannotated. Additional analyses of these genomes are therefore required to determine whether they possess cell wall polysaccharide biosynthetic loci. We also acknowledge that the annotations of the genes within these loci in the analysed reference genomes may not be complete and represents a limitation of this analysis; however, the widespread nature of the precursor synthesis genes supports the hypothesis of their ubiquitous presence across the genus.

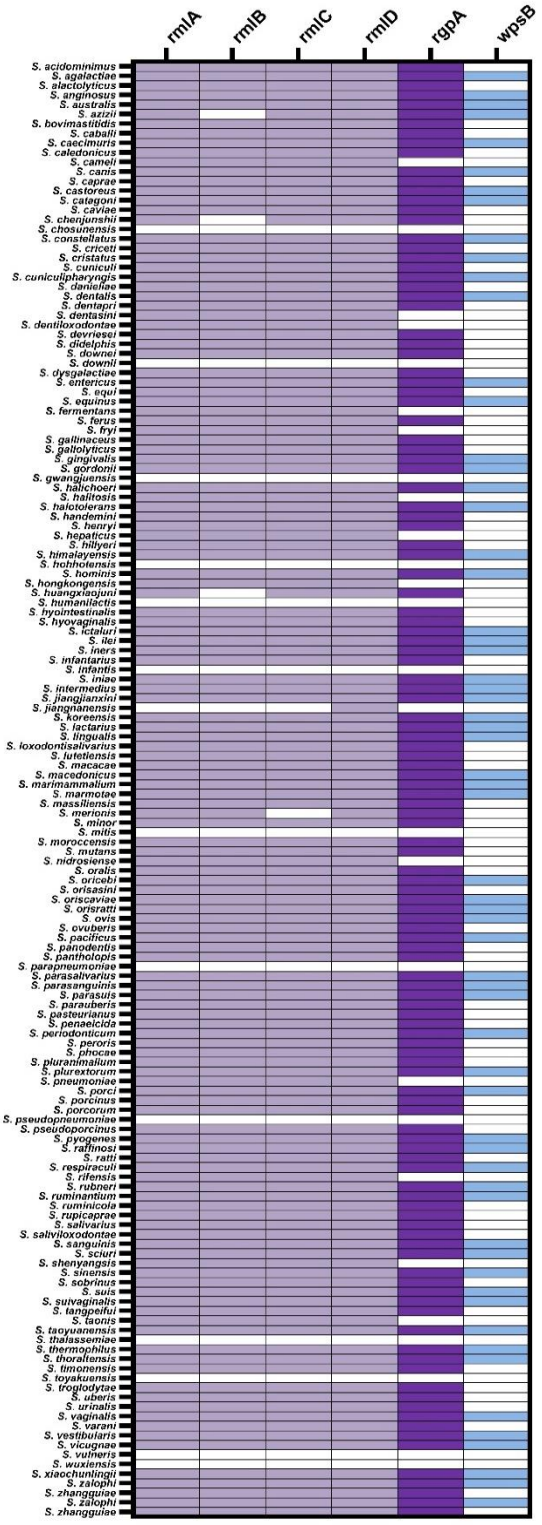


Fig. 2. Presence or absence of genes associated with RGP precursor synthesis, rhamnan backbone biosynthesis, and side chain biosynthesis in 150 high-quality *Streptococcus* reference genomes. Presence of genes associated with RGP precursor synthesis (*rmlA, rmlB, rmlC, rmlD*) is indicated by a pale purple colour, rhamnan backbone biosynthesis (*rgpA*) by a darker purple, and side chain biosynthesis (DUF2304/*wpsB*) by a light blue. The absence of a particular target gene is indicated by white annotation in the figure.

Exploring the diversity of *rgp* loci of salivarius group streptococci

Among the eleven identified streptococcal groups, members of the salivarius group represent approximately one third of genomes analysed (Fig. 1). This paired with previous studies describing the genetic and structural diversity of RGPs in *S. thermophilus*, prompted a deeper interrogation of the diversity of the *rgp* loci of members of the broader salivarius group as a proxy for the *Streptococcus* genus. The *rgp* loci derived from 223 high-quality, complete genomes or scaffolds of members of the salivarius group available in the NCBI database (13 of which were sequenced as part of this study) were identified and compared using hierarchical clustering (HCL) analysis (Supplementary Table S2). This incorporated the *rgp* loci of *S. raffinosi* (n=4), *S. salivarius* (n=39), *S. thermophilus* (n=168) and *S. vestibularis* (n=12). Through this HCL analysis, 14 *rgp* genotypes were identified (designated here as *rgp*1-14) based on the presence of distinct gene families within the established genotypes (Fig. 3A).

The rhamnan and side chain biosynthetic modules within the *rgp* loci of *S. thermophilus* have been used as a basis to define three backbone (Bt1-3) and five side chain (termed the variable side chain, Vt1-5) sub-genotypes to classify strains of this species with a high degree of resolution (38). At a higher level, the unique combinations of distinct Vt and Bt sub-genotypes has allowed the identification of seven *rgp* genotypes (*rgp*1-7). In the present study, we sought to evaluate the genetic diversity of these loci and their component Bt and Vt sub-genotypes among the broader salivarius group. The organisation of the gene clusters was also evaluated to establish if the lineages of distinct genotypes exist or if there is an apparent shared evolutionary pathway of these loci across different members of the salivarius group.

All four member species were observed to harbour the locus with a similar gene organisation as described for *S. thermophilus* (38,43). The HCL analysis established that *rgp*1 is a dominant genotype among the salivarius group loci incorporating 60 strains with members from all four species, i.e. *S. raffinosi* (n=2), *S. salivarius* (n=14), *S. thermophilus* (n=36) and *S. vestibularis* (n=8) (Fig. 3A & B). The gene content across all members of the *rgp*1 genotype is highly similar with a Vt1 and Bt1 sub-genotype of their respective backbone- and side chain-associated modules (Fig. 3C). It is noteworthy that there appear to be minor gene insertion/deletion events at strain level (Fig. 3C). The *rgp*2 genotype (representing Vt1 and Bt2 sub-genotypes) includes predominantly *S. thermophilus* loci and with those of two *S. salivarius* strains (Fig. 3B). The *rgp*3 is a *S. thermophilus* specific genotype with a small number of strains (n=4) that possess a unique combination of variable and backbone genotype (Vt3 and Bt2) (Fig. 4A).

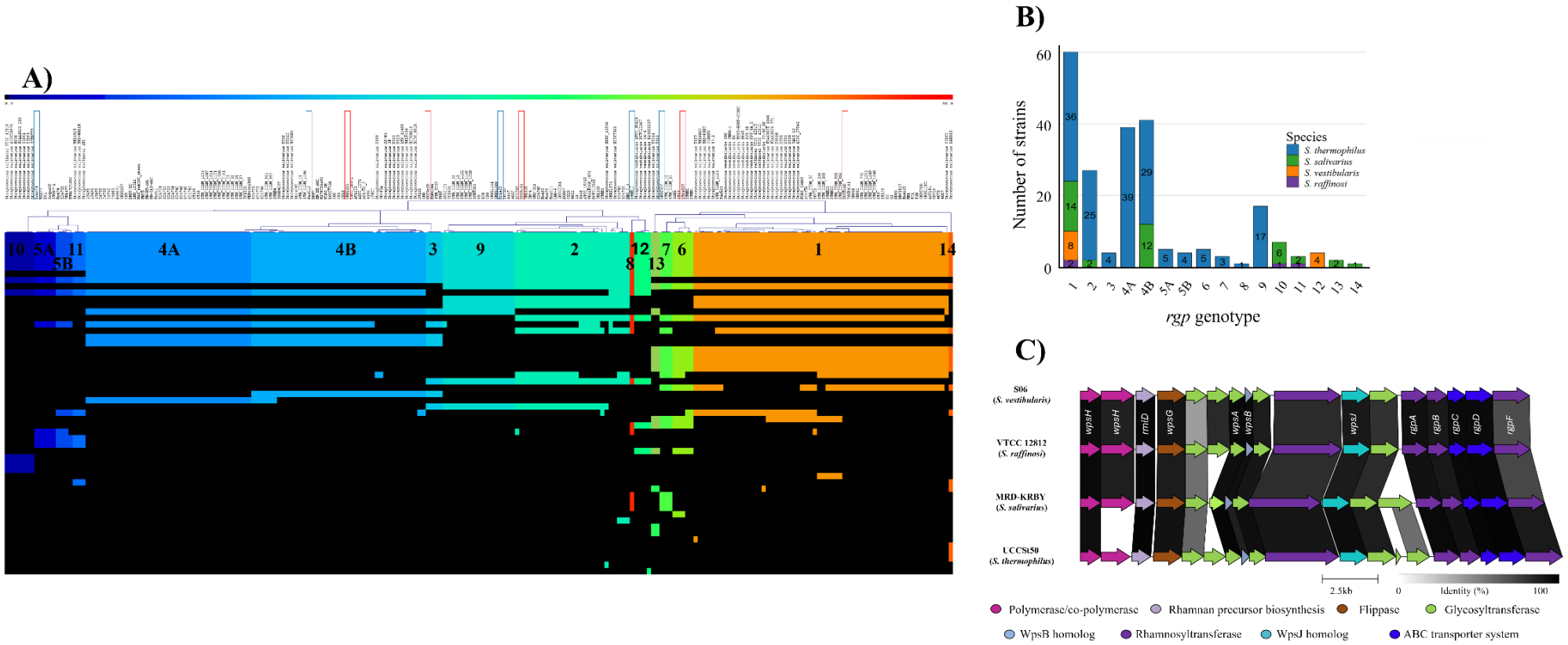


Fig. 3. **A)** HCL analysis of the *rgp* loci of 4 *S. raffinosi* strains, 39 *S. salivarius* strains, 168 *S. thermophilus* strains and 12 *S. vestibularis* strains. The heatmap was generated based on the presence (colour) or absence (black) of gene families. The assigned genetically distinct *rgp* genotypes are indicated by colour and an internal text number. Strains subjected to biochemical structural analysis either previously (38,39,47) (red) or as part of this study (blue) are indicated by a box. **B)** The distribution of all salivarius group strains analysed in this study (n=223), colour-coded by species. **C)** Schematic overview and comparative analysis of the *rgp*1 loci from salivarius group strains with representatives of each salivarius group species. Gene names (written above VTCC 12812 were proposed based on comparisons to *Lactococcus lactis cwps* clusters, including HHpred and TMHMM analysis (57). The predicted functions of genes are colour-coded and described at the base of the figure. Regions of homology are joined by blocks of different shades of grey to black and generated using CAGECAT (58).

Based on the current analysis, *rgp*4 loci (with a combination of Vt3 and Bt3 genotypes) includes two overall subtypes based on the presence of a distinct glycosyltransferase-encoding gene that is lacking in other members of the genotype, and which is predicted to encode a WpsJ (side chain transferase) within the side chain associated module (Fig. 4A). Therefore, these loci are now classified as *rgp*4A and 4B. Furthermore, *rgp*4A genotype appears to be *S. thermophilus* specific, while the *rgp*4B genotype includes representatives of *S. salivarius* (n=12) in addition to *S. thermophilus* isolates (n=29). The *rgp*4 genotype represents the largest of the salivarius group genotypes (total of n=80), consistent with previous reports (38,46). The *rgp*5 genotype is also now recognised to possess two subtypes (*rgp*5A and 5B) based on the presence of a distinct glycosyltransferase-encoding gene in members of the *rgp*5B genotype in comparison to members of *rgp*5A genotype. Members of the *rgp*5 genotype incorporate a side chain module that lacks any unique gene content and is therefore classified as Vtx while they all possess a newly identified backbone sub-genotype Bt4 (Fig. 4A). Notably, *rgp*5-7 genotypic groups harbour *S. thermophilus* loci only (Fig. 3B). Of the newly identified *rgp* genotypes, *rgp*8 & 9 members are all currently identified as *S. thermophilus* strains while *rgp*10 and *rgp*11 members belong to both *S. salivarius* and *S. raffinosi*; *rgp*12, *rgp*13 and *rgp*14 include isolates of *S. vestibularis* (*rgp*12) or *S. salivarius* (*rgp*13, *rgp*14) exclusively. These findings indicate that there are several *rgp* genotypes that are represented in multiple species, while there are also examples of species-specific *rgp* genotypes. In addition to the existing Bt and Vt sub-genotypes that are currently defined, we have identified three and six additional Bt and Vt sub-genotypes, respectively (Fig. 4A & Supplementary Table S3).

The *rgp* loci of all members of the salivarius group exhibit an identical gene context, i.e. it is located downstream of genes associated with the 30S ribosomal subunit, a DNA primase-encoding gene and *rpoD*, which encodes an RNA polymerase sigma factor. The genomic position of the *rgp* locus is also identical in all strains relative to *rpoD,* notwithstanding genomic inversions in certain strains (Fig. 4B). Notably, the organisation of the *rgp* clusters of all analysed strains of the salivarius group is identical, i.e. the 5' end of the gene cluster harbours the genes associated with the biosynthesis of the side chain structure, while the 3' region of the cluster is associated with the rhamnan backbone component (Fig. 4A).

To evaluate, as would be expected, if the assigned genotypes (or sub-genotypes) are directly linked to unique chemical structures, we elucidated the RGP structures of five streptococcal strains representing distinct and previously uncharacterised *rgp* genotypes. The selected genotypes include the following novel combinations of Bt and Vt sub-genotypes: VtxBt4 (*rgp*5), Vt2Bt2 (*rgp*9), Vt3Bt3 (*rgp*4), Vt5Bt1 (*rgp*7), Vt9Bt1 (*rgp*8). With the availability of the chemical structures of nine *rgp* genotypes (*rgp*1–9) through this and previous studies (Fig. 4C), we analysed and predicted the major structural features of unresolved *rgp* types (*rgp*10–14) as described in the ensuing section.

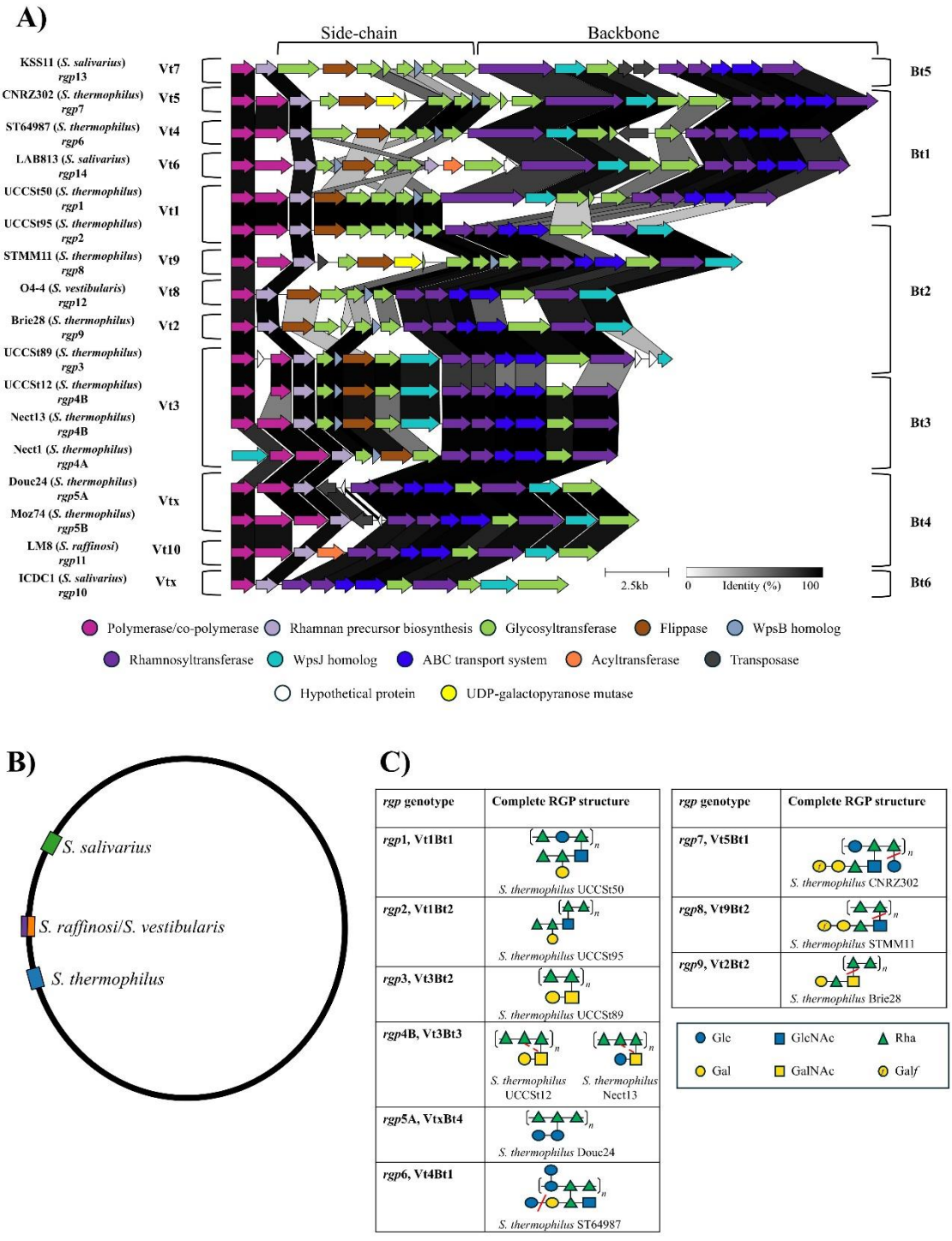


Fig. 4. A) - Schematic overview and comparative analysis of the *rgp* loci from salivarius group strains representative of distinct *rgp* genotypes (*rgp*1-14). Numbers in parentheses following strain names correspond to the *rgp* genotypes defined by HCL analysis (Fig. 3A). Regions of homology are joined by blocks of different shades of grey to black and generated using CAGECAT (58). The predicted functions of genes are colour-coded and described at the base of the figure. **B)**

Schematic representation of the genomic position of the *rgp* loci in the salivarius group species. Boxes are colour coded by species and according to Fig. 3B. **C)** Complete RGP structures of distinct Vt and Bt genotype combinations representing nine *rgp* genotypes (*rgp*1-9). Monosaccharide symbols are based on those indicated by the Standard Nomenclature for Glycans (SNFG). A red line indicates structural modifications that may or may not be present: the Bt1 structure may or may not contain a glucose residue; the Vt3 structure may be a disaccharide containing either β-Gal or a β-Glc residues and may be attached to the rhamnose backbone at different linkage points; the Vt4 side chain may be a tri- or tetrasaccharide; the Vt5 side chain may or may not have an additional glucose monosaccharide modification; the representative *rgp*8 strain, STMM11 may have a side chain component (~ 30 %); the selected *rgp*9 strain, Brie28, has limited branching with a side chain structure (~ 20 %).

Structural elucidation of RGPs representing previously uncharacterised *rgp* genotypes

To date, the RGP chemical structures of strains representing *rgp*1 (UCCSt50) (47), *rgp*2 (UCCSt95), *rgp*3 (UCCSt89), *rgp*4B (UCCSt12) (38), and *rgp*6 (ST64987) (39) have been elucidated. The rhamnose-rich RGPs of these salivarius group streptococci are structurally diverse heteropolysaccharides, typically composed of a polyrhamnose backbone decorated with strain-specific side chains that may contain N-acetyl glucosamine (GlcNAc), N-acetyl galactosamine (GalNAc), glucose (Glc), galactose (Gal), and/or rhamnose (Rha) (12,44,59). Structural diversity arises from the variation in monosaccharide composition, the number of and glycosidic linkage between monosaccharides in the repeating unit or side chain decoration, and the branching patterns.

In the present study, we significantly expand the salivarius group RGP structural repertoire by elucidating the RGP structures from *S. thermophilus* strains representing *rgp*5A (Douc24), *rgp*7 (CNRZ302), *rgp*8 (STMM11), and *rgp*9 (Brie28) genotypes. An additional *rgp*4B strain (Nect13) was also analysed in the context of this study as it contained a glycosyltransferase-encoding gene involved in RGP side chain biosynthesis whose product shares only 65 % amino acid similarity with its equivalent in UCCSt12 (Fig. 4A) (38). To summarise, the rhamnan backbone structure of two of the five strains is composed of a rhamnose disaccharide with -2-α-Rha-3-α-Rha- subunits (*rgp*8 and *rgp*9 strains), while the remaining three strains possess trisaccharide backbone subunits composed of -2-α-Rha-3-α-Rha-2-α-Rha- (*rgp*4B and *rgp*5A) or -6-α-Glc-2-α-Rha-2-α-Rha- (*rgp*7) (Table 1; Supplementary Table S4-S8, Supplementary Fig. S1 – S11). The side chain structures of each assessed strain representing the five genotypes were distinct

as predicted based on the unique glycosyltransferase-encoding gene content of each identified genotype (Fig. 3A). The side chain structures differ in composition and length, ranging from disaccharides to tetrasaccharides and are primarily composed of Glc, Gal, Rha and/or GlcNAc/GalNAc. Consistent with previous studies, four of the five side chain structures contain either a GlcNAc or GalNAc directly linked to the rhamnan backbone structure (Fig. 3C). The exception to this general observation is Douc24 representing *rgp*5A, whose side chain structure does not commence with either of these amino sugars and which is consistent with the unique gene composition of its associated gene cluster.

Table 1 – Chemical structures of the RGP of S. thermophilus strains Nect13, UCCSt12, Douc24, CNRZ302, STMM11 and Brie28. The rhamnan backbone subunit and its attachment point to the rhamnan backbone is highlighted in grey.

Strain	***rgp* genotype**	**RGP structure**
Nect13	*rgp*4B	[β-Glc-3-β-GalNAc-3-]-2-α-Rha-2-α-Rha-3-α-Rha--2-α-Rha-[β-Glc-3-β-GalNAc-3-]-2-α-Rha-3-α-Rha-
UCCSt12 (38)	*rgp*4B	[β-Gal-3-β-GalNAc-3-]-2-α-Rha-2-α-Rha-3-α-Rha--2-α-Rha-[β-Gal-3-β-GalNAc-3-]-2-α-Rha-3-α-Rha-
Douc24	*rgp*5A	-2-α-Rha-2-α-Rha-[α-Glc-6-α-Glc-2-]-3-α-Rha-
CNRZ302	*rgp*7	[β-Galf-3-α-Gal-2-α-Rha-3-β-GlcNAc-3-]-2-α-Rha-[α-Glc-4-]-2-α-Rha-6-α-Glc-
STMM11	*rgp*8	[β-Gal*f*-3-α-Gal-2-α-Rha-3-β-GlcNAc-4-]-2-α-Rha-3-α-Rha-
Brie28	*rgp*9	[β-Gal-4-α-Rha-3-β-GalNAc-4-]-2-α-Rha-3-α-Rha-

Establishing a genotype-to-structure relationship for salivarius group RGPs

It has been shown in previous studies that the *cwps* genotype of *Lactococcus lactis/cremoris* strains correlates with the biochemical structure of the polysaccharide (60). Similar genotype-to-structure relationships have more recently been explored in the dairy-associated species *Pseudolactococcus raffinolactis* and *Pseudolactococcus laudensis* (61). An equivalent framework has not yet been established for streptococcal species.

By integrating comparative genomics with biochemical characterisation of currently elucidated RGP structures, we established a foundation for the reliable prediction of RGP structural features in *S. thermophilus* and related members of the salivarius group. Our analyses indicate that their key RGP structural features, including oligosaccharide chain length, monosaccharide composition, and the presence of distinctive modifications such as galactofuranose (Gal*f*), can be inferred directly from the *rgp* locus gene content despite exhibiting substantial genetic diversity among *rgp* clusters. Extending this approach across the salivarius group suggests that the conserved modular organisation of *rgp* loci (i.e. side chain module and backbone module) and shared gene content enables a reasonably accurate prediction of the backbone and side chain subunit structure. This framework, therefore, provides a basis for predicting RGP structural similarities as well as structural diversity across the salivarius group, and beyond.

The rhamnan backbone component of streptococcal RGPs is encoded by a conserved region of the *rgp* locus characterised by the presence of well-defined rhamnosyltransferases (RgpA, B and F) together with an ABC-type transport system (encoded by *rgpC* and *rgpD*), whose encoding genes are typically co-located

in the specific arrangement *rgpABCDF* (Fig. 4A). In addition, a WpsJ homologue is frequently located in this region of the locus and additional glycosyltransferase-encoding genes may be located up- or down-stream of the *rgpABCDF* module (Fig. 4A). RgpE homologues (encoded by *rgpE*) have also been identified in the streptococcal *rgp* cluster, in the backbone-associated region, and are proposed to 'cap' the rhamnan polymer, thus determining chain length in *L. lactis* (57,60). In *S. mutans*, RgpE contributes to the formation of glucose side chains after the core rhamnose chain is assembled (18,62). Despite this overall conservation, expansion of the HCL and comparative genomic analysis of *rgp* loci to include all known salivarius group species, the present study enabled identification of three additional backbone genotypes (Bt4, 5 and 6), reflecting further genetic and potential structural diversity within the rhamnan biosynthetic framework. This analysis highlights that variation in the backbone-associated *rgp* modules is associated with structural diversity, with rhamnan subunits ranging in length from di- to tetrasaccharides (Fig. 4C).

Variation in rhamnan backbone architecture appears to correlate with differences in the glycosyltransferase-encoding gene content of the backbone-associated *rgp* modules. The presence of *rgp*A and *rgp*B across all identified backbone genotypes supports their proposed role in catalysing rhamnan backbone assembly in salivarius group species, consistent with the observation that all structurally characterised streptococcal RGPs contain a minimum of two rhamnose residues. Variation in backbone length, ranging from di- to tetrasaccharide repeating units, may be associated with the presence of additional glycosyltransferase-encoding genes within the backbone-associated module, including RgpE, which may influence rhamnan chain elongation. This is reflected in the consistent association of

disaccharide backbones with Bt2 sub-genotypes and trisaccharide backbones with Bt3 sub-genotypes (Fig. 4C). Furthermore, the presence of an additional glycosyltransferase-encoding gene within the backbone region in all Bt1 sub-genotypes is consistent with the incorporation of a glucose residue into the rhamnan component (Supplementary Fig. S12).

Extending these observations to the remaining backbone genotypes enables tentative predictions of their rhamnan structures. For example, the representative Bt5-type strain lacks the glycosyltransferase-encoding gene predicted to incorporate glucose into Bt1-associated structures, suggesting a backbone structure composed primarily of rhamnose residues for this particular strain. Additionally, Bt4 strains encode an additional glycosyltransferase within the backbone-associated module, which may contribute to structural variation relative to other genotypes. Similarly, the Bt6-type strain analysed in this study, harbours two additional glycosyltransferase-encoding genes within the backbone region, indicating potential for further modification of the backbone structure. However, these predictions require experimental validation.

The 5′ variable region of the *rgp* locus encodes the biosynthetic machinery for the synthesis and assembly of the RGP side chain and displays greater organisational and gene content diversity compared to the 3’ portion of the *rgp* locus (Fig. 4A). This observed diversity is linked to the identification of six additional variable genotypes in the current study (Vt6-10, including a Vtx genotype; Supplementary Table S3), further highlighting the complexity of these genetic clusters. The first glycosyltransferase-encoding gene within this region is highly conserved and shares sequence similarity with *rgpI* of *S. mutans*, which has been proposed to encode a

regulatory protein influencing glucose side chain branching frequency (18,44). However, TMHMM and HHpred analyses also indicate similarity to polysaccharide polymerases described in *L. lactis* (57), suggesting a role in polysaccharide assembly. While *rmlD* homologues, encoding the enzyme responsible for the final step in the dTDP-L-rhamnose biosynthetic pathway were identified within the 5' portion of *rgp* cluster of all analysed strains, *rmlA* to -*C* are located outside of the *rgp* locus (38). Additionally, *wpsA* homologues, which encode a glycosyltransferase associated with the initiation of side chain biosynthesis in *L. lactis* and *S. thermophilus* (47,57) were also identified in this region. The leftward end of the modules of the analysed *rgp* loci also display considerable intergroup variation, particularly in the glycosyltransferase-encoding gene content (Fig. 4A).

Side chain biosynthesis is initiated by WpsA, the priming glycosyltransferase and its associated activator (WpsB), through the addition of a GlcNAc or a GalNAc moiety to the lipid carrier undecaprenyl-phosphate, as detailed previously (47,57). This feature is mirrored by the conservation of WpsA homologues across all analysed *rgp* genotypes, with the exception of *rgp*5, *rgp*11 and *rgp*13 type strains which lack the majority of side chain synthesis-associated genes (Fig. 4A). Strains initiating side chain biosynthesis with a GlcNAc moiety (Vt1, 4, 5, 8 and 9-type loci), possess WpsA homologues closely related to those found in *L. lactis* MG1363, sharing ~ 52 % sequence identity. Conversely, strains possessing a *rgp* Vt2, 3 and 7-type share a reduced similarity (< 40 %) and possess a different first monosaccharide in their side chain structure, i.e. a GalNAc vs GlcNAc moiety. In this study, by extending the gene compositional analysis to include all salivarius group members, we identified a third WpsA variant present in the only classified *rgp*14 type strain that exhibits a Vt6 sub-genotype. This variant has low sequence

similarity with all other WpsA homologues in the representative strains presented in this study (Fig. 4A) and shares ~ 47 % identity with that of the *L. lactis* MG1363 WpsA. In this study, we confirm that the backbone-linked monosaccharide in the side chain units for Vt5 and Vt9 type strains is a GlcNAc residue, while those for Vt2 type strains are GalNAc residues (Fig. 4C). The lack of a *wpsA* homolog (and a seemingly minimal side chain module) in the *rgp*5, 10, and 11-type strains is reflected in the non-canonical side chain structure of the *rgp*5A strain, in which the backbone structure is instead decorated with two glucose residues. In this case, the glucose disaccharide side chain is predicted to be added by the additional glycosyltransferases encoded within the backbone module (Fig. 4A; Supplementary Fig. S12).

A unique (predicted) acyltransferase-encoding gene is present in all *rgp*11 (*S. salivarius* and *S. raffinosi*) type strains and in the single representative *rgp*14 (*S. salivarius*) type strain (Fig. 4A). The presence of an UDP-galactopyranose mutase correlates to the presence of a Gal*f* residue in side chain structure in the CWPS of several *L. lactis* strains (C-type strains (60)). In this study, the same observation is made for *rgp*7 and *rgp*8 type strains that harbour a UDP-galactopyranose mutase-encoding gene in their respective *rgp* loci and indeed appear to have undergone a pyranosyl to furanosyl tautomerisation in the corresponding galactose moieties in their respective side chain structures.

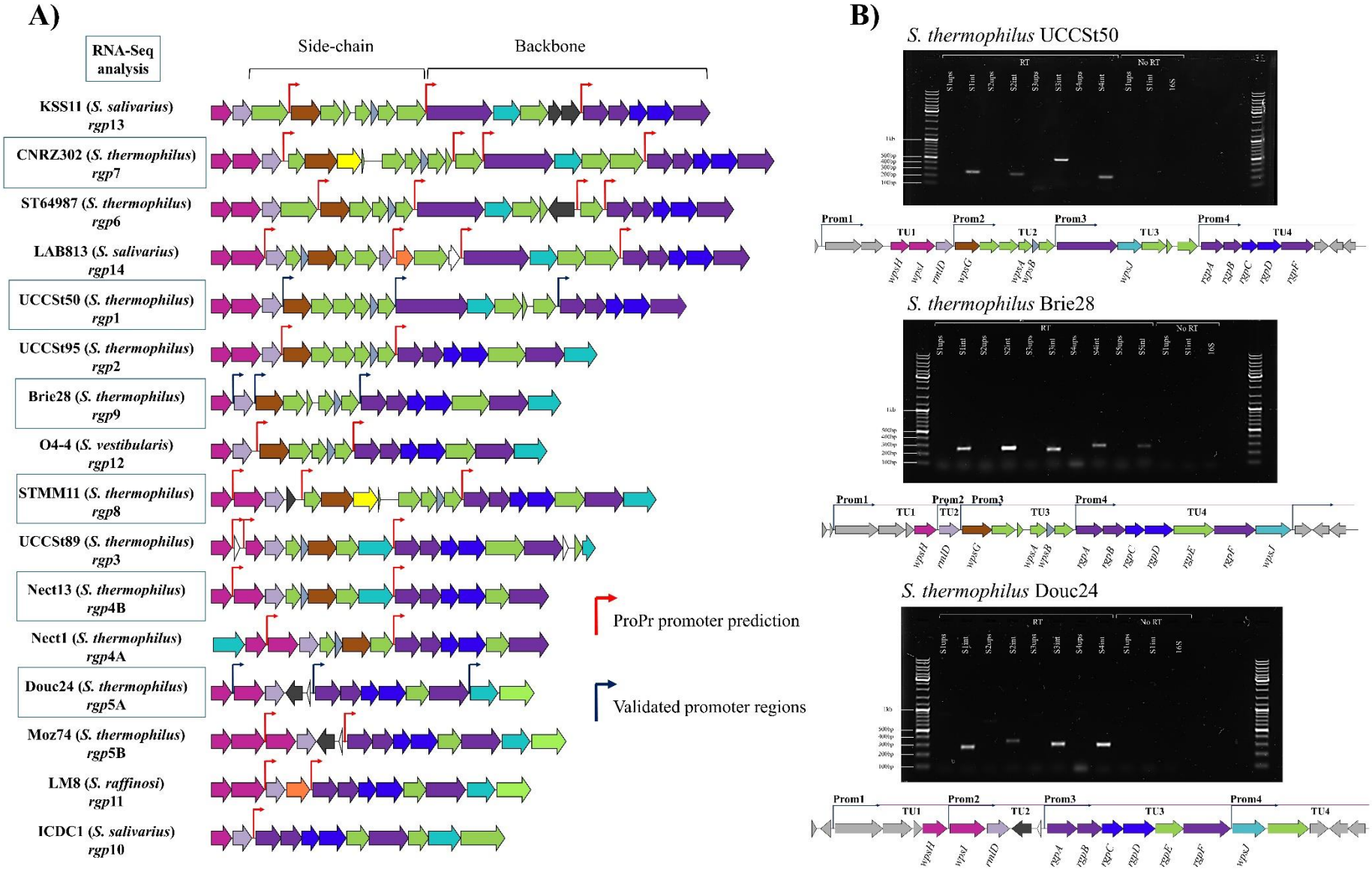
Collectively, these observations support the establishment of a genotype-to-structure relationship for streptococcal RGPs, in which variation in *rgp* gene content can be used to predict key structural features of both the backbone and side chain components.

Transcriptional units within the *rgp* loci underpin their modular nature & unique organisation

Lactococcus lactis/cremoris has been a useful paradigm and comparator for the functional analysis of *S. thermophilus rgp* loci (57,60). However, while the order of the genes present in the lactococcal *cwps* clusters are organised into two to three discrete modules that follow the order of the biosynthetic process, the organisation of *S. thermophilus rgp* gene clusters does not appear to be linear, i.e. follow the order of biosynthesis (38). Comparative analysis of the *rgp* clusters of salivarius group species revealed a similar organisation to that of *S. thermophilus* (Fig. 4A). Therefore, to define the modular organisation and to identify the transcriptional units that are present in these loci, RNA-Seq analysis was performed on six streptococcal strains representing *rgp*1, 4B, 5A, 7, 8 and 9 genotypes. In parallel, an *in silico* prediction of the promoters within the *rgp* loci of these strains as well as representatives of *rgp*10-14 was performed (Fig. 5A). At the 5' end of the *rgp* locus associated with the biosynthesis of the side chain component, the number of transcriptional units varies between the analysed strains with between one and three promoter regions identified (Fig. 5A). This region of the *rgp* locus is associated with the highest variability and with a number of observed insertion/deletions, which are likely achieved through homologous recombination events. It is notable; however, that where apparent insertions have occurred (or *rgp* genotype-specific genes or gene sets), each such "insertion" is associated with its own (predicted) promoter. In contrast, the rightward end of the locus, which is associated with rhamnan backbone biosynthesis, appears less diverse in terms of their predicted promoters, with one or two transcriptional units and with promoters typically located upstream of the first rhamnosyl-/glycosyl-transferase-encoding gene within

the module (in the case of *rgp*1, 7, 13, 14) or upstream of *wpsJ* (in the case of *rgp*5A) and upstream of *rgpA* (in all *rgp* types) (Fig. 5A). The discrete modules within these loci incorporate functions including glycosyltransferase activity and may be associated with the addition of specific side chain structures/moieties.

The genetic plasticity of the *rgp* loci across the salivarius group is evident and an excellent paradigm for horizontal gene transfer between members of this group. Interestingly, the *rgp* loci of representative members of *rgp*10 are predicted to harbour a single promoter upstream of *rgpA* and which is reminiscent of the (pseudo)lactococcal *cwps* genotype E (61). To validate the RNA-Seq data and promoter predictions, PCR-based assays were performed on three *S. thermophilus* strains representing *rgp*1 (UCCSt50), *rgp*5A (Douc24) and *rgp*9 (Brie28) genotypes, using cDNA generated from RNA extracts of the strain at mid-log and targeting each predicted promoter. Using a combination of primers located upstream (as a negative control) of the predicted promoter and within the identified transcriptional units, the transcriptional units were confirmed while PCRs using the upstream primer did not yield an amplicon confirming the absence of host DNA in the RNA (and subsequent cDNA) preparation (Fig. 5B).


A)
RNA-Seq analysis
Side-chain
Backbone
KSS11 (S. salivarius) rgp13
CNRZ302 (S. thermophilus) rgp7
ST64987 (S. thermophilus) rgp6
LAB813 (S. salivarius) rgp14
UCCSt50 (S. thermophilus) rgp1
UCCSt95 (S. thermophilus) rgp2
Brie28 (S. thermophilus) rgp9
O4-4 (S. vestibularis) rgp12
STMM11 (S. thermophilus) rgp8
UCCSt89 (S. thermophilus) rgp3
Nect13 (S. thermophilus) rgp4B
Nect1 (S. thermophilus) rgp4A
Douc24 (S. thermophilus) rgp5A
Moz74 (S. thermophilus) rgp5B
LM8 (S. raffinosi) rgp11
ICDC1 (S. salivarius) rgp10
ProPr promoter prediction
Validated promoter regions
B)
S. thermophilus UCCSt50
RT
No RT
Prom1
Prom2
Prom3
Prom4
TU1
TU2
TU3
TU4
wpsH
wpsI
rmlD
wpsG
wpsA
wpsB
wpsJ
rgpA
rgpB
rgpC
rgpD
rgpF
S. thermophilus Brie28
rgpE
wpsJ
S. thermophilus Douc24

1

Fig. 5. **A)** Schematic representation of the salivarius group *rgp* gene clusters with predicted (red arrow) (63) and validated promoter regions (blue arrow). **B)** Gel electrophoresis summarising the RT-PCR results for the transcriptional units of the *rgp* gene clusters of *S. thermophilus* strains UCCSt50, Brie28 and Douc24. Total RNA was purified from each strain and used for cDNA synthesis ("RT"). To identify the transcriptional units of the *rgp* clusters, PCRs were performed using primer pairs consisting of an "ups" primer positioned upstream the predicted promoter region and an internal primer located within the downstream transcriptional unit ("int"). Gel lane labels (e.g., S1ups, S1int, S2ups, etc) indicate the specific upstream/internal primer combinations used for each strain (Supplementary Table S9). No-RT controls ("No-RT") generated from the same RNA preparations were amplified with 16S primers and S1 primer sets to confirm the absence of genomic DNA contamination. Below each gel electrophoresis image is a schematic representation of the corresponding *rgp* gene cluster of each *S. thermophilus* strain highlighting the deduced transcriptional units present. The identified promoter regions (Prom) are shown by blue arrows, and the identified transcriptional units (TU) are represented by purple lines. Numbers on the left-hand side of each gel electrophoresis image correspond to the size of the DNA ladder used (Thermo Scientific GeneRuler DNA Ladder Mix)

Discussion

The genus *Streptococcus* was among the first bacterial taxa to be described in the late nineteenth century and today encompasses more than 100 recognised species occupying a broad range of ecological niches. Despite extensive study, the taxonomy and phylogeny of this genus remain highly dynamic, driven by continual expansion of genome sequence databases and the increasing availability of metagenome-assembled genomes (MAGs) from diverse environments. Recent large-scale phylogenetic studies have proposed between eight and nine major streptococcal groups depending on the genomes included and the analytical approaches employed (6,8,64). In the present study, we sought to provide an updated and comprehensive overview of streptococcal phylogeny using the largest collection of high-quality streptococcal genomes analysed to date, while simultaneously exploring the diversity, organisation, and functional significance of *rgp* loci using the salivarius group streptococci as a paradigm. By integrating large-scale phylogenomic analysis with comparative genomics, structural elucidation, and transcriptional mapping, we demonstrate that variation within *rgp* loci underpins diversity in RGP architecture. Furthermore, our findings establish genotype-to-structure relationships for salivarius group RGPs and reveal that these loci are highly dynamic, modular genetic units for RGP biosynthesis that most likely are shaped by ecological adaptation and selective pressures acting on the bacterial cell surface.

The phylogenomic analysis performed in this study, encompassing almost 5,000 high-quality streptococcal genomes and MAGs, represents the most comprehensive dataset examined to date and supports the classification of the genus into eleven phylogenetic groups. Importantly, the inclusion of large numbers of MAGs enabled

representation of previously uncultured or poorly characterised taxa, highlighting the substantial unresolved diversity that remains within the *Streptococcus* genus. The strong representation of the salivarius group observed within the pangenome analysis (Fig. 1) combined with the knowledge that members of this group are among the earliest human colonisers led us to apply its members as a model taxonomic group for investigating *rgp* locus presence, diversity and organisation.

Cell envelope-associated polysaccharides, and RGPs in particular, are increasingly recognised as central components of streptococcal biology. However, despite their widespread distribution across streptococci, their genetic organisation and biochemical structures have been characterised for only a limited number of species, including *S. thermophilus*, *S. mutans* and *S. pyogenes*. This comparatively narrow representation likely reflects the biological importance of these cell wall structures in such species, where RGPs have been linked directly to bacteriophage susceptibility (17,40,45,47), cell division (14,21), and virulence (24,65), respectively. The present study substantially expands this understanding by demonstrating that *rgp* loci are a widespread feature across the majority of streptococci and that all currently recognised salivarius group species harbour *rgp* loci with conserved modular organisation.

One of the most significant findings of this study is the extent of *rgp* locus diversity across the salivarius group. Previous studies had defined seven *rgp* genotypes within *S. thermophilus* (38), while in the present study, we identified an additional seven distinct *rgp* genotypes. Furthermore, we identified multiple novel backbone-associated genetic region genotypes (Bt) and variable side chain-associated genotypes (Vt) distributed across *S. thermophilus*, *S. salivarius*, *S. vestibularis* and *S. raffinosi*. While several genotypes were shared between species, many appear to

be species-specific, suggesting that both horizontal gene transfer and niche-specific evolutionary trajectories may contribute to *rgp* diversification. The presence of conserved locus organisation and genomic positioning across all salivarius group species indicates that the *rgp* locus likely represents an ancestral and functionally important component of the cell envelope. However, the extensive variation typically observed within the side chain-associated regions suggests that selective pressures primarily target surface-exposed structural features while preserving the conserved biosynthetic machinery required for rhamnan backbone assembly.

The newly identified *rgp* genotypes present in oral/human-associated salivarius group species may also reflect adaptation to distinct ecological niches. Several *rgp* loci identified exclusively in *S. salivarius*, *S. vestibularis* or *S. raffinosi* display reduced side chain-associated modules or encoded additional putative modification enzymes, including membrane-bound acyltransferase 3 (AT3)-family proteins. In many bacterial species, the acylation of surface sugars is mediated by membrane-bound acyltransferase-3 (AT3) proteins, which transfer activated acyl groups to extracellular polysaccharide substrates and influence host colonisation, environmental survival, and other surface properties (66). Similarly, glycerol phosphate modifications of RGPs in *S. mutans* and *S. pyogenes* contribute to altered cell surface properties and host interactions (21,25). Although the function of putative acyltransferases identified here has not been experimentally validated, the absence of these genes among dairy-associated *S. thermophilus* strains and presence in oral-associated salivarius group species suggests that these modifications may contribute to host adaptation or ecological fitness within the oral cavity. Further biochemical and mutational analyses is required to determine the direct contribution

of these genes in RGP modifications and how these modifications may influence cell surface biology.

Through structural elucidation of five previously uncharacterised RGPs, this study substantially expands the known structural repertoire of streptococcal RGPs. Importantly, comparative analysis demonstrated a strong correlation between *rgp* locus gene content and experimentally resolved RGP structures, enabling a degree of prediction of key structural features from genomic information. Variation within glycosyltransferase-encoding genes and priming glycosyltransferase homologues was shown to strongly correlate with side chain composition and the identity of the first side chain monosaccharide. Furthermore, the presence of UDP-galactopyranose mutase homologues correlated with galactofuranose-containing RGP structures. These observations extend the potential genotype-to-structure relationships previously established in *L. lactis* to streptococci and provide a framework for predicting RGP diversity across the genus.

Collectively, this study provides the most extensive comparative analysis of streptococcal *rgp* loci and RGP structures reported to date and significantly expands current understanding of the diversity, evolution, and functional significance of these important, multifunctional, cell envelope-associated polysaccharides. By integrating phylogenetic analysis, comparative genomics, structural elucidation, and transcriptional mapping, we establish a framework linking *rgp* gene content to polysaccharide architecture across the salivarius group streptococci. Our findings demonstrate that *rgp* loci are highly dynamic and modular biosynthetic clusters whose diversity likely reflects adaptation to distinct ecological niches and selective pressures acting on the bacterial cell surface. The predictive genotype-to-structure relationships established in this study provide a foundation for future exploration

of streptococcal cell wall diversity, including its roles in phage biology, host interactions, and niche adaptation across the entire genus.

Materials and Methods

Phylogenetic analysis

Initially 13,762 *Streptococcus* genomes and MAGs were retrieved from a diverse set of databases: SPIRE (a Searchable Planetary-scale mIcrobiome Resource) (50), curatedFoodMetagenomicData (cFMD) (52), NCBI Genome database, and MGnify Genome catalogues (Unified Human Gastrointestinal Genome (UHGG) collection and Skin biome collection) (49,51). CheckM2 (67) was used to filter for high quality genomes (>90% complete and <5% contamination) to generate a final dataset of 4,966 genome sequences consisting of high-quality MAGs (2,737) and genomes of isolated streptococcal strains (2,229). The genomes were taxonomically classified using GTDB-TK (53) v2.7.2 classify workflow with GTDB-TK reference database (version 232). The pangenome toolbox PIRATE v1.0.5 (Pangenome Iterative Refinement and Threshold Evaluation) (68) was used to generate a pangenome, including *Lactococcus lactis* ATCC 19435 (GenBank Accession number NZ_LKLC00000000) as outgroup. Single-copy core genes (present in more than 90% of genomes) were concatenated in a gene alignment. The alignment was trimmed with ClipKIT v2.4.1 (kpi mode) to remove gaps, and successively by only including the top 10% of bases ranked by entropy. IQ-TREE (69) v2.4.0 (-fast) was used for tree inference, and the resulting tree was visualised using iTOL v7.6 software (https://itol.embl.de/). Phylogenetic groups were defined using the R package ‘ape’ (70) with 1 as the threshold for number of substitutions per site in the 10% most variable positions of the core genome and a given group consisting of a minimum of 10 genomes.

Automated search for genes associated with RGP synthesis

To assess the prevalence of the *rgp* gene cluster within the *Streptococcus* genus, 150 high-quality (determined by CheckM2) *Streptococcus* reference genomes representing available streptococcal species were retrieved from NCBI Genome Datasets (https://www.ncbi.nlm.nih.gov/datasets/genome/). All genomes were reoriented to start with *dnaA* using dnaapler (71) v1.3.0 and subsequently annotated with Bakta (72) v1.12.0. A custom python script was developed to identify conserved genes associated with RGP synthesis. Specifically, the target genes/gene products were associated with rhamnose-nucleotide precursor synthesis (*rmlA-D*), rhamnan structure synthesis (*rgpA*), and variable side chain synthesis (*wpsB*). The target gene names and annotated products for each of the gene targets are as follows: *rmlA* (rfbA; glucose-1-phosphate thymidylyltransferase), *rmlB* (rfbB; dTDP-glucose 4,6-dehydratase), *rmlC* (rfbC, rmlC, rmlC-like protein; dTDP-4-dehydrorhamnose 3,5-epimerase; dTDP-4-keto-6-deoxyglucose 3,5-epimerase; dTDP-4-keto-6-deoxy-D-glucose epimerase), *rmlD* (rfbD, dTDP-4-dehydrorhamnose reductase), *rgpA* (cps2T, beta-1,4-rhamnosyltransferase), and *wpsB* (DUF2304).

Streptococcal strains

S. thermophilus strains were routinely grown from single colonies or bacterial stocks maintained at -70 °C (25 % [wt/vol] glycerol) in M17 broth (Sigma, USA) supplemented with 0.5 % lactose (Sigma, USA) at 42 °C.

DNA extraction and sequencing

Genomic DNA of *S. thermophilus* strains Brie28, Douc24, Moz74, Moz76, Moz77, Moz83, Moz111, Nect1, Nect13, Roque89, Strac42, UCCSt89 and Vach60 (Supplementary Table S2) was extracted using a modified protocol of the Nucleobond AXG 100 with Buffer set III (Macherey-Nagel, Germany) as described previously (45).

Bacterial genome sequencing was performed by Plasmidsaurus (USA) using Oxford Nanopore Technology (UK) long-read sequencing. Long-read sequencing was conducted on a PromethION P24 platform with R10.4.1 flow cells, and libraries were prepared using the Rapid Barcoding Kit 96 V14. Basecalling was performed using 'dorato super-accurate' basecalling with default Q10 quality filtering. Sequences were subsequently annotated using Bakta with default parameters (72).

Identification and comparative analysis of the *rgp* loci of streptococcal strains

The *rgp* gene clusters of streptococcal strains applied in the HCL and comparative analysis in this study were identified as described previously between two conserved genes, i.e. the genes that encode the 30S ribosomal subunit (5' flank) and the Rex transcriptional regulator (3' flank) (38,41). Fasta files of these regions were extracted from 223 high-quality complete genomes or scaffolds of members of the salivarius group available in the NCBI database (Supplementary Table S2). For consistency in *rgp* cluster comparisons, the fasta files from each *rgp* were reannotated using the Prokka v1.14.6 pipeline (73) to generate GFF3 files. These output files were input into the Roary pangenome pipeline (https://github.com/sanger-pathogens/Roary) (74). Protein families were defined

using Roary with default parameters and a minimum BLASTp identity threshold of 50 % (-i 50) and paralog splitting was disabled (-s). A binary presence-absence matrix of clustered proteins was generated and analysed using two-way hierarchical clustering (HCL) within the multi-experiment viewer (MeV v4.9) (75). Genotypes were assigned according to previously established classifications (38) or, where necessary, based on newly resolved groups identified through HCL and comparative sequence analyses.

Detailed annotation of the *rgp* loci of the strains in Fig. 2C and Fig. 3A was performed using HHpred for structural predictions (76), while transmembrane helices were detected using the TMHMM server DeepTMHMM (77) v1.0 available at https://services.healthtech.dtu.dk/services/DeepTMHMM-1.0/. Predicted functions of genes were proposed based on comparisons to *Lactococcus lactis cwps* clusters, including the HHpred and TMHMM analysis (57).

Preparation and structural analysis of the RGPs of *S. thermophilus* strains

S. thermophilus strains Brie28, CNRZ302, Douc24, Nect13 and STMM11 were grown overnight in M17 (Sigma, USA) supplemented with 0.5 % lactose (Sigma, USA) (1 % total lactose concentration) at 42 °C. Culture volumes ranged from 4 to 8 litres, and cells were harvested by centrifugation at 4,400× *g* at 4 °C for 30 minutes. The resulting cell pellets were washed twice with ice-cold sterile distilled water, first with 300 mL per culture, then pooled and washed again in 40 mL, and subsequently stored at –20 °C.

CWPSs (RGP) were extracted from the cell pellet with consecutive treatments with 0.01 M and 0.1 M HCl (100°C, 20 min) after pre-extraction with cold 5 % TCA (48 h, 5°C). The HCl extracts were deproteinated by addition of TCA (up to 5 %),

dialysed, freeze-dried and purified on Sephadex G-50 column, as described previously (39). HCl preparations had identical monosaccharide composition and ^{1}H NMR spectra. Unless otherwise specified, the preparations of 0.01 M HCl extractions were used for structural analysis of RGPs.

Monosaccharide, methylation analysis was carried out as described previously (39,60). Gel filtration and anion-exchange chromatography, alkaline deacylation, deamination and NMR spectroscopy were performed as described previously (38). Reverse-phase HPLC was performed on Agilent Zorbax C18 column (250 × 9 mm) in solvent A (0.1 % TFA v/v) and B 80 % MeCN (v/v), 3 ml/min, from 3 % B 10 min, then gradient to 100 % B over 30 min. with a UV detector at 220 nm.

RNA extraction and sequencing

S. thermophilus strains Brie28, CNRZ302, Douc24, Nect13, STMM11 and UCCSt50 were grown in M17 (Sigma, USA) supplemented with 0.5 % lactose (Sigma, USA) to mid-log phase (OD_{600} between 0.4 and 0.7) before harvesting by centrifugation. Cell pellets were resuspended in DNA/RNA shield (Zymo research, USA) and stored at -80 °C. Biological duplicates were prepared for all samples. Frozen cell pellets were sent to the University of Groningen for RNA extraction, rRNA depletion, library construction and RNA sequencing. RNA extraction and sequencing were performed as previously described (78). Ribosomal RNA was depleted using the RiboCop rRNA Depletion Kit (Lexogen, Austria) and sequencing libraries were prepared using the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs, USA). Libraries were sequenced on an Illumina NextSeq 1000 platform. Sequence quality was assessed using FastQC (Babraham Bioinformatics, UK) and reads were mapped to the corresponding

reference genomes using Bowtie2 with default settings. SAM files were converted to BAM format using SAMtools. Gene counts were generated using featureCounts following SAM/BAM file processing with SAMtools.

RNA-Seq analysis and transcriptional unit prediction

To identify transcriptional units within the *rgp* biosynthetic clusters, RNA-Seq read alignments were manually inspected and viewed using the Artemis genome browser (v16) (79). The coverage was viewed in Artemis by layering sorted BAM files from the two independent RNA-Seq samples with genome annotation files in GFF format. Coverage profiles across the *rgp* loci from the six strains were examined to identify transcriptional boundaries. Putative transcriptional start and termination sites were inferred by identifying regions where RNA-Seq coverage showed a sharp increase in read depth followed by sustained coverage across coding regions and a subsequent abrupt drop, consistent with the boundaries of a transcriptional unit. Additionally, promoters located within the *rgp* locus were predicted using the ProPr promoter prediction tool using default settings (Promoters in intergenic regions and auto detect best model) (63). Promoter predictions were compared with RNA-Seq coverage profiles to refine the positions of putative transcription start sites within the *rgp* clusters of the six *S. thermophilus* strains (Brie28, CNRZ302, Douc24, Nect13, STMM11 and UCCSt50).

RNA purification and validation of predicted promoters

Primers targeting regions upstream of predicted promoters and within the corresponding transcriptional units were designed (Supplementary Table S9) and synthesised (Integrated DNA Technologies, USA). *S. thermophilus* strains Brie28, CNRZ302, Douc24, and UCCSt50 were grown in LM17 medium to mid-log phase in biological triplicate. For each culture, 1 mL of cells was harvested and resuspended in 2 mL RNAprotect Bacteria Reagent (Qiagen, Germany) as per the manufacturer's instructions.

Total RNA was purified using the RNeasy Mini Kit (Qiagen, Germany) according to the manufacturer's protocols 4 and 7, with the following modifications: the lysis incubation was extended to 30 minutes, the DNase I treatment was performed at 37 °C for 30 minutes, and an additional post-purification DNase I treatment (Roche, Switzerland) was carried out at 37 °C for 20 minutes. The reaction was terminated by the addition of 0.2 M EDTA and incubating at 75 °C for 10 minutes. Complementary DNA (cDNA) was synthesized from 16 µL of purified RNA using the LunaScript RT SuperMix (New England Biolabs, USA) according to the manufacturer's instructions. Each replicate of each strain was then used as the template for PCR amplification of the predicted promoter regions to define the transcriptional units of the *rgp* loci. PCRs were performed using DreamTaq PCR master Mix (Thermo Fisher Scientific, USA) with the following cycling parameters: initial denaturation at 95 °C for 3 min; 30 cycles of 95 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min 30 s; and a final extension at 72 °C for 10 min.

Data availability

Several *S. thermophilus* strain genomes were sequenced to completion using long reads sequencing approach (Oxford Nanopore Technology, ONT) as part of this study. The genome sequences of *S. thermophilus* Brie28, Douc24, Moz74, Moz76, Moz77, Moz83, Moz111, Nect1, Nect13, Roque89, Strac42, UCCSt89, and Vach60 have been deposited in GenBank under BioProject accession number PRJNA1405548. RNA-Seq raw reads (.fastq files) were deposited to the Sequence Read Archive (SRA) under the BioProject accession number PRJNA1463093.

Supplementary Material

Supplementary Table S1 – CH3_Supplementary_TableS1.xlsx

A link to the 4,966 high-quality genomes used in the pangenome analysis, including Genome or MAG identifier, the database the genome was retrieved from, the GTDB classification, the species name, genome size, GC content and phylogenetic group

Supplementary Table S2. Streptococcus strain genomes (n=223) used in the HCL analysis presented in this study and their associated genome assembly level and NCBI accession number.

	Strain	Species	Assembly level in NCBI	Accession number	Notes
1	1A	*Streptococcus thermophilus*	Complete	CP065384	
2	1F8CT	*Streptococcus thermophilus*	Chromosome	CM003138	
3	4021	*Streptococcus thermophilus*	Complete	CP065493	
4	4052	*Streptococcus thermophilus*	Complete	CP065476	
5	4067	*Streptococcus thermophilus*	Complete	CP065496	
6	4078	*Streptococcus thermophilus*	Complete	CP065477	
7	4134	*Streptococcus thermophilus*	Complete	CP065478	
8	4145	*Streptococcus thermophilus*	Complete	CP065492	
9	4147	*Streptococcus thermophilus*	Complete	CP065504	
10	13496	*Streptococcus thermophilus*	Complete	CP061025	
11	13498	*Streptococcus thermophilus*	Complete	CP061024.	
12	13499	*Streptococcus thermophilus*	Complete	CP061023	
13	24738	*Streptococcus thermophilus*	Complete	CP061022	
14	24739	*Streptococcus thermophilus*	Complete	CP061021	
15	24740	*Streptococcus thermophilus*	Complete	CP061020	
16	24853	*Streptococcus thermophilus*	Complete	CP061019	
17	90728	*Streptococcus thermophilus*	Complete	CP065479	
18	90729	*Streptococcus thermophilus*	Complete	CP065480	
19	90730	*Streptococcus thermophilus*	Complete	CP065481	
20	ACA-DC_2	*Streptococcus thermophilus*	Chromosome	LT604076	
21	APC151	*Streptococcus thermophilus*	Complete	CP019935	
22	ASCC_1275	*Streptococcus thermophilus*	Complete	CP006819	
23	ATCC_19258	*Streptococcus thermophilus*	Complete	CP038020	Whole genome annotated using Prokka
24	AVA116	*Streptococcus thermophilus*	Complete	CP065498	
25	AVA1121	*Streptococcus thermophilus*	Complete	CP065488	
26	B59671	*Streptococcus thermophilus*	Complete	CP022547	
27	Brie28	*Streptococcus thermophilus*	Complete	JBTWWX000000000	Sequenced in the context of this study

28	CH8	*Streptococcus thermophilus*	Complete	CP089060	
29	CICC_20372	*Streptococcus thermophilus*	Complete	CP132900	
30	CNRZ302	*Streptococcus thermophilus*	Complete	CP065489	
31	CNRZ385	*Streptococcus thermophilus*	Complete	CP065495	
32	CNRZ760	*Streptococcus thermophilus*	Complete	CP065482	
33	CNRZ887	*Streptococcus thermophilus*	Complete	CP065491	
34	CNRZ1066	*Streptococcus thermophilus*	Complete	CP000024	
35	CNRZ1151	*Streptococcus thermophilus*	Complete	CP065483	
36	CNRZ1202	*Streptococcus thermophilus*	Complete	CP065506	
37	CNRZ1575	*Streptococcus thermophilus*	Complete	CP065490	
38	CS5	*Streptococcus thermophilus*	Complete	CP028896	
39	CS6	*Streptococcus thermophilus*	Complete	CP050870	
40	CS8	*Streptococcus thermophilus*	Complete	CP016439	
41	CS9	*Streptococcus thermophilus*	Complete	CP030927	
42	CS18	*Streptococcus thermophilus*	Complete	CP030928	
43	CS20	*Streptococcus thermophilus*	Complete	CP030250	Whole genome annotated using Prokka
44	DGCC7710	*Streptococcus thermophilus*	Complete	CP123436	
45	DMST-H2	*Streptococcus thermophilus*	Complete	CP063275	Whole genome annotated using Prokka
46	Douc24	*Streptococcus thermophilus*	Complete	JBTWWW000000000	Sequenced in the context of this study
47	EG007	*Streptococcus thermophilus*	Complete	CP069275	
48	EPS	*Streptococcus thermophilus*	Complete	CP025400	Whole genome annotated using Prokka
49	EU01	*Streptococcus thermophilus*	Complete	CP047191	
50	F1B	*Streptococcus thermophilus*	Complete	CP194006	
51	F2A	*Streptococcus thermophilus*	Complete	CP194005	
52	FDAARGOS_1574	*Streptococcus thermophilus*	Complete	CP086001	
53	GABA	*Streptococcus thermophilus*	Complete	CP025399	Whole genome annotated using Prokka
54	IDCC2201	*Streptococcus thermophilus*	Complete	CP035306	
55	JIM_8232	*Streptococcus thermophilus*	Complete	FR875178	
56	KLDS_31003	*Streptococcus thermophilus*	Complete	CP016877	
57	KLDS_SM	*Streptococcus thermophilus*	Complete	CP016026	
58	LMD-9	*Streptococcus thermophilus*	Complete	CP000419	
59	LMG_18311	*Streptococcus thermophilus*	Complete	CP000023	

60	M17PTZA496	*Streptococcus thermophilus*	Chromosome	M002372	
61	MAG_rmk202_sterm	*Streptococcus thermophilus*	Complete	CP046134	
62	MM1	*Streptococcus thermophilus*	Complete	CP065484	
63	MM20	*Streptococcus thermophilus*	Complete	CP065485	
64	MN-BM-A01	*Streptococcus thermophilus*	Complete	CP012588	
65	MN-BM-A02	*Streptococcus thermophilus*	Complete	CP010999	
66	MN-ZLW-002	*Streptococcus thermophilus*	Complete	CP003499	
67	Moz74	*Streptococcus thermophilus*	Complete	JBTWWV000000000	Sequenced in the context of this study
68	Moz76	*Streptococcus thermophilus*	Complete	JBTWWU000000000	Sequenced in the context of this study
69	Moz77	*Streptococcus thermophilus*	Complete	JBTWWT000000000	Sequenced in the context of this study
70	Moz83	*Streptococcus thermophilus*	Complete	JBTWWS000000000	Sequenced in the context of this study
71	Moz109	*Streptococcus thermophilus*	Complete	CP075363	
72	Moz111	*Streptococcus thermophilus*	Complete	JBTWWR000000000	Sequenced in the context of this study
73	MRD6044	*Streptococcus thermophilus*	Complete	CP174173	Whole genome annotated using Prokka
74	MTH17CL396	*Streptococcus thermophilus*	Chromosome	CM002371	
75	N4L	*Streptococcus thermophilus*	Chromosome	LS974444	
76	NBC5527	*Streptococcus thermophilus*	Complete	CP176540	Whole genome annotated using Prokka
77	NCTC12958	*Streptococcus thermophilus*	Chromosome	LS483339	
78	ND03	*Streptococcus thermophilus*	Complete	CP002340	
79	ND07	*Streptococcus thermophilus*	Complete	CP016394	
80	Nect1	*Streptococcus thermophilus*	Complete	JBTWWQ000000000	Sequenced in the context of this study
81	Nect13	*Streptococcus thermophilus*	Complete	JBTWWP000000000	Sequenced in the context of this study
82	NWC_2_1	*Streptococcus thermophilus*	Complete	CP031021	
83	R1	*Streptococcus thermophilus*	Complete	CP065486	
84	Roque89	*Streptococcus thermophilus*	Complete	JBTWWO000000000	Sequenced in the context of this study
85	RR	*Streptococcus thermophilus*	Complete	CP065788	
86	S9	*Streptococcus thermophilus*	Complete	CP013939	
87	S24724	*Streptococcus thermophilus*	Complete	CP072443	
88	S24726	*Streptococcus thermophilus*	Complete	CP072442	
89	S24728	*Streptococcus thermophilus*	Complete	CP072441	
90	S24730	*Streptococcus thermophilus*	Complete	CP072440	
91	S24732	*Streptococcus thermophilus*	Complete	CP072439	

92	S24733	*Streptococcus thermophilus*	Complete	CP072438	
93	S24734	*Streptococcus thermophilus*	Complete	CP072437	
94	S24735	*Streptococcus thermophilus*	Complete	CP072436	
95	S24736	*Streptococcus thermophilus*	Complete	CP072435	
96	S24742	*Streptococcus thermophilus*	Complete	CP072434	
97	S24743	*Streptococcus thermophilus*	Complete	CP072433	
98	S24745	*Streptococcus thermophilus*	Complete	CP072431	
99	S24746	*Streptococcus thermophilus*	Complete	CP072430	
100	S24747	*Streptococcus thermophilus*	Complete	CP072429	
101	S24748	*Streptococcus thermophilus*	Complete	CP072778	
102	S24749	*Streptococcus thermophilus*	Complete	CP072428	
103	S24750	*Streptococcus thermophilus*	Complete	CP072427	
104	SCB0351	*Streptococcus thermophilus*	Complete	CP142105	
105	SMQ-301	*Streptococcus thermophilus*	Complete	CP011217	
106	ST3	*Streptococcus thermophilus*	Complete	CP017064	
107	ST19	*Streptococcus thermophilus*	Complete	CP065487	
108	ST55	*Streptococcus thermophilus*	Complete	CP065502	
109	ST057-1	*Streptococcus thermophilus*	Complete	CP102797	Whole genome annotated using Prokka
110	ST106	*Streptococcus thermophilus*	Complete	CP031881	
111	ST109	*Streptococcus thermophilus*	Complete	CP031545	
112	ST128	*Streptococcus thermophilus*	Complete	CP065500	
113	ST64987	*Streptococcus thermophilus*	Complete	CP049053	
114	STH_CIRM_16	*Streptococcus thermophilus*	Chromosome	LR822006	
115	STH_CIRM_18	*Streptococcus thermophilus*	Chromosome	LR822008	
116	STH_CIRM_19	*Streptococcus thermophilus*	Complete	LR822009	
117	STH_CIRM_23	*Streptococcus thermophilus*	Chromosome	LR822011	
118	STH_CIRM_29	*Streptococcus thermophilus*	Chromosome	LR822010	
119	STH_CIRM_30	*Streptococcus thermophilus*	Chromosome	LR822012	
120	STH_CIRM_32	*Streptococcus thermophilus*	Chromosome	LR822013	
121	STH_CIRM_36	*Streptococcus thermophilus*	Chromosome	LR822014	
122	STH_CIRM_65	*Streptococcus thermophilus*	Chromosome	LR822015	
123	STH_CIRM_67	*Streptococcus thermophilus*	Chromosome	LR824002	

124	STH_CIRM_336	*Streptococcus thermophilus*	Chromosome	LR822017	
125	STH_CIRM_368	*Streptococcus thermophilus*	Chromosome	LR822023	
126	STH_CIRM_772	*Streptococcus thermophilus*	Chromosome	LR822019	
127	STH_CIRM_956	*Streptococcus thermophilus*	Chromosome	LR822020	
128	STH_CIRM_961	*Streptococcus thermophilus*	Chromosome	LR822025	
129	STH_CIRM_967	*Streptococcus thermophilus*	Chromosome	LR822026	
130	STH_CIRM_998	*Streptococcus thermophilus*	Chromosome	LR822027	
131	STH_CIRM_1035	*Streptococcus thermophilus*	Chromosome	LR822029	
132	STH_CIRM_1046	*Streptococcus thermophilus*	Chromosome	LR822030	
133	STH_CIRM_1047	*Streptococcus thermophilus*	Chromosome	LR822031	
134	STH_CIRM_1048	*Streptococcus thermophilus*	Chromosome	LR822033	
135	STH_CIRM_1049	*Streptococcus thermophilus*	Chromosome	LR822034	
136	STH_CIRM_1050	*Streptococcus thermophilus*	Chromosome	LR822032	
137	STH_CIRM_1051	*Streptococcus thermophilus*	Chromosome	LR822035	
138	STH_CIRM_1055	*Streptococcus thermophilus*	Chromosome	LR822036	
139	STH_CIRM_1116	*Streptococcus thermophilus*	Chromosome	LR822039	
140	STH_CIRM_1121	*Streptococcus thermophilus*	Chromosome	LR822037	
141	STH_CIRM_1122	*Streptococcus thermophilus*	Chromosome	LR822041	
142	STH_CIRM_1125	*Streptococcus thermophilus*	Chromosome	LR822040	
143	STH_CIRM_1358	*Streptococcus thermophilus*	Chromosome	LR822042	
144	STH_CIRM_2101	*Streptococcus thermophilus*	Chromosome	LR822043	
145	STMM1	*Streptococcus thermophilus*	Complete	CP125881	
146	STMM2	*Streptococcus thermophilus*	Complete	CP125763	
147	STMM3	*Streptococcus thermophilus*	Complete	CP125764	
148	STMM4	*Streptococcus thermophilus*	Complete	CP125765	
149	STMM11	*Streptococcus thermophilus*	Complete	CP125766	
150	STMM22	*Streptococcus thermophilus*	Complete	CP125767	
151	STMM25	*Streptococcus thermophilus*	Complete	CP125768	
152	Strac42	*Streptococcus thermophilus*	Complete	JBTWWN000000000	Sequenced in the context of this study
153	TCI125	*Streptococcus thermophilus*	Complete	CP150875	
154	TCI633	*Streptococcus thermophilus*	Complete	CP113237	
155	TH-4	*Streptococcus thermophilus*	Complete	CP102538	

156	TH982	*Streptococcus thermophilus*	Chromosome	CM003136	
157	TH985	*Streptococcus thermophilus*	Chromosome	CM003139	
158	TH1435	*Streptococcus thermophilus*	Chromosome	CM002369	
159	TH1436	*Streptococcus thermophilus*	Chromosome	CM002370	
160	TH1477	*Streptococcus thermophilus*	Chromosome	CM003135	
161	TK-P3A	*Streptococcus thermophilus*	Complete	CP045596	
162	TSGB4141	*Streptococcus thermophilus*	Complete	CP116772	
163	TSGB4237	*Streptococcus thermophilus*	Complete	CP145130	
164	UCCSt50	*Streptococcus thermophilus*	Complete	CP194174	
165	UCCSt89	*Streptococcus thermophilus*	Complete	JBTWWM000000000	Sequenced in the context of this study
166	UCCSt95	*Streptococcus thermophilus*	Complete	CP101646	
167	Vach60	*Streptococcus thermophilus*	Complete	JBTWWL000000000	Sequenced in the context of this study
168	VHProbi R08	*Streptococcus thermophilus*	Complete	CP094945	
169	57.1	*Streptococcus salivarius*	Complete	CP002888	
170	ATCC 27945	*Streptococcus salivarius*	Complete	CP015282	
171	ATCC25975	*Streptococcus salivarius*	Complete	CP015283	
172	CCHSS3	*Streptococcus salivarius*	Complete	FR873481	
173	DB-B5	*Streptococcus salivarius*	Complete	CP054153	
174	E636	*Streptococcus salivarius*	Complete	CP190755	
175	FDAARGOS_259	*Streptococcus salivarius*	Complete	CP020451	
176	FDAARGOS_771	*Streptococcus salivarius*	Complete	CP053998	
177	FDAARGOS_1045	*Streptococcus salivarius*	Complete	CP066093	
178	HSISS4	*Streptococcus salivarius*	Complete	CP013216	
179	ICDC1	*Streptococcus salivarius*	Complete	CP018186	
180	ICDC2	*Streptococcus salivarius*	Complete	CP018187	
181	ICDC3	*Streptococcus salivarius*	Complete	CP018189	
182	JF	*Streptococcus salivarius*	Complete	CP014144	
183	JIM8777	*Streptococcus salivarius*	Complete	FR873482	
184	KSS1	*Streptococcus salivarius*	Complete	CP145856	
185	KSS2	*Streptococcus salivarius*	Complete	CP145860	
186	KSS3	*Streptococcus salivarius*	Complete	CP145861	
187	KSS4	*Streptococcus salivarius*	Complete	CP145862	

188	KSS5	*Streptococcus salivarius*	Complete	CP150847	
189	KSS6	*Streptococcus salivarius*	Complete	CP145863	
190	KSS7	*Streptococcus salivarius*	Complete	CP145864	
191	KSS8	*Streptococcus salivarius*	Complete	CP145865	
192	KSS9	*Streptococcus salivarius*	Complete	CP145866	
193	KSS10	*Streptococcus salivarius*	Complete	CP145867	
194	KSS11	*Streptococcus salivarius*	Complete	CP145858	
195	KSS12	*Streptococcus salivarius*	Complete	CP145868	
196	LAB813	*Streptococcus salivarius*	Complete	CP040804	
197	LMG 11489	*Streptococcus salivarius*	Complete	CP133476	
198	MRD1919	*Streptococcus salivarius*	Complete	CP177181	Whole genome annotated using Prokka
199	MRD4434	*Streptococcus salivarius*	Chromosome	CP174172	Whole genome annotated using Prokka
200	MRD6060	*Streptococcus salivarius*	Complete	CP177182	Whole genome annotated using Prokka
201	MRD-KRBY	*Streptococcus salivarius*	Complete	CP177180	Whole genome annotated using Prokka
202	MRD-NRLLH	*Streptococcus salivarius*	Complete	CP177179	Whole genome annotated using Prokka
203	NBRC 13956	*Streptococcus salivarius*	Complete	AP031488	
204	NCTC_8618	*Streptococcus salivarius*	Complete	CP009913	
205	NCTC7366	*Streptococcus salivarius*	Chromosome	LS483366	
206	NCTC8618	*Streptococcus salivarius*	Chromosome	LR134274	
207	SALI-10	*Streptococcus salivarius*	Complete	CP090007	
208	22-06 S6	*Streptococcus vestibularis*	Scaffold	LXZX00000000	*rgp* cluster found in LXZX01000005
209	ATCC 49124	*Streptococcus vestibularis*	Scaffold	AEVI00000000	*rgp* cluster found in GL831112 AEVI01000000
210	B13320257	*Streptococcus vestibularis*	Scaffold	JAKUUZ000000000	*rgp* cluster found in JAKUUZ010000024
211	DP3_6A_2	*Streptococcus vestibularis*	Scaffold	JAQMFT000000000	*rgp* cluster found in JAQMFT010000101

212	DP3_9B	*Streptococcus vestibularis*	Scaffold	JAQMFV000000000	*rgp* cluster found in JAQMFV010000125
213	E653	*Streptococcus vestibularis*	Complete	CP185269	
214	MGYG-HGUT-02302	*Streptococcus vestibularis*	Scaffold	CABMFU000000000	*rgp* cluster found in CABMFU010000005
215	N4_129_000G1_dasN4_129_000G1_concoct_5	*Streptococcus vestibularis*	Scaffold	JAWHJX000000000	*rgp* cluster found in JAWHJX010000004
216	NCTC12167	*Streptococcus vestibularis*	Chromosome	LR134275	
217	O4-4	*Streptococcus vestibularis*	Scaffold	JALDVX000000000	*rgp* cluster found in JALDVX010000006
218	OM08-1	*Streptococcus vestibularis*	Scaffold	QSTK00000000	*rgp* cluster found in QSTK01000005
219	S06	*Streptococcus vestibularis*	Scaffold	JALDVO000000000	*rgp* cluster found in JALDVO010000009
220	LM8	*Streptococcus raffinosi*	Scaffold	JBOZIT000000000	JBOZIT010000005
221	VTCC 12812	*Streptococcus raffinosi*	Scaffold	JASUZV000000000	JASUZV010000002
222	VTCC 12813	*Streptococcus raffinosi*	Scaffold	JASZXY000000000	JASZXY010000001
223	VTCC 12814	*Streptococcus raffinosi*	Scaffold	JASZXZ000000000	JASZXZ010000006

Supplementary Table S3. Summary of *rgp* genotypes identified in this study. For each strain, the variable side chain (Vt) and rhamnan backbone (Bt) types encoded within the *rgp* locus are indicated, along with the resulting Vt–Bt combination that defines the *rgp* genotype. This overview highlights the modular organisation of the *rgp* locus and the range of backbone–side chain combinations observed across the analysed salivarius group strains.

Strain	**Species**	***rgp***	**Vt**	**Bt**	**VtBt**	**Previously classified /source**
UCCSt50	*S. thermophilus*	1	Vt1	Bt1	Vt1Bt1	(38)
UCCSt95	*S. thermophilus*	2	Vt1	Bt2	Vt1Bt2	(38)
UCCSt89	*S. thermophilus*	3	Vt3	Bt2	Vt3Bt2	(38)
Nect1	*S. thermophilus*	4A	Vt3	Bt3	Vt3Bt3	This study
Nect13	*S. thermophilus*	4B	Vt3	Bt3	Vt3Bt3	This study
Douc24	*S. thermophilus*	5A	Vtx	Bt4	VtxBt4	This study
Moz74	*S. thermophilus*	5B	Vtx	Bt4	VtxBt4	This study
ST64987	*S. thermophilus*	6	Vt4	Bt1	Vt4Bt1	(38)
CNRZ302	*S. thermophilus*	7	Vt5	Bt1	Vt5Bt1	This study/(38)
STMM11	*S. thermophilus*	8	Vt9	Bt2	Vt9Bt2	This study
Brie28	*S. thermophilus*	9	Vt2	Bt2	Vt2Bt2	This study/(38)
ICDC1	*S. salivarius*	10	Vtx	Bt6	VtxBt6	This study
LM8	*S. raffinosi*	11	Vt10	Bt4	Vt10Bt4	This study
O4-4	*S. vestibularis*	12	Vt8	Bt2	Vt8Bt2	This study
KSS11	*S. salivarius*	13	Vt7	Bt5	Vt7Bt5	This study
LAB813	*S. salivarius*	14	Vt6	Bt1	Vt6Bt1	This study

Supplementary Table S4. NMR data for *S. thermophilus* strain UCCSt12 RGP degradation products (600 MHz, 25 °C, δ ppm).

Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
β-Gal E OS1	H	4.54	3.60	3.66	3.94	3.71	3.77; 3.77
	C	104.5	71.9	73.7	69.7	76.2	62.1
anhTal D OS1	H	5.08	3.93	4.40	4.41	4.15	3.78; 3.84
	C	91.0	83.0	82.5	73.2	82.3	61.2
Gral3d B 2-19	H	5.12	3.64	3.72; 3.85			
	C	90.3	81.1	60.4			
α-Rha A 2-19	H	5.04	4.30	3.93	3.52	3.85	1.29
	C	103.2	71.2	81.1	72.2	70.6	17.8
α-Rha C 2-19	H	4.97	4.10	3.88	3.54	3.98	1.28
	C	100.7	71.3	79.6	72.5	70.3	17.8
β-GalNAc D 2-19	H	4.66	3.95	3.75	3.94	3.68	3.77; 3.82
and 2-20	C	104.5	53.8	72.2	69.0	76.2	62.3
Gral3d K 2-20	H	5.08	3.64	3.70; 3.85			
	C	90.4	81.2	60.7			
α-Rha L 2-20	H	5.20	4.07	3.79	3.45	3.70	1.27
	C	102.9	71.1	71.2	73.3	70.4	17.8
α-Rha M 2-20	H	5.05	4.26	3.95	3.54	3.96	1.28
	C	100.1	78.0	81.2	72.4	70.4	17.8

Supplementary Table S5. ^{1}H and ^{13}C NMR data (δ, ppm, D_2O, 35°C, 500 MHz) for the RGP from *Streptococcus thermophilus* strain Douc24.

Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
α-Rha A	H	5.28	4.05	3.81	3.53	3.73	1.34
	C	101.6	78.8	71.3	73.3	70.5	18.0
α-Rha B	H	5.13	4.17	3.99	3.71	3.79	1.30
	C	100.7	78.3	75.7	73.6	71.1	17.8
α-Rha C	H	5.12	4.13	3.91	3.47	3.71	1.28
	C	102.0	79.5	71.0	73.5	70.5	17.8
α-Glc D	H	5.07	3.56	3.74	3.68	4.00	3.70; 4.13
	C	99.6	72.4	74.2	70.2	72.2	66.2
α-Glc E	H	4.98	3.56	3.74	3.44	3.70	3.77; 3.85
	C	99.2	72.8	74.2	70.7	73.1	61.7

Supplementary Table S6. ^{1}H and ^{13}C NMR data (δ, ppm, D_2O, 40°C, 600 MHz) for the polysaccharide from *Streptococcus thermophilus* CNRZ302 (PS) and its derivatives DOS and DPS (25 °C). NOE: A1:G3; B1:F2; C1:F4; D1:K3; E1:B1,2; F1:E4,6; G1:D1,2; K1:B3. HMBC: A1:G3; B1:F2; C1:F4; D1:K3; F1:E6; G1:D2; K1:B3.

Sugar		1	2	3	4	5	6
β-Galf A PS	H	5.23	4.20	4.07	4.06	3.84	3.66; 3.70
	C	110.1	82.6	78.0	84.1	71.9	63.9
β-Galf A DOS		5.23	4.20	4.07	4.05	3.84	3.66; 3.70
		110.2	82.6	77.9	83.9	71.8	63.8
α-Rha B PS	H	5.09	4.29	3.95	3.57	3.81	1.33
	C	100.7	77.9	79.7	72.8	71.1	17.8
α-Rha B DPS	H	5.10	4.14	3.92	3.52	3.81	1.31

	C	100.8	78.2	70.6	73.1	70.7	17.8
α-Glc C PS	H	5.07	3.57	3.71	3.46	3.99	3.78; 3.84
	C	100.7	72.8	74.1	70.7	73.1	61.6
α-Glc C DPS	H	5.08	3.56	3.70	3.45	4.01	3.79; 3.84
	C	100.7	72.6	73.9	70.6	73.0	61.5
α-Rha D PS	H	5.02	3.89	3.85	3.52	4.01	1.27
	C	99.4	77.6	70.8	73.1	70.4	17.7
α-Rha D DOS		5.09	4.03	3.90	3.55	3.87	1.32
		98.1	77.0	70.7	73.1	70.4	17.6
α-Glc E PS	H	4.98	3.52	3.71	3.59	4.22	3.94; 3.94
	C	99.3	72.6	74.1	70.1	71.5	67.2
α-Glc E DPS	H	5.02	3.56	3.76	3.48	4.18	3.72; 3.95
	C	99.3	72.6	73.9	70.6	72.0	67.5
α-Rha F PS	H	4.93	4.07	4.05	3.53	3.90	1.44
	C	100.4	79.3	69.8	82.7	69.1	18.6
α-Rha F DPS	H	4.94	4.05	4.05	3.55	3.89	1.40
	C	99.9	80.2	69.7	82.4	69.2	18.1
α-Gal G PS	H	4.92	3.95	4.00	4.17	4.23	3.73; 3.73
	C	98.8	68.5	78.1	70.3	72.0	62.0
α-Gal G DOS		5.05	3.94	4.01	4.16	4.27	3.72; 3.72
		98.6	68.4	78.0	70.3	71.9	62.0
α-Gal H	H	4.90	3.83	3.92	4.02	4.20	3.73; 3.73
	C	98.8	69.5	70.5	70.3	72.1	62.0
β-GlcNAc K PS	H	4.74	3.74	3.67	3.49	3.47	3.76; 3.96
	C	103.9	57.0	82.5	70.1	77.0	62.1

anh-Man K DOS		5.09	3.84	4.22	4.18	3.95	3.68; 3.77
		90.7	84.6	85.1	76.5	84.2	61.7

Supplementary Table S7. ^{1}H and ^{13}C NMR data (δ, ppm, D_2O, 40°C, 600 MHz) for the PS1, DPS and OS1 from *Streptococcus thermophilus* strain STMM11. NAc: 2.09/23.6 ppm.

Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
β-Gal*f* A PS1	H	5.23	4.20	4.08	4.06	3.84	3.67; 3.70
	C	110.1	82.6	78.0	84.0	71.9	63.9
β-Gal*f* A OS1	H	5.23	4.20	4.08	4.06	3.84	3.67; 3.71
	C	110.1	82.6	77.9	84.0	71.9	63.9
-2-α-Rha B PS1, DPS	H	5.19	4.08	3.95	3.51	3.83	1.32
	C	101.9	79.2	71.1	73.2	70.4	18.0
-2-α-Rha C,D PS1	H	5.02	3.88	3.83	3.51	4.01	1.27
	C	99.3	77.5	70.8	73.2	70.4	17.8
-2-α-Rha C OS1	H	5.09	4.02	3.90	3.56	3.86	1.32
	C	98.2	77.2	70.9	73.1	70.5	17.7
-3-α-Rha G DPS	H	4.96	4.16	3.84	3.57	3.77	1.29
	C	103.1	71.1	78.6	72.7	70.5	17.9
-3-α-Rha P PS1	H	4.93	4.16	3.84	3.57	3.77	1.29
	C	103.3	71.1	78.6	72.7	70.5	17.9
-2,4-α-Rha X PS1	H	5.22	4.01	4.03	3.63	3.83	1.37
	C	101.5	80.0	71.2	81.7	68.6	18.5
-3-α-Gal K PS1	H	4.93	3.96	4.00	4.17	4.23	3.73; 3.73
	C	98.7	68.5	78.1	70.3	71.9	62.0
-3-α-Gal K DPS		5.05	3.95	4.00	4.16	4.27	3.73; 3.73
		98.8	68.4	78.1	70.3	71.9	62.0
α-Gal L PS1	H	4.91	3.83	3.93	4.02	4.21	3.73; 3.73
	C	98.7	69.4	70.4	70.3	72.1	62.0

β-GlcNAc N PS1	H	4.78	3.81	3.66	3.57	3.40	3.80; 3.90
	C	102.4	57.0	82.5	69.4	76.8	61.6
anh-Man N OS1		5.10	3.85	4.23	4.18	3.96	3.68; 3.77
		90.7	84.9	85.1	76.8	84.4	61.8

Supplementary Table S8. ^{1}H and ^{13}C NMR data (δ, ppm, D_2O, 45°C, 600 MHz) for the polysaccharide from *Streptococcus thermophilus* Brie28. NOE: A1:B3; A'1:B'3; B1:A2; B'1:A'2; G1:H3; H1:A'4; K1:G4.

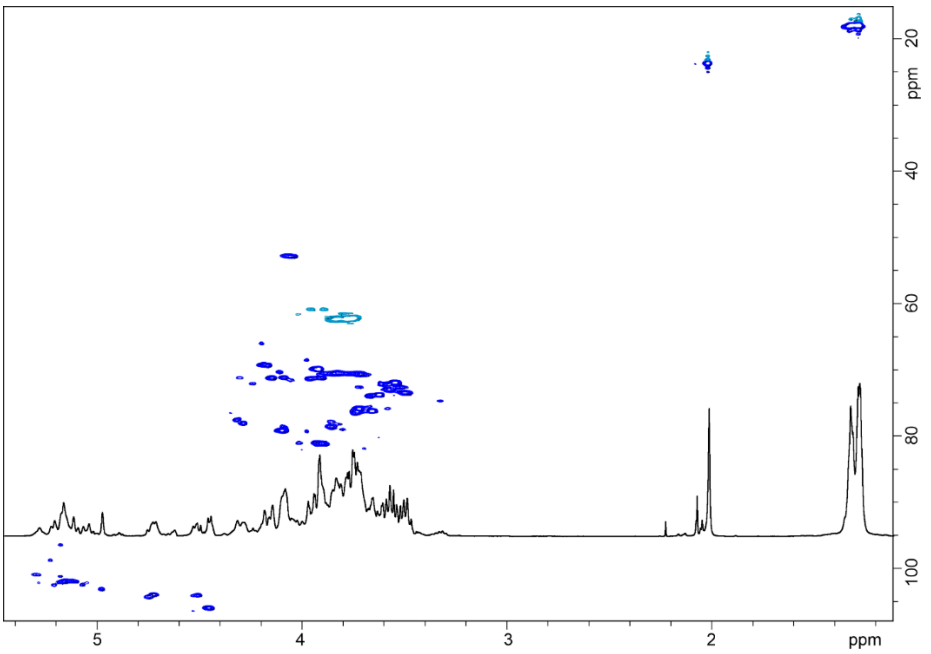
Sugar		1	2	3	4	5	6
α-Rha A DPS	H	5.20	4.09	3.96	3.51	3.84	1.33
	C	102.0	79.2	71.2	73.4	70.5	18.0
α-Rha A PS	H	5.20	4.09	3.96	3.51	3.84	1.33
	C	102.0	79.2	71.2	73.4	70.5	18.0
α-Rha A' PS	H	5.22	4.03	4.06	3.66	3.85	1.38
	C	101.7	79.7	71.3	81.6	68.8	18.4
α-Rha B DPS	H	4.97	4.17	3.86	3.57	3.77	1.29
	C	103.3	71.1	78.8	72.9	70.6	17.9
α-Rha B PS	H	4.97	4.15	3.86	3.57	3.77	1.29
	C	103.3	71.1	78.8	72.9	70.6	17.9
α-Rha G DOS	H	4.96	4.01	4.06	3.72	3.98	1.37
	C	101.9	71.3	71.3	81.8	69.0	18.1
α-Rha G PS	H	4.90	3.86	4.03	3.70	3.89	1.35
	C	103.3	71.6	71.3	82.0	69.0	18.2
anh-Tal H DOS	H	5.05	3.86	4.27	4.32	4.11	3.76; 3.83
	C	91.3	82.6	80.7	72.6	82.5	61.3
β-GalNAc H PS	H	4.77	4.00	3.78	4.01	3.62	3.77; 3.77

	C	102.9	53.1	80.7	68.7	75.9	61.9
β-Gal K DOS		4.66	3.55	3.67	3.92	3.69	3.76; 3.79
		104.8	72.8	73.9	69.7	76.3	62.0
β-Gal K PS		4.65	3.55	3.67	3.93	3.68	3.77; 3.77
		104.9	72.9	74.0	69.8	76.4	61.9

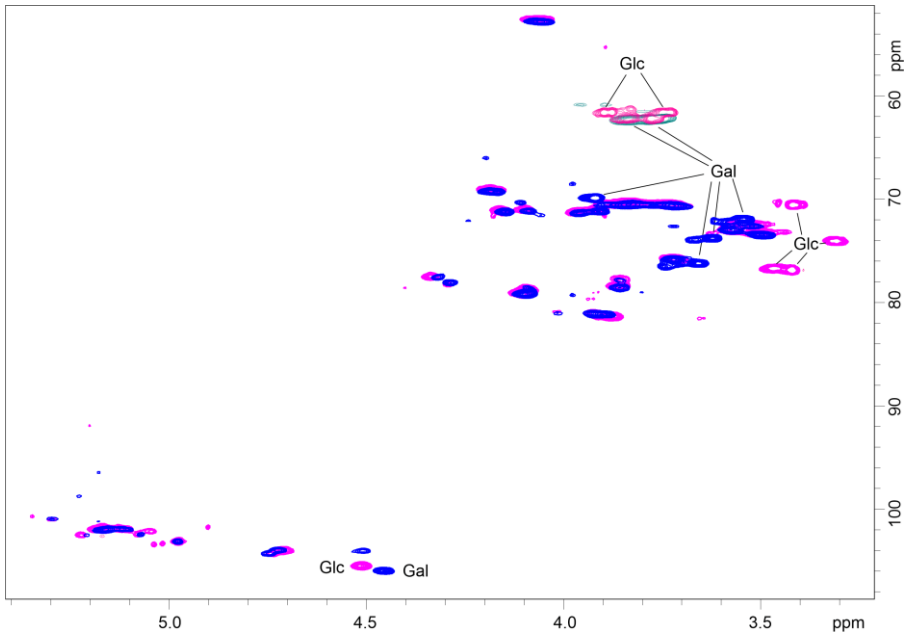
Supplementary Table S9. List of primers designed and used in this study located upstream of the predicted promoter and within the identified transcriptional units

Promoter region	**Sequence (5' – 3')**	**Outcome**	**Product size (bp)**
UCCSt50			
Prom1			
S1upsF	gacgtcctaagtttcatgtgtcttt	No Product	570bp
S1intF	gagcagtaccaatatcgtcgac	Product	268bp
S1R	ttgccttggcacatctatactaactg		
Prom2			
S2upsF	gctaaggctaaggcgacag	No Product	580bp
S2intF	gcctatgcgattgcgaatctctt	Product	247bp
S2R	cttgaaaaacatccgataaagcttc		
Prom3			
S3upsF	gtcaagaagatttgtatatgtatcaac	No product	882bp
S3intF	cagtgtctgcgaagtgtttttaatca	Product	521 bp
S3R	gcttggttgatataccgctttag		
Prom4			
S4upsF	catggggatgtcataccaatca	No product	548bp
S4intF	gagagatgtgacgaatggaggta	Product	233bp
S4R	cctagctttggaggattgacc		
Douc24			
Prom1			
S1upsF	S1upsF	No product	874
S1intDF	gattgataaagctttaatagaagatattaagag	Product	283
S1DR	ttgtaactgcaatacctaagcggtctg		
Prom2			
S2upsDF	ggaatccaactcttagccttaggtatc	No product	712
S2intDF	gaacaaccagtaagatcattagctgat	Product	348
S2DR	ccttcaaccatagtaagtctaggac		
Prom3			
S3upsDF	gcttgaatctcgatatgagtgtcagtt	No product	829
S3intDF	ggacctgctcgcgttattgcatat	Product	303
S3DR	tataggtttcaattcctggattatc		

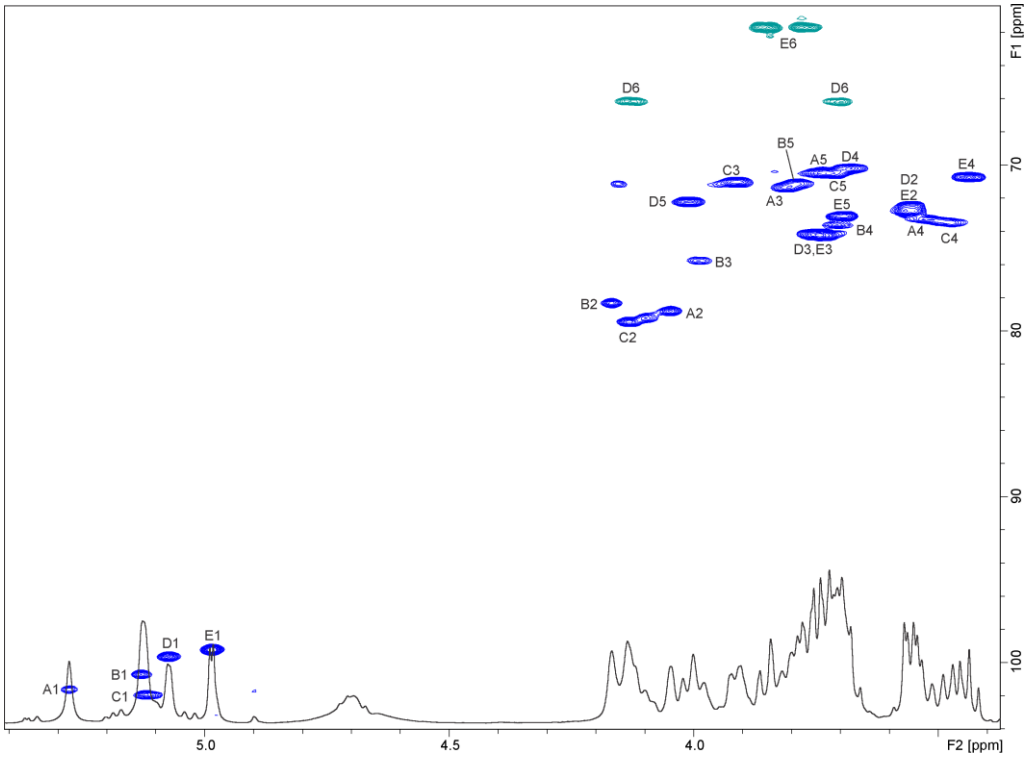
Prom4			
S4upsDF	cacgccatttgtagataacgttgtattgaa	No product	939
S4intDF	ctggtctgattgcaacaatgacgaa	Product	299
S4DR	agctaactcggattttagaacctggag		
Brie28			
Prom1			
S1upsBF	gtgatggaccagattcttcagcttgat	No product	469
S1intF	S1intF	Product	270
S1R	S1R		
Prom2			
S2upsBF	ccaagaaacggccggtgtatattg	No product	692
S2intBF	catggtgcaacgcttgtttatatttca	Product	261
S2BR	taacgactgtcagcgtcttgtgag		
Prom3			
S3upsBF	gtcaaatgatgcagcagaagatacaac	No product	724
S2intF	S2intF	Product	246
S2R	S2R		
Prom4			
S4upsBF	cttggtataaaattggtagtatgaatggtt	No product	706
S4intBF	cgagtatcaccgaatatcaagtaccat	Product	286
S4BR	ggattgatatagaattgaccgcccatc		
Prom5			
S5upsBF	cggcattgctgaaactgttacgag	No product	673
S5intBF	actaagtctgagtcggttctatct	Product	266
S5BR	caccagatgctcctgaggc		



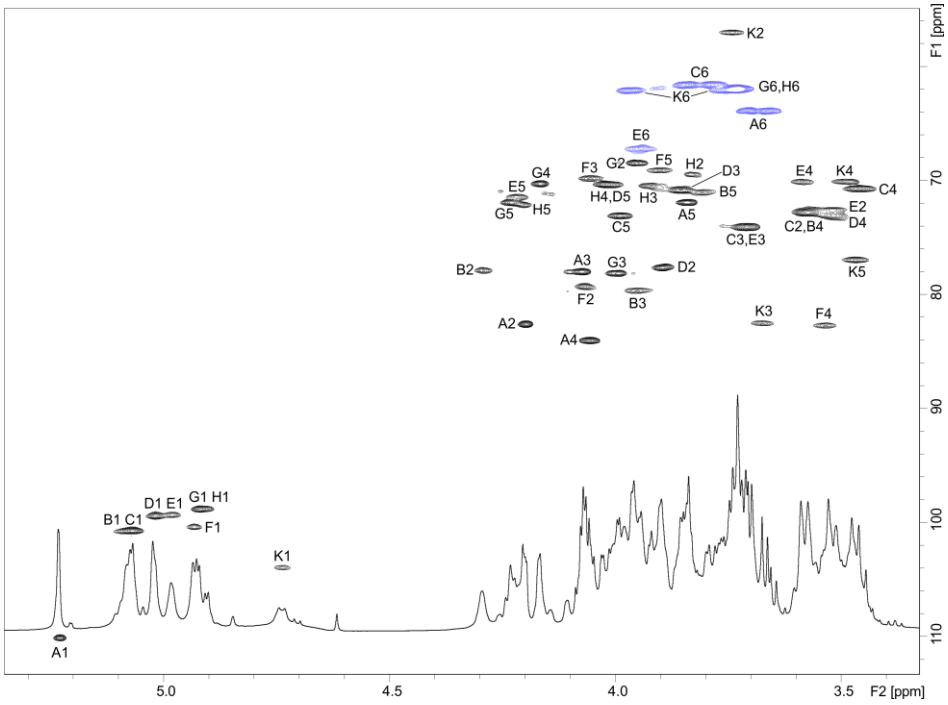
Supplementary Fig. S1. ^{1}H-^{13}C HSQC spectrum of RGP from *S. thermophilus* UCCSt12. The Nect13 structure was not analysed in full detail, it was determined to be the same as UCCSt12 with a terminal glucose instead of galactose.



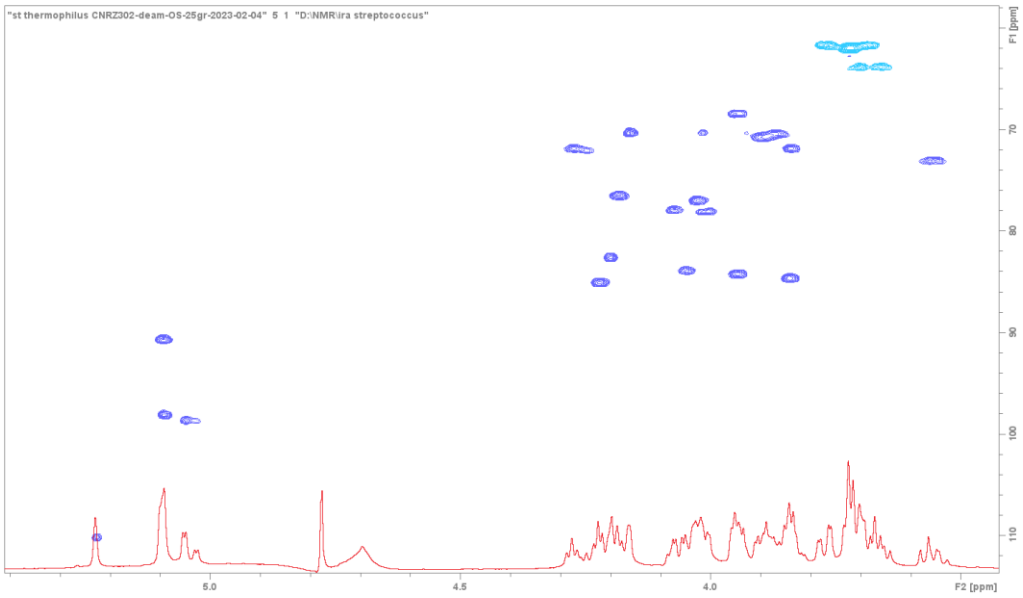
Supplementary Fig. S2. Overlap of UCCSt12 (blue-green) and Nect13 (magenta-red) HSQC spectra. Sugar differences are indicated.



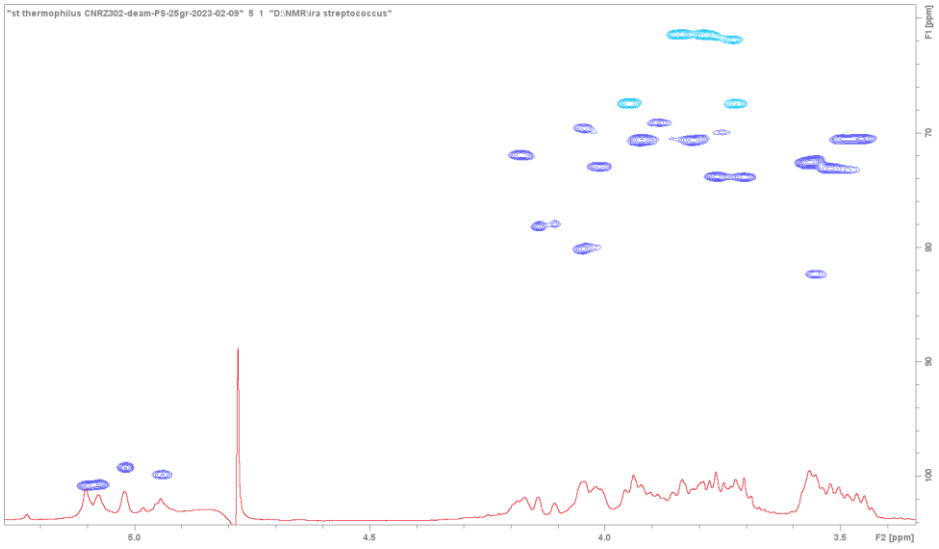
Supplementary Fig. S3. ^{1}H-^{13}C HSQC spectrum of the rhamnan from *S. thermophilus* strain Douc24, D_2O, 35 °C, 500 MHz.



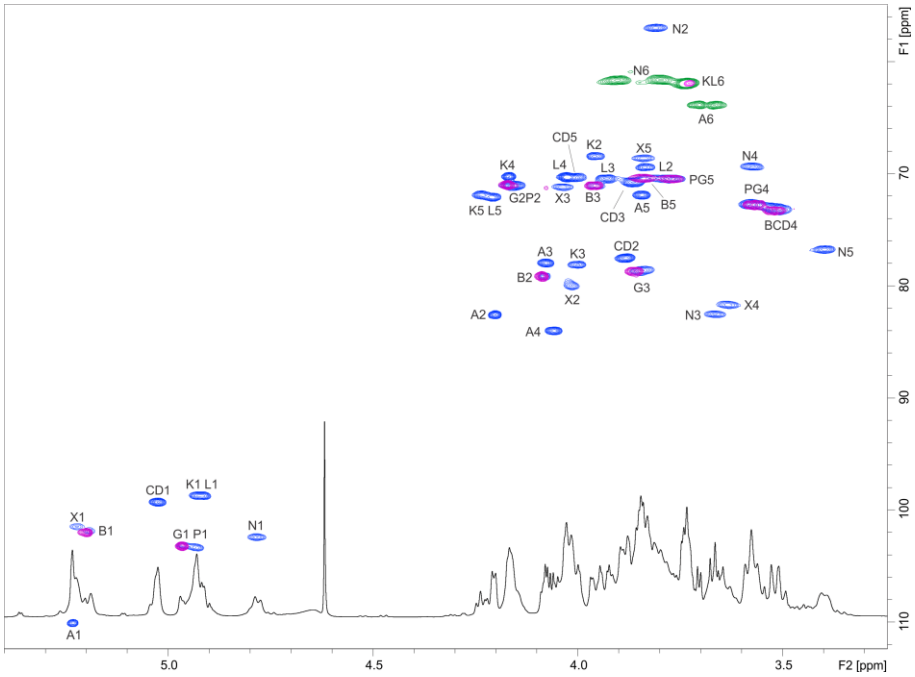
Supplementary Fig. S4. ^{1}H-^{13}C HSQC spectrum of the polysaccharide from *Streptococcus thermophilus* CNRZ302.



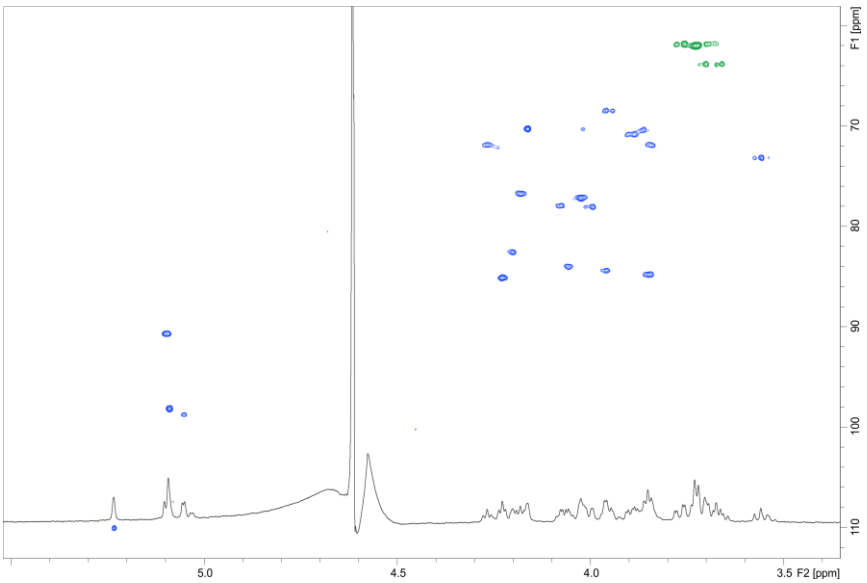
Supplementary Fig. S5. ^{1}H-^{13}C HSQC spectrum of the DOS from *Streptococcus thermophilus* CNRZ302.



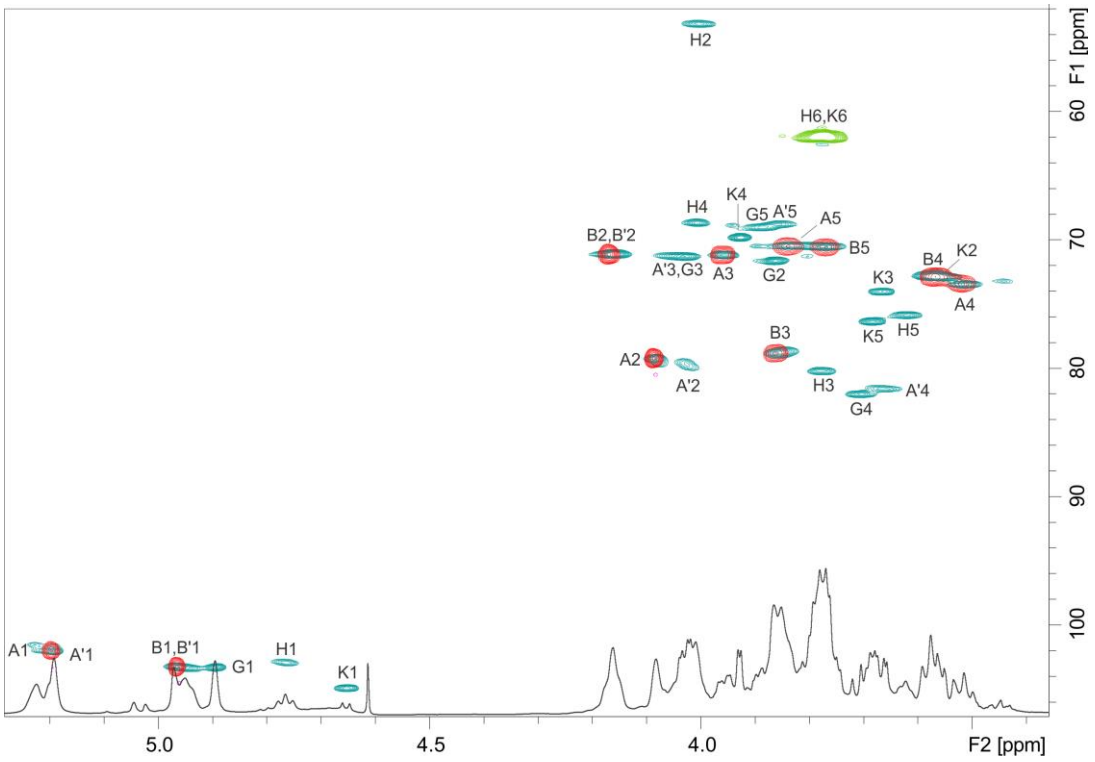
Supplementary Fig. S6. ^{1}H-^{13}C HSQC spectrum of the DPS from *Streptococcus thermophilus* CNRZ302.



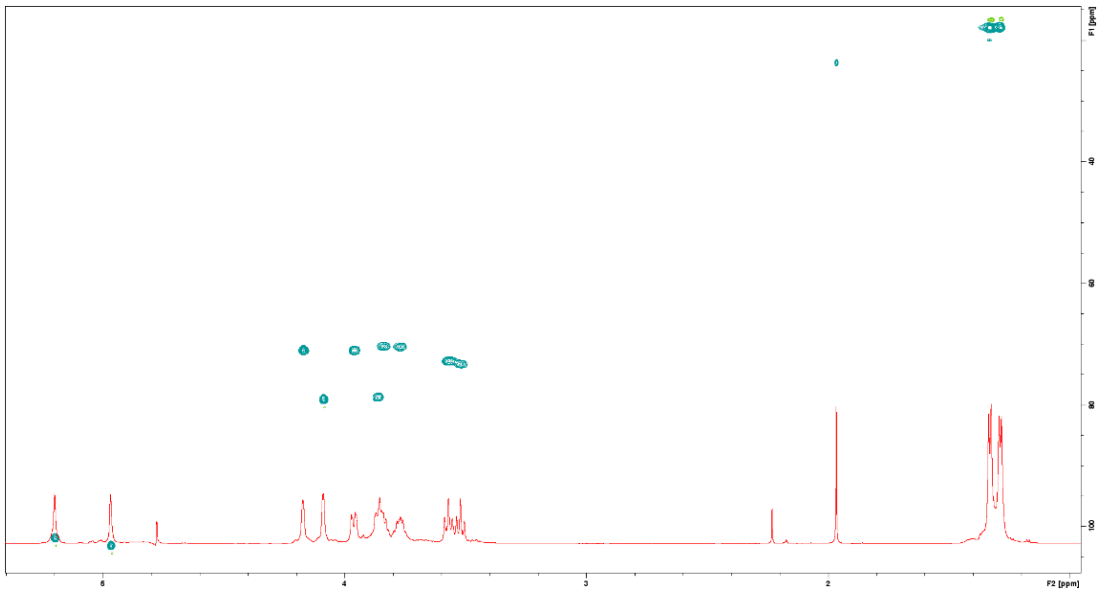
Supplementary Fig. S7. ^{1}H-^{13}C HSQC spectrum of the PS1 (blue-green) and deaminated PS (magenta) from *S. thermophilus* strain STMM11, D_2O, 40 °C, 600 MHz.



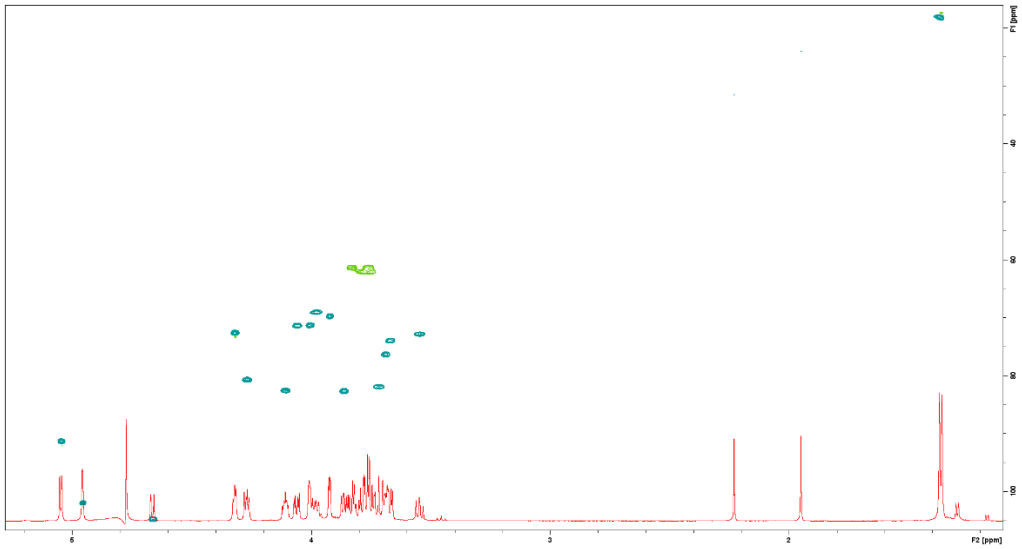
Supplementary Fig. S8. ^{1}H-^{13}C HSQC spectrum of the OS mixture (most of the OS2 signals are not visible because of low intensity) from *S. thermophilus* strain STMM11 (D_2O, 25 °C, 600 MHz).



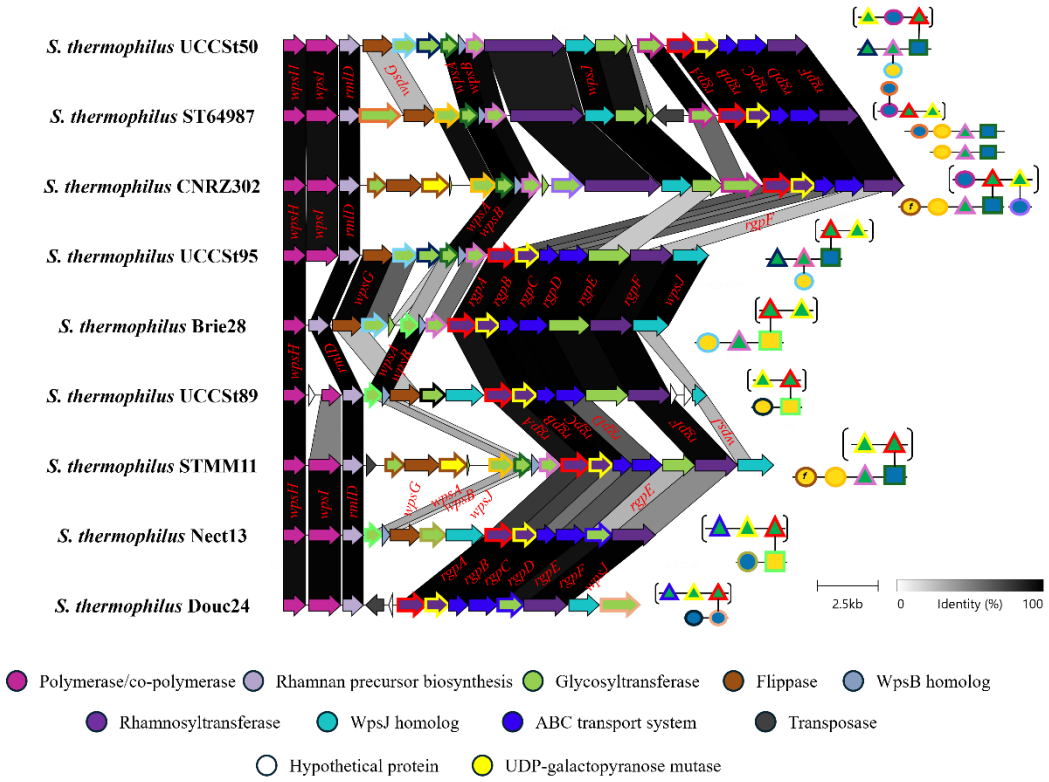
Supplementary Fig. S9. ^{1}H-^{13}C HSQC spectra of the polysaccharide from *Streptococcus thermophilus* Brie28 (green-cyan) and DPS (red).



Supplementary Fig. S10. ^{1}H-^{13}C HSQC spectra of the DPS from *Streptococcus thermophilus* Brie28.



Supplementary Fig. S11. H-^{13}C HSQC spectra of the DOS from *Streptococcus thermophilus* Brie28.



Supplementary Fig. S12. Linking *rgp* genetic content to RGP structures in representative *S. thermophilus* strains. Schematic overview of representative *rgp* clusters with their corresponding RGP structures. Genes which encode products that are putatively involved in the synthesis or modification of the RGP structure are outlined in a distinct colour, with matching colours used to indicate the potentially associated monosaccharide or structural feature within the polysaccharide. The predicted functions of genes are colour-coded and described at the base of the figure. Regions of homology are joined by blocks of different shades of grey to black and generated using CAGECAT (58). Monosaccharide symbols follow the Standard Nomenclature for Glycans (SNFG).

References

1. Facklam R. What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. Clin Microbiol Rev. 2002 Oct;15(4):613–30. doi:10.1128/CMR.15.4.613-630.2002

2. Doern CD, Carey-Ann BD. It's Not Easy Being Green: the Viridans Group Streptococci, with a Focus on Pediatric Clinical Manifestations. J Clin Microbiol. 2010 Nov;48(11):3829–35. doi:10.1128/JCM.01563-10

3. Thompson CC, Emmel VE, Fonseca EL, Marin MA, Vicente ACP. Streptococcal taxonomy based on genome sequence analyses. F1000Res. 2013 Mar 1;2:67. doi:10.12688/f1000research.2-67.v1

4. Richards VP, Palmer SR, Pavinski Bitar PD, Qin X, Weinstock GM, Highlander SK, et al. Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. Genome Biol Evol. 2014 Apr;6(4):741–53. doi:10.1093/gbe/evu048 PubMed PMID: 24625962; PubMed Central PMCID: PMC4007547.

5. Patel S, Gupta RS. Robust demarcation of fourteen different species groups within the genus *Streptococcus* based on genome-based phylogenies and molecular signatures. Infection, Genetics and Evolution. 2018 Dec;66:130–51. doi:10.1016/j.meegid.2018.09.020

6. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. Genome Res. 2020 Nov;30(11):1667–79. doi:10.1101/gr.260828.120

7. Gao XY, Zhi XY, Li HW, Klenk HP, Li WJ. Comparative Genomics of the Bacterial Genus *Streptococcus* Illuminates Evolutionary Implications of Species Groups. Reid SD, editor. PLoS ONE. 2014 Jun 30;9(6):e101229. doi:10.1371/journal.pone.0101229

8. Teng JLL, Ma Y, Chen JHK, Luo R, Foo CH, Li TT, et al. *Streptococcus oriscaviae* sp. nov. Infection Associated with Guinea Pigs. Kovac J, editor. Microbiol Spectr. 2022 Jun 29;10(3):e00014-22. doi:10.1128/spectrum.00014-22

9. Delorme C, Abraham AL, Renault P, Guédon E. Genomics of *Streptococcus salivarius*, a major human commensal. Infection, Genetics and Evolution. 2015 Jul;33:381–92. doi:10.1016/j.meegid.2014.10.001

10. Nguyen HV, Trinh ATV, Bui LNH, Hoang ATL, Tran QTL, Trinh TT. *Streptococcus raffinosi* sp. nov., isolated from human breast milk samples. International Journal of Systematic and Evolutionary Microbiology. 2024 Jul 3;74(7). doi:10.1099/ijsem.0.006442

11. Mistou MY, Sutcliffe IC, Van Sorge NM. Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria. Bitter W, editor. FEMS Microbiology Reviews. 2016 Jul;40(4):464–79. doi:10.1093/femsre/fuw006

12. Guérin H, Kulakauskas S, Chapot-Chartier MP. Structural variations and roles of rhamnose-rich cell wall polysaccharides in Gram-positive bacteria. Journal of Biological Chemistry. 2022 Oct;298(10):102488. doi:10.1016/j.jbc.2022.102488

13. Lavelle K, Sinderen DV, Mahony J. Cell wall polysaccharides of Gram positive ovococcoid bacteria and their role as bacteriophage receptors. Computational and Structural Biotechnology Journal. 2021;19:4018–31. doi:10.1016/j.csbj.2021.07.011

14. De A, Liao S, Bitoun JP, Roth R, Beatty WL, Wu H, et al. Deficiency of RgpG Causes Major Defects in Cell Division and Biofilm Formation, and Deficiency of LytR-CpsA-Psr Family Proteins Leads to Accumulation of Cell Wall Antigens in Culture Medium by *Streptococcus mutans*. Nojiri H, editor. Appl Environ Microbiol. 2017 Sep;83(17):e00928-17. doi:10.1128/AEM.00928-17

15. Kovacs CJ, Faustoferri RC, Bischer AP, Quivey RG. *Streptococcus mutans* requires mature rhamnose-glucose polysaccharides for proper pathophysiology, morphogenesis and cellular division. Molecular Microbiology. 2019 Sep;112(3):944–59. doi:10.1111/mmi.14330

16. Rush JS, Zamakhaeva S, Murner NR, Deng P, Morris AJ, Kenner CW, et al. Structure and mechanism of biosynthesis of *Streptococcus mutans* cell wall polysaccharide. Nat Commun. 2025 Jan 22;16(1):954. doi:10.1038/s41467-025-56205-1

17. Shibata Y, Yamashita Y, Van Der Ploeg JR. The serotype-specific glucose side chain of rhamnose–glucose polysaccharides is essential for adsorption of bacteriophage M102 to *Streptococcus mutans*. FEMS Microbiology Letters. 2009 May;294(1):68–73. doi:10.1111/j.1574-6968.2009.01546.x

18. Ozaki K, Shibata Y, Yamashita Y, Nakano Y, Tsuda H, Koga T. A novel mechanism for glucose side-chain formation in rhamnose-glucose polysaccharide synthesis. FEBS Letters. 2002 Dec 4;532(1–2):159–63. doi:10.1016/S0014-5793(02)03661-X

19. Yamashita Y, Shibata Y, Nakano Y, Tsuda H, Kido N, Ohta M, et al. A Novel Gene Required for Rhamnose-Glucose Polysaccharide Synthesis in *Streptococcus*

mutans. J Bacteriol. 1999 Oct 15;181(20):6556–9. doi:10.1128/JB.181.20.6556-6559.1999

20. Tsukioka Y, Yamashita Y, Oho T, Nakano Y, Koga T. Biological function of the dTDP-rhamnose synthesis pathway in *Streptococcus mutans*. J Bacteriol. 1997 Feb;179(4):1126–34. doi:10.1128/jb.179.4.1126-1134.1997

21. Zamakhaeva S, Chaton CT, Rush JS, Ajay Castro S, Kenner CW, Yarawsky AE, et al. Modification of cell wall polysaccharide guides cell division in *Streptococcus mutans*. Nat Chem Biol. 2021 Aug;17(8):878–87. doi:10.1038/s41589-021-00803-9

22. Yamashita Y, Tsukioka Y, Tomihisa K, Nakano Y, Koga T. Genes Involved in Cell Wall Localization and Side Chain Formation of Rhamnose-Glucose Polysaccharide in *Streptococcus mutans*. J Bacteriol. 1998 Nov;180(21):5803–7. doi:10.1128/JB.180.21.5803-5807.1998

23. Rush JS, Edgar RJ, Deng P, Chen J, Zhu H, Van Sorge NM, et al. The molecular mechanism of N-acetylglucosamine side-chain attachment to the Lancefield group A carbohydrate in *Streptococcus pyogenes*. Journal of Biological Chemistry. 2017 Nov;292(47):19441–57. doi:10.1074/jbc.M117.815910

24. van Sorge NM, Cole JN, Kuipers K, Henningham A, Aziz RK, Kasirer-Friede A, et al. The Classical Lancefield Antigen of Group A *Streptococcus* Is a Virulence Determinant with Implications for Vaccine Design. Cell Host & Microbe. 2014 Jun;15(6):729–40. doi:10.1016/j.chom.2014.05.009

25. Edgar RJ, Van Hensbergen VP, Ruda A, Turner AG, Deng P, Le Breton Y, et al. Discovery of glycerol phosphate modification on streptococcal rhamnose

polysaccharides. Nat Chem Biol. 2019 May;15(5):463–71. doi:10.1038/s41589-019-0251-4

26. Van Der Beek SL, Zorzoli A, Çanak E, Chapman RN, Lucas K, Meyer BH, et al. Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification. Molecular Microbiology. 2019 Apr;111(4):951–64. doi:10.1111/mmi.14197

27. Van Der Beek SL, Le Breton Y, Ferenbach AT, Chapman RN, Van Aalten DMF, Navratilova I, et al. GACA is essential for Group A *Streptococcus* and defines a new class of monomeric d TDP -4-dehydrorhamnose reductases (RMLD). Molecular Microbiology. 2015 Dec;98(5):946–62. doi:10.1111/mmi.13169

28. Rush JS, Parajuli P, Ruda A, Li J, Pohane AA, Zamakhaeva S, et al. PplD is a de-N-acetylase of the cell wall linkage unit of streptococcal rhamnopolysaccharides. Nat Commun. 2022 Feb 1;13(1):590. doi:10.1038/s41467-022-28257-0

29. Henningham A, Davies MR, Uchiyama S, Van Sorge NM, Lund S, Chen KT, et al. Virulence Role of the GlcNAc Side Chain of the Lancefield Cell Wall Carbohydrate Antigen in Non-M1-Serotype Group A *Streptococcus*. Biswas I, editor. mBio. 2018 Mar 7;9(1):e02294-17. doi:10.1128/mBio.02294-17

30. Kabanova A, Margarit I, Berti F, Romano MR, Grandi G, Bensi G, et al. Evaluation of a Group A *Streptococcus* synthetic oligosaccharide as vaccine candidate. Vaccine. 2010 Dec;29(1):104–14. doi:10.1016/j.vaccine.2010.09.018

31. Caliot É, Dramsi S, Chapot-Chartier MP, Courtin P, Kulakauskas S, Péchoux C, et al. Role of the Group B Antigen of *Streptococcus agalactiae*: A Peptidoglycan-Anchored Polysaccharide Involved in Cell Wall Biogenesis. Wessels

MR, editor. PLoS Pathog. 2012 Jun 14;8(6):e1002756. doi:10.1371/journal.ppat.1002756

32. Wang Z, Enotarpi J, Buffi G, Pezzicoli A, Gstöttner CJ, Nicolardi S, et al. Chemical Synthesis and Immunological Evaluation of Fragments of the Multiantennary Group-Specific Polysaccharide of Group B *Streptococcus*. JACS Au. 2022 Jul 25;2(7):1724–35. doi:10.1021/jacsau.2c00302

33. Zorzoli A, Meyer BH, Adair E, Torgov VI, Veselovsky VV, Danilov LL, et al. Group A, B, C, and G *Streptococcus* Lancefield antigen biosynthesis is initiated by a conserved α-d-GlcNAc-β-1,4-l-rhamnosyltransferase. Journal of Biological Chemistry. 2019 Oct;294(42):15237–56. doi:10.1074/jbc.RA119.009894

34. Beaussart A, Péchoux C, Trieu-Cuot P, Hols P, Mistou MY, Dufrêne YF. Molecular mapping of the cell wall polysaccharides of the human pathogen *Streptococcus agalactiae*. Nanoscale. 2014 Oct 31;6(24):14820–7. doi:10.1039/C4NR05280C

35. Michon F, Brisson JR, Dell A, Kasper DL, Jennings HJ. Multiantennary group-specific polysaccharide of group B *streptococcus*. Biochemistry. 1988 Jul 12;27(14):5341–51. doi:10.1021/bi00414a059

36. Sutcliffe IC, Black GW, Harrington DJ. Bioinformatic insights into the biosynthesis of the Group B carbohydrate in *Streptococcus agalactiae*. Microbiology. 2008 May 1;154(5):1354–63. doi:10.1099/mic.0.2007/014522-0

37. Deng L, Kasper DL, Krick TP, Wessels MR. Characterization of the Linkage between the Type III Capsular Polysaccharide and the Bacterial Cell Wall of Group B *Streptococcus*. Journal of Biological Chemistry. 2000 Mar;275(11):7497–504. doi:10.1074/jbc.275.11.7497

38. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. Bipartite *rgp* Locus Diversity in *Streptococcus thermophilus* Corresponds to Backbone and Side Chain Differences of Its Rhamnose-Containing Cell Wall Polysaccharide. Appl Environ Microbiol. 2022 Dec 13;88(23):e0150422. doi:10.1128/aem.01504-22 PubMed PMID: 36350137; PubMed Central PMCID: PMC9746298.

39. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Molecular Microbiology. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494

40. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. Cell Wall Glycans Mediate Recognition of the Dairy Bacterium *Streptococcus thermophilus* by Bacteriophages. Dudley EG, editor. Appl Environ Microbiol. 2018 Dec;84(23):e01847-18. doi:10.1128/AEM.01847-18

41. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

42. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. Dairy lactococcal and streptococcal phage–host interactions: an industrial perspective in an evolving phage landscape. FEMS Microbiology Reviews. 2020 Nov 24;44(6):909–32. doi:10.1093/femsre/fuaa048

43. Hols P, Hancy F, Fontaine L, Grossiord B, Prozzi D, Leblondbourget N, et al. New insights in the molecular biology and physiology of revealed by

comparative genomics. FEMS Microbiology Reviews. 2005 Aug;29(3):435–63. doi:10.1016/j.femsre.2005.04.008

44. Kampff Z, Van Sinderen D, Mahony J. Cell wall polysaccharides of streptococci: A genetic and structural perspective. Biotechnology Advances. 2023 Dec;69:108279. doi:10.1016/j.biotechadv.2023.108279

45. Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, Van Sinderen D, Mahony J. The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Elkins CA, editor. Appl Environ Microbiol. 2025 Sep 12;e01238-25. doi:10.1128/aem.01238-25

46. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in *Streptococcus thermophilus* bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

47. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Appl Environ Microbiol. 2022 Jan 11;88(1):e0172321. doi:10.1128/AEM.01723-21 PubMed PMID: 34669424; PubMed Central PMCID: PMC8752142.

48. Broendum SS, Williams DE, Hayes BK, Kraus F, Fodor J, Clifton BE, et al. High avidity drives the interaction between the streptococcal C1 phage endolysin, PlyC, with the cell surface carbohydrates of Group A *Streptococcus*. Molecular Microbiology. 2021 Aug;116(2):397–415. doi:10.1111/mmi.14719

49. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research. 2019 Nov 7;gkz1035. doi:10.1093/nar/gkz1035

50. Schmidt TSB, Fullam A, Ferretti P, Orakov A, Maistrenko OM, Ruscheweyh HJ, et al. SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Research. 2024 Jan 5;52(D1):D777–83. doi:10.1093/nar/gkad943

51. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021 Jan;39(1):105–14. doi:10.1038/s41587-020-0603-3

52. Carlino N, Blanco-Míguez A, Punčochář M, Mengoni C, Pinto F, Tatti A, et al. Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. Cell. 2024 Oct;187(20):5775-5795.e15. doi:10.1016/j.cell.2024.07.039

53. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Borgwardt K, editor. Bioinformatics. 2022 Nov 30;38(23):5315–6. doi:10.1093/bioinformatics/btac672

54. Jauneikaite E, Tocheva AS, Jefferies JMC, Gladstone RA, Faust SN, Christodoulides M, et al. Current methods for capsular typing of *Streptococcus pneumoniae*. Journal of Microbiological Methods. 2015 Jun;113:41–9. doi:10.1016/j.mimet.2015.03.006

55. Vollmer W, Massidda O, Tomasz A. The Cell Wall of *Streptococcus pneumoniae*. Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Braunstein M, Rood JI, editors. Microbiol Spectr. 2019 May 31;7(3):7.3.19. doi:10.1128/microbiolspec.GPP3-0018-2018

56. Eletu SD, Sheppard CL, Rose S, Smith K, Andrews N, Lim WS, et al. Re-validation and update of an extended-specificity multiplex assay for detection of *Streptococcus pneumoniae* capsular serotype/serogroup-specific antigen and cell-wall polysaccharide in urine specimens. Access Microbiology. 2020 Mar 1;2(3). doi:10.1099/acmi.0.000094

57. Theodorou I, Courtin P, Palussière S, Kulakauskas S, Bidnenko E, Péchoux C, et al. A dual-chain assembly pathway generates the high structural diversity of cell-wall polysaccharides in *Lactococcus lactis*. Journal of Biological Chemistry. 2019 Nov;294(46):17612–25. doi:10.1074/jbc.RA119.009957

58. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Robinson P, editor. Bioinformatics. 2021 Aug 25;37(16):2473–5. doi:10.1093/bioinformatics/btab007

59. Lavelle K, McDonnell B, Fitzgerald G, van Sinderen D, Mahony J. Bacteriophage-host interactions in *Streptococcus thermophilus* and their impact on co-evolutionary processes. FEMS Microbiology Reviews. 2023 Jul 5;47(4):fuad032. doi:10.1093/femsre/fuad032

60. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, et al. The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. Mol Microbiol. 2020 Oct;114(4):582–96. doi:10.1111/mmi.14561 PubMed PMID: 32515029.

61. Soto-Serrano A, Sadovskaya I, Vinogradov E, Li W, Yu J, White K, et al. Expanding the Lactococcal Cell Wall Polysaccharide Paradigm: Novel Structures and Metabolic Pathways in the Emerging Dairy Species *Pseudolactococcus*

laudensis and *Pseudolactococcus raffinolactis*. MicrobiologyOpen. 2025 Dec;14(6):e70133. doi:10.1002/mbo3.70133

62. Shibata Y, Ozaki K, Seki M, Kawato T, Tanaka H, Nakano Y, et al. Analysis of Loci Required for Determination of Serotype Antigenicity in *Streptococcus mutans* and Its Clinical Utilization. J Clin Microbiol. 2003 Sep;41(9):4107–12. doi:10.1128/JCM.41.9.4107-4112.2003

63. De Jong A, Pietersma H, Cordes M, Kuipers OP, Kok J. PePPER: a webserver for prediction of prokaryote promoter elements and regulons. BMC Genomics. 2012 Dec;13(1):299. doi:10.1186/1471-2164-13-299

64. Pan Y, An H, Fu T, Zhao S, Zhang C, Xiao G, et al. Characterization of *Streptococcus pluranimalium* from a cattle with mastitis by whole genome sequencing and functional validation. BMC Microbiol. 2018 Dec;18(1):182. doi:10.1186/s12866-018-1327-0

65. Kovacs CJ, Faustoferri RC, Quivey RG. RgpF Is Required for Maintenance of Stress Tolerance and Virulence in *Streptococcus mutans*. O'Toole G, editor. J Bacteriol. 2017 Dec 15;199(24). doi:10.1128/JB.00497-17

66. Pearson C, Tindall S, Potts JR, Thomas GH, Van Der Woude MW. Diverse functions for acyltransferase-3 proteins in the modification of bacterial cell surfaces: This article is part of the Bacterial Cell Envelopes collection. Microbiology. 2022 Mar 1;168(3). doi:10.1099/mic.0.001146

67. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023 Aug;20(8):1203–12. doi:10.1038/s41592-023-01940-w

68. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. GigaScience. 2019 Oct 1;8(10):giz119. doi:10.1093/gigascience/giz119

69. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015 Jan;32(1):268–74. doi:10.1093/molbev/msu300

70. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Schwartz R, editor. Bioinformatics. 2019 Feb 1;35(3):526–8. doi:10.1093/bioinformatics/bty633

71. Bouras G, Grigson SR, Papudeshi B, Mallawaarachchi V, Roach MJ. Dnaapler: A tool to reorient circular microbialgenomes. JOSS. 2024 Jan 11;9(93):5968. doi:10.21105/joss.05968

72. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microbial Genomics. 2021 Nov 30;7(11). doi:10.1099/mgen.0.000685

73. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9. doi:10.1093/bioinformatics/btu153

74. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3. doi:10.1093/bioinformatics/btv421

75. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: A Free, Open-Source System for Microarray Data Management and Analysis. BioTechniques. 2003 Feb;34(2):374–8. doi:10.2144/03342mt01

76. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology. 2018 Jul;430(15):2237–43. doi:10.1016/j.jmb.2017.12.007

77. Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks [Internet]. 2022 [cited 2026 Mar 16]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.04.08.487609 doi:10.1101/2022.04.08.487609

78. Friess L, Van Sinderen D, Lee C. A CRISPRi Gene Regulation System for Bifidobacteria. Microbial Biotechnology. 2025 Nov;18(11):e70260. doi:10.1111/1751-7915.70260

79. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012 Feb 15;28(4):464–9. doi:10.1093/bioinformatics/btr703

Chapter IV

Functional evaluation of rhamnose-glucose polysaccharide biosynthesis and its impact on phage receptor requirements

Contributions: Irina Sadovskaya and Evgeny Vinogradov: Biochemical analysis and NMR spectroscopy.

Abstract

Bacteriophage infection of *Streptococcus thermophilus* remains a major challenge in industrial dairy fermentations. The narrow host range typically exhibited by individual *S. thermophilus*-infecting phages is, at least in part, attributed to structural variation in cell surface saccharidic receptors, including rhamnose-glucose polysaccharides (RGP). In the current study, structural requirements underpinning RGP-mediated phage binding were investigated employing *S. thermophilus* strain UCCSt50, which is susceptible to phages belonging to the *Brussowvirus*, P738 and *Piorkowskivirus* genera. Comparative analysis of bacteriophage-insensitive mutants (BIMs) demonstrates that mutations within the *rgp* locus impact RGP biosynthesis to varying extents, resulting in altered phage susceptibility profiles. The RGP structures of two BIMs were elucidated, providing further insight into the minimal structural requirements for the initial phage-host interaction involving strain UCCSt50. Structural analysis demonstrated that complete loss of the RGP side chain abolished susceptibility to phages SW13 and P738, while variants with partial truncation (a single sugar residue) of the side chain retained susceptibility to SW13 infection, yet were insensitive to P738. These findings indicate that P738 exhibits a substantially greater dependence on an intact side chain architecture, while SW13 can tolerate considerable variation in side chain composition and length. Additionally, expression of a heterologous priming glycosyltransferase in a side chain-deficient derivative of *S. thermophilus* UCCSt50 restored susceptibility to both P738 and SW13, suggesting a likely unlocking of the biosynthesis of the RGP side chain structure. Collectively, these findings demonstrate that subtle variation in RGP side chain structures differently impacts the host adsorption capability of *S. thermophilus* phages.

Introduction

Bacteriophage infection of *Streptococcus thermophilus* remains one of the principal causes of acidification delays and product inconsistencies associated with thermophilic dairy fermentations (1–3). Despite the extensive diversity of phages infecting *S. thermophilus*, individual phages typically exhibit a narrow host range, often infecting only a single strain or a limited subset of strains within the species (2,4,5). This restricted host range is generally attributed to the diversity of cell surface-associated saccharidic structures that serve as receptors for many phages of this bacterial species (6–10). Cell wall polysaccharide structures present on the cell surface of *S. thermophilus* include rhamnose-glucose polysaccharides (RGPs) and exopolysaccharides (EPSs). Members of the *Piorkowskivirus* genus are described to recognise EPS-associated host receptors (6–8), whereas members of the *Brussowvirus* and P738 genera recognise the RGP as the primary receptor (6,9,10).

It is proposed that the highly specific nature of these interactions corresponds to the variability in the monosaccharide composition and architecture of these cell surface polysaccharides. For example, a recent study investigating phage-insensitive derivatives of *S. thermophilus* UCCSt50 suggested that P738 phages require the complete tetrasaccharide side chain structure, or a specific component of this structure, for adsorption (10). In contrast, *Brussowvirus* phage SW13 was shown to infect strains possessing a seemingly shortened side chain structure (tri- or di-saccharide). While these observations established that members of distinct phage groups exhibit different specificity or dependency on specific RGP structural features, the broader contribution of side chain modifications to phage infection profiles remains unresolved. Therefore, this study aims to further investigate the

functional role of RGP side chain biosynthesis in mediating susceptibility to members of genetically distinct dairy streptococcal phage genera.

The *rgp* locus, which is universally present all *S. thermophilus* chromosomes examined to date, encodes the biosynthetic machinery for RGP production. This biosynthetic cluster is typically 20–30 kb in length and displays a modular arrangement with two functional modules regions. The 3' region of the loci encodes proteins associated with rhamnan backbone synthesis including the well characterised rhamnosyltransferases (RgpA, RgpB, RgpF) and an ABC-type polysaccharide transporter system (RgpC and RgpD), while the 5' region encodes the enzymatic machinery to synthesize the decorative side chain attached to the backbone structure (10,11). The notable strain level genetic diversity in each module has enabled the identification of seven major *rgp* genotypes (*rgp*1–7) that embody different combinations of the three backbone genotypes (Bt1–Bt3) and five variable side chain genotypes (Vt1–Vt5) (11). The *S. thermophilus* RGP structure similarly consists of a rhamnan backbone together with a more structurally diverse side chain. Previous structural analyses have demonstrated substantial diversity in side chain composition and organisation between strains, being correspondingly reflected by variation in glycosyltransferase-encoding gene content within the variable region (11). As the side chain structure is believed to be exposed at the bacterial cell surface, it has been proposed to represent the primary interaction site for phage adsorption and receptor recognition (12). Therefore, mutations affecting side chain biosynthesis may alter phage susceptibility by modifying or eliminating specific saccharidic components or affecting "docking" points required for successful infection.

S. thermophilus strain UCCSt50 represents an unusual dairy streptococcal strain as it is susceptible to infection by three phages in our collection belonging to three distinct genera, namely *Brussowvirus*, P738 and *Piorkowskivirus*. In this study, we exploited the multi-genus phage susceptibility profile of *S. thermophilus* UCCSt50 to investigate how structural variation in the RGP influences phage recognition and infectivity by distinct phages possessing different receptor-binding specificities. Using a combination of genetically distinct bacteriophage-insensitive mutants, structural analysis and heterologous glycosyltransferase expression, the current study aims to further investigate the structural determinants underpinning phage specificity in *S. thermophilus*.

Results

P738-resistant mutants of *S. thermophilus* UCCSt50 exhibit distinct infection profiles

S. thermophilus UCCSt50 was previously demonstrated to be infected by members of the *Brussowvirus* and P738 phage genera (10), while we recently established that it is also infected by phage SW16, which is a member of the *Piorkowskivirus* genus (2). Previous studies have elucidated that *Brussowvirus* SW13 and the namesake of its genus, P738, recognise an RGP receptor, whereas members of the *Piorkowskivirus* genus are reported to recognise an EPS receptor (6–10).

In the present study, we exploited a collection of spontaneous phage-insensitive derivatives or BIMs (bacteriophage-insensitive mutants) of *S. thermophilus* UCCSt50 that had previously been generated following challenge with either phage P738 (10) or SW13 (9). These BIMs were assessed for the efficiency of plaquing of three phages, i.e. P738, SW13 and SW16, relative to the wild type strain. Among the tested BIMs (named here as B1, F1B, F1A, F2E and F22D) three different infection sensitivity profile types were observed: (a) those that were completely resistant to SW13 and P738 and with reduced sensitivity to SW16; (b) those that were resistant to P738 and displayed reduced sensitivity to SW13 and SW16 and; (c) those that were completely resistant to P738 and SW13, yet exhibiting increased sensitivity to SW16. Therefore, a representative of each infection sensitivity phenotype was selected for further analysis to further our understanding of the role of the RGP in the recognition and binding processes of P738 and SW13, i.e. mutants F1A, F2E and F22D (Table 1). To establish if genes within the *rgp* locus or other

regions underpinned the observed phage-resistance phenotypes for mutants F1A, F2E and F22D, their complete genome sequences were elucidated.

Table 1. Efficiency of plaquing (EOP) of P738, SW13 and SW16 on *S. thermophilus* UCCSt50 and its' P738-resistant derivatives.

	Efficiency of plaquing of phage:		
Strain/derivative name	**P738**	**SW13**	**SW16**
UCCSt50	1	1	1
UCCSt50-F1A	$\leq 5.5 \times 10^{-6}$	$\leq 8.2 \times 10^{-6}$	0.4 ± 0.31
UCCSt50-F2E	$\leq 5.5 \times 10^{-6}$	0.45 ± 0.23	0.45 ± 0.07
UCCSt50-F22D	$\leq 5.5 \times 10^{-6}$	$\leq 8.2 \times 10^{-6}$	4.26 ± 0.24
UCCSt50-B1	$\leq 5.5 \times 10^{-6}$	$\leq 8.2 \times 10^{-6}$	4.34 ± 0.17
UCCSt50-F1B	$\leq 5.5 \times 10^{-6}$	0.63 ± 0.12	0.49 ± 0.09

Phage-resistance phenotype is elicited through distinct mutations in the *rgp* locus

The complete genome sequences of mutant strains F1A, F2E and F22D were elucidated in this study and compared to the genome obtained from their WT derivative strain UCCSt50. Subsequently, the CRISPR spacer arrays were identified using CRISPRCasFinder (13) and the spacer sequences were extracted and compared to those of the parent strain. The three mutants were observed to possess largely identical CRISPR spacer loci relative to the parent strain and with no spacer acquisition observed, and they were therefore deemed to be non-CRISPR-mediated BIMs (Supplementary Table S1). Of note, BIM F2E possessed a shortened CRISPR 3 array containing 17 spacers compared to the 20 spacers present in the parent strain and the other BIMs (Supplementary Table S1). Comparison of the arrays revealed that spacers 17–19 of the parental strain were absent in F2E, while the terminal spacer of the parental array was retained. As these deleted spacers were located at the distal end of the array and did not display sequence similarity (Blastn analysis) to phages P738, SW13 or SW16, the observed contraction is unlikely to have contributed to the phage-resistance phenotype of F2E.

Subsequent SNP analysis revealed that mutant F1A possesses five SNPs (single nucleotide polymorphism of which one is located in a gene within the *rgp* locus (Table 2). This latter SNP in F1A represents a nonsense mutation resulting in the incorporation of a stop codon in *orf06995*$_{UCCSt50}$, which is predicted to encode the WpsJ side chain transferase based on its predicted membrane topology and using the annotations of the established model strain *Lactococcus cremoris* NZ9000 (14). Mutant F2E was identified to possess three SNPs with one located in a gene in the

rgp locus (leading to an amino acid substitution in *orf*07005$_{UCCSt50}$), which is predicted to encode a glycosyltransferase (Fig. 1). Mutant F22D was shown to carry six SNPs (Table 2); however, no SNPs were identified within the *rgp* cluster that could account for the observed phage-resistance phenotype. Through nucleotide alignment of both the *rgp* and *eps* loci, a single nucleotide (guanine) deletion was identified within the priming glycosyltransferase-encoding gene (*orf07015*$_{UCCSt50}$) located in the *rgp* cluster (Fig. 1), resulting in a frameshift and a truncated protein product.

Table 2. Summary of mutations and deletions identified in the genomes of *S. thermophilus* UCCSt50 BIMs.

BIM	SNP position	Base Modification	Amino Acid Changes	UCCSt50 locus tag	Predicted function
F1A	374,768	c → t	Pro^{251} → Ser^{251}	02085	ACP S-mannosyltransferase
	590,420	c → g	Ser^{81} → Arg^{81}	03195	Ferrous iron transport protein
	652,902	g → a	Gly^{56} → Glu^{56}	03520	HPr (Ser) kinase/phosphatase
	775,422	c → g	His^{58} → Gln^{58}	04145	Bifunctional DnaQ family exonuclease
	1,338,607	c → t	Gln^{238} → stop	**06995**	**DUF2142 – glycosyltransferase**
F2E	1,342,676	c → g	His^{209} → Gln^{209}	**07005**	**Glycosyltransferase**
	1,403,163	a → c	Asp^{468} → Ala^{468}	07320	APC family permease
	1,703,041	g → c	Gln^{494} → His^{494}	08880	DNA-directed RNA polymerase subunit
F22D	559,000	g → a	Ala^{141} – Thr^{141}	03015	trhA PAQR family membrane homeostasis protein
	581,786	t → c	Phe^{324} – Leu^{324}	03145	Aminoacyltransferase
	830,546	g → t	Ala^{161} – Ser^{161}	04430	transcription repressor NadR
	846,116	a → t	Tyr^{317} – Phe^{317}	04505	lepA translation elongation factor 4
	925,880	g → a	Gly^{112} – Arg^{112}	04920	Hypothetical protein
	1,757,441	t → a	Ile^{121} – Ile^{121}	09150	50S ribosomal protein L6
	1,343,878	g → -	Frameshift	**07015**	**Priming glycosyltransferase**

To establish whether the identified mutations within genes located in the *rgp* cluster are responsible for the observed phage resistance phenotypes, the native genes from the parent strain, UCCSt50, were cloned into the high copy expression vector pNZ44 (15) and introduced into BIMs F1A and F2E. It has previously been demonstrated that introduction of the native $orf07020_{UCCSt50}$ gene resulted in the restoration of P738 phage sensitivity in BIM F1B (EOP = 0.35 ± 0.33), confirming the role of $orf07020_{UCCSt50}$ in the P738 infection process (10). Similarly, complementation of BIMs F2E and F1A with the corresponding wild-type $orf07005_{UCCSt50}$ and $orf06995_{UCCSt50}$ genes, respectively, restored the phage susceptibility phenotype (Table 3). In the present study, the transformation efficiency of BIM F22D was below detectable limits and impeded our genetic complementation efforts; however, given the proven phenotype restoration in BIM B1, which has a mutation in the same priming glycosyltransferase-encoding gene, the role and importance of $orf07015_{UCCSt50}$ in RGP synthesis and phage interactions is inferred (9).

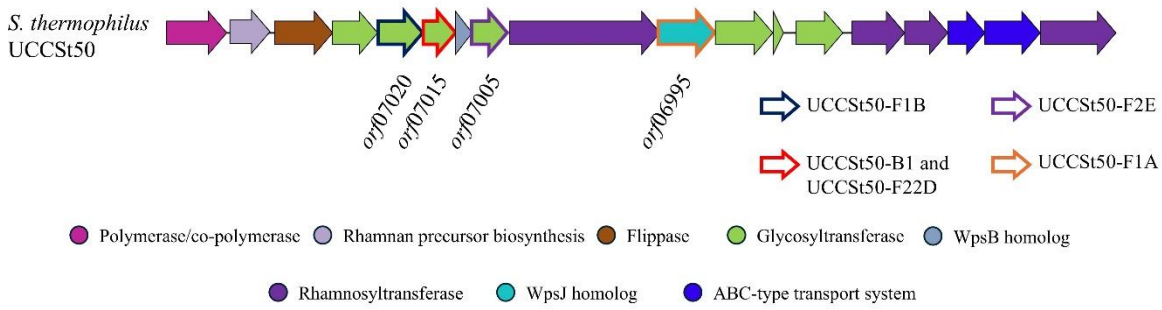


Fig. 1. Schematic representation of the *rgp* locus of *S. thermophilus* UCCSt50 highlighting SNPs identified in BIMs from this study and previous studies (9,10). Genes where mutations were identified are colour-coded according to the BIM in which they were identified and indicated in the key.

Table 3. Relative efficiencies of plaquing (EOP) values of a panel of bacteriophages (P738, SW13 and SW16) on UCCSt50, derived BIMs and complemented strains.

Strain/derivative name	Efficiency of plaquing of phage: P738	SW13	SW16	Source of strain
St50	1	1	1	(9)
St50::pNZ44	1.06 ± 0.11	1.1 ± 0.13	1.1 ± 0.11	(10)
UCCSt50-B1	≤ 5.5×10^{-6}	≤ 8.2×10^{-6}	4.34 ± 0.17	(9)
B1::pNZ44	≤ 5.5×10^{-6}	≤ 8.2×10^{-6}	1.33 ± 0.03	(9)
B1::pNZ44-07015	0.79 ±0.05	0.86 ± 0.06	9.16 ± 0.09	(9)
UCCSt50-F1B	≤ 5.5×10^{-6}	0.63 ± 0.12	0.49 ± 0.09	(10)
F1B::pNZ44	≤ 5.5×10^{-6}	0.54 ± 0.15	0.34 ± 0.14	(10)
F1B::pNZ44-07020	0.48 ± 0.15	0.83 ± 0.12	1.18 ± 0.02	(10)
UCCSt50-F1A	≤ 5.5×10^{-6}	≤ 8.2×10^{-6}	0.4 ± 0.31	(10)
F1A::pNZ44-06995	0.62 ± 0.15	0.85 ± 0.14	5.29 ± 0.04	This study
UCCSt50-F2E	≤ 5.5×10^{-6}	0.45 ± 0.23	0.45 ± 0.07	(10)
F2E::pNZ44-07005	0.78 ± 0.15	0.74 ± 0.1	1.19 ± 0.11	This study
UCCSt50-F22D	≤ 5.5×10^{-6}	≤ 8.2×10^{-6}	4.26 ± 0.24	(10)

RGP-associated mutations impact side chain biosynthesis & transfer to the rhamnan backbone

The distinct sensitivity profiles of each of the three phages on the generated BIMs likely reflects the diverse interactions and specific receptor requirements of each phage on the same host strain. To determine the functional and structural impact of the identified mutations, the chemical composition and structure of two of the mutants (F1A and F2E) were determined (Fig. 2; Supplementary Fig. S1 & S2). For BIM F1A, which harbours a mutation in *orf6995*$_{UCCSt50}$, the rhamnan backbone structure (–2-α-Rha-2-α-Rha-6-α-Glc-) was observed to be identical to that of the parent strain. In contrast, the side chain structure was not detected in the major

fraction of the extracted RGP material from BIM F1A indicating that very low amounts if any RGP side chain was produced (<10 % branching). Transmembrane domain predictions of the product of *orf6995*$_{UCCSt50}$ identified 11 transmembrane domains. Protein domain searches using Pfam and HHPred identified the gene product as containing a DUF2142 domain of a GT-C fold-containing glycosyltransferase, respectively. The predicted membrane topology and domain profile of this protein are consistent with those of the lactococcal WpsJ, which is demonstrated to transfer the lactococcal cell wall polysaccharide side chain on to the rhamnan backbone structure (14). Based on the protein characteristics and the absence of the RGP side chain structure in BIM F1A, we propose that the product of *orf6995*$_{UCCSt50}$ is equivalent to the WpsJ RGP side chain transferase of *Lactococcus* (14).

BIM F2E harbours a mutation in *orf07005*$_{UCCSt50}$, which is a gene located within the side chain biosynthetic region of the *rgp* locus (Table 2 & Fig. 1). Compositional and methylation analysis of RGP extracts of F2E identified the expected rhamnan backbone structure (–2-α-Rha-2-α-Rha-6-α-Glc-) with a single N-acetylglucosamine (GlcNAc) residue attached in ~60 % of the repeating unit structures (Fig. 2). The remainder of the tetrasaccharide side chain structure (present in WT UCCSt50) was not observed in the major fraction of the RGP extract of this strain. As the GlcNAc branch was determined by methylation analysis, the branch point and glycosidic linkage can only be predicted to be identical to that of the parent strain (Fig. 2). The product of this gene is predicted to be a glycosyltransferase belonging to the GT2 family. This is consistent with the characteristics of several of the lactococcal side chain subunit glycosyltransferases (e.g. WpsC and WpsD) (14). Our previous work has identified *orf07015*$_{UCCSt50}$ as

the priming glycosyltransferase responsible for the transfer of the GlcNAc moiety on to the rhamnan backbone (9). Therefore, based on the predicted protein characteristics, the presence of a monosaccharidic side chain structure attached to the rhamnan backbone, we propose that the product of *orf07005*$_{UCCSt50}$ functions as a glycosyltransferase that is required for the attachment of the second monosaccharide (i.e. -α-Rha-) to the GlcNAc moiety of the RGP side chain structure.

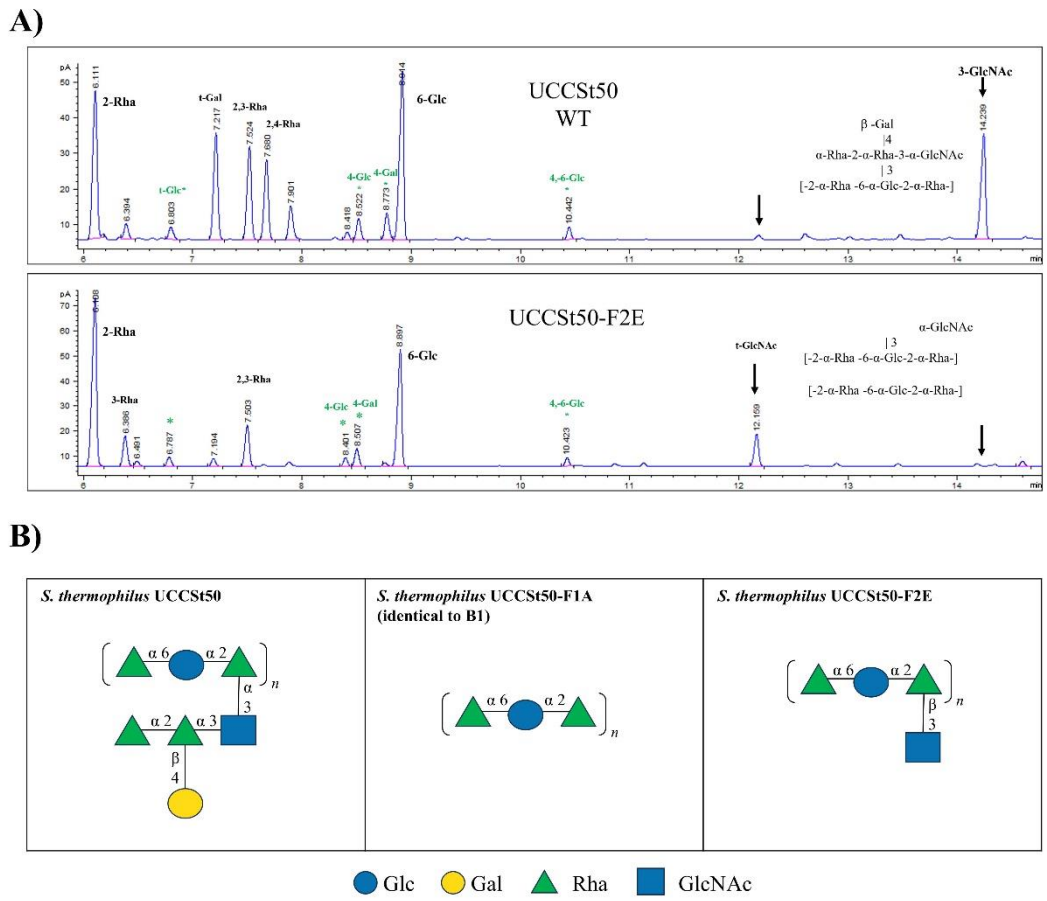


Fig. 2. A) Methylation analysis profiles of the RGP preparations of the wild-type UCCSt50 and BIM F2E. Peaks corresponding to GlcNAc are indicated with arrows. The GlcNAc branch was determined by methylation analysis therefore the branch point and glycosidic linkage can only be predicted. The 2,4-linked Rha and t-Gal are absent in F2E. The peaks labelled in green and an asterisk represent sugars most probably corresponding to the exopolysaccharide (EPS) structure. **B)** Structures of the RGP from *S. thermophilus* strain UCCSt50 (9) and its phage-resistant derivatives F1A and F2E. Monosaccharide symbols are based on those indicated by the Standard Nomenclature for Glycans (SNFG).

Side chain synthesis can be restored through expression of a heterologous priming glycosyltransferase

Previous studies have demonstrated that the priming glycosyltransferase encoded within the *rgp* locus initiates biosynthesis of the side chain component of the RGP through transfer of the first monosaccharide residue to the undecaprenyl-phosphate lipid carrier (9). In *S. thermophilus* UCCSt50, the first residue incorporated into the side chain structure is GlcNAc (9,10). In contrast, structural analysis of the RGP produced by *S. thermophilus* Brie28 (16) revealed that side chain biosynthesis is initiated with a N-acetylgalactosamine (GalNAc) residue (reported in Chapter III). Therefore, we sought to determine if expression of a heterologous priming glycosyltransferase from strain Brie28 can restore side chain biosynthesis in the side chain deficient derivative B1 and, if so, what impact this would have on the susceptibility profile of the resulting strain to phages recognising the UCCSt50 and Brie28 cell surface receptors.

To investigate this, the priming glycosyltransferase-encoding gene from *S. thermophilus* Brie28 was cloned and expressed in the side chain deficient derivative B1 (9). The priming glycosyltransferase of Brie28 shares ~ 45 % amino acid sequence identity with the corresponding glycosyltransferase of UCCSt50. The resulting derivative was subsequently challenged with phages that infect either UCCSt50 (i.e. phages P738, SW13 and SW16) or Brie28 (i.e. phage SW11).

Expression of the Brie28-derived priming glycosyltransferase (B1::Brie28GT) restored susceptibility of the UCCSt50 BIM B1 to both P738 and SW13, while no alteration in the infectivity profile of SW16 was observed (Table 4). Given the previously established requirement of P738 for the complete UCCSt50 side chain

structure, restoration of phage susceptibility likely reflects recovery of a compatible side chain architecture, although this requires further investigation. These findings suggest that introduction of the heterologous priming glycosyltransferase permits restoration of side chain biosynthesis in the B1 background, potentially through an "unlocking" mechanism whereby the downstream glycosyltransferases and associated biosynthetic machinery can extend the initiated structure irrespective of the precise identity of the priming residue (GlcNAc or GalNAc). Alternatively, it indicates that phage P738 can bind to the RGP side chain structure that incorporates either a GalNAc or GlcNAc residue in the first position in the side chain subunit. Restoration of phage susceptibility was specific to the UCCSt50-associated phages P738 and SW13, as the engineered derivative did not become susceptible to phage SW11, which infects the *S. thermophilus* strain Brie28 (Table 4).

Table 4. Infection profiles of bacteriophages P738, SW13, SW16 and SW11 against *S. thermophilus* strains UCCSt50, B1, B1::Brie28GT and Brie28. Infection phenotypes were determined by plaque assay and are represented as presence (green) or absence (white) of infection/plaque formation.

Strain	P738	SW13	SW16	SW11
UCCSt50				
B1				
B1::Brie28GT				
Brie28				

Discussion

Our findings expand current understanding of rhamnose-glucose polysaccharide (RGP) biosynthesis in *S. thermophilus* by combining comparative phage sensitivity profiling, (comparative) genome analysis and structural characterisation of bacteriophage-insensitive mutants (BIMs). While previous studies established the importance of RGP as a receptor for members of the *Brussowvirus* and P738 genera (9,10) and have also defined the diversity of *rgp* loci and their corresponding structures (11), the precise functions of many genes involved in side chain biosynthesis have remained unresolved. Through the analysis of BIMs possessing distinct phage infection phenotypes, the present study enabled functional assignment of several genes within the *rgp* locus, providing further insights into the structural requirements underpinning infection by RGP-recognising phages P738 and SW13.

The modular organisation of the *S. thermophilus rgp* locus has previously been described and resembles the bipartite architecture of lactococcal *cwps* loci, which encode a rhamnan backbone together with a structurally variable surface-exposed side chain structure (PSP) (17). In *L. lactis*, extensive genetic and biochemical analyses have enabled the assignment of specific functions to many of the proteins involved in side chain biosynthesis and assembly (14). In contrast, despite an increasing number of structurally resolved RGPs and *rgp* loci available for *S. thermophilus*, direct experimental evidence supporting the roles of individual genes within the cluster has remained limited. The BIMs characterised in this study present a series of defined genetic mutations that facilitate the functional dissection of the UCCSt50 *rgp* locus while also creating a parallel with the established lactococcal model of cell wall polysaccharide (CWPS) biosynthesis.

The side chain deficient derivatives B1 and F22D support the assignment of *orf07015*$_{UCCSt50}$ as the priming glycosyltransferase responsible for initiating side chain biosynthesis. Previous structural analysis of BIM B1 revealed the complete absence of the side chain structure in this derivative (9), while the present study identified an independent mutation affecting the same gene in BIM F22D. The recurrence of mutations within this gene of the locus among independently isolated BIMs strongly supports its central role in side chain initiation (and side chain function as a phage receptor) and is consistent with the proposed function of WpsA homologues in lactococcal CWPS biosynthesis. Loss of this activity would be expected to prevent assembly of the side chain repeating unit and consequently abolish production of the surface-exposed component (side chain) of the RGP.

Structural analysis of BIM F2E provided additional insight into the sequential assembly of the side chain structure. Mutation of *orf07005*$_{UCCSt50}$ resulted in an RGP in which approximately 60% of repeating units carried a single GlcNAc residue attached to the rhamnan backbone, whereas the remaining units lacked a side chain entirely. Given that *orf07015*$_{UCCSt50}$ has previously been assigned as the priming glycosyltransferase responsible for transfer of the initial GlcNAc residue to the lipid carrier undecaprenyl-phosphate (Und-P), producing Und-P-GlcNAc, which serves as the foundation on which the side chain structure is assembled (9,10), these findings suggest that *orf07005*$_{UCCSt50}$ functions after this step and is required for the addition of the second monosaccharide of the side chain structure. Based on the predicted GT2 family glycosyltransferase domain and the observed structure resolved in this study, the product of *orf07005*$_{UCCSt50}$ is proposed to catalyse incorporation of the first rhamnose residue following side chain initiation in the *S. thermophilus* UCCSt50 model strain. It has previously been proposed that

orf07020$_{UCCSt50}$ is a glycosyltransferase-encoding gene responsible for the addition of the terminal rhamnose residue to the side chain structure (10). These assignments represent the first direct functional links between a specific *S. thermophilus* glycosyltransferase and a defined structural feature of the RGP.

Similarly, structural analysis of BIM F1A enabled the functional assignment of *orf06995*$_{UCCSt50}$. Although the rhamnan backbone structure remained intact, the side chain structure was largely absent from the extracted polysaccharide (the major structure representing ~ 90 % of units). The predicted membrane topology and domain architecture of the encoded protein closely resemble those of the lactococcal WpsJ side chain transferase, which catalyses transfer of the mature side chain structure onto the rhamnan backbone (14). The absence of detectable side chain structures in F1A therefore supports the hypothesis that *orf06995*$_{UCCSt50}$ performs an equivalent function in *S. thermophilus*. Collectively, the mutations characterised in this and previous studies allow the construction of a preliminary model of side chain biosynthesis in UCCSt50, whereby *orf07015*$_{UCCSt50}$ initiates side chain assembly, *orf07005*$_{UCCSt50}$ contributes to elongation of the side chain structure, *orf07020*$_{UCCSt50}$ participates in subsequent side chain additions, and *orf06995*$_{UCCSt50}$ mediates transfer of the side chain component onto the rhamnan backbone.

Beyond their contribution to understanding RGP biosynthesis, the BIMs included in this study, provided a unique opportunity to investigate the structural requirements of phages recognising the same polysaccharide receptor. Previous work demonstrated that P738 requires an intact RGP side chain structure (or a specific component of this structure) for infection, in comparison SW13 remained capable of infecting the partially truncated derivative F1B (10). The inclusion of

additional BIMs in the present study extends these observations and further refines the minimal structural requirements for infection by each phage. Strains lacking the side chain structure entirely were resistant to both P738 and SW13, indicating that the side chain component contains the essential receptor determinants recognised by both phages. However, SW13 retained the ability to infect strains possessing substantially reduced side chain structures, including derivatives containing only a partial side chain architecture with a single monosaccharidic branch. In these assays, the P738 remained unable to infect any BIM and could only infect the wild-type strain, UCCSt50. These findings demonstrate that the two phages exhibit substantially different dependencies on specific components within the RGP side chain despite recognising the same general receptor structure. Despite the apparent tolerance of SW13 towards variation in side chain composition, this phage nevertheless retains a relatively narrow host range among *S. thermophilus* strains (2). This observation suggests that additional structural determinants within the RGP architecture, such as specific glycosidic linkages or phage “docking” points, may underpin phage recognition and adsorption. Interestingly, several RGP-associated mutations also altered susceptibility to the EPS-binding phage SW16. While the mechanistic basis of this observation remains unclear, it may reflect broader changes in cell surface architecture resulting from disruption of RGP biosynthesis. Although further work is required to determine whether direct regulatory or structural interplay exist between RGP and EPS biosynthesis and expression in *S. thermophilus*, these findings suggest that alterations in one surface polysaccharide may have consequences extending beyond the immediate receptor itself.

The heterologous expression of the Brie28 priming glycosyltransferase provided an additional perspective on the flexibility of RGP biosynthesis. Restoration of susceptibility to both P738 and SW13 in the side chain deficient BIM B1 suggests that the introduced enzyme was capable of re-establishing a functional side chain biosynthetic pathway despite sharing only moderate sequence identity with the native UCCSt50 enzyme. While the precise structure produced by the engineered derivative remains unresolved, restoration of P738 sensitivity strongly suggests the generation of a side chain architecture that is compatible with this particular phage adsorption. These findings raise the possibility that the biosynthetic pathway encoded by the *rgp* locus might tolerate variation in the nature of the initiating monosaccharide, thereby allowing reconstruction of a phage-compatible receptor structure. This also opens the possibility of engineering alternative side chain structures to further investigate their role in the phage-host interactome. However, it remains unclear whether receptor restoration reflects flexibility within the RGP biosynthetic pathway, phage receptor recognition of specific sugar residues, or a combination of both.

Collectively, the findings presented in this study establish functional roles for several previously uncharacterised genes within the *S. thermophilus rgp* locus while simultaneously refining our understanding of the structural requirements underpinning phage adsorption. By linking specific genetic mutations to defined alterations in RGP structure and phage susceptibility, this work provides a framework for the continued functional annotation of *rgp* loci and further demonstrates how even subtle variation in cell surface polysaccharide architecture can influence host range and receptor specificity.

Materials and Methods

Bacteria, growth conditions & media

S. thermophilus strains were routinely grown from single colonies or bacterial stocks maintained at -70 °C (25 % [wt/vol] glycerol) in M17 broth (Sigma, USA) supplemented with 0.5 % lactose (Sigma, USA) at 42 °C. For *S. thermophilus* strains containing plasmids (pNZ44 or pNZ44-acrIIA6), chloramphenicol (Sigma, USA) was added to LM17 at a final concentration of 3 µg/mL in agar and 5 µg/mL in broth.

Bacteriophage-insensitive mutant isolation and characterisation

The bacteriophage-insensitive mutants (BIMs) used in this study (F1A, F2E and F22D) were generated as previously described (10) through exposure of *S. thermophilus* UCCSt50 to phage P738, followed by isolation of surviving phage-resistant derivatives.

Genomic DNA of the three BIMs was extracted using a modified protocol of the Nucleobond AXG 100 with Buffer set III (Macherey-Nagel, Germany) as described previously (10). Bacterial genome sequencing was performed by Plasmidsaurus (USA) using Oxford Nanopore Technology (UK) long-read sequencing. Long-read sequencing was conducted on a PromethION P24 platform with R10.4.1 flow cells, and libraries were prepared using the Rapid Barcoding Kit 96 V14. Basecalling was performed using ‘dorato super-accurate’ basecalling with default Q10 quality filtering. Sequences were subsequently annotated using Bakta with default parameters (18).

To verify that no spacer acquisitions had been made to the BIMs CRISPR (1, 2, and 3) loci spacer content or number, all sequenced genomes were uploaded to CRISPRCasFinder (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index) (13), and CRISPR spacer arrays were analysed. A single nucleotide polymorphism (SNP) analysis of F1A, F2E and F22D genome sequences against the genome of the parent strain UCCSt50 was performed using the Bowtie2 alignment tool (19) and SAMtools (20). Further manual annotations and functional predictions of ORFs associated with the *rgp* cluster were performed using a combination of TMHMM (21), Pfam (22), and HHpred (23) analyses.

BIM complementation

SNPs were identified as described above and summarised in Table 2. The native genes from the parent strain UCCSt50 were amplified with Phusion Hot Start II High-Fidelity PCR Master Mix (Thermo Fisher Scientific, USA) using primer combinations listed in Supplementary Table S2. The resulting amplicons were purified using the GenElute PCR Clean-Up Kit (Sigma, USA). Enzymatic digestion and ligation of amplicons and the pNZ44 cloning vector were performed as described previously (10). Briefly, 10 µL of the ligation mixture was added to 100 µL of competent cells of either BIM F1A or F2E, which were prepared as per previous methods (10). The mixture was held on ice for 30 minutes before a heat shock was applied at 42 °C for 45 seconds followed by holding on ice for 2 minutes. The mixture was subsequently transferred to a pre-chilled electroporation cuvette before applying a single 2.0 kV pulse followed by the addition of 900 µL pre-warmed HJL recovery media (3 % tryptone, 1 % yeast extract, 0.5 % KH_2PO_4, 0.2 % beef extract, 0.5 % lactose, 20 mM $CaCl_2$, and 200 mM $MgCl_2$) and incubating

(in the electroporation cuvette) at 42 °C for a minimum of 4 hours. Cells were plated on LM17 agar supplemented with 3 μg/mL chloramphenicol and incubated at 42 °C for 48–72 hours. Positive transformants were verified to contain the recombinant plasmid(s) by colony-based PCR using pNZ44_F and pNZ44_R primers (Supplementary Table S2).

Heterologous glycosyltransferase expression

The priming glycosyltransferase from the *S. thermophilus* strain Brie28 (16) (GenBank accession number JBTWWX000000000) was amplified with Phusion Hot Start II High-Fidelity PCR Master Mix (Thermo Fisher Scientific, USA) using Brie28GT specific primer combinations listed in Supplementary Table S2 and Brie28 genomic DNA (extracted as described above). The resulting amplicons were purified using the GenElute PCR Clean-Up Kit (Sigma, USA). Enzymatic digest, ligation and transformation procedures were performed as described above. The integrity of the construct was validated by amplifying Sanger sequencing (Genewiz, Germany) using primers pNZ44_F and pNZ44_R (Supplementary Table S2).

Efficiency of Plaquing

The efficiency of plaquing (EOP) of phages P738 (Accession number: MK911750) (24), SW13 (Accession number: MH892362) (2) and SW16 (Accession number: MH892350) (2) on the host strain UCCSt50, BIMs, complemented strains and control strains was determined using the standard overlay method as described in detail previously (10,25). The relative EOP of phages (Table 3) was determined by dividing the observed titre of the phage on a given BIM/complement/control host by that on the parent strain, UCCSt50. All assays were performed in at least

duplicate. For quantitative analysis, averages and standard deviations were calculated in MS Excel using a Student's *t*-test.

Preparation and structural analysis of RGP of *S. thermophilus* F1A and F2E

Approximately 8 L cultures of *S. thermophilus* strains F1A and F2E were grown overnight in M17 (Sigma, USA) supplemented with 0.5% lactose (Sigma, USA) (1% total lactose concentration) at 42°C. Cells were harvested by centrifugation at 4,400 × *g* at 4°C for 30 min. The resulting cell pellets were washed twice with ice-cold sterile distilled water, first with 300 mL per culture, then pooled and washed again in 40 mL before storing the cell pellets at −20°C.

Cell wall polysaccharides (RGPs) were extracted from cell pellets following treatments with cold 5% trichloroacetic acid (TCA) and hot diluted HCl and purified on a Sephadex G-50 column, as described previously (8, 10). It was further purified by reverse-phase HPLC using an Agilent Zorbax C18 column (250 × 9 mm) in solvent A (0.1% TFA, vol/vol) and B (80% MeCN, vol/vol) at a flow rate of 3 mL/minute, starting from 3% B for 10 minutes then gradient to 100% B over 30 minutes. The RGP purified by HPLC was used for NMR analysis as described previously (10) using a Bruker AVANCE III 600 MHz (1H) spectrometer with a 5 mm Z-gradient probe with acetone internal reference (2.225 ppm for 1H and 31.45 ppm for 13C) using standard pulse sequences cosygpprqf (gCOSY), mlevphpr (TOCSY, mixing time 120 ms), roesyphpr (ROESY, mixing time 500 ms), hsqcedetgp (HSQC), hsqcetgpml (HSQC-TOCSY, 100 ms TOCSY delay), and hmbcgplpndqf (HMBC, 70 ms long range transfer delay).

Monosaccharide composition and methylation analyses were performed as described previously (17).

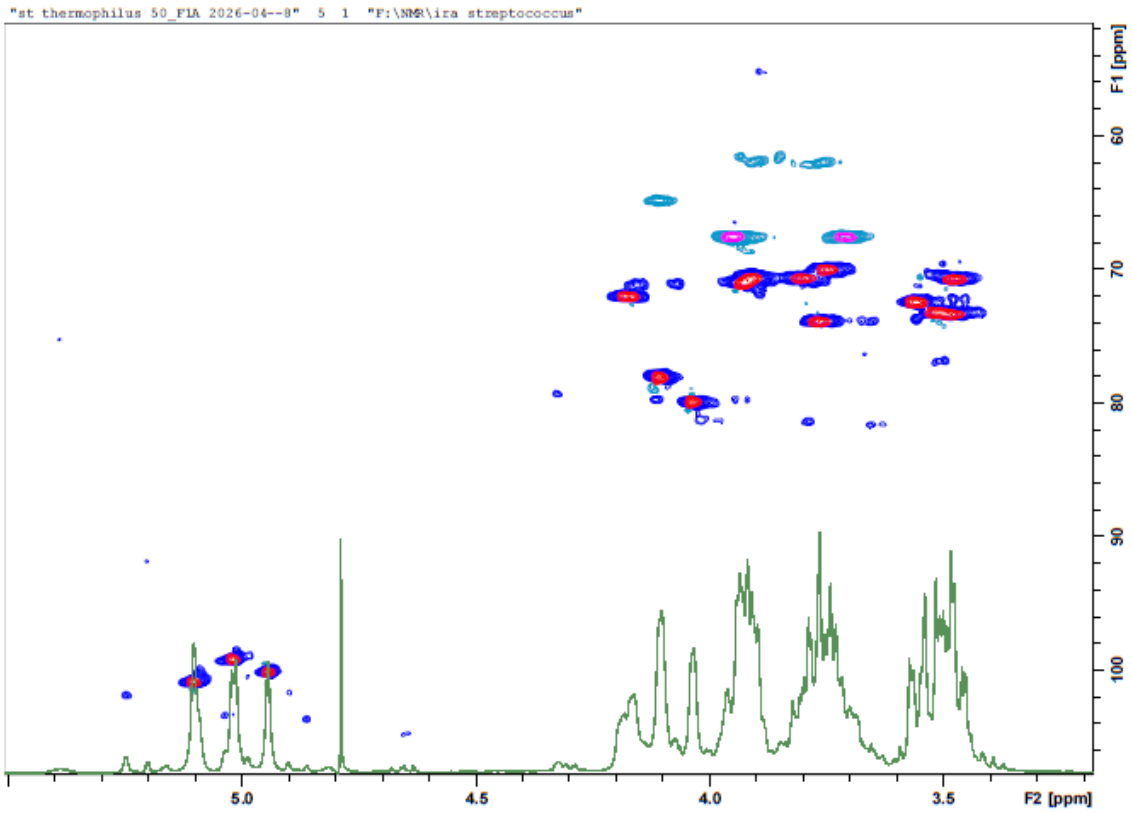
Supplementary Material

Supplementary Table S1. CRISPR loci and spacer array profiles of *S. thermophilus* UCCSt50 and BIMs F1A, F2E and F22D

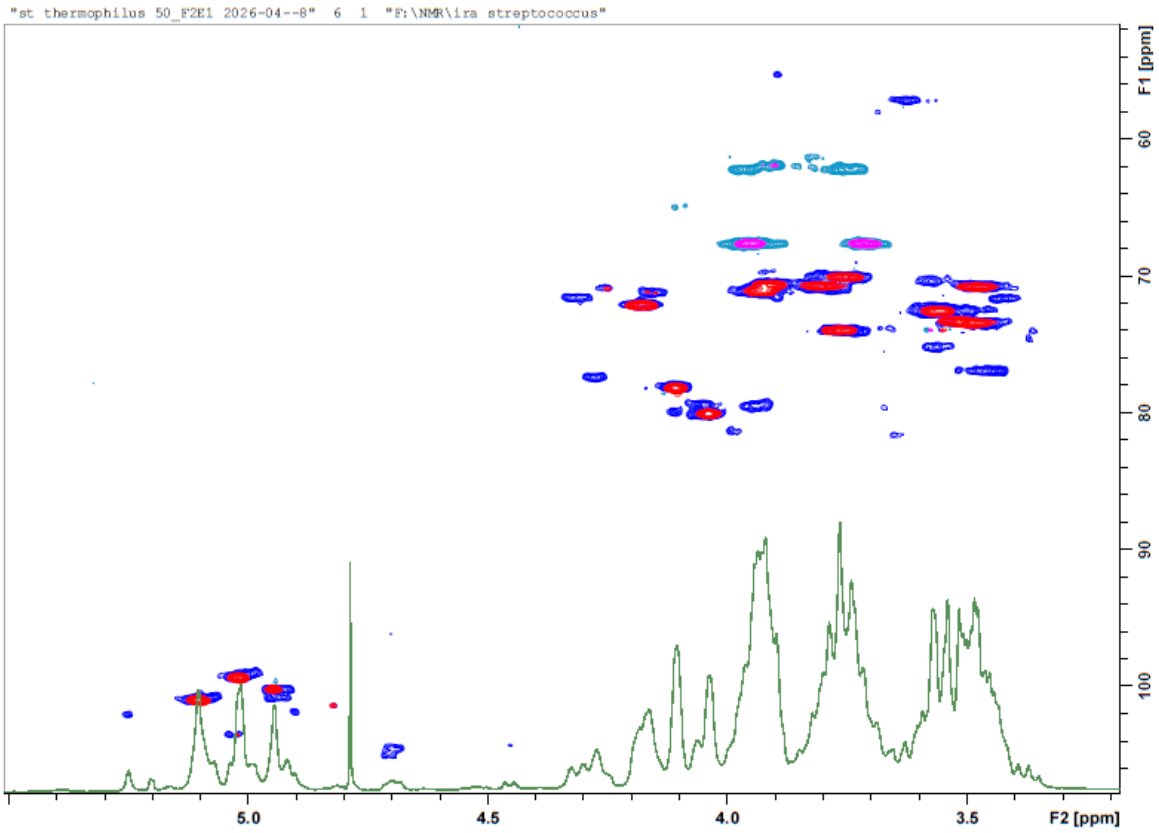
Strain	C1 (bp)	C1 (spacers)	C2 (bp)	C2 (spacers)	C3 (bp)	C3 (spacers)
UCCSt50	2,409	36	101	1	1,358	20
F1A	2,409	36	101	1	1,358	20
F2E	2,409	36	101	1	1,160	17
F22D	2,409	36	101	1	1,358	20

Supplementary Table S2. Primers used in this study.

Primer name	Sequence (5' – 3')	Amplicon size (bp)	Source
F1A_06995_F	aaaaaaccatggaaggaggcactcacttgaacagaattaaacaaataaaagttg	1308	This
F1A_06995_R	aaaaaactgcagttatgcaaatacccccccagtaat		study
F2E_07005_F	aaaaaaccatggaaggaggcactcacatgaaagataatcatacttgggtg	777	This
F2E_07005_R	aaaaaactgcagctatccattcgtcttctgactt		study
pNZ44_F	ctaatgtcactaacctgccccg	Template	(8)
pNZ44_R	gctttatcaactgctgct	dependent	
Brie28GT_F	aaaaaaccatggaaggaggcactcacatgtcagatttactcattattattcc	1027	This
Brie28GT_R	aaaaaactgcagttatttatctttccctctcaaggc		study



Supplementary Fig. S1. ^{1}H-^{13}C HSQC and ^{1}H NMR spectra of the RGP from the BIM F1A.



Supplementary Fig. S2. ^{1}H-^{13}C HSQC and ^{1}H NMR spectra of the RGP from the BIM F2E.

References

1. Ortiz Charneco G, De Waal PP, Van Rijswijck IMH, Van Peij NNME, Van Sinderen D, Mahony J. Bacteriophages in the Dairy Industry: A Problem Solved? Annu Rev Food Sci Technol. 2023 Mar 27;14(1):367–85. doi:10.1146/annurev-food-060721-015928

2. Lavelle K, Martinez I, Neve H, Lugli GA, Franz CMAP, Ventura M, et al. Biodiversity of *Streptococcus thermophilus* Phages in Global Dairy Fermentations. Viruses. 2018 Oct 22;10(10):577. doi:10.3390/v10100577

3. Hanemaaijer L, Kelleher P, Neve H, Franz C, De Waal P, Van Peij N, et al. Biodiversity of Phages Infecting the Dairy Bacterium *Streptococcus thermophilus*. Microorganisms. 2021 Aug 27;9(9):1822. doi:10.3390/microorganisms9091822

4. Szymczak P, Janzen T, Neves AR, Kot W, Hansen LH, Lametsch R, et al. Novel Variants of *Streptococcus thermophilus* Bacteriophages Are Indicative of Genetic Recombination among Phages from Different Bacterial Species. Dudley EG, editor. Appl Environ Microbiol. 2017 Mar;83(5):e02748-16. doi:10.1128/AEM.02748-16

5. Duplessis M, Lévesque CM, Moineau S. Characterization of *Streptococcus thermophilus* Host Range Phage Mutants. Appl Environ Microbiol. 2006 Apr;72(4):3036–41. doi:10.1128/AEM.72.4.3036-3041.2006

6. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. Cell Wall Glycans Mediate Recognition of the Dairy Bacterium *Streptococcus thermophilus* by Bacteriophages. Appl Environ Microbiol. 2018 Dec 1;84(23):e01847-18. doi:10.1128/AEM.01847-18 PubMed PMID: 30242010; PubMed Central PMCID: PMC6238053.

7. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in *Streptococcus thermophilus* bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

8. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Mol Microbiol. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494 PubMed PMID: 32073719.

9. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Appl Environ Microbiol. 2022 Jan 11;88(1):e0172321. doi:10.1128/AEM.01723-21 PubMed PMID: 34669424; PubMed Central PMCID: PMC8752142.

10. Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, Van Sinderen D, Mahony J. The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Elkins CA, editor. Appl Environ Microbiol. 2025 Sep 12;e01238-25. doi:10.1128/aem.01238-25

11. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. Bipartite *rgp* Locus Diversity in *Streptococcus thermophilus* Corresponds to Backbone and Side Chain Differences of Its Rhamnose-Containing Cell Wall Polysaccharide. Ercolini D, editor. Appl Environ Microbiol. 2022 Dec 13;88(23):e01504-22. doi:10.1128/aem.01504-22

12. Lavelle K, McDonnell B, Fitzgerald G, van Sinderen D, Mahony J. Bacteriophage-host interactions in *Streptococcus thermophilus* and their impact on co-evolutionary processes. FEMS Microbiology Reviews. 2023 Jul 5;47(4):fuad032. doi:10.1093/femsre/fuad032

13. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research. 2018 Jul 2;46(W1):W246–51. doi:10.1093/nar/gky425

14. Theodorou I, Courtin P, Palussière S, Kulakauskas S, Bidnenko E, Péchoux C, et al. A dual-chain assembly pathway generates the high structural diversity of cell-wall polysaccharides in *Lactococcus lactis*. Journal of Biological Chemistry. 2019 Nov;294(46):17612–25. doi:10.1074/jbc.RA119.009957

15. McGrath S, Fitzgerald GF, Van Sinderen D. Improvement and Optimization of Two Engineered Phage Resistance Mechanisms in *Lactococcus lactis*. Appl Environ Microbiol. 2001 Feb;67(2):608–16. doi:10.1128/AEM.67.2.608-616.2001

16. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

17. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, et al. The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. Molecular Microbiology. 2020 Oct;114(4):582–96. doi:10.1111/mmi.14561

18. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-

free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microbial Genomics. 2021 Nov 30;7(11). doi:10.1099/mgen.0.000685

19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9. doi:10.1038/nmeth.1923

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9. doi:10.1093/bioinformatics/btp352

21. Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks [Internet]. 2022 [cited 2026 Mar 16]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.04.08.487609 doi:10.1101/2022.04.08.487609

22. Sonnhammer ELL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins. 1997 Jul;28(3):405–20. doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L

23. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology. 2018 Jul;430(15):2237–43. doi:10.1016/j.jmb.2017.12.007

24. Philippe C, Levesque S, Dion MB, Tremblay DM, Horvath P, Lüth N, et al. Novel Genus of Phages Infecting *Streptococcus thermophilus*: Genomic and Morphological Characterization. Ercolini D, editor. Appl Environ Microbiol. 2020 Jun 17;86(13):e00227-20. doi:10.1128/AEM.00227-20

25. Lillehaug D. An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. J Appl Microbiol. 1997 Jun;83(1):85–90. doi:10.1046/j.1365-2672.1997.00193.x

Chapter V

Genotype to biochemical structure correlations among exopolysaccharides and *eps* loci of *Streptococcus thermophilus*

Contributions: Irina Sadovskaya and Evgeny Vinogradov: Biochemical analysis and NMR spectroscopy.

Abstract:

Exopolysaccharides (EPSs) produced by *Streptococcus thermophilus* contribute significantly to the rheological, textural and sensory properties of fermented dairy products and are increasingly recognised for their potential health-associated benefits. Despite the extensive structural diversity reported among *S. thermophilus* EPSs, direct comparison between experimentally resolved EPS structures and the corresponding *eps* genotype classification framework has remained limited due to the lack of publicly available genome sequence data for many industrial starter strains. In the present study, we investigated the relationship between *eps* locus diversity and EPS structure in *S. thermophilus* through comparative genomics and structural analysis. Hierarchical clustering (HCL) analysis of 182 *eps* loci extracted from publicly available genomes and previously characterised strains revealed extensive genetic diversity across the species while also identifying several dominant genotypes. Structural elucidation of the EPSs produced by *S. thermophilus* strains UCCSt50, Nect13 and STMM11 demonstrated that each strain produced a distinct repeating-unit structure corresponding to *eps* genotypes H, D and C_2, respectively. Comparative analysis revealed that multiple independently characterised strains possessing highly similar *eps* loci also produced identical or highly similar EPS structures. Collectively, comparative analysis of currently available EPS structures identified 11 distinct EPS structures among *S. thermophilus* strains with an *eps* genotype assigned to five of these structures. These findings establish a comparative framework linking *eps* genotype to EPS structure and support the existence of predictive genotype-to-structure relationships among technologically important dairy streptococci. Furthermore, this work expands current understanding of EPS diversity within *S. thermophilus* and provides a

foundation for future studies investigating the biological and technological significance of these cell wall polysaccharides.

Introduction:

Streptococcus thermophilus is one of the most extensively employed commercial starter cultures, being widely used in the manufacture of yoghurts, various cheeses and other fermented dairy products. Its long-established technological value and industrial use stems not only from its rapid acidification of milk, but also from the ability of many *S. thermophilus* strains to produce exopolysaccharides (EPSs). EPSs of *S. thermophilus* are high molecular weight, surface-associated or secreted heteropolysaccharides that significantly contribute to texture, viscosity, flavour perception, and overall sensory quality of fermented dairy products (1,2). Beyond their technological roles, EPSs produced by *S. thermophilus* are gaining attention for their potential contributions to human health (3,4). Although the probiotic status of *S. thermophilus* is primarily based on its lactose-metabolising capabilities (1,3–5), numerous bioactivities have been attributed to its EPS, including inhibition of pathogen colonisation and biofilm formation, immune modulation, anti-inflammatory effects, and prebiotic functions (4). Importantly, many of these effects may persist even in the absence of viable cells, positioning EPS as promising "postbiotic" molecules (6). Taken together, these attributes not only underscore the industrial relevance of *S. thermophilus*, but also the growing recognition that detailed characterisation of its EPS and other saccharidic structures will be central to future applications in both food technology and health-associated contexts.

Industrial interest in EPS-producing starter culture-associated strains increased substantially following the recognition that certain yoghurt-associated cultures generate products with desirable texture and viscosity properties without the requirement for additional stabilisers or hydrocolloids (7–9). As consumer demand has shifted towards 'clean-label' fermented foods with fewer additives, EPS

produced by various *S. thermophilus* strains has become increasingly attractive as a naturally produced texturising agent capable of improving product quality while maintaining straightforward ingredient formulae (10,11). As a result, the genetic and structural basis of EPS biosynthesis in *S. thermophilus* and other LAB species has been the focus of considerable research efforts (1,12,13). Comparative genomic studies have revealed extensive diversity among *S. thermophilus eps* loci, with multiple distinct *eps* cluster types identified across industrial and dairy-associated strains (14–17). Similarly, structural analyses have demonstrated substantial variation in EPS repeating unit composition, linkage arrangement and branching architecture between strains (18–21). Although previous studies have suggested that the biochemical and structural diversity of *S. thermophilus* EPSs is associated with the observed genetic diversity within the corresponding *eps* loci, direct comparison between experimentally resolved EPS structures and *eps* genotype classification has not been undertaken. The above-mentioned studies have collectively established EPS as one of the principal contributors to the technological performance of industrial starter cultures and have reinforced the importance of structural elucidation studies for understanding the functional properties of dairy-associated polysaccharides.

To date, more than 30 EPS structures produced by *S. thermophilus* have been structurally resolved, revealing diversity in repeating unit composition, linkage arrangement and branching architecture (19,21–36). In the current study, we have collated and compared the EPS structural and genetic data that is available for 36 *S. thermophilus* strains. Our analysis indicates that many independently characterised strains produce identical or highly similar EPS structures and suggests the existence of dominant EPS structural groups among previously analysed

industrial dairy isolates. Previous studies have proposed that conservation or variation within the corresponding *eps* loci may account for these shared or distinct structural features among strains (23,31,35). Additionally, recent comparative genomic analysis of the *eps* loci of 23 *S. thermophilus* strains identified ten genotypes (*eps* A–J) (15). The majority of the *S. thermophilus* strains from which EPS structures have been resolved represent industrial starter cultures. Consequently, the complete genome sequences of these strains are rarely available in public databases, reflecting their proprietary and commercial value. This disconnect between available EPS structural data and corresponding genetic information has limited the direct comparison between EPS structure and *eps* genotype classification. Establishing predictive genotype-to-structure (or structure to genotype) relationships therefore represents an important step towards understanding the technological functionality and diversity of EPSs produced by *S. thermophilus*.

In the present study, the *eps* loci and EPS structures of *S. thermophilus* strains UCCSt50 (37,38), Nect13 (15) and STMM11 (39) were investigated to establish a relationship between *eps* genotype classification and experimentally resolved EPS structure. This enabled genotype assignment of structurally characterised strains and facilitated direct comparison between conserved *eps* loci and corresponding EPS repeating unit structures. Collectively, this study establishes a comparative framework linking *eps* locus diversity to EPS structure and advances the predictive interpretation of technologically important EPSs produced by *S. thermophilus* strains, significantly expanding the current understanding of streptococcal cell wall polysaccharides.

Results:

Hierarchical clustering reveals diversity among *S. thermophilus eps* loci and enables genotype assignment of structurally characterised strains

The *eps* locus, which is located on the *S. thermophilus* chromosome, encodes the EPS biosynthetic machinery, is approximately 15-30 kb in length. The genes that make up the *eps* locus encode proteins involved in regulation of EPS production, sugar transfer, EPS subunit polymerisation, and export of the saccharide subunits which are subsequently polymerised on the external surface of the cell (15,16,40). Despite gene synteny across strains of *S. thermophilus*, particularly in the *eps* locus termini, there is substantial variation in the central regions of the cluster, which typically represent glycosyltransferase-encoding genes (15,16,41). Previous studies investigating this genetic diversity within the *eps* loci identified 15 distinct *eps* cluster types described based on sequence similarity of *eps*-associated genes (14). More recently the genetic diversity observed across these loci has facilitated hierarchical clustering analyses of 24 industrial *S. thermophilus* strains that identified six major *eps* genotypes, termed *eps* A–F (40), while subsequent examination of a larger collection of dairy streptococcal isolates (28 *S. thermophilus* strains total) expanded this classification to include four additional genotypes, resulting in the currently recognised *eps* A–J classification framework employed in this study (15). To investigate the current genetic diversity of *eps* loci within *S. thermophilus*, hierarchical clustering (HCL) analysis was performed on 182 *eps* loci extracted from publicly available, complete *S. thermophilus* genome assemblies, *eps* cluster sequences of strains previously employed as reference strains in the *eps* genotyping analysis (15), and available partial *eps* cluster sequences from strains

for which EPS structures have also been resolved (Fig. 1) (Supplementary Table S1).

HCL analysis demonstrated that the reference strains representing the established *eps* A–J genotypes were distributed across the clustering analysis, highlighting the extensive genetic diversity previously reported among *S. thermophilus eps* loci in smaller datasets and which provided the basis of the A-J genotype-based classification system (14–16,40). The analysis also revealed additional undefined regions of diversity outside the established classification framework suggesting that the current classification system does not capture the full extent of diversity present in the *S. thermophilus eps* loci. However, further analysis is required to validate the extension of the *eps* genotyping system to include additional genotypes or sub genotypes. HCL analysis also revealed several localised genotype-associated cluster segments, where multiple strains grouped closely around specific reference genotypes, indicating conservation of gene content and overall *eps* locus organisation among these strains. Strains within these genotype-associated clusters were subsequently assigned to the corresponding *eps* genotype and colour coded accordingly (Fig. 1). Several genotypes were well represented within the analysis including *eps* H- (n=28), D- (n=22) and C_1-associated groups (n=25). However, there appears to be a degree of heterogeneity within the C_1 genotype, which may warrant an additional sub-genotype, although additional studies including detailed genetic and structural analyses would be required to validate this.

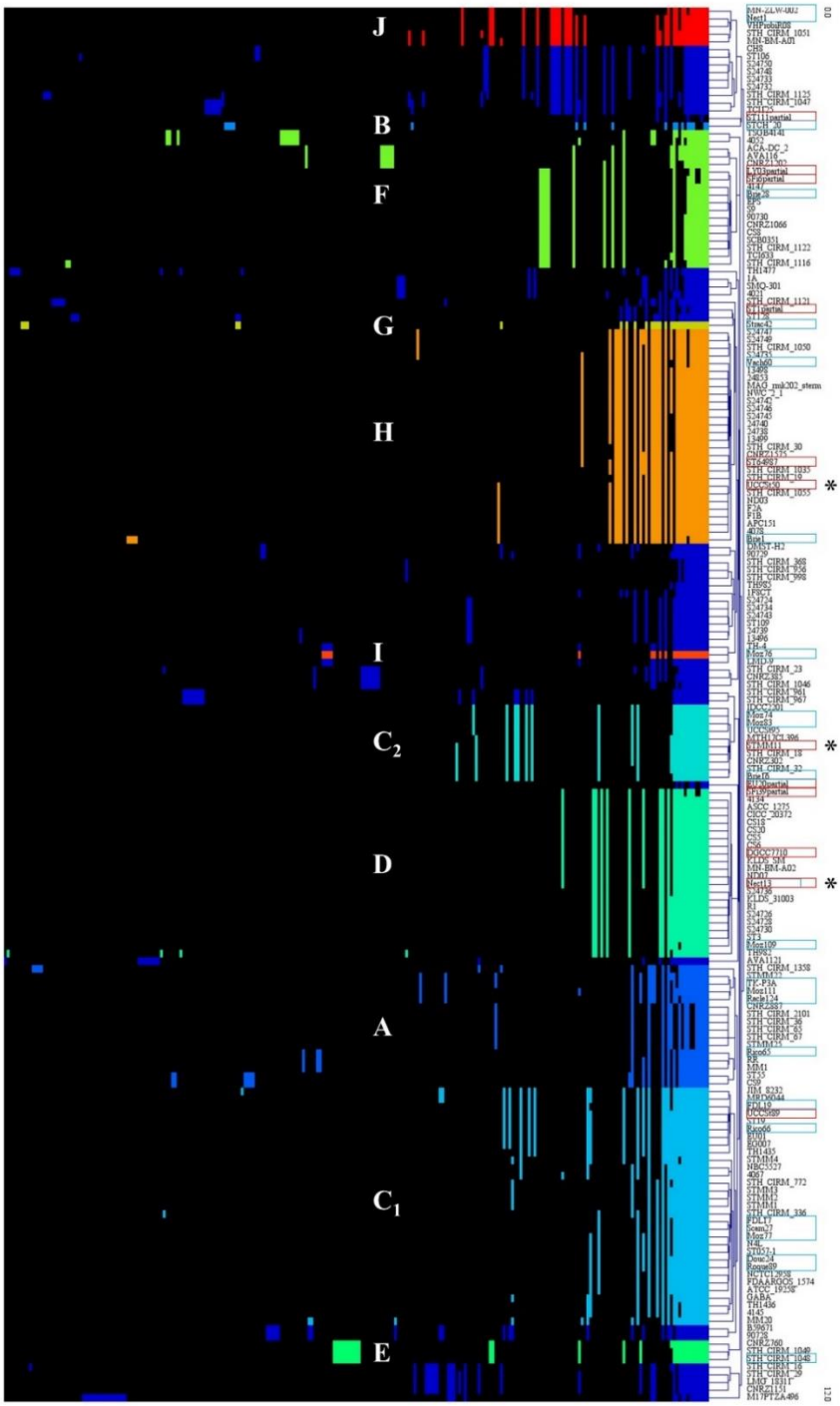


Fig. 1. HCL analysis of genes from *eps* loci of 182 *S. thermophilus* strains. The heatmap was generated based on protein family presence (colour) or absence (black). Reference strains representing the previously established *eps* A–J genotypes are indicated by blue boxes. Strains for which both *eps* locus sequences (complete or partial) and EPS structures are available are indicated by red boxes. Strains for which the EPS structure was elucidated in the present study are marked with an asterisk (*).

Our HCL analysis furthermore reveals that *eps* loci of dairy streptococci exhibit substantial diversity across the species and creates a foundation to investigate the relationship between *eps* genotype and the associated EPS structures. The analysis enabled genotype assignment of several publicly available *S. thermophilus* strains, including strains where EPS structural data is also available. Three structurally characterised strains (SFi39, ASCC 1275 and DGCC 7710) clustered within the *eps* D genotype, while ST64987 was assigned to the *eps* H genotype and LY03 and SFi6 grouped within the *eps* F-associated region (Fig. 1). The observed conservation of *eps* locus genetic content among strains within the *eps* D- and H-associated clusters suggests that strains sharing a common *eps* genotype produce a similar EPS structure. Therefore, *S. thermophilus* strains UCCSt50 (*eps* H) and Nect13 (*eps* D) were selected for structural analysis to determine whether their experimentally resolved EPS structures corresponded to those previously reported for other members of the same predicted genotypes. In contrast, STMM11 was selected as a representative of the comparatively underrepresented *eps* C_2 genotype, for which currently no experimentally resolved EPS structure is available, thereby facilitating structural comparison between this strain and other genotype associated structures or the assignment of a distinct structure to the C_2 genotype.

Representative *S. thermophilus eps* genotypes exhibit distinct EPS structures

The EPS structures of strains UCCSt50, Nect13 and STMM11, representing *eps* genotypes H, D and C_2, respectively, were elucidated in this study. Each strain was shown to produce a structurally distinct repeating unit differing in monosaccharide composition, linkage arrangement and branching architecture (Fig. 2). The major monosaccharide components of all three EPSs were determined to be galactose (Gal) and glucose (Glc). The EPS produced by *S. thermophilus* UCCSt50 was shown to be composed of branched pentasaccharide repeating units containing galactose (Gal) and glucose (Glc) residues (Fig. 3A). The precise linkage arrangement could not be fully determined, with two possible linkage variants existing; E-C, D-B (I) or E-B, D-C (II) linkages (Fig. 3A; Supplementary Table S2; Supplementary Fig. S1). The NMR chemical shifts closely corresponded to those previously reported for the EPS produced by *S. thermophilus* ST64987 (27) and *Lactobacillus delbrueckii* subsp. *Bulgaricus* 291 (42).

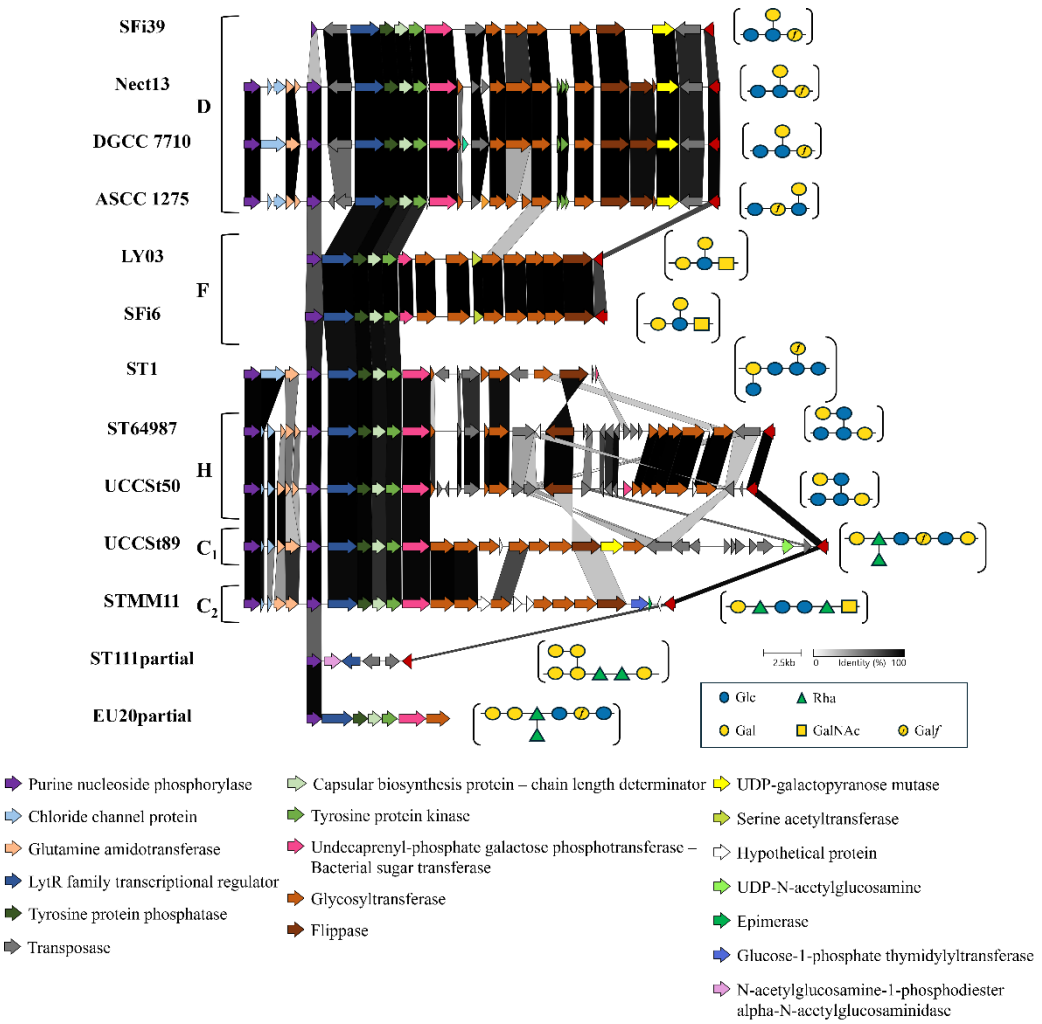


Fig. 2. Schematic overview and comparative analysis of the *eps* loci and associated EPS structures among structurally characterised *S. thermophilus* strains. Predicted gene functions are colour-coded and indicated below the figure. Conserved regions of sequence similarity between *eps* loci are connected by shaded regions generated using CAGECAT (43). Schematic representations of the associated EPS structures are shown on the right. Monosaccharide symbols are presented according to the Standard Nomenclature for Glycans (SNFG).

The EPS isolated from *S. thermophilus* STMM11 consists of a linear hexasaccharide repeating unit containing Gal, Glc, rhamnose (Rha) and *N*-acetylglucosamine (GlcNAc) residues (Fig. 3B; Supplementary Table S3; Supplementary Fig. S2). The branched tetrasaccharide repeating unit of the EPS isolated from *S. thermophilus* strain Nect13 contains galactofuranose (Gal*f*) in addition to Glc and Gal (Fig. 3C; Supplementary Table S4; Supplementary Fig. S3). Detailed 2D NMR analysis of the EPS for each strain is presented in the supplementary materials (Supplementary Tables S2 – S4; Supplementary Fig. S1 – S3).

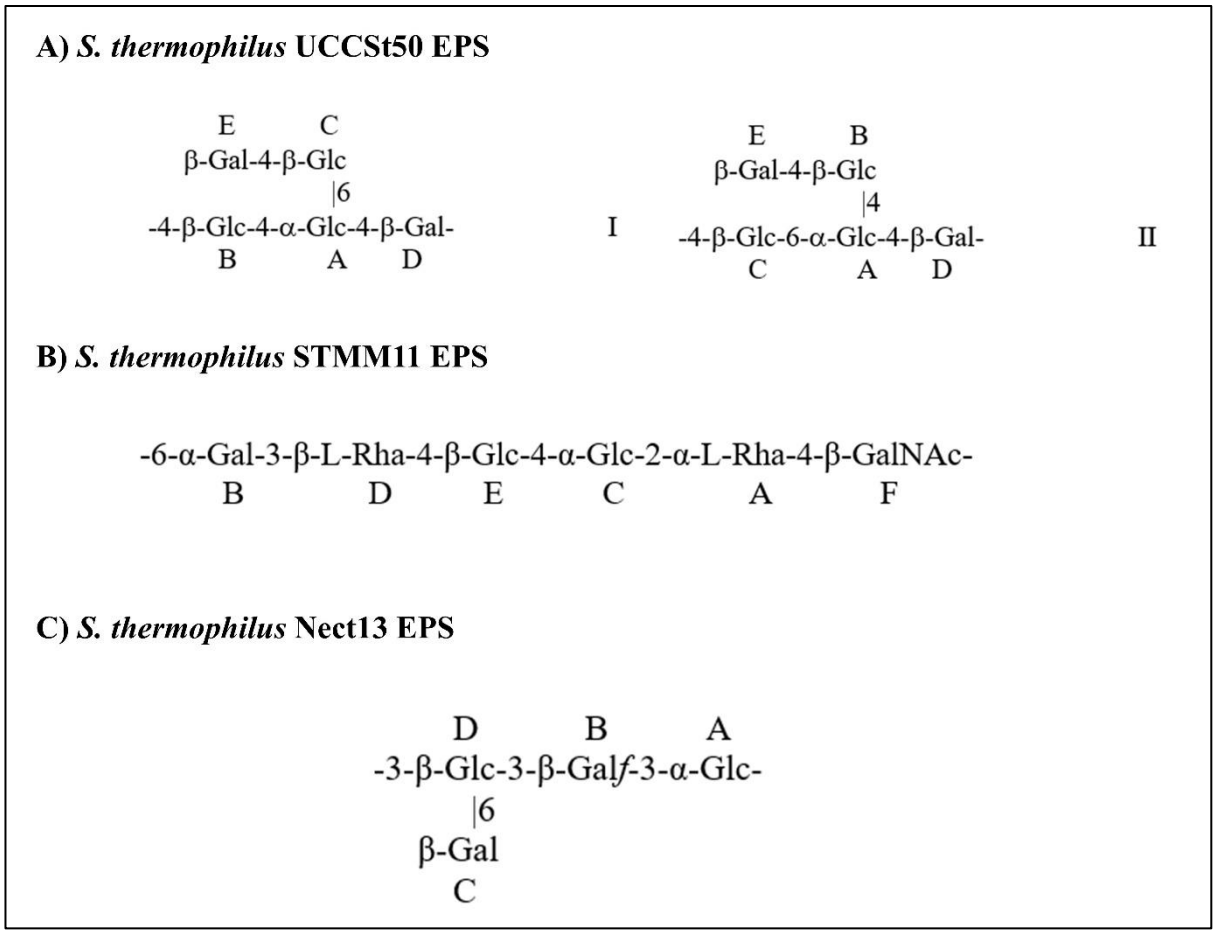


Fig. 3. Chemical structures of the EPS from *S. thermophilus* **A)** UCCSt50 **B)** STMM11 **C)** Nect13.

Linking *eps* genotype to structure in *S. thermophilus*

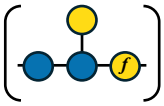
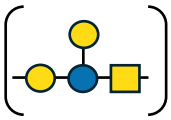
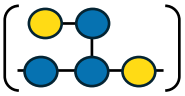
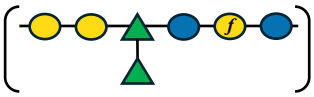
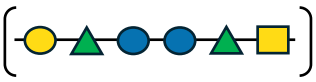
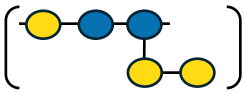
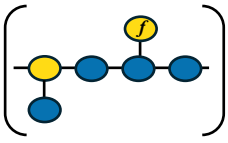
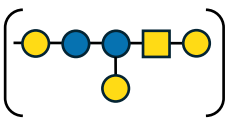
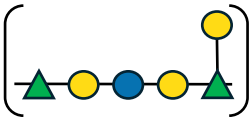
The genome sequences (complete or partial) of 13 of the 36 *S. thermophilus* strains for which EPS structures have been resolved are publicly available (Supplementary Table S5). As mentioned above, based on HCL and comparative analyses (Figs. 1 & 2) strains SFi39, Nect13, DGCC 7710 and ASCC 1275, have been assigned to the *eps* D genotype in this study. Each of these strains was also observed to produce an identical EPS structure, consistent with their shared genotype (Fig. 2 & Table 1). Furthermore, while genome sequence data for strains XJ53 (23), STD (25), CH101 (25), STCth-9204 (19), SY89 (22), and SY102 (22) are not available, their identical EPS structure suggests that they share highly similar if not identical *eps* loci, although this remains to be established.

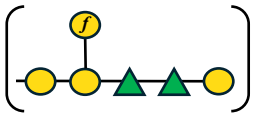
Similarly, LY03 (25) and SFi6 (29) have been classified as *eps* F-type strains in this study and share an identical EPS structure (Fig. 2 & Table 1). *S. thermophilus* strains ST64987 and UCCSt50, classified in this study as members of the *eps* H genotype, also possess highly similar *eps* loci (Fig. 2) and produce highly similar EPS structures. A structurally distinct Gal/Glc-containing EPS was reported for *S. thermophilus* THS (30); however, differences in both monosaccharide composition and glycosidic linkage organisation distinguished this polymer from the *eps* H-associated structural group. An additional distinct EPS structure was resolved for *S. thermophilus* OR 901 (31), ST111 (25), Rs (44) and Sts (44) consisting of Gal and Rha residues. Furthermore, an identical branched EPS repeating unit structure was identified for three independently characterised *S. thermophilus* strains, DCGG 7698 (19), EU20 (22) and UCCSt89 (33), and could be classified as C_1-type strains. A related but structurally distinct complex EPS structure was observed for *S. thermophilus* ST1 (34) and GST-6 (35). A distinct EPS structure containing less

frequently incorporated residues such as GlcNAc was reported for *S. thermophilus* S-3 (36), while the EPS structures reported for SFi12 (24) and S3 (32) also appear to be structurally distinct from all other currently resolved *S. thermophilus* EPS structures. The EPS structure resolved for STMM11 in the present study, is structurally distinct from all previously reported *S. thermophilus* EPSs currently available in the literature.

Collectively, comparative analysis of currently available, structurally resolved *S. thermophilus* EPS polymers identified 11 distinct EPS structures. Integration of comparative *eps* locus analysis with the established *eps* genotyping framework enabled assignment of experimentally resolved EPS structures to five eps genotypes (*eps* D, F, H, C_1 and C_2) (Table 1), including the first structurally characterised representative of the *eps* C_2 genotype resolved in this study. These findings support a strong association between *eps* genotype and EPS structural conservation among several genotype-associated groups. However, due to the limited availability of corresponding genome sequence data for many structurally characterised *S. thermophilus* strains, direct genotype-to-structure relationships remain unresolved for a proportion of currently available EPS structures (Table 1). Furthermore, the extensive genetic diversity observed throughout the HCL analysis suggests that additional uncharacterised genotype-to-structure relationships likely remain to be identified.

Table 1. Comparative analysis of the 11 structurally resolved EPS repeating unit structures produced by *S. thermophilus* strains and their associated *eps* genotype assignment. Conserved EPS structural groups were assigned to *eps* genotypes based on HCL, comparative *eps* locus analysis, and structural similarity to strains possessing publicly available *eps* locus sequences. Representative strains possessing each conserved EPS structure are indicated. Strains where the EPS structure was elucidated in the present study are indicated with an asterisk (*). Unassigned structures represent EPSs for which corresponding genome sequence data were unavailable or insufficient for *eps* genotype assignment.

Assigned *eps* genotype	**Associated conserved EPS structure**	**Representative strains**
eps D		Nect13*, ASCC 1275 (21), DGCC 7710 (19), XJ53 (23), SFi39 (24), STD (25), CH101 (25), STCth-9204 (19), SY89 (22), and SY102 (22)
eps F		DGCC 7785 (19), LY03 (25), IMOD1/2/3 (22), NCFB 859 (22), '21'(22), SFi20 (26), ST-10255y (19), ST4239 (19), CNCMI 733 (28) and SFi6 (29)
eps H		UCCSt50*, ST64987 (27) and ST-143 (19)
eps C_1		DCGG 7698 (19) , EU20partial (22) and UCCSt89 (33))
eps C_2		STMM11*
Unassigned		THS (30)
Unassigned		ST1 (34) and GST-6 (35)
Unassigned		S-3 (36)
Unassigned		SFi12 (24)

Unassigned		S3 (32)
Unassigned		OR 901 (31), ST111partial (25), Rs (44) and Sts (44)

Discussion:

The findings presented in this study significantly expand current understanding of EPS genetic and structural diversity within members of the *S. thermophilus* taxon by integrating compositional and structural analysis with the recently established *eps* genotyping framework (15). While structural heterogeneity was observed among the EPSs produced by strains UCCSt50, Nect13 and STMM11, comparative analysis of previously resolved EPS structures revealed that several independently characterised *S. thermophilus* strains repeatedly produced identical or highly similar EPS structures. Importantly, where complete genome sequences or sufficient *eps* locus sequence data were publicly available, strains possessing conserved EPS structures also exhibited highly similar *eps* loci and were therefore assigned to the same *eps* genotypes. Furthermore, the genotype-associated groups identified through HCL analysis enabled putative *eps* genotype assignment of several previously characterised strains for which experimentally resolved EPS structures were available in the absence of publicly accessible genome sequence data. These observations suggest that future prediction of EPS structural features may be possible directly from *eps* locus information and glycosyltransferase-associated gene content as publicly available *S. thermophilus* genome datasets continue to expand. These findings support a strong relationship between *eps* genotype classification and experimentally resolved EPS structures in *S. thermophilus*. Furthermore, these observations are consistent with the genotype-to-structure relationships established for the RGP, suggesting that conserved genetic organisation underpins the biosynthesis of both major saccharidic structures present on the surface of *S. thermophilus* (EPS and RGP).

EPS produced by *S. thermophilus* strains are predominantly heteropolysaccharides that are primarily composed of glucose, galactose, and rhamnose (19), and on occasion may also contain *N*-acetylglucosamine, *N*-acetylgalactosamine, ribose and fucose (21,27,30,32,34). The compositional diversity observed among *S. thermophilus* EPS structures is reflected in the genetic diversity present in their corresponding *eps* loci. In particular, substantial variation exists within genes encoding glycosyltransferases, including galactosyltransferases, glucosyltransferases, rhamnosyltransferases and *N*-acetylglucosamine transferases, which together are presumed to determine the monosaccharide composition and linkage characteristics of the complete polysaccharide structure (16).

One of the most notable observations is the repeated identification of a relatively small number of dominant EPS structures across multiple independently characterised strains. In particular, highly conserved EPS structures corresponding to the EPS D- (19,21–25), F- (19,22,25,26,28,29) and H-associated (19,27) groups have been repeatedly reported throughout the literature and in this study (D and H). This observed structural convergence may reflect an industrial selection bias within the available EPS dataset presented in this study. We propose that structural elucidation studies have predominantly focused on technologically important dairy strains exhibiting desirable fermentation-associated phenotypes, including enhanced viscosity, ropiness, texture formation and improved rheological properties during the fermentation process and, in particular, during yoghurt manufacture. As a result, strains selected for EPS structural characterisation in previous studies are unlikely to represent the true extent of the diversity of *S. thermophilus* EPS structures and may instead disproportionately reflect strains producing technologically favourable polymers. The repeated occurrence of

specific EPS structures across industrial dairy isolates may therefore indicate that certain *eps* genotypes are associated with beneficial physicochemical properties during dairy fermentation as it has been previously reported that the composition or combination of EPS types, rather than EPS quantity or yield alone, is one of the major contributors to the rheology, texture, and microstructural properties of yoghurt (7). From an industrial perspective, these observations raise the possibility that genotype-based screening approaches can be employed to identify strains capable of producing technologically desirable EPSs. Rather than relying exclusively on extensive downstream physicochemical characterisation, PCR- or sequence-based approaches directed towards specific *eps* genotypes may provide a rapid strategy for preliminary identification of strains with favourable fermentation-associated phenotypes. However, as EPS functionality is also influenced by factors including EPS yield, polymer length, degree of branching, molecular mass and environmental conditions (7), further studies linking *eps* genotype to EPS structures and detailed rheological analyses will be required before robust genotype-to-function relationships can be established.

Structural analysis of *S. thermophilus* EPS is highly complex and technically challenging. It is continually reported that EPS yields vary significantly between strains and are strongly influenced by cultivation conditions, carbon source, medium composition and growth phase, often resulting in insufficient or heterogeneous material for downstream analysis (21,45–47). Additionally, EPS is known to occur in multiple physical states on the cell surface, where it can be attached to the outer surface in a loose or tightly associated state or secreted into the extracellular matrix (referred to as free-EPS) (2,18), each state contributing to the difficulty or ease of extraction and subsequent structural analysis (48,49). As a

result, crude EPS preparations frequently contain mixtures of polymers with distinct molecular weights and linkage patterns, requiring extensive purification efforts to facilitate NMR-based structural resolution (21). Following purification, full characterisation typically requires multiple analytical approaches, including monosaccharide composition analysis, glycosyl-linkage analysis and NMR spectroscopy, reflecting the intrinsic complexity of these polymers (19,21,34).While *S. thermophilus* EPSs play a vital role in modifying the rheological properties and texture of fermented dairy products, accurately characterising their saccharidic structures remains a significant challenge, largely due to high strain-to-strain variability in EPS production and the technical difficulties associated with EPS purification and analysis.

The *eps* loci of *S. thermophilus* display substantial genetic diversity. This observation is consistent with previous studies describing EPS biosynthetic regions as highly variable chromosomal loci frequently associated with insertion sequences, recombination events, transposases and other mobile genetic elements (50,51). Such genomic organisation likely facilitates horizontal gene transfer and recombination-mediated diversification of glycosyltransferase-associated gene content, thereby contributing to the extensive genetic and structural diversity observed among *S. thermophilus* EPSs (51). The diversity observed among the RGPs and EPSs may also reflect selective pressures acting on these surface-associated polymers within dairy (fermentation) environments, including adaptation to fermentation conditions, interactions with other starter culture bacteria and phage predation.

Collectively, our findings establish a comparative framework linking *eps* locus diversity with experimentally resolved EPS structures in *S. thermophilus*. By

integrating hierarchical clustering, comparative genomics and structural elucidation approaches, this study demonstrates that conserved *eps* locus organisation is strongly associated with conserved EPS structures. Furthermore, identification of dominant genotype-associated EPS structural groups among industrial dairy isolates highlights the potential industrial relevance of specific *eps* genotypes and supports the future application of genotype-based approaches for the selection of technologically desirable starter cultures. Importantly, the extensive diversity observed among both *eps* loci and EPS structures also indicates that substantial genotype-to-structure relationships remain unresolved across the species. These findings therefore provide a foundation for future studies integrating large-scale comparative genomics with detailed EPS structural and functional analyses to further understand the biological and technological significance of EPS diversity in *S. thermophilus*.

Materials and Methods

Preparation and structural analysis of EPS of *S. thermophilus* strains

Approximately 8 litres of *S. thermophilus* strains, UCCSt50, Nect13 and STMM11 were grown overnight at 42 °C from single colonies in M17 broth (Sigma, USA) with 1 % (UCCSt50 and STMM11) or 5 % (Nect13) total lactose concentration (Sigma, USA). Cells were harvested by centrifugation at 4,400 × *g* at 4 °C for 30 minutes before the cell pellets were washed twice with ice-cold sterile distilled H_2O and stored at −20 C until required for further processing.

For *S. thermophilus* UCCSt50 and STMM11 cells were subjected to consecutive acid extractions as described previously (27). The crude extracts were purified on Sephadex G-50 column. The preparations had similar NMR spectra and were therefore pooled together. The EPS were separated from RGPs by reverse-phase HPLC on Zorbax C18 column. The UCCSt50 EPS was further re- purified by anion-exchange chromatography on a Hitrap Q column.

For *S. thermophilus* Nect13, the cell pellet was extracted by autoclaving in water (115 °C, 30 min; deproteinated with trichloroacetic acid (TCA, final concentration 5%), dialysed and freeze-dried to obtain the crude extract, which was purified on Sephadex G-50 column. High molecular weight material was further purified by reverse-phase HPLC on Zorbax C18 column yielding pure EPS.

General and analytical methods

Gel-permeation and ion-exchange chromatography, monosaccharide and methylation analysis and NMR spectroscopy were performed as described previously (52). Gel-permeation chromatography was performed using Sephadex G-50 (GE Healthcare, 2.6 × 80 cm, and 1 × 40 cm), Biogel P-2 (Biorad, 2.6 ×

60 cm), and Sephadex G-15 columns. Ion-exchange chromatography employed HiTrap Q and HiTrap SP columns (GE Healthcare). Monosaccharide composition and methylation analyses were performed by GC–MS using a Trace GC ULTRA system (Thermo Scientific). Reverse-phase HPLC was performed on Zorbax C18 column, eluted with a gradient 2%-80% MeCN in 0.1 % TFA, with UV detection at 220 nm. NMR experiments were carried out on a Varian INOVA 500 MHz spectrometer using standard pulse sequences including gCOSY, TOCSY, NOESY, ROESY, HSQC and HMBC as described previously (52).

eps locus extraction and genotype assignment

The *eps* gene clusters of streptococcal strains included in this study, where the complete genome sequences were available, were identified as previously described (15) between two conserved flanking genes, i.e. the genes that encode the chloride channel/purine nucleoside phosphorylase (5' flank) and VanZ (3' flank). Amino acid fasta files of these regions were extracted from 182 *S. thermophilus* genomes including 167 high-quality publicly available genomes at the chromosome or complete genome assembly, nine reference strains where the genomes were available at the contig assembly level and six *eps* cluster sequences deposited in NCBI as *eps* cluster sequence data (Supplementary Table S5) using Artemis (53). The fasta files from each *rgp* were reannotated using the Prokka v1.14.6 pipeline (54) to generate gff files. These output files were input into the Roary pipeline (55) to calculate the pangenome. Protein families were defined using Roary with default parameters and a minimum BLASTp identity threshold of 50 % (-i 50) and do not split paralogs settings (-s). A binary presence-absence matrix of clustered proteins was generated and analysed using two-way hierarchical clustering (HCL) within

the multi-experiment viewer (MeV v4.9) (56). Reference strains representing *eps* A-J genotypes (15) were presented in different colours.

Identification of previously described EPS structures

Previously published *S. thermophilus* EPS structures were compiled through manual literature searches (19,21,23,27,33,35,36) in combination with 'the EPS database' available at https://epsdatabase.cermav.cnrs.fr/database.html and using '*Streptococcus thermophilus*' as the organism search term (22,24–26,28–32,34,44,57). The reported EPS structures are shown in this study exactly as they appeared in their original publications.

Supplementary Material

Supplementary Table S1. *Streptococcus thermophilus* genomes and associated NCBI accession numbers used in the HCL analysis of this study

Strain	Species	Assembly level in NCBI	Accession number
4021	*Streptococcus thermophilus*	Complete	CP065493
4052	*Streptococcus thermophilus*	Complete	CP065476
4067	*Streptococcus thermophilus*	Complete	CP065496
4078	*Streptococcus thermophilus*	Complete	CP065477
4134	*Streptococcus thermophilus*	Complete	CP065478
4145	*Streptococcus thermophilus*	Complete	CP065492
4147	*Streptococcus thermophilus*	Complete	CP065504
13496	*Streptococcus thermophilus*	Complete	CP061025
13498	*Streptococcus thermophilus*	Complete	CP061024.
13499	*Streptococcus thermophilus*	Complete	CP061023
24738	*Streptococcus thermophilus*	Complete	CP061022
24739	*Streptococcus thermophilus*	Complete	CP061021
24740	*Streptococcus thermophilus*	Complete	CP061020
24853	*Streptococcus thermophilus*	Complete	CP061019
90728	*Streptococcus thermophilus*	Complete	CP065479
90729	*Streptococcus thermophilus*	Complete	CP065480
90730	*Streptococcus thermophilus*	Complete	CP065481
1A	*Streptococcus thermophilus*	Complete	CP065384
1F8CT	*Streptococcus thermophilus*	Chromosome	CM003138
ACA-DC_2	*Streptococcus thermophilus*	Chromosome	LT604076
APC151	*Streptococcus thermophilus*	Complete	CP019935
ASCC_1275	*Streptococcus thermophilus*	Complete	CP006819
ATCC_19258	*Streptococcus thermophilus*	Complete	CP038020
AVA116	*Streptococcus thermophilus*	Complete	CP065498
AVA1121	*Streptococcus thermophilus*	Complete	CP065488
B59671	*Streptococcus thermophilus*	Complete	CP022547
Brie1	*Streptococcus thermophilus*	Contig	JAHBRI000000000
Brie16	*Streptococcus thermophilus*	Contig	JAHBRO000000000
Brie28	*Streptococcus thermophilus*	Complete	JBTWWX000000000
CH8	*Streptococcus thermophilus*	Complete	CP089060
CICC_20372	*Streptococcus thermophilus*	Complete	CP132900
CNRZ302	*Streptococcus thermophilus*	Complete	CP065489
CNRZ385	*Streptococcus thermophilus*	Complete	CP065495
CNRZ760	*Streptococcus thermophilus*	Complete	CP065482
CNRZ887	*Streptococcus thermophilus*	Complete	CP065491
CNRZ1066	*Streptococcus thermophilus*	Complete	CP000024
CNRZ1151	*Streptococcus thermophilus*	Complete	CP065483
CNRZ1202	*Streptococcus thermophilus*	Complete	CP065506

CNRZ1575	*Streptococcus thermophilus*	Complete	CP065490
CS5	*Streptococcus thermophilus*	Complete	CP028896
CS6	*Streptococcus thermophilus*	Complete	CP050870
CS8	*Streptococcus thermophilus*	Complete	CP016439
CS9	*Streptococcus thermophilus*	Complete	CP030927
CS18	*Streptococcus thermophilus*	Complete	CP030928
CS20	*Streptococcus thermophilus*	Complete	CP030250
DGCC7710	*Streptococcus thermophilus*	Complete	CP123436
DMST-H2	*Streptococcus thermophilus*	Complete	CP063275
Douc24	*Streptococcus thermophilus*	Complete	JBTWWW000000000
EG007	*Streptococcus thermophilus*	Complete	CP069275
EPS	*Streptococcus thermophilus*	Complete	CP025400
EU01	*Streptococcus thermophilus*	Complete	CP047191
EU20	*Streptococcus thermophilus*	*eps* gene cluster	DQ393661
F1B	*Streptococcus thermophilus*	Complete	CP194006
F2A	*Streptococcus thermophilus*	Complete	CP194005
FDAARGOS_1574	*Streptococcus thermophilus*	Complete	CP086001
FDL17	*Streptococcus thermophilus*	Contig	JAHRBF000000000
FDL19	*Streptococcus thermophilus*	Contig	JAHRBE000000000
GABA	*Streptococcus thermophilus*	Complete	CP025399
IDCC2201	*Streptococcus thermophilus*	Complete	CP035306
JIM_8232	*Streptococcus thermophilus*	Complete	FR875178
KLDS_31003	*Streptococcus thermophilus*	Complete	CP016877
KLDS_SM	*Streptococcus thermophilus*	Complete	CP016026
LMD-9	*Streptococcus thermophilus*	Complete	CP000419
LMG_18311	*Streptococcus thermophilus*	Complete	CP000023
LY03	*Streptococcus thermophilus*	*eps* gene cluster	DQ393658
M17PTZA496	*Streptococcus thermophilus*	Chromosome	M002372
MAG_rmk202_sterm	*Streptococcus thermophilus*	Complete	CP046134
MM1	*Streptococcus thermophilus*	Complete	CP065484
MM20	*Streptococcus thermophilus*	Complete	CP065485
MN-BM-A01	*Streptococcus thermophilus*	Complete	CP012588
MN-BM-A02	*Streptococcus thermophilus*	Complete	CP010999
MN-ZLW-002	*Streptococcus thermophilus*	Complete	CP003499
Moz109	*Streptococcus thermophilus*	Complete	CP075363
Moz111	*Streptococcus thermophilus*	Complete	JBTWWR000000000
Moz74	*Streptococcus thermophilus*	Complete	JBTWWV000000000
Moz76	*Streptococcus thermophilus*	Complete	JBTWWU000000000
Moz77	*Streptococcus thermophilus*	Complete	JBTWWT000000000
Moz83	*Streptococcus thermophilus*	Complete	JBTWWS000000000
MRD6044	*Streptococcus thermophilus*	Complete	CP174173
MTH17CL396	*Streptococcus thermophilus*	Chromosome	CM002371
N4L	*Streptococcus thermophilus*	Chromosome	LS974444
NBC5527	*Streptococcus thermophilus*	Complete	CP176540

NCTC12958	*Streptococcus thermophilus*	Chromosome	LS483339
ND03	*Streptococcus thermophilus*	Complete	CP002340
ND07	*Streptococcus thermophilus*	Complete	CP016394
Nect1	*Streptococcus thermophilus*	Complete	JBTWWQ000000000
Nect13	*Streptococcus thermophilus*	Complete	JBTWWP000000000
NWC_2_1	*Streptococcus thermophilus*	Complete	CP031021
R1	*Streptococcus thermophilus*	Complete	CP065486
Racle124	*Streptococcus thermophilus*	Contig	JAHDUS000000000
Rico65	*Streptococcus thermophilus*	Contig	JAHBRL000000000
Rico66	*Streptococcus thermophilus*	Contig	JAHDUN000000000
Roque89	*Streptococcus thermophilus*	Complete	JBTWWO000000000
RR	*Streptococcus thermophilus*	Complete	CP065788
S24724	*Streptococcus thermophilus*	Complete	CP072443
S24726	*Streptococcus thermophilus*	Complete	CP072442
S24728	*Streptococcus thermophilus*	Complete	CP072441
S24730	*Streptococcus thermophilus*	Complete	CP072440
S24732	*Streptococcus thermophilus*	Complete	CP072439
S24733	*Streptococcus thermophilus*	Complete	CP072438
S24734	*Streptococcus thermophilus*	Complete	CP072437
S24735	*Streptococcus thermophilus*	Complete	CP072436
S24736	*Streptococcus thermophilus*	Complete	CP072435
S24742	*Streptococcus thermophilus*	Complete	CP072434
S24743	*Streptococcus thermophilus*	Complete	CP072433
S24745	*Streptococcus thermophilus*	Complete	CP072431
S24746	*Streptococcus thermophilus*	Complete	CP072430
S24747	*Streptococcus thermophilus*	Complete	CP072429
S24748	*Streptococcus thermophilus*	Complete	CP072778
S24749	*Streptococcus thermophilus*	Complete	CP072428
S24750	*Streptococcus thermophilus*	Complete	CP072427
S9	*Streptococcus thermophilus*	Complete	CP013939
Scam27	*Streptococcus thermophilus*	Contig	JAHDUQ000000000
SCB0351	*Streptococcus thermophilus*	Complete	CP142105
SFi39	*Streptococcus thermophilus*	*eps* gene cluster	AF373595
SFi6	*Streptococcus thermophilus*	*eps* gene cluster	U40830
SMQ-301	*Streptococcus thermophilus*	Complete	CP011217
ST057-1	*Streptococcus thermophilus*	Complete	CP102797
ST106	*Streptococcus thermophilus*	Complete	CP031881
ST109	*Streptococcus thermophilus*	Complete	CP031545
ST111	*Streptococcus thermophilus*	*eps* gene cluster	DQ393662
ST128	*Streptococcus thermophilus*	Complete	CP065500
ST19	*Streptococcus thermophilus*	Complete	CP065487
ST1	*Streptococcus thermophilus*	Contig	JAIZBX000000000
ST3	*Streptococcus thermophilus*	Complete	CP017064
ST55	*Streptococcus thermophilus*	Complete	CP065502

ST64987	*Streptococcus thermophilus*	Complete	CP049053
STCH_20	*Streptococcus thermophilus*	*eps* gene cluster	MK483556
STH_CIRM_1035	*Streptococcus thermophilus*	Chromosome	LR822029
STH_CIRM_1046	*Streptococcus thermophilus*	Chromosome	LR822030
STH_CIRM_1047	*Streptococcus thermophilus*	Chromosome	LR822031
STH_CIRM_1048	*Streptococcus thermophilus*	Chromosome	LR822033
STH_CIRM_1049	*Streptococcus thermophilus*	Chromosome	LR822034
STH_CIRM_1050	*Streptococcus thermophilus*	Chromosome	LR822032
STH_CIRM_1051	*Streptococcus thermophilus*	Chromosome	LR822035
STH_CIRM_1055	*Streptococcus thermophilus*	Chromosome	LR822036
STH_CIRM_1116	*Streptococcus thermophilus*	Chromosome	LR822039
STH_CIRM_1121	*Streptococcus thermophilus*	Chromosome	LR822037
STH_CIRM_1122	*Streptococcus thermophilus*	Chromosome	LR822041
STH_CIRM_1125	*Streptococcus thermophilus*	Chromosome	LR822040
STH_CIRM_1358	*Streptococcus thermophilus*	Chromosome	LR822042
STH_CIRM_16	*Streptococcus thermophilus*	Chromosome	LR822006
STH_CIRM_18	*Streptococcus thermophilus*	Chromosome	LR822008
STH_CIRM_19	*Streptococcus thermophilus*	Complete	LR822009
STH_CIRM_2101	*Streptococcus thermophilus*	Chromosome	LR822043
STH_CIRM_23	*Streptococcus thermophilus*	Chromosome	LR822011
STH_CIRM_29	*Streptococcus thermophilus*	Chromosome	LR822010
STH_CIRM_30	*Streptococcus thermophilus*	Chromosome	LR822012
STH_CIRM_32	*Streptococcus thermophilus*	Chromosome	LR822013
STH_CIRM_36	*Streptococcus thermophilus*	Chromosome	LR822014
STH_CIRM_336	*Streptococcus thermophilus*	Chromosome	LR822017
STH_CIRM_368	*Streptococcus thermophilus*	Chromosome	LR822023
STH_CIRM_65	*Streptococcus thermophilus*	Chromosome	LR822015
STH_CIRM_67	*Streptococcus thermophilus*	Chromosome	LR824002
STH_CIRM_772	*Streptococcus thermophilus*	Chromosome	LR822019
STH_CIRM_956	*Streptococcus thermophilus*	Chromosome	LR822020
STH_CIRM_961	*Streptococcus thermophilus*	Chromosome	LR822025
STH_CIRM_967	*Streptococcus thermophilus*	Chromosome	LR822026
STH_CIRM_998	*Streptococcus thermophilus*	Chromosome	LR822027
STMM1	*Streptococcus thermophilus*	Complete	CP125881
STMM2	*Streptococcus thermophilus*	Complete	CP125763
STMM3	*Streptococcus thermophilus*	Complete	CP125764
STMM4	*Streptococcus thermophilus*	Complete	CP125765
STMM11	*Streptococcus thermophilus*	Complete	CP125766
STMM22	*Streptococcus thermophilus*	Complete	CP125767
STMM25	*Streptococcus thermophilus*	Complete	CP125768
Strac42	*Streptococcus thermophilus*	Complete	JBTWWN000000000
TCI125	*Streptococcus thermophilus*	Complete	CP150875
TCI633	*Streptococcus thermophilus*	Complete	CP113237
TH-4	*Streptococcus thermophilus*	Complete	CP102538

TH982	*Streptococcus thermophilus*	Chromosome	CM003136
TH985	*Streptococcus thermophilus*	Chromosome	CM003139
TH1435	*Streptococcus thermophilus*	Chromosome	CM002369
TH1436	*Streptococcus thermophilus*	Chromosome	CM002370
TH1477	*Streptococcus thermophilus*	Chromosome	CM003135
TK-P3A	*Streptococcus thermophilus*	Complete	CP045596
TSGB4141	*Streptococcus thermophilus*	Complete	CP116772
UCCSt50	*Streptococcus thermophilus*	Complete	CP194174
UCCSt89	*Streptococcus thermophilus*	Complete	JBTWWM000000000
UCCSt95	*Streptococcus thermophilus*	Complete	CP101646
Vach60	*Streptococcus thermophilus*	Complete	JBTWWL000000000
VHProbi R08	*Streptococcus thermophilus*	Complete	CP094945

Supplementary Table S2. ^{1}H and ^{13}C NMR data for *S. thermophilus* UCCSt50 EPS polysaccharide (600 MHz, 25 °C, ppm).

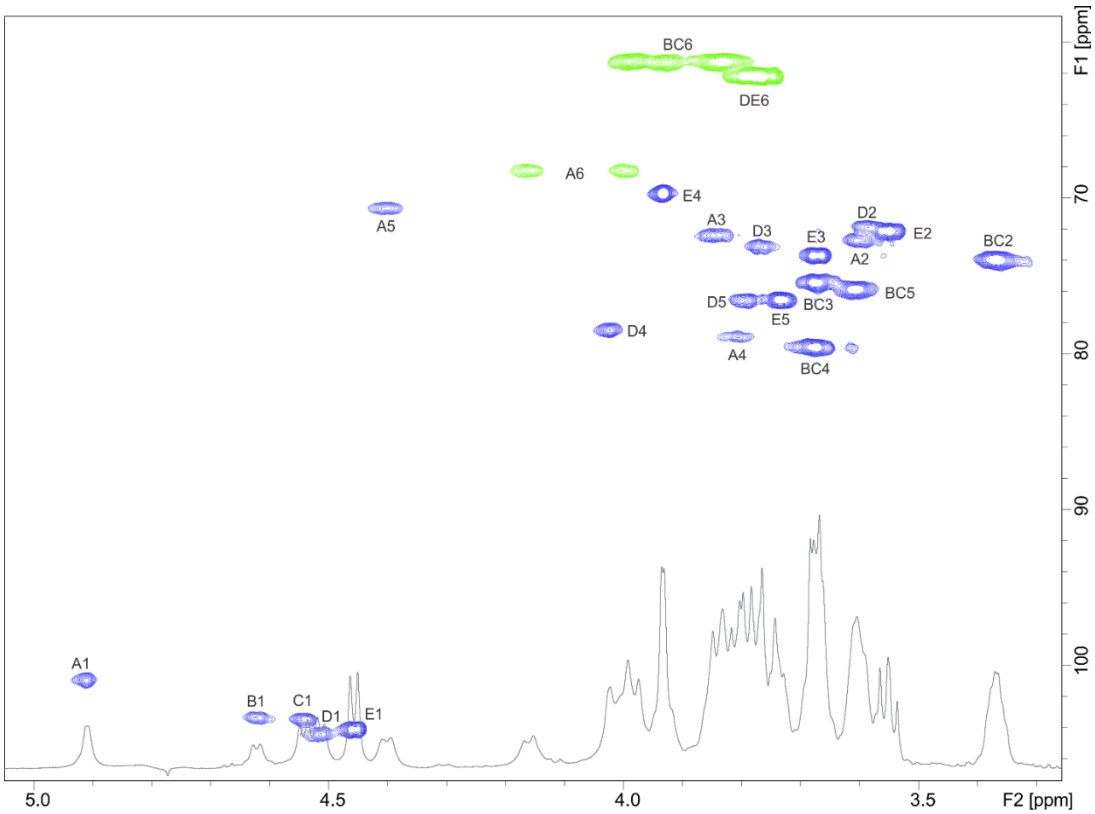
Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
α-Glc A	H	4.91	3.60	3.85	3.81	4.40	4.00; 4.16
	C	100.9	72.7	72.4	78.9	70.6	68.3
β-Glc B	H	4.62	3.37	3.68	3.68	3.61	3.82; 3.98
	C	103.4	74.0	75.5	79.6	75.9	61.3
β-Glc C	H	4.54	3.37	3.68	3.68	3.61	3.82; 3.98
	C	103.5	74.0	75.5	79.6	75.9	61.3
β-Gal D	H	4.51	3.59	3.77	4.02	3.80	3.78; 3.78
	C	104.4	71.9	73.1	78.5	76.6	62.1
β-Gal E	H	4.45	3.55	3.67	3.93	3.73	3.78; 3.78
	C	104.1	72.1	73.7	69.7	76.5	62.1

Supplementary Table S3. ^{1}H and ^{13}C NMR data (δ, ppm, D_2O, 35°C, 500 MHz) for the EPS from *S. thermophilus* strain STMM11. NAc:2.04/23.6 ppm.

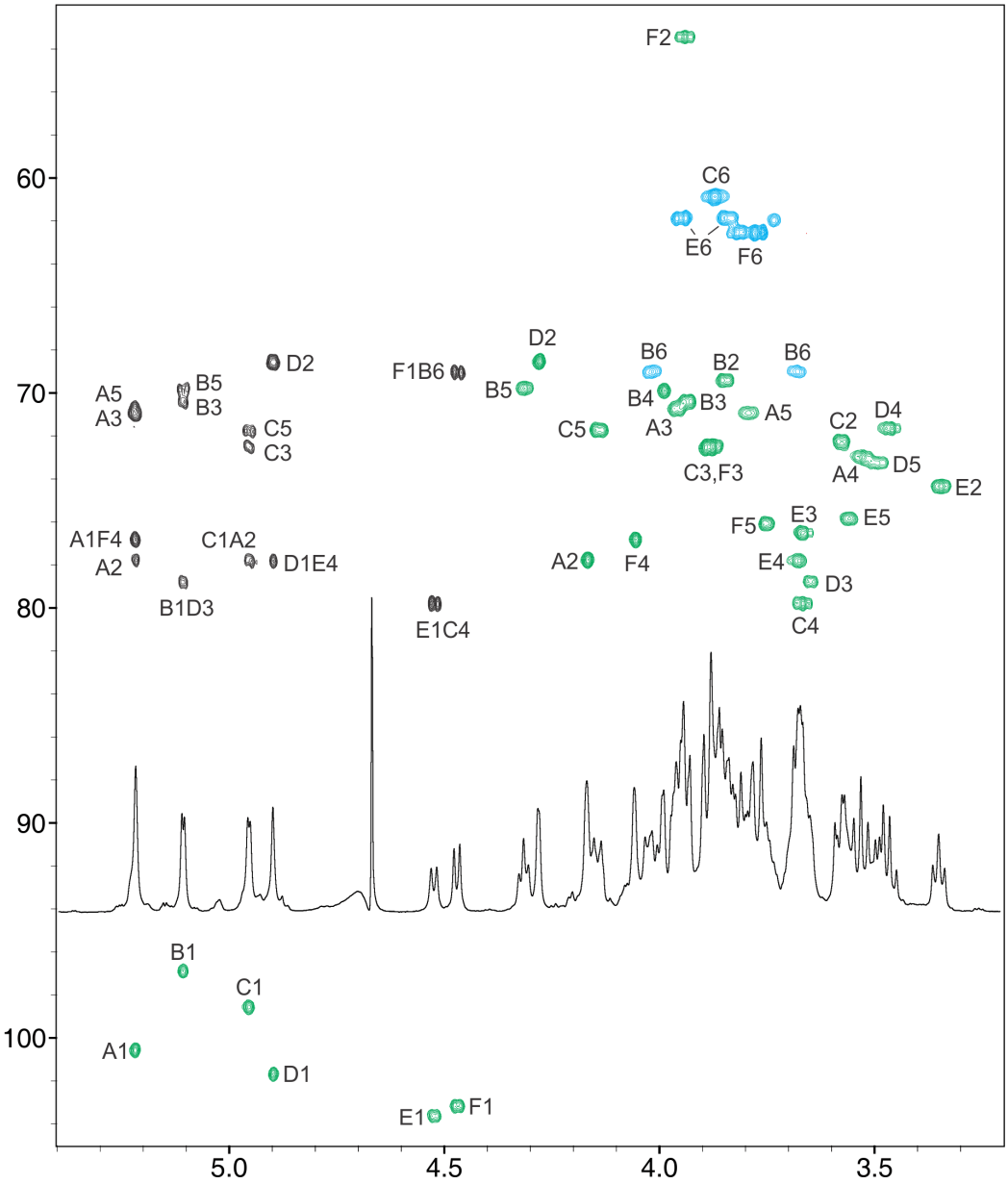
Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
α-Rha A	H	5.22	4.17	3.96	3.52	3.79	1.32
	C	100.5	77.83	70.7	73.0	70.9	17.8
α-Gal B	H	5.11	3.85	3.93	3.99	4.31	3.68; 4.02
	C	96.9	69.4	70.4	69.9	69.8	69.0
α-Glc C	H	4.96	3.58	3.88	3.67	4.14	3.87; 3.87
	C	98.6	72.3	72.5	79.8	71.7	60.9
β-Rha D	H	4.90	4.28	3.65	3.46	3.50	1.36
	C	101.7	68.5	78.9	71.6	73.2	18.0
β-Glc E	H	4.52	3.35	3.67	3.68	3.56	3.84; 3.95
	C	103.6	74.3	76.5	77.8	75.9	61.9
β-GalNAc F	H	4.47	3.94	3.88	4.06	3.75	3.77; 3.82
	C	103.2	53.4	72.5	76.8	76.1	62.6

Supplementary Table S4. ^{1}H and ^{13}C NMR data (δ, ppm, D_2O, 25°C, 500 MHz) for the EPS from *S. thermophilus* strain Nect13.

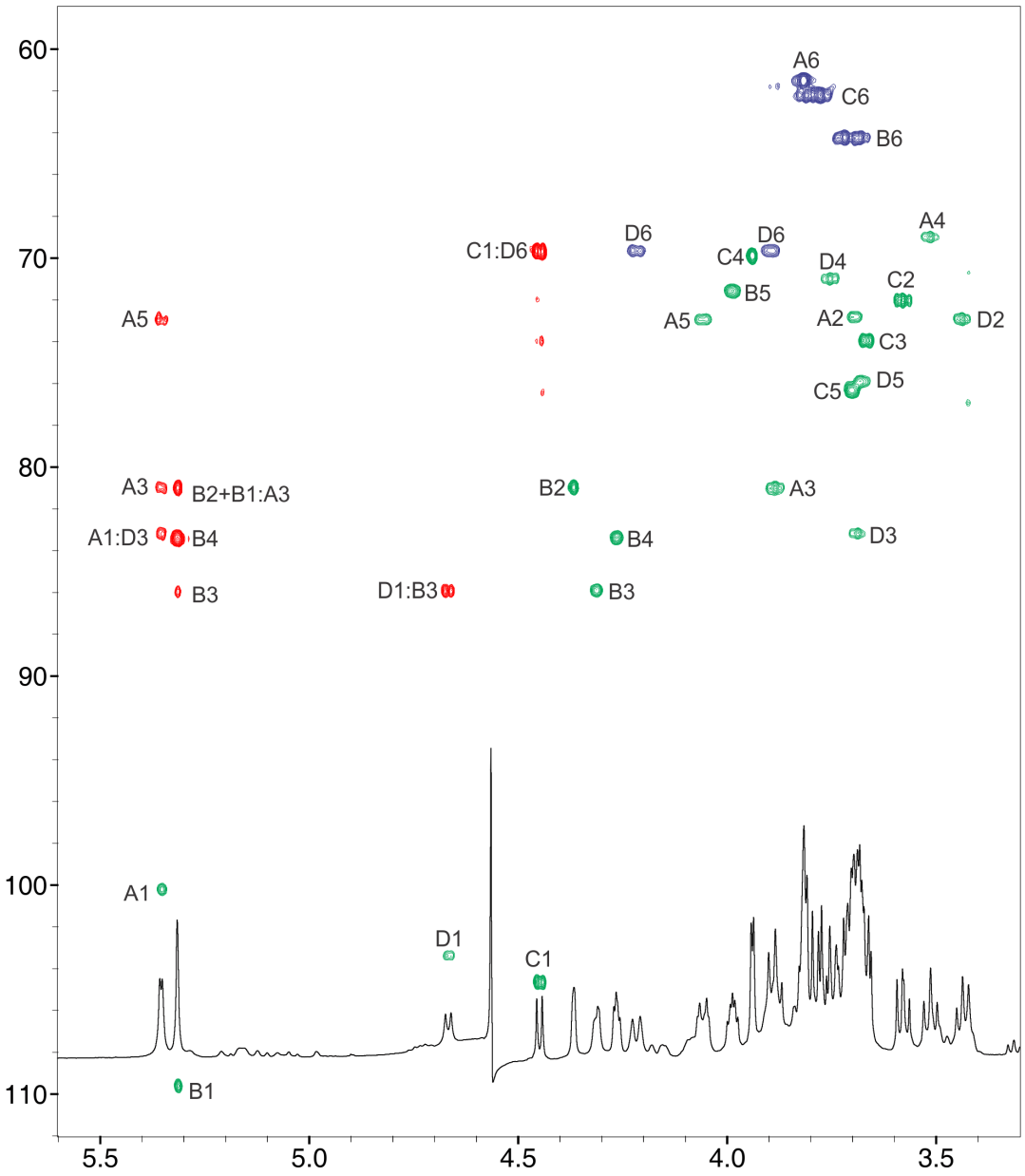
Sugar		H/C 1	H/C 2	H/C 3	H/C 4	H/C 5	H/C 6
α-Glc A	H	5.36	3.68	3.89	3.51	4.07	3.81; 3.81
	C	100.1	72.8	80.7	68.7	72.8	61.3
β-Galf B	H	5.31	4.36	4.31	4.25	3.99	3.68; 3.72
	C	109.6	81.0	85.9	83.3	71.5	64.2
β-Gal C	H	4.67	3.57	3.67	3.93	3.70	3.77; 3.80
	C	103.4	71.9	73.8	69.8	76.3	62.1
β-Glc D	H	4.44	3.42	3.68	3.76	3.68	3.88; 4.23
	C	104.6	72.9	82.8	70.9	75.8	69.5



Supplementary Fig. S1. ^{1}H-^{13}C HSQC and ^{1}H NMR spectra of the EPS from *S. thermophilus* UCCSt50.



Supplementary Fig. S2. ^{1}H-^{13}C HSQC (blue-green) and a fragment of HMBC (black) spectra of the PS2 from *S. thermophilus* strain STMM11 (35 °C, 600 MHz).



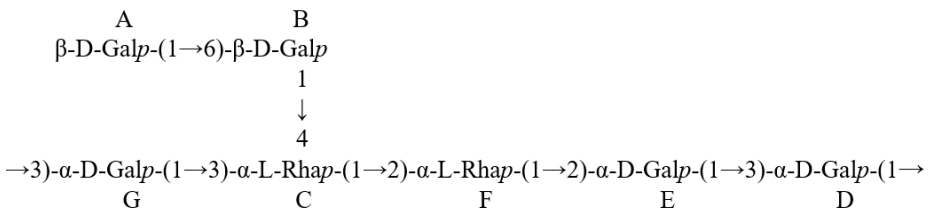
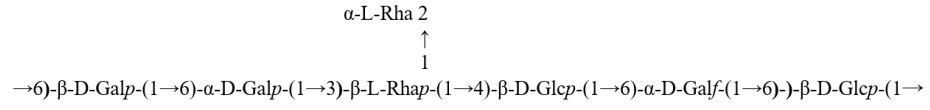
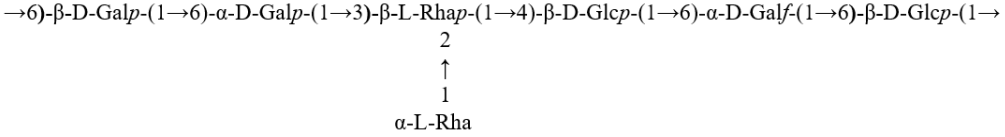
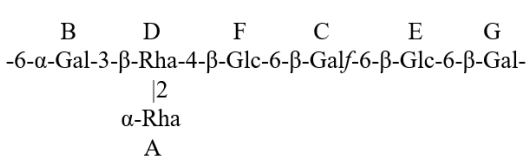
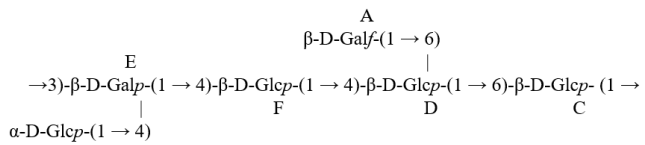
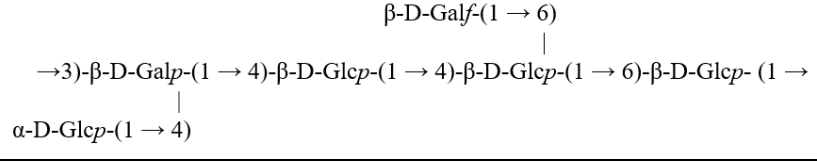
Supplementary Fig. S3. ^{1}H-^{13}C HSQC (blue-green), ^{1}H and a fragment of HMBC (red) spectra of the EPS from *S. thermophilus* strain Nect13 (45 °C, 600 MHz).

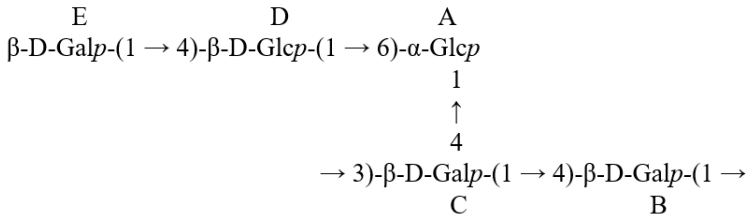
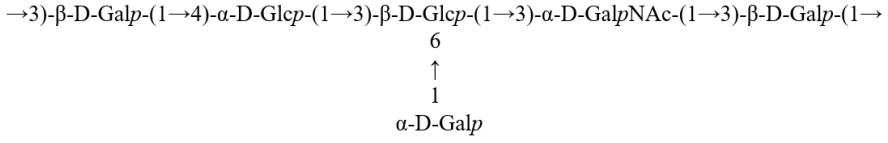
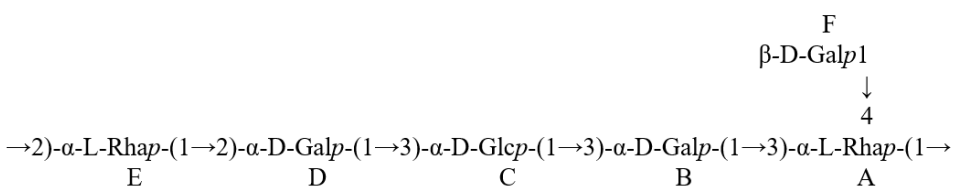
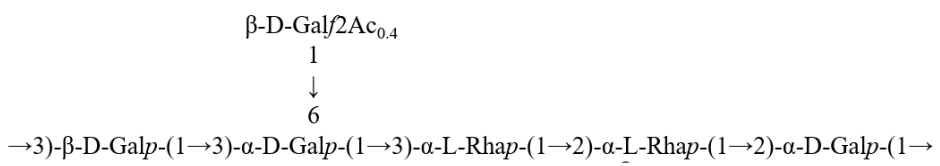
Supplementary Table S5. Published structurally resolved *Streptococcus thermophilus* EPS structures and associated genome availability. Structures are presented as originally reported in the corresponding publications.

Strain name	EPS repeating unit structure	Genome availability	Reference
SY89 and SY102	β-D-Gal*p* 1 ↓ 6 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Gal*f*-(1 →	Not available	(22)
XJ53	B β-D-Gal*p*1 ↓ 6 → 3)-α-D-Glc*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Gal*f*-(1 → D C A	Not available	(23)
SFi39	β-D-Gal*p*1 ↓ 6 → 3)-α-D-Glc*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Gal*f*-(1 →	20kb *eps* cluster available NCBI accession number AF373595	Structure – (24) *eps* cluster - (58)
STD and CH101	β-D-Gal*p* 1 ↓ 6 → 3)-α-D-Glc*p*(1 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Gal*f*-(1 →	Not available	(25)
DGCC 7710	β-D-Gal*p* 1 ↑ 6 → 3)-α-D-Glc*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Gal*f*-(1 →	Shotgun sequencing - AWVZ00000000 Complete genome – CP123436	Structure – (19) Genome shotgun - (59) CP123436 – unpublished

ASCC 1275	Structure 1 A C B 3)-α-D-Glc*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-β-D-Gal*f*-(1 → 6 ↑ 1 β-D-Gal*p* F D Structure 2 4)-β-D-Gal*p*-(1 → E Structure 3 6)-β-D-Glc*p*-(1 →	Complete genome - CP006819	Structure – (21) Genome - (17)
STCth-9204	→3)-β-Gal*f*-(1 → 3)-α-Glc-(1 → 3)-β-Glc-(1 → 6 ↑ 1 β-Gal	Not available	(19)
DGCC 7785	α-D-Gal*p* 1 ↑ 6 3)-β-D-Gal*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-α-D-GalpNAc-(1 →	Not available	(19)
LY03	α-D-Gal*p* 1 ↓ 6 → 3)-β-D-Gal*p*(1 → 3)-β-D-Glc*p*-(1 → 3)-α-D-Gal*p*NAc-(1 →	15kb *eps* gene cluster available - DQ393658.1	Structure - (25) *eps* cluster - (14)
IMOD1/2/3, NCFB 859 and '21'	α-D-Gal*p* 1 ↓ 6 → 3)-β-D-Gal*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-α-D-Gal*p*NAc-(1 →	Not available	(22)
SFi20	A B N → 3)-β-D-Gal*p*-(1 → 3)-β-D-Glc*p*-(1 → 3)-α-D-Gal*p*NAc-(1 → 6 ↑ 1 α-D-Gal*p*1 X	Not available	(26)

CNCMI 733	→ 3)-β-d-Galp-(1 → 3)-[α-d-Galp-(1 → 6)]-β-d-Glcp-(1 → 3)-α-d-GalpNAc-(1 →	Not available	(28)
SFi6	→ 3)-β-D-Gal*p*-(1 → 3)-[α-D-Gal*p*-(1 → 6)]-β-D-Glc*p*-(1 → 3)-α-D-Gal*p*NAc-(1 →	15kb *eps* gene cluster available - U40830.1	(29)
ST-10255y and ST4239	→ 3)-α-GalNAc-(1 → 3)-β-Gal-(1 → 3)-β-Glc-(1 → 6 ↑ 1 α-Gal	Not available	(19)
ST64987	B A D -4-β-Glc-4-α-Glc-4-β-Gal- |6 β-Gal-4-β-Glc F C	Complete genome – CP049053	(27)
ST-143	→4)-β-Glc-(1 → 4)-α-Glc-(1 → 4)-β-Gal-(1 → 6 ↑ 1 β-Glc 4 ↑ 1 β-Gal	Not available	(19)
OR 901	β-D-Gal$_p$-(1 → 6)-β-D-Gal$_p$-(1 → 4) | →2)-α-D-Gal$_p$-(1 → 3)-α-D-Gal$_p$-(1 → 3)-α-D-Gal$_p$-(1 → 3)-α-L-Rha$_p$-(1 →2)-α-L-Rha$_p$-(1→	Not available	(31)
ST111partial	β-D-Gal*p*-(1 → 6)-β-D-Gal*p* 4 ↓ 1 →3)-α-DGal*p*-(1 → 3)-α-DGal*p*-(1 → 3)-α-LRha*p*-(1 → 2)-α-LRha*p*-(1 →2)-α-DGal*p*-	5kb *eps* cluster available DQ393662	Structure - (25) *eps* cluster - unpublished

Rs and Sts		Not available	(44)
DGCC 7698		Not available	(19)
EU20partial		7kb *eps* gene cluster available - DQ393660	Structure - (22) *eps* cluster - (14)
UCCSt89		Shotgun sequencing – JANFMW000000000	(33)
ST1		Shotgun sequencing - JAIZBX000000000.1	Structure - (34) Genome - unpublished
GST-6		Not available	(35)

THS		Not available	(30)
S-3		Not available	(36)
SFi12		Not available	(24)
S3		Not available	(32)

References

1. Jurášková D, Ribeiro SC, Silva CCG. Exopolysaccharides Produced by Lactic Acid Bacteria: From Biosynthesis to Health-Promoting Properties. Foods. 2022 Jan 8;11(2):156. doi:10.3390/foods11020156

2. Jurášková D, Pires VC, Ribeiro SC, Ferreira SS, Bernardo F, Evtyugin D, et al. Exopolysaccharide (EPS)-Producing *Streptococcus thermophilus*: Functional and Probiotic Potential. Foods. 2025 Aug 28;14(17):3013. doi:10.3390/foods14173013

3. Kapse N, Pisu V, Dhakephalkar T, Margale P, Shetty D, Wagh S, et al. Unveiling the Probiotic Potential of *Streptococcus thermophilus* MCC0200: Insights from In Vitro Studies Corroborated with Genome Analysis. Microorganisms. 2024 Feb 7;12(2):347. doi:10.3390/microorganisms12020347

4. Angelin J, Kavitha M. Exopolysaccharides from probiotic bacteria and their health potential. International Journal of Biological Macromolecules. 2020 Nov;162:853–65. doi:10.1016/j.ijbiomac.2020.06.190

5. Martinović A, Cocuzzi R, Arioli S, Mora D. *Streptococcus thermophilus*: To Survive, or Not to Survive the Gastrointestinal Tract, That Is the Question! Nutrients. 2020 Jul 22;12(8):2175. doi:10.3390/nu12082175

6. Allouche R, Hafeez Z, Dary-Mourot A, Genay M, Miclo L. *Streptococcus thermophilus*: A Source of Postbiotics Displaying Anti-Inflammatory Effects in THP 1 Macrophages. Molecules. 2024 Mar 30;29(7):1552. doi:10.3390/molecules29071552

7. Mende S, Rohm H, Jaros D. Influence of exopolysaccharides on the structure, texture, stability and sensory properties of yoghurt and related products. International Dairy Journal. 2016 Jan;52:57–71. doi:10.1016/j.idairyj.2015.08.002

8. Tiwari S, Kavitake D, Devi PB, Halady Shetty P. Bacterial exopolysaccharides for improvement of technological, functional and rheological properties of yoghurt. International Journal of Biological Macromolecules. 2021 Jul;183:1585–95. doi:10.1016/j.ijbiomac.2021.05.140

9. Zhang L, Folkenberg DM, Amigo JM, Ipsen R. Effect of exopolysaccharide-producing starter cultures and post-fermentation mechanical treatment on textural properties and microstructure of low fat yoghurt. International Dairy Journal. 2016 Feb;53:10–9. doi:10.1016/j.idairyj.2015.09.008

10. Hernández-Figueroa RH, López-Malo A, Mani-López E. Lactic Acid Bacteria-Derived Exopolysaccharides: Dual Roles as Functional Ingredients and Fermentation Agents in Food Applications. Fermentation. 2025 Sep 17;11(9):538. doi:10.3390/fermentation11090538

11. Chauhan K, Rao A. Clean-label alternatives for food preservation: An emerging trend. Heliyon. 2024 Aug;10(16):e35815. doi:10.1016/j.heliyon.2024.e35815

12. Korcz E, Varga L. Exopolysaccharides from lactic acid bacteria: Techno-functional application in the food industry. Trends in Food Science & Technology. 2021 Apr;110:375–84. doi:10.1016/j.tifs.2021.02.014

13. Lynch KM, Zannini E, Coffey A, Arendt EK. Lactic Acid Bacteria Exopolysaccharides in Foods and Beverages: Isolation, Properties,

Characterization, and Health Benefits. Annu Rev Food Sci Technol. 2018 Mar 25;9(1):155–76. doi:10.1146/annurev-food-030117-012537

14. De Vuyst L, Weckx S, Ravyts F, Herman L, Leroy F. New insights into the exopolysaccharide production of *Streptococcus thermophilus*. International Dairy Journal. 2011 Sep;21(9):586–91. doi:10.1016/j.idairyj.2011.03.016

15. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

16. Cui Y, Jiang X, Hao M, Qu X, Hu T. New advances in exopolysaccharides production of *Streptococcus thermophilus*. Arch Microbiol. 2017 Aug;199(6):799–809. doi:10.1007/s00203-017-1366-1

17. Wu Q, Tun HM, Leung FCC, Shah NP. Genomic insights into high exopolysaccharide-producing dairy starter bacterium *Streptococcus thermophilus* ASCC 1275. Sci Rep. 2014 May 15;4(1):4974. doi:10.1038/srep04974

18. Broadbent JR, McMahon DJ, Welker DL, Oberg CJ, Moineau S. Biochemistry, Genetics, and Applications of Exopolysaccharide Production in *Streptococcus thermophilus*: A Review. Journal of Dairy Science. 2003 Feb;86(2):407–23. doi:10.3168/jds.S0022-0302(03)73619-4

19. Pachekrepapol U, Lucey JA, Gong Y, Naran R, Azadi P. Characterization of the chemical structures and physical properties of exopolysaccharides produced by various *Streptococcus thermophilus* strains. Journal of Dairy Science. 2017 May;100(5):3424–35. doi:10.3168/jds.2016-12125

20. Xiao M, Nie J, Fu X, Zhang L, Mou H, Niu Q. *Streptococcus thermophilus*: A next-generation chassis for customizing exopolysaccharide with rare sugar. Food Research International. 2026 Feb;226:118150. doi:10.1016/j.foodres.2025.118150

21. Padmanabhan A, Shah NP. Structural characterization of exopolysaccharide from *Streptococcus thermophilus* ASCC 1275. Journal of Dairy Science. 2020 Aug;103(8):6830–42. doi:10.3168/jds.2019-17439

22. Marshall VM, Laws AP, Gu Y, Levander F, Radstrom P, De Vuyst L, et al. Exopolysaccharide-producing strains of thermophilic lactic acid bacteria cluster into groups according to their EPS structure. Lett Appl Microbiol. 2001 Jun;32(6):433–7. doi:10.1046/j.1472-765X.2001.00937.x

23. Xia W, Han J, Zhu S, Wang Y, Zhang W, Wu Z. Structural elucidation of the exopolysaccharide from *Streptococcus thermophilus* XJ53 and the effect of its molecular weight on immune activity. International Journal of Biological Macromolecules. 2023 Mar;230:123177. doi:10.1016/j.ijbiomac.2023.123177

24. Lemoine J, Chirat F, Wieruszeski JM, Strecker G, Favre N, Neeser JR. Structural characterization of the exocellular polysaccharides produced by *Streptococcus thermophilus* SFi39 and SFi12. Appl Environ Microbiol. 1997 Sep;63(9):3512–8. doi:10.1128/aem.63.9.3512-3518.1997

25. De Vuyst L, Zamfir M, Mozzi F, Adriany T, Marshall V, Degeest B, et al. Exopolysaccharide-producing *Streptococcus thermophilus* strains as functional starter cultures in the production of fermented milks. International Dairy Journal. 2003 Jan;13(8):707–17. doi:10.1016/S0958-6946(03)00105-5

26. Navarini L, Abatangelo A, Bertocchi C, Conti E, Bosco M, Picotti F. Isolation and characterization of the exopolysaccharide produced by *Streptococcus*

thermophilus SFi20. International Journal of Biological Macromolecules. 2001 Mar;28(3):219–26. doi:10.1016/S0141-8130(01)00118-0

27. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Molecular Microbiology. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494

28. Doco T, Wieruszeski JM, Fournet B, Carcano D, Ramos P, Loones A. Structure of an exocellular polysaccharide produced by *Streptococcus thermophilus*. Carbohydrate Research. 1990 May;198(2):313–21. doi:10.1016/0008-6215(90)84301-A

29. Stingele F, Neeser JR, Mollet B. Identification and characterization of the eps (Exopolysaccharide) gene cluster from *Streptococcus thermophilus* Sfi6. J Bacteriol. 1996 Mar;178(6):1680–90. doi:10.1128/jb.178.6.1680-1690.1996

30. Nordmark EL, Yang Z, Huttunen E, Widmalm G. Structural Studies of an Exopolysaccharide Produced by *Streptococcus thermophilus* THS. Biomacromolecules. 2005 Jan 1;6(1):105–8. doi:10.1021/bm0496514

31. Bubb WA, Urashima T, Fujiwara R, Shinnai T, Ariga H. Structural characterisation of the exocellular polysaccharide produced by *Streptococcus thermophilus* OR 901. Carbohydrate Research. 1997 Jun;301(1–2):41–50. doi:10.1016/S0008-6215(97)00083-9

32. Faber EJ, Van Den Haak MJ, Kamerling JP, Vliegenthart JFG. Structure of the exopolysaccharide produced by *Streptococcus thermophilus* S3. Carbohydrate Research. 2001 Mar;331(2):173–82. doi:10.1016/S0008-6215(01)00013-1

33. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. Bipartite *rgp* Locus Diversity in *Streptococcus thermophilus* Corresponds to Backbone and Side Chain Differences of Its Rhamnose-Containing Cell Wall Polysaccharide. Ercolini D, editor. Appl Environ Microbiol. 2022 Dec 13;88(23):e01504-22. doi:10.1128/aem.01504-22

34. Säwén E, Huttunen E, Zhang X, Yang Z, Widmalm G. Structural analysis of the exopolysaccharide produced by *Streptococcus thermophilus* ST1 solely by NMR spectroscopy. J Biomol NMR. 2010 Jun;47(2):125–34. doi:10.1007/s10858-010-9413-0

35. Zhang J, Cao Y, Wang J, Guo X, Zheng Y, Zhao W, et al. Physicochemical characteristics and bioactivities of the exopolysaccharide and its sulphated polymer from *Streptococcus thermophilus* GST-6. Carbohydrate Polymers. 2016 Aug;146:368–75. doi:10.1016/j.carbpol.2016.03.063

36. Xu Z, Guo Q, Zhang H, Xiong Z, Zhang X, Ai L. Structural characterisation of EPS of *Streptococcus thermophilus* S-3 and its application in milk fermentation. International Journal of Biological Macromolecules. 2021 May;178:263–9. doi:10.1016/j.ijbiomac.2021.02.173

37. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Alexandre G, editor. Appl Environ Microbiol. 2022 Jan 11;88(1):e01723-21. doi:10.1128/AEM.01723-21

38. Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, Van Sinderen D, Mahony J. The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Elkins CA, editor.

Appl Environ Microbiol. 2025 Oct 22;91(10):e01238-25. doi:10.1128/aem.01238-25

39. Ruta S, Murray M, Kampff Z, McDonnell B, Lugli GA, Ventura M, et al. Microbial Ecology of Pecorino Siciliano PDO Cheese Production Systems. Fermentation. 2023 Jun 29;9(7):620. doi:10.3390/fermentation9070620

40. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in *Streptococcus thermophilus* bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

41. Alexandraki V, Kazou M, Blom J, Pot B, Papadimitriou K, Tsakalidou E. Comparative Genomics of *Streptococcus thermophilus* Support Important Traits Concerning the Evolution, Biology and Technological Properties of the Species. Front Microbiol. 2019 Dec 20;10:2916. doi:10.3389/fmicb.2019.02916

42. Faber EJ, Kamerling JP, Vliegenthart JFG. Structure of the extracellular polysaccharide produced by *Lactobacillus delbrueckii* subsp. *bulgaricus* 291. Carbohydrate Research. 2001 Mar;331(2):183–94. doi:10.1016/S0008-6215(01)00012-X

43. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Robinson P, editor. Bioinformatics. 2021 Aug 25;37(16):2473–5. doi:10.1093/bioinformatics/btab007

44. Faber EJ, Zoon P, Kamerling JP, Vliegenthart JFG. The exopolysaccharides produced by *Streptococcus thermophilus* Rs and Sts have the same repeating unit but differ in viscosity of their milk cultures. Carbohydrate Research. 1998 Aug;310(4):269–76. doi:10.1016/S0008-6215(98)00189-X

45. Wa Y, Chanyi RM, Nguyen HTH, Gu R, Day L, Altermann E. Extracellular Polysaccharide Extraction from *Streptococcus thermophilus* in Fermented Milk. Cocolin L, editor. Microbiol Spectr. 2022 Apr 27;10(2):e02280-21. doi:10.1128/spectrum.02280-21

46. Zhang T, Zhang C, Li S, Zhang Y, Yang Z. Growth and exopolysaccharide production by *Streptococcus thermophilus* ST1 in skim milk. Braz J Microbiol. 2011 Dec;42(4):1470–8. doi:10.1590/S1517-83822011000400033

47. Sørensen HM, Rochfort KD, Maye S, MacLeod G, Brabazon D, Loscher C, et al. Exopolysaccharides of Lactic Acid Bacteria: Production, Purification and Health Benefits towards Functional Food. Nutrients. 2022 Jul 18;14(14):2938. doi:10.3390/nu14142938

48. Nachtigall C, Surber G, Herbi F, Wefers D, Jaros D, Rohm H. Production and molecular structure of heteropolysaccharides from two lactic acid bacteria. Carbohydrate Polymers. 2020 May;236:116019. doi:10.1016/j.carbpol.2020.116019

49. Mende S, Dong T, Rathemacher A, Rohm H, Jaros D. Physicochemical characterisation of the exopolysaccharides of *Streptococcus thermophilus* ST -143. Int J of Food Sci Tech. 2014 May;49(5):1254–63. doi:10.1111/ijfs.12505

50. Bourgoin F, Pluvinet A, Gintz B, Decaris B, Guédon G. Are horizontal transfers involved in the evolution of the *Streptococcus thermophilus* exopolysaccharide synthesis loci? Gene. 1999 Jun;233(1–2):151–61. doi:10.1016/S0378-1119(99)00144-4

51. Cui Y, Xu T, Qu X, Hu T, Jiang X, Zhao C. New Insights into Various Production Characteristics of *Streptococcus thermophilus* Strains. IJMS. 2016 Oct 12;17(10):1701. doi:10.3390/ijms17101701

52. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, et al. The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. Molecular Microbiology. 2020 Oct;114(4):582–96. doi:10.1111/mmi.14561

53. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012 Feb 15;28(4):464–9. doi:10.1093/bioinformatics/btr703

54. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9. doi:10.1093/bioinformatics/btu153

55. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3. doi:10.1093/bioinformatics/btv421

56. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: A Free, Open-Source System for Microarray Data Management and Analysis. BioTechniques. 2003 Feb;34(2):374–8. doi:10.2144/03342mt01

57. Vaningelgem F, Meulen RVD, Zamfir M, Adriany T, Laws AP, Vuyst LD. *Streptococcus thermophilus* ST 111 produces a stable high-molecular-mass exopolysaccharide in milk-based medium. International Dairy Journal. 2004 Oct;14(10):857–64. doi:10.1016/j.idairyj.2004.03.007

58. Germond J, Delley M, D'Amico N, Vincent SJF. Heterologous expression and characterization of the exopolysaccharide from *Streptococcus thermophilus* Sfi39. European Journal of Biochemistry. 2001 Oct;268(19):5149–56. doi:10.1046/j.0014-2956.2001.02450.x

59. Barrangou R, Coûté-Monvoisin AC, Stahl B, Chavichvily I, Damange F, Romero DA, et al. Genomic impact of CRISPR immunization against bacteriophages. Biochemical Society Transactions. 2013 Dec 1;41(6):1383–91. doi:10.1042/BST20130160

Chapter VI

General Discussion

General discussion and future perspectives

Streptococcus thermophilus is one of the most commercially important lactic acid bacterial species employed in food fermentations and is used extensively in the manufacture of yoghurt, hard cheeses and a variety of other fermented dairy products globally (1). Its long history of safe use in food production has resulted in its designation as both a “Generally Recognized as Safe” (GRAS) microorganism by the United States Food and Drug Administration and a “Qualified Presumption of Safety” (QPS) organism by the European Food Safety Authority (2). The technological success of *S. thermophilus* is largely attributable to its rapid metabolism of lactose, efficient acidification of milk and production of metabolites that contribute to the flavour perception, texture and rheological properties of fermented foods (3). In particular, extracellular and cell-associated polysaccharides produced by this species play important roles in determining the viscosity, mouthfeel and overall quality of many fermented dairy products (4–7).

Despite its extensive industrial application, *S. thermophilus* is considered a relatively young bacterial species in evolutionary terms, having emerged relatively recently (estimated 3,000–30,000 years ago) from a commensal ancestor of the salivarius group (8,9). Phylogenetically, *S. thermophilus* belongs to the salivarius group of the viridans streptococci, which also includes *Streptococcus salivarius*, *Streptococcus vestibularis* (8), and the most recently recognised species, *Streptococcus raffinosi* (10). Although now recognised as a distinct species, evidence suggests that *S. thermophilus* emerged relatively recently through adaptation of a salivarius-like ancestor to the nutrient-rich dairy environment. This evolutionary transition has been accompanied by genome decay, with pseudogenes typically accounting for approximately 10 % of the predicted genes in *S.*

thermophilus genomes, and horizontal gene transfer events that collectively facilitated its specialisation for growth in milk (3,11–13).

While the domestication of *S. thermophilus* has yielded strains with highly desirable technological properties, it has not diminished the persistent threat posed by bacteriophage infection. Indeed, phages remain one of the principal causes of fermentation delays in industrial dairy processes, where infection can result in delayed acidification, reduced product quality and subsequent economic losses (14,15). The outcome of these phage-host interactions is frequently determined by cell surface-associated structures that function as phage receptors. Among these, exopolysaccharides (EPSs) and rhamnose-glucose polysaccharides (RGPs) represent the structurally diverse and biologically important cell wall polysaccharides (CWPS) produced by *S. thermophilus*. Therefore, understanding the molecular basis of phage-host interactions remains a major objective within dairy-associated microbiology. The studies presented throughout this thesis sought to address several outstanding questions relating to the genetic diversity, composition/structure and biological functions of the RGP and EPS in *S. thermophilus*. Through comparative genomics, structural characterisation, transcriptional analyses and functional genetic studies, this work establishes direct links between polysaccharide genotype, biochemical structure and phage susceptibility. Collectively, the findings presented here demonstrate that CWPS diversity is a major determinant of phage-host interactions in *S. thermophilus* and provides a framework for understanding and predicting these interactions.

One of the principal findings of this thesis was the demonstration that RGPs are a widespread and conserved feature of the *Streptococcus* genus. Although RGP structures have been extensively characterised in a limited number of model

streptococcal species, including *Streptococcus mutans*, *Streptococcus pyogenes*, *Streptococcus agalactiae* and *S. thermophilus* (16), their broader distribution throughout the genus has remained poorly defined. The findings presented in Chapter III revealed that genes associated with RGP biosynthesis are conserved across multiple streptococcal species, indicating that these structures likely represent a fundamental component of the streptococcal cell wall. Furthermore, while substantial diversity was observed within individual *rgp* loci, their overall organisation, chromosomal location and core biosynthetic functions remained conserved among members of the salivarius group. The widespread presence of these loci raises important questions regarding the biological significance of RGPs in the broader context of streptococci. The synthesis of complex cell wall polysaccharides represents a considerable metabolic investment and therefore implies the existence of strong selective pressures maintaining these structures throughout streptococcal evolution. In pathogenic streptococci, RGPs have been implicated in cell division, maintenance of cell wall integrity, biofilm formation, host interactions and immune recognition, and virulence (17–24). While many of these functions remain less well defined in dairy-associated streptococci, the conservation of *rgp* loci among both pathogenic and non-pathogenic species in this genus suggests that these polysaccharides fulfil biological roles extending far beyond the phage-host interactions explored throughout this thesis. The presence of conserved precursor synthesis genes, backbone-associated biosynthetic functions and side chain biosynthetic modules across phylogenetically diverse streptococci suggests that maintenance of the RGP structure is likely important for core aspects of cell envelope organisation and cellular physiology. Therefore, the salivarius group may represent a useful model system through which broader questions

relating to the evolution, biosynthesis and biological functions of streptococcal CWPSs can be explored. This may be particularly relevant in the context of oral streptococci, which are among the earliest bacterial colonisers of humans and play important roles in the establishment, succession and maintenance of microbial communities within the oral cavity (25). The findings presented in Chapter III demonstrated that genes associated with RGP biosynthesis are conserved not only among members of the salivarius group but also within several oral streptococcal species. In particular, species such as *Streptococcus anginosus*, *Streptococcus gordonii*, *Streptococcus sanguinis* and *Streptococcus parasanguinis* represent attractive candidates for future investigation, as they possess the conserved RGP biosynthesis-associated genes identified in this thesis, are well represented within publicly available genome collections (Chapter III) and are established colonisers of the human oral cavity (26). Furthermore, while often regarded as commensal members of the oral microbiota, several oral streptococci have the ability to cause disease, namely, bacteraemia and endocarditis, as has been documented for *S. parasanguinis* (27), *S. gordonii* (28) and *S. sanguinis* (29). Given their dual roles as early colonisers of the oral cavity and opportunistic pathogens, these species provide an attractive system through which the biological significance of CWPSs, including RGPs, may be further explored. In particular, it would be interesting to determine whether the structural diversity, biosynthetic organisation and functional roles of RGPs (and EPSs) identified throughout this thesis are conserved in these organisms and whether such polysaccharides contribute to processes associated with oral colonisation, biofilm formation and host interactions.

Findings from Chapter III and V demonstrate that the observed genetic diversity within the *rgp* and *eps* loci directly underpin the structural diversity observed

among streptococcal CWPSs and that this is primarily attributable to the unique glycosyltransferase-encoding gene content within these loci. While previous studies have highlighted the existence of considerable variation within both loci (14,30–35), direct links between genotype and polysaccharide structure remained limited due to the lack of strains for which genomic and structural data were available. Through the structural elucidation of previously uncharacterised RGP and EPS structures, together with hierarchical clustering and comparative analyses of their corresponding biosynthetic loci, this thesis provides new insights into the relationship between polysaccharide genotype and biochemical structure. In particular, the identification of multiple distinct *rgp* genotypes within the salivarius group and the assignment of experimentally resolved EPS structures to established *eps* genotypes provides an important foundation for the development of genotype-to-structure frameworks in *S. thermophilus*. The establishment of such frameworks represents an important step towards a predictive understanding of CWPS biosynthesis in this genus. Similar approaches have been used in *Lactococcus lactis*, where comparative genomic analyses have facilitated the prediction of CWPS structures (36) and informed studies of phage-host interactions (37). In contrast, equivalent frameworks for streptococci remain comparatively underdeveloped despite the growing availability of genome sequence data. The findings presented in this thesis suggest that the increasing integration of comparative genomics, structural biology and functional assignment of CWPS biosynthesis genes may allow polysaccharide structures to be inferred directly from genomic information. Such an approach would significantly accelerate the characterisation of streptococcal cell surface diversity and facilitate the identification of structural

features associated with important biological traits including host colonisation, immune recognition and phage susceptibility.

Although advances have been made in the genetic manipulation of *S. thermophilus* through the development of natural transformation systems (38), electroporation-based methodologies (39) and more recently CRISPR-Cas-assisted genome editing approaches (40), functional interrogation of complex biosynthetic loci remains challenging. Difficulties associated with transformation efficiency, host defence systems (mainly restriction and modification systems) and the labour-intensive nature of generating and validating multiple mutants continue to limit lab-based functional studies in this species (41). Furthermore, the efficiency of genetic manipulation can vary considerably between *S. thermophilus* strains. While some strains exhibit high levels of natural competence, others remain poorly transformable and frequently require strain-specific optimisation of competence induction conditions (41,42). As increasingly sophisticated genetic engineering methodologies become available for *S. thermophilus* and related streptococci, it should become more feasible to investigate the functions of individual CWPS biosynthesis genes using targeted genetic approaches.

Beyond establishing relationships between polysaccharide genotype and structure, the findings presented throughout this thesis provide important insights into the genetic basis of RGP biosynthesis. The generation and characterisation of bacteriophage-insensitive mutants (BIMs) possessing mutations within distinct genes of the *rgp* locus provided important insights into the genetic basis of RGP side chain biosynthesis in *S. thermophilus*. Structural analyses of the associated polysaccharides produced by these derivatives revealed distinct modifications to the side chain architecture, ranging from complete loss of the side chain structure

to partially truncated intermediates (di or tri-saccharide for BIM F1B) and derivatives retaining only a single side chain residue. When considered alongside transcriptional analyses of the *rgp* locus of UCCSt50 and the restoration of side chain biosynthesis through heterologous complementation, these findings enabled the construction of a preliminary model of RGP side chain assembly and provided experimental evidence for the functions of several genes within the *rgp* locus of the UCCSt50 strain. These studies demonstrate how coupling genetics and structural biology can move investigations of streptococcal CWPS beyond proposed models (30) towards experimentally validated pathways. While the functional assignments established in this thesis are specific to *S. thermophilus*, homologous genes are widely distributed among *rgp* loci throughout the *Streptococcus* genus (16). These findings may aid the interpretation of analogous biosynthetic pathways in other streptococci, although experimental validation will be required to determine the extent to which these functions are conserved. The framework established here therefore provides a foundation for future studies seeking to resolve RGP biosynthesis beyond the salivarius group.

One of the most significant findings of the presented thesis was the demonstration that structural variation within the RGP directly influences phage susceptibility in *S. thermophilus*. While previous studies, including Chapter II of this thesis, had established that CWPSs function as receptors for several dairy streptococcal phages (18,20,43,44), the precise structural features required for phage adsorption remained poorly understood. Through the identification of the RGP as the receptor for phage P738 and the subsequent characterisation of BIMs possessing structurally altered RGPs, the studies presented here demonstrate that distinct phages recognising the same polysaccharide can possess considerably different receptor

requirements. We propose that phage adsorption is more nuanced than the presence or absence of the receptor but instead reflects specific molecular interactions between phage receptor-binding proteins and defined structural features or "docking points" on the host cell surface. The distinct infection phenotypes observed for phages P738 and SW13 provide a compelling example of this proposal. While complete loss of the RGP side chain abolished infection by both phages, derivatives possessing substantially truncated side chain structures remained susceptible to SW13 while displaying complete resistance to P738. Such observations may help explain how closely related phages can display different host ranges despite recognising the same general receptor. This highlights the importance of considering receptor architecture, rather than receptor identity alone, when investigating phage-host interactions in future studies.

The findings presented throughout this thesis have also substantially advanced our understanding of receptor utilisation among phages infecting *S. thermophilus*. Through the identification of the RGP as the receptor for phage P738 and the functional characterisation of RGP-dependent infection, this thesis contributes to a growing body of evidence demonstrating that CWPSs are central determinants of phage susceptibility in *S. thermophilus*. Our current understanding of these interactions is that members of the *Brussowvirus* and P738 genera utilise the RGP as a receptor, while *Piorkowskivirus* phages recognise distinct EPS structures (18,20,43,44). A substantial proportion of the five phage genera infecting *S. thermophilus* can now be linked to defined saccharidic receptors, representing a significant understanding of the phage-host interactome. Despite this progress, important knowledge gaps remain. In particular, the receptors utilised by members of the *Moineauvirus* and *Vansinderenvirus* genera have been proposed (31,33)

however are yet to be experimentally validated. Resolving these outstanding interactions represents an important future research direction and will be necessary to establish a complete receptor framework for phages infecting *S. thermophilus*. Furthermore, recent analyses of receptor-binding proteins have revealed the existence of distinct subgroups within both the *Moineauvirus* and *Brussowvirus* genera (45), raising the possibility that phages belonging to different subgroups may recognise distinct receptors or receptor variants. Such observations suggest that receptor usage among dairy streptococcal phages may be more complex than currently appreciated and highlight the need for continued integration of comparative genomics, structural biology and experimental receptor validation studies.

The findings presented in this thesis demonstrate that CWPSs represent dynamic determinants of phage susceptibility. These findings provide a framework through which phage-host interactions may be interpreted and, ultimately, predicted from genomic information. Such predictive approaches have long represented a major objective of dairy phage research and may facilitate the rational selection of starter cultures possessing desirable technological characteristics while simultaneously reducing susceptibility to problematic phages. The work presented therefore represents an important step towards a more mechanistic understanding of phage-host interactions in *S. thermophilus* and provides a foundation for future studies investigating similar interactions in the wider *Streptococcus* genus.

Despite these advances, several important questions remain unresolved. Addressing the knowledge gaps regarding the receptor recognised by *Moineauvirus* and *Vansinderenvirus* phages for will likely require the continued integration of approaches employed throughout this thesis, including BIM generation while also

exploring additional approaches including targeted mutagenesis, adsorption assays and fluorescent receptor-binding protein analyses. Furthermore, while the findings presented here demonstrate that distinct phages recognising the same polysaccharide can possess different minimal structural requirements, the precise molecular basis of these interactions remains unknown. Future studies should seek to define the minimal receptor requirements necessary for phage adsorption, including the contribution of individual sugar residues, glycosidic linkages, side chain length and the three-dimensional presentation of these interactions. The ability to engineer altered RGP structures through targeted genetic manipulation, provides a promising route through which such questions may be addressed. The increasing availability of advanced structural biology approaches presents exciting opportunities for investigating phage-host interactions. The application of AlphaFold-based structural modelling (46,47) coupled with cryo-electron images may facilitate the identification of the "docking points" described in this thesis and provide mechanistic insights into the binding specificity of individual phages. Similarly, the continued development of online bioinformatic tools will enable an increased understanding of RGP and EPS biosynthesis.

Finally, while this thesis has explored the diversity of CWPSs within *S. thermophilus*, these structures likely represent only a fraction of the broader polysaccharide diversity present throughout the *Streptococcus* genus. While this thesis establishes a framework for interpreting *rgp* and *eps* diversity within *S. thermophilus* and its closely related species, comparatively little is known regarding the structures encoded by many *rgp* and *eps* loci present in other oral, animal-associated and pathogenic streptococcal species. Whether the genotype-to-structure relationships and biological functions described throughout this thesis are

conserved across the wider *Streptococcus* genus remains unclear. Expanding these analyses beyond the salivarius group therefore represents an important future research direction and will help determine the extent to which the findings presented here are representative of the entire genus.

Streptococcus was among the earliest bacterial genera to be described and, over its ~150-year history, has been extensively studied in the context of human health, pathogenesis and food fermentations. During this time, our understanding of the genus has progressed from simple observations of "small organisms as found in either isolated or arranged in pairs, sometimes in chains" by Theodor Billroth in 1874 (48) to detailed insights into the prevalence, diversity and ecological significance of its constituent species. Advances in genomics, structural biology and molecular microbiology have revealed an extraordinary diversity of streptococci occupying a wide range of ecological niches and performing roles that range from beneficial commensalism and food fermentation to opportunistic and invasive disease. The findings presented throughout this thesis further contribute to this expanding body of knowledge by providing new insights into the distribution, diversity and biological significance of streptococcal CWPSs. The continued expansion of publicly available genome collections, together with advances in bioinformatics, structural biology and functional genetic approaches, now places us in a unique position to address questions that would have been unimaginable only a decade ago. These advances provide opportunities to further unravel the complexity of this bacterial genus and continue building upon more than a century of discovery.

References

1. Iyer R, Tomar SK, Uma Maheswari T, Singh R. Streptococcus thermophilus strains: Multifunctional lactic acid bacteria. International Dairy Journal. 2010 Mar;20(3):133–41. doi:10.1016/j.idairyj.2009.10.005

2. Martinović A, Cocuzzi R, Arioli S, Mora D. *Streptococcus thermophilus*: To Survive, or Not to Survive the Gastrointestinal Tract, That Is the Question! Nutrients. 2020 Jul 22;12(8):2175. doi:10.3390/nu12082175

3. Alexandraki V, Kazou M, Blom J, Pot B, Papadimitriou K, Tsakalidou E. Comparative Genomics of *Streptococcus thermophilus* Support Important Traits Concerning the Evolution, Biology and Technological Properties of the Species. Front Microbiol. 2019 Dec 20;10:2916. doi:10.3389/fmicb.2019.02916

4. Broadbent JR, McMahon DJ, Welker DL, Oberg CJ, Moineau S. Biochemistry, Genetics, and Applications of Exopolysaccharide Production in *Streptococcus thermophilus*: A Review. Journal of Dairy Science. 2003 Feb;86(2):407–23. doi:10.3168/jds.S0022-0302(03)73619-4

5. De Vuyst L, Zamfir M, Mozzi F, Adriany T, Marshall V, Degeest B, et al. Exopolysaccharide-producing *Streptococcus thermophilus* strains as functional starter cultures in the production of fermented milks. International Dairy Journal. 2003 Jan;13(8):707–17. doi:10.1016/S0958-6946(03)00105-5

6. Jurášková D, Pires VC, Ribeiro SC, Ferreira SS, Bernardo F, Evtyugin D, et al. Exopolysaccharide (EPS)-Producing *Streptococcus thermophilus*: Functional and Probiotic Potential. Foods. 2025 Aug 28;14(17):3013. doi:10.3390/foods14173013

7. Korcz E, Varga L. Exopolysaccharides from lactic acid bacteria: Techno-functional application in the food industry. Trends in Food Science & Technology. 2021 Apr;110:375–84. doi:10.1016/j.tifs.2021.02.014

8. Delorme C, Abraham AL, Renault P, Guédon E. Genomics of *Streptococcus salivarius*, a major human commensal. Infection, Genetics and Evolution. 2015 Jul;33:381–92. doi:10.1016/j.meegid.2014.10.001

9. Roux, E., Nicolas, A., Valence, F. *et al.* The genomic basis of the *Streptococcus thermophilus* health-promoting properties. BMC Genomics 23, 210 (2022). https://doi.org/10.1186/s12864-022-08459-y

10. Nguyen HV, Trinh ATV, Bui LNH, Hoang ATL, Tran QTL, Trinh TT. *Streptococcus raffinosi* sp. nov., isolated from human breast milk samples. International Journal of Systematic and Evolutionary Microbiology. 2024 Jul 3;74(7). doi:10.1099/ijsem.0.006442

11. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, et al. Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. Nat Biotechnol. 2004 Dec;22(12):1554–8. doi:10.1038/nbt1034

12. Hatmaker EA, Riley LA, O'Dell KB, Papanek B, Graveley BR, Garrett SC, et al. Complete Genome Sequence of Industrial Dairy Strain *Streptococcus thermophilus* DGCC 7710. Genome Announc. 2018 Feb 8;6(6):e01587-17. doi:10.1128/genomeA.01587-17

13. Hols P, Hancy F, Fontaine L, Grossiord B, Prozzi D, Leblondbourget N, et al. New insights in the molecular biology and physiology of revealed by

comparative genomics. FEMS Microbiology Reviews. 2005 Aug;29(3):435–63. doi:10.1016/j.femsre.2005.04.008

14. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. Dairy lactococcal and streptococcal phage–host interactions: an industrial perspective in an evolving phage landscape. FEMS Microbiology Reviews. 2020 Nov 24;44(6):909–32. doi:10.1093/femsre/fuaa048

15. Garneau JE, Moineau S. Bacteriophages of lactic acid bacteria and their impact on milk fermentations. Microb Cell Fact. 2011;10(Suppl 1):S20. doi:10.1186/1475-2859-10-S1-S20

16. Kampff Z, Van Sinderen D, Mahony J. Cell wall polysaccharides of streptococci: A genetic and structural perspective. Biotechnology Advances. 2023 Dec;69:108279. doi:10.1016/j.biotechadv.2023.108279

17. Shibata Y, Yamashita Y, Van Der Ploeg JR. The serotype-specific glucose side chain of rhamnose–glucose polysaccharides is essential for adsorption of bacteriophage M102 to *Streptococcus mutans*. FEMS Microbiology Letters. 2009 May;294(1):68–73. doi:10.1111/j.1574-6968.2009.01546.x

18. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. Cell Wall Glycans Mediate Recognition of the Dairy Bacterium *Streptococcus thermophilus* by Bacteriophages. Dudley EG, editor. Appl Environ Microbiol. 2018 Dec;84(23):e01847-18. doi:10.1128/AEM.01847-18

19. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Appl Environ Microbiol. 2022 Jan

11;88(1):e0172321. doi:10.1128/AEM.01723-21 PubMed PMID: 34669424; PubMed Central PMCID: PMC8752142.

20. Kampff Z, Sadovskaya I, Vinogradov E, McDonnell B, Van Sinderen D, Mahony J. The cell surface-associated rhamnose-glucose polysaccharide represents the receptor of *Streptococcus thermophilus* bacteriophage P738. Elkins CA, editor. Appl Environ Microbiol. 2025 Oct 22;91(10):e01238-25. doi:10.1128/aem.01238-25

21. Zamakhaeva S, Chaton CT, Rush JS, Ajay Castro S, Kenner CW, Yarawsky AE, et al. Modification of cell wall polysaccharide guides cell division in *Streptococcus mutans*. Nat Chem Biol. 2021 Aug;17(8):878–87. doi:10.1038/s41589-021-00803-9

22. De A, Liao S, Bitoun JP, Roth R, Beatty WL, Wu H, et al. Deficiency of RgpG Causes Major Defects in Cell Division and Biofilm Formation, and Deficiency of LytR-CpsA-Psr Family Proteins Leads to Accumulation of Cell Wall Antigens in Culture Medium by *Streptococcus mutans*. Nojiri H, editor. Appl Environ Microbiol. 2017 Sep;83(17):e00928-17. doi:10.1128/AEM.00928-17

23. van Sorge NM, Cole JN, Kuipers K, Henningham A, Aziz RK, Kasirer-Friede A, et al. The Classical Lancefield Antigen of Group A *Streptococcus* Is a Virulence Determinant with Implications for Vaccine Design. Cell Host & Microbe. 2014 Jun;15(6):729–40. doi:10.1016/j.chom.2014.05.009

24. Kovacs CJ, Faustoferri RC, Bischer AP, Quivey RG. *Streptococcus mutans* requires mature rhamnose-glucose polysaccharides for proper pathophysiology, morphogenesis and cellular division. Molecular Microbiology. 2019 Sep;112(3):944–59. doi:10.1111/mmi.14330

25. Abranches J, Zeng L, Kajfasz JK, Palmer SR, Chakraborty B, Wen ZT, et al. Biology of Oral Streptococci. Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Braunstein M, Rood JI, editors. Microbiol Spectr. 2018 Sep 7;6(5):6.5.11. doi:10.1128/microbiolspec.GPP3-0042-2018

26. Baty JJ, Stoner SN, Scoffield JA. Oral Commensal Streptococci: Gatekeepers of the Oral Cavity. O'Toole G, editor. J Bacteriol. 2022 Nov 15;204(11):e00257-22. doi:10.1128/jb.00257-22

27. Wang J, Hu W, Zhang R, Jin F, Hu J, Bai Q, et al. Meningitis due to *Streptococcus parasanguinis* after percutaneous radiofrequency treatment for trigeminal neuralgia: A case report. Science Progress. 2023 Apr;106(2):00368504231170302. doi:10.1177/00368504231170302

28. Mosailova N, Truong J, Dietrich T, Ashurst J. *Streptococcus gordonii* : A Rare Cause of Infective Endocarditis. Case Reports in Infectious Diseases. 2019 Jun 12;2019:1–2. doi:10.1155/2019/7127848

29. Zhu B, Macleod LC, Kitten T, Xu P. *Streptococcus Sanguinis* Biofilm Formation & Interaction with Oral Pathogens. Future Microbiol. 2018 Jun;13(8):915–32. doi:10.2217/fmb-2018-0043

30. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. Bipartite *rgp* Locus Diversity in *Streptococcus thermophilus* Corresponds to Backbone and Side Chain Differences of Its Rhamnose-Containing Cell Wall Polysaccharide. Ercolini D, editor. Appl Environ Microbiol. 2022 Dec 13;88(23):e01504-22. doi:10.1128/aem.01504-22

31. Szymczak P, Rau MH, Monteiro JM, Pinho MG, Filipe SR, Vogensen FK, et al. A comparative genomics approach for identifying host-range determinants in

Streptococcus thermophilus bacteriophages. Sci Rep. 2019 May 29;9(1):7991. doi:10.1038/s41598-019-44481-z

32. De Vuyst L, Weckx S, Ravyts F, Herman L, Leroy F. New insights into the exopolysaccharide production of *Streptococcus thermophilus*. International Dairy Journal. 2011 Sep;21(9):586–91. doi:10.1016/j.idairyj.2011.03.016

33. Parlindungan E, McDonnell B, Lugli GA, Ventura M, Van Sinderen D, Mahony J. Dairy streptococcal cell wall and exopolysaccharide genome diversity. Microbial Genomics. 2022 Apr 26;8(4). doi:10.1099/mgen.0.000803

34. Cui Y, Jiang X, Hao M, Qu X, Hu T. New advances in exopolysaccharides production of *Streptococcus thermophilus*. Arch Microbiol. 2017 Aug;199(6):799–809. doi:10.1007/s00203-017-1366-1

35. Wu Q, Tun HM, Leung FCC, Shah NP. Genomic insights into high exopolysaccharide-producing dairy starter bacterium *Streptococcus thermophilus* ASCC 1275. Sci Rep. 2014 May 15;4(1):4974. doi:10.1038/srep04974

36. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, et al. The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. Molecular Microbiology. 2020 Oct;114(4):582–96. doi:10.1111/mmi.14561

37. White K, Eraclio G, McDonnell B, Lugli GA, Crowley T, Ventura M, et al. Lactococcal phage–host profiling through binding studies between cell wall polysaccharide types and *Skunavirus* receptor-binding proteins. Microbial Genomics. 2025 Apr 28;11(4). doi:10.1099/mgen.0.001395

38. Gardan R, Besset C, Gitton C, Guillot A, Fontaine L, Hols P, et al. Extracellular Life Cycle of ComS, the Competence-Stimulating Peptide of *Streptococcus thermophilus*. J Bacteriol. 2013 Apr 15;195(8):1845–55. doi:10.1128/JB.02196-12

39. Kong L, Xiong Z, Xia Y, Ai L. High-efficiency transformation of *Streptococcus thermophilus* using electroporation. J Sci Food Agric. 2021 Dec;101(15):6578–85. doi:10.1002/jsfa.11292

40. Kong L, Song X, Xia Y, Ai L, Xiong Z. Construction of a CRISPR/nCas9-assisted genome editing system for exopolysaccharide biosynthesis in *Streptococcus thermophilus*. Food Research International. 2022 Aug;158:111550. doi:10.1016/j.foodres.2022.111550

41. Rozanov AS, Shaposhnikov LA, Bondarenko KD, Sazonov AE. Advances in Genetic Transformation of Lactic Acid Bacteria: Overcoming Barriers and Enhancing Plasmid Tools. IJMS. 2025 Sep 19;26(18):9146. doi:10.3390/ijms26189146

42. Dandoy D, Fremaux C, Henry De Frahan M, Horvath P, Boyaval P, Hols P, et al. The fast milk acidifying phenotype of *Streptococcus thermophilus* can be acquired by natural transformation of the genomic island encoding the cell-envelope proteinase PrtS. Microb Cell Fact. 2011 Dec;10(S1):S21. doi:10.1186/1475-2859-10-S1-S21

43. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, et al. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. Molecular Microbiology. 2020 Jul;114(1):31–45. doi:10.1111/mmi.14494

44. Lavelle K, Sadovskaya I, Vinogradov E, Kelleher P, Lugli GA, Ventura M, et al. *Brussowvirus* SW13 Requires a Cell Surface-Associated Polysaccharide To Recognize Its *Streptococcus thermophilus* Host. Alexandre G, editor. Appl Environ Microbiol. 2022 Jan 11;88(1):e01723-21. doi:10.1128/AEM.01723-21

45. White K, Szafran JD, McDonnell B, Chawande G, Faherty-McGrath L, Lugli GA, et al. Diversity and prevalence of dairy streptococcal phages in two European dairy fermentation facilities over a one-year period. Microbial Genomics. 2026 May 26;12(5). doi:10.1099/mgen.0.001729

46. Baptista D, Gomez-Lucas L, Jänes J, Krogan NJ, Amorim MJ, Ivarsson Y, et al. AlphaFold models of host-pathogen interactions elucidate the prevalence and structural modes of molecular mimicry [Internet]. 2025 [cited 2026 Jun 9]. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.06.04.657796 doi:10.1101/2025.06.04.657796

47. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9. doi:10.1038/s41586-021-03819-2

48. Etymologia: *Streptococcus*. Emerg Infect Dis. 2016 Nov;22(11):1977–1977. doi:10.3201/eid2211.ET2211

Acknowledgements

Firstly, I would like to wholeheartedly thank Jenn and Douwe. Sometimes I get flashbacks to my interview for this PhD and I still have no idea how you agreed to give me the opportunity to work with ye. I am incredibly grateful for that initial vote of confidence and for all the support, guidance, patience and encouragement that followed. Douwe, thank you for your insight, advice and willingness to support (and sometimes challenge) my thinking. Sometimes a “what does this mean?” comment is exactly what was needed. Thank you for these small pushes and encouragement throughout.

To Jenn in particular, thank you for always making time for me, for your enthusiasm, honesty and endless support. Your guidance has shaped not only this thesis but also the scientist I have become. Although the voodoo doll might have a few bruises, I’m beyond proud of this thesis and the body of work it represents, and that is in no small part due to your continual confidence in my abilities. What is remarkable is that, alongside all that you do for your students and colleagues, you have also found the time and energy to champion initiatives that support women in STEM, equality and inclusion. These efforts have made the lab (for me) a completely welcoming and supportive place to work, and they have certainly made me feel more comfortable and confident during my time here. As a woman in science, having role models and advocates like you matters more than you probably realise. This one's for the girls.

To my amazing scientific collaborators, Irina Sadovskaya and Evgeny Vinogradov that I’ve had the opportunity to work with. Even though we have never met in person, you have both been such an important part of my PhD journey. Your

generosity in sharing your expertise and your time, along with the care, detail and kindness of your responses, meant that I never hesitated to send an email or ask a question. It has been a privilege to work with you both. I've learned so much through our collaboration, and your contributions have shaped both the direction and strength of my PhD. Thank you, sincerely, for being such wonderful collaborators. It has meant a great deal to me.

To everyone who has worked beside me in labs 335–340, past and present, thank you for creating such a welcoming, supportive and enjoyable environment to work in. I have been incredibly fortunate to spend my PhD surrounded by such a talented, generous and genuinely fun group of people. Whether it was scientific discussions, troubleshooting experiments, sharing advice or providing a much-needed distraction, you all helped make the lab a genuinely nice place to be. A special acknowledgment goes to anyone who took the time to bake treats, organise lab outings or just fancied a pint after a long week. Your efforts were often exactly what was needed. It's been a pleasure celebrating milestones with ye.

To my family, I don't know if there are words to explain what your support has meant to me, but I'll try. To Lucy and Joe, yes it's about cheese. Even going for a PhD is the most middle child thing I have ever done. Thank you both for not even trying to understand what this thesis is about but understanding it was important to me. Thank you for mocking me, I love ye too. To my dad, Barry, whose pride in his three children is seen and heard by everyone. Thank you for checking in with me and encouraging me every step of the way. To my mom, Tracy, whose love has carried me to where I am today. Thank you for Sundays and coffees and texts. Thank you for supporting and encouraging my choices, even when I wasn't sure where they might lead me. Thank you for giving me the confidence to continue my

education and the freedom to figure things out along the way. You and dad have always been on my side, and I hope you know that means a lot to me. To Walt, thank you for the wagging tail, for always being happy to see me, for the many pets and for hanging in there. I love you dearly, my sweet boy.

Finally, to Gill, my entire world. Thank you. Thank you for the late nights, the grumpy moods, the practice presentations, and for loving me through everything. Thank you for your patience, your support and believing in me. Thank you for waiting for me. I cannot wait to say yes to you and grow old with you. The best is yet to come. I love you.

To me, past, present and future – be proud, be excited, enjoy this, you deserve it.