CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Outsmarting phages to generate better dairy products
Authors	White, Kelsey
Publication date	2024
Original Citation	White, K. E. 2024. Outsmarting phages to generate better dairy products. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2024, Kelsey White. - https://creativecommons.org/licenses/by-nc-sa/4.0/
Download date	2026-11-15 13:54:32
Item downloaded from	https://hdl.handle.net/10468/17033



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



Outsmarting phages to generate better dairy products

Thesis presented by

Kelsey Elizabeth White, BSc

for the degree of

Doctor of Philosophy

University College Cork

School of Microbiology

Head of School: Dr David Clarke

Supervisors: Prof. Douwe van Sinderen and Prof. Jennifer Mahony

2024

Table of Contents

Declaration iii

List of Publications v

Abbreviations vi

Abstract ix

Chapter 1. General Introduction 1

Chapter 2. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures 66

Chapter 3. Profiling phage-host interactions between *Skunavirus* receptor binding proteins and lactococcal cell wall polysaccharide structures 130

Chapter 4. Biodiversity of skunaviruses isolated from dairy fermentation factories employing mesophilic undefined starter cultures 196

Chapter 5. Genomic analysis and technological assessment of lactococcal strains isolated from undefined mesophilic starter cultures 271

Chapter 6. General discussion and future perspectives 327

Acknowledgements 334

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.

Signed: …….

Kelsey White

This thesis is dedicated to the memory of my mom (Liz) and brother (Brian).

Inár gcroí go deo.

List of Publications

(**Chapter 2**) White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* **90**:e02152-23.

White K, Eraclio G, Lugli GA, Ventura M, Mahony J, Bello FD, van Sinderen D. 2024. A Metagenomics Approach to Enumerate Bacteriophages in a Food Niche, p. 185–199. *In* Tumban, E (ed), Bacteriophages: Methods and Protocols. Springer US, New York, NY.

Casey E, McDonnell B, White K, Stamou P, Crowley T, O'Neill I, Lavelle K, Hayes S, Lugli GA, Arboleya S, James K, Ventura M, Martinez I, Gueimonde M, Dal Bello F, Nally K, Mahony J, van Sinderen D. 2022. Needle in a Whey-Stack: PhRACS as a Discovery Tool for Unknown Phage-Host Combinations. *mBio* **13**:e03334-21.

(**Part of Chapter 1**) White K, Yu JH, Eraclio G, Dal Bello F, Nauta A, Mahony J, van Sinderen D. 2022. Bacteriophage-host interactions as a platform to establish the role of phages in modulating the microbial composition of fermented foods. *Microbiome Res Rep* **1**:3.

Abbreviations

aa: Amino acid

ABC: ATP binding cassette

Abi: Abortive infection

ANOVA: Analysis of variance

Ara: Arabinose

BIM: Bacteriophage-insensitive mutant

BLAST: Basic local alignment search tool

BppA: Accessory baseplate protein

BREX: Bacteriophage exclusion

CBM: Carbohydrate binding module

CFU: Colony forming unit

CsCl: Cesium chloride

C_T: Cycle threshold

CWPS: Cell wall polysaccharides

Dit: Distal tail protein

dsDNA: Double-stranded DNA

EOP: Efficiency of plaquing

EPS: Exopolysaccharides

FU: (Arbitrary) fluorescence unit

Gal: Galactose

Gal*f*: Galactofuranose

GalNAc: N-acetyl-galactosamine

GFP: Green fluorescent protein

GlcNAc: N-acetyl-glucosamine

Gro: Glycerol

HCL: Hierarchical clustering

HHpred: Homology detection and structure production by HMM-HMM comparison

^{His}**GFP-RBP**: polyhistidine-tagged green fluorescent protein- receptor binding protein fusion proteins

LAB: Lactic acid bacteria

lac: Lactose utilization gene cluster

LCA: Last common ancestor

LTA: Lipoteichoic acid

MCL: Markov Clustering

MTase: Methyltransferase

MTP: Major tail protein

NCBI: National Centre for Biotechnology Information

NPS: Neck passage structure

NT: Nucleotide

OD: Optical density

ORF: Open reading frame

PDC: Phage defense candidate

PFU: Plaque forming units

PHASTEST: PHAge Search Tool with Enhanced Sequence Translation

PhRACS: Phage RBP-activated cell sorting

PSP: Polysaccharide pellicle

qPCR: Quantitative polymerase chain reaction

RBP: Receptor binding protein

Rha: Rhamnose

RM: Restriction modification

RSM: Reconstituted skimmed milk

SDS-PAGE: sodium dodecyl sulfate–polyacrylamide gel electrophoresis

SNFG: Symbols Nomenclature for Glycans

Tal: Tail associated lysin

TEM: Transmission electron microscopy

TMP: Tape measure protein

TpeX: Major tail extension protein

Abstract

The dairy fermentation industry relies on the activity of various lactic acid bacteria in starter cultures to acidify milk and produce a variety of fermented dairy products. Undefined mesophilic starter cultures are predominantly composed of many strains (in an unknown composition) belonging to species such as *Lactococcus lactis* and *Lactococcus cremoris*. Bacteriophage infection of the bacterial strains within these starter cultures are one of the most significant and persistent threats to successful dairy fermentations. Delayed or disrupted fermentations, as a result of phage attacks, may cause costly disruptions resulting in unnecessary waste and a loss of product. One of the most problematic and abundant lactococcal phage genera is *Skunavirus*. Most lactococcal phages, including skunaviruses, commence infection with the binding of a phage-encoded receptor binding protein (RBP) to a surface-exposed host receptor, which in the case of skunaviruses is a cell wall polysaccharide (CWPS). To ensure the manufacture of consistent and high-quality fermentation end products, it is essential that the composition and integrity of starter cultures is maintained. Therefore, this thesis aims to investigate phage-host interactions between skunaviruses and their lactococcal hosts pertinent to dairy fermentation facilities utilising undefined mesophilic starter cultures.

Chapter 2 of this thesis outlines a combined culture-based analysis and metagenomics investigation to evaluate the composition of two mesophilic undefined starter cultures. A novel qPCR-based genotype detection assay, capable of discerning nine distinct lactococcal genotypes, was devised and used to examine compositional changes of an undefined starter culture. These

analytical approaches facilitated a multifaceted assessment of starter culture compositional stability during milk fermentation, which has become an important quality control aspect due to increasing demand for reliable and high-quality dairy products.

In Chapter 3, a phageome analysis of whey samples originating from dairy fermentation factories employing mesophilic undefined starter cultures was conducted to investigate the phage prevalence and diversity in the samples. Particular focus was placed on investigating the range of *Skunavirus* RBP genotypes and associated RBP-CWPS binding abilities present in these whey phageomes. These findings allowed for the refinement and expansion of the *Skunavirus* RBP grouping systems thereby vastly expanding our knowledge on lactococcal *Skunavirus* RBP diversity and CWPS specificity.

Chapter 4 of this thesis details the diversity of skunaviruses present in the same whey samples used for the phageome analysis described in Chapter 3 through culture-dependent phage screening, targeted isolation based on RBP specificity, and phage RBP-activated cell sorting (PhRACS). Through this investigation, the genomes of 26 newly isolated *Skunavirus* members were sequenced and analysed, with a specific focus placed on surveying genes encoding auxiliary carbohydrate binding module-containing proteins. Host range analysis of the skunaviruses revealed a clear correlation between the encoded RBP phylogroups and *cwps* genotypes of corresponding host strains. In addition, phageome-derived *Skunavirus* contigs were analysed to further determine the diversity of phages present in the dairy fermentation facilities.

Chapter 5 assesses the strain-level diversity of lactococcal strains originating from mesophilic undefined starter cultures. Culture-dependent approaches were used to assess the technological attributes, such as acidification, proteolytic activity, and phage sensitivity, of these isolates. Genomic analysis of sequenced lactococcal strains was focused on industrially relevant genomic elements such as *cwps* biosynthesis gene clusters, plasmids, and phage-defense systems. In addition, a genomic analysis of strains belonging to the emerging dairy-associated species *Lactococcus laudensis* was carried out.

In summation, both the bacterial composition of starter cultures as well as the phage diversity of corresponding whey samples were thoroughly investigated, generating comprehensive insights into the complex microbial community of these mesophilic fermentation systems. Understanding the intricate phage-host interactions occurring in the environment in which mesophilic fermentation systems are extensively applied will aid in efforts to development and maintain robust dairy starter culture systems for sustainable food production practices.

Chapter 1. General Introduction

Part of this review has been published in:

White K, Yu JH, Eraclio G, Dal Bello F, Nauta A, Mahony J, van Sinderen D. 2022. Bacteriophage-host interactions as a platform to establish the role of phages in modulating the microbial composition of fermented foods. *Microbiome Res Rep* **1**:3.

SUMMARY

Mesophilic dairy fermentation relies on the activity of robust starter cultures, which are commonly comprised of lactic acid bacteria such as *Lactococcus lactis* and *Lactococcus cremoris*. Bacteriophage infection represents a persistent threat to dairy fermentations and may cause delayed or failed fermentations. To develop robust starter cultures, it is important to establish a clear understanding of how phages interact with and modulate the composition of these complex microbial communities. Both culture-dependent and -independent methods have been instrumental in defining lactococcal phage-host interactions. This knowledge needs to be integrated, interpreted and expanded in order to obtain a full understanding of the overall complexity of such interactions pertinent to fermented foods through a combination of culturomics, metagenomics, and phageomics. With such knowledge it is believed that factory-specific detection and monitoring systems may be developed to ensure robust and reliable fermentation practices. In this review we explore and discuss current knowledge surrounding phage-host interactions between *L. lactis* and *L. cremoris* and their corresponding phages with a particular focus on skunaviruses, the impact of virulent phages on the microbial composition of starter cultures, host-encoded phage defense systems, and methods to define bacterial and phage populations within complex food systems.

INTRODUCTION

Lactic acid bacteria (LAB) have been employed in dairy fermentations for thousands of years. Some of the most commonly applied dairy LAB include particular members of the genera *Lactococcus, Leuconostoc, Pediococcus,* and *Streptococcus* (1). LAB produce lactic acid following catabolism of the principal (mammalian) milk-associated carbohydrate, i.e. lactose. As LAB continue to grow and metabolize, pH of their medium will decrease, thereby causing gradual coagulation of caseins (milk protein), which ultimately results in the generation of curds and whey (2). The curds are separated from the whey and processed using approaches that are specific to the desired dairy fermentation product (2). While the original dairy fermentations would have occurred spontaneously, over the past century dairy fermentations have been much more controlled and at the same time scaled up to the current multi-billion dollar/euro industry (3, 4). Spontaneous fermentations have all but been replaced by the use of exogenously added starter cultures, nonetheless, traditional and artisanal production systems, commonly relying on the activity of indigenous microbiota, are still used in the production of certain dairy products (5).

Starter cultures can be classified as undefined (also known as mixed strain starters or MSS) or defined (also known as defined strain starters or DSS) systems (4). The abundance, composition and number of strains present in undefined starter cultures are unknown (6), whereas defined starter cultures typically consist of a small number of strains (typically two to six) that are consciously combined to achieve products with specific physical and organoleptic properties (6). Although the exact composition of undefined

starter cultures is unknown, the particular species making up these cultures is typically determined by the fermentation temperature, requiring the use of either thermophilic or mesophilic starter bacteria (Table 1). Thermophilic fermentations (~37 to 45 °C) typically include *Streptococcus thermophilus* strains and may incorporate strains of particular *Lactobacillus* species (7). The lower temperature or mesophilic fermentations (~20 to 32 °C) instead rely generally on *Lactococcus lactis* and *Lactococcus cremoris* strains (occasionally incorporating strains of *Leuconostoc* and/or *Lacticaseibacillus* species). Significant research aimed at characterising the genetic and technological attributes of the dairy-adapted lactococcal and streptococcal species has been undertaken (8–17). Most notably perhaps are the research efforts aimed at solving the persistent problem of bacteriophages that continue to plague the dairy fermentation industry.

Bacteriophages or phages are ubiquitous and co-inhabit any ecological niche where bacteria exist (18). Therefore, it is perhaps unsurprising that they are an ever-present nuisance in a non-sterile dairy fermentation environment employing deliberate high-density, yet low-diversity (lactic acid) bacterial cultivations (19). In mesophilic fermentations, lactococcal phages are the most common cause of failed or delayed fermentations. As a result, lactococcal phages and their interactions with their hosts have been the focus of decades of research. For this reason, phage-host interactions between the lactococcal species *L. lactis* and *L. cremoris* and their associated phages will be the main focus of this review.

Table 1. Typical composition and applications of different starter cultures.

Starter Type	Typical composition	Applications	References
Artisanal	Undefined composition of mesophilic or thermophilic bacteria, occasionally fungi	Regional products e.g., fermented meats, vegetable fermentation, Caciocavallo, Pecorino, Vastedda cheeses	(5, 20, 21)
Mixed strain starter (MSS)	Undefined composition of mesophilic or thermophilic bacteria	Fermented products e.g., Gouda, Edam cheeses	(4, 22–24)
Defined strain starter (DSS)	Defined composition of typically, mesophilic bacteria e.g., *Lactococcus lactis or cremoris*	Cheddar style cheeses	(25, 26)
Mesophilic	*L. lactis, L. cremoris, Leuconostoc, Lacticaseibacillus casei, L. rhamnosus*	Cheddar, Gouda, Edam, Camembert, etc.	(7, 22, 23, 26, 27)
Thermophilic	*Streptococcus thermophilus, Lactobacillus helveticus, Lb. delbrueckii* subsp. *bulgaricus*	Grana, Pecorino, Mozzarella, Emmentaler, Sbrinz, Gruyère, yoghurt, etc.	(7, 26, 28, 29)

ROLE OF *LACTOCOCCUS* SPECIES IN DAIRY STARTER CULTURES

Currently, 26 bacterial species (according to NCBI taxonomy browser as of September 2024) are considered to belong to the genus Lactococcus, represented by Gram positive ovococcoid bacteria. Four of these species may be considered emerging species as the taxonomic classification of these species have not yet been validate according to the International Code of Nomenclature of Prokaryotes (8). Just five of these species have been isolated from milk: *L. cremoris, L. lactis*, *Lactococcus laudensis, Lactococcus raffinolactis,* and *Lactococcus hircilactis*. The biotechnological functionalities of the *L. cremoris* and *L. lactis* species in dairy fermentations have been extensively investigated in recent decades, while information regarding the emerging species is (understandably) sparse by comparison. However, there

is sufficient evidence to conclude that certain members of these three latter species have evolved to grow in milk. For example, an *L. raffinolactis* strain was isolated from raw milk (30), while another strain was retrieved from a dairy starter culture and was found to encompass homologues of *lac* (required for lactose utilisation) and *opp* (allowing oligopeptide uptake) operons (30). *L. hirclilactis* and *L. laudensis* strains have also been assessed for their ability to grow in milk to assess the role they may play in a diary fermentation system (31). Currently, very few genome sequences of members of *L. raffinolactis*, *L. hirclilactis* and *L. laudensis* are publicly available. Nonetheless, we may expect that, as more strains of these species will be isolated and their genomes sequenced, we will gain a better understanding of their genetic and metabolic diversity. As of yet no phages have been reported for strains of these three species—likely a reflection of their low abundance and limited research surrounding them, rather than an absence of infecting phages.

L. lactis and *L. cremoris* are the most extensively applied in industrial mesophilic dairy fermentations, likely due to their generally recognised as safe (GRAS) status (32). *L. cremoris* had originally been classified as a subspecies of *L. lactis*, however the two *L. lactis* subspecies, i.e. subsp. *lactis* and *cremoris*, were recently elevated to species (and designated *L. lactis* and *L. cremoris*, respectively) level following multi-locus sequence comparisons of the 16S rRNA, *rpoB, recA,* and *pheS* genes (33). Their primary role in dairy fermentation is to acidify milk through the production of lactic acid and to contribute to the product flavour profile through the action of proteases, peptidases, lipases and other, sometimes strain specific, metabolic activities. As an example of the latter, the citrate-fermenting *L. lactis* biovar *diacetylactis*

strains contributes to the flavour and texture of certain dairy products through the production of carbon dioxide, diacetyl, and acetoin (14, 34). Despite their global application in starter cultures, *L. lactis* and *L. cremoris* are believed to have originated from a plant-based environmental source (16, 35). The adaptation of these species to the dairy niche seems to have resulted in significant genome redundancy and decay (36). These genetic adjustments seem to have been facilitated by various mobile elements such as prophages, insertion elements/transposons and plasmids (16, 37, 38). *L. lactis* and *L. cremoris* chromosomes (excluding plasmids) typically range in size from ~2.2 to 2.5 Mb. Dairy *L. lactis* and *L. cremoris* strains have also been found to harbour up to twelve plasmids, varying widely in size from 0.87 Kbp to 193 Kbp (10, 39). Plasmids may be transferred between strains through conjugation or (co-)mobilisation and in certain cases transduction (11, 40). A lactococcal strain's plasmid complement typically exceeds 150 kb and encompasses genes/operons required for technologically relevant functions such as lactose metabolism, casein degradation, citrate utilisation, bacteriocin production, exopolysaccharide (EPS) production, conjugation, and peptide and amino acid uptake (16, 41, 42). These plasmids have also been found to encode a number of phage defense systems to combat the pervasive threat of lactococcal phages such as restriction modification (RM) and abortive infection (Abi) systems (11, 42–44). However, before a phage can infect a lactococcal cell it has to recognize and bind to its host-encode cell surface receptor.

Cell wall polysaccharide structures

Phage infection begins with the recognition and binding of a phage encoded anti-receptor to the surface-exposed host receptor and in the case of many lactococcal phages this is constituted by the cell wall polysaccharide (CWPS) (45–49). The *cwps* gene cluster, which is responsible for CWPS biosynthesis, of given lactococcal strain typically encompass a DNA region of 25-30 kbp (46, 48, 50, 51). The ever-increasing number of available lactococcal genome sequences has revealed the genetic diversity of the *cwps* clusters, while biochemical analysis of the CWPS from selected strains has unveiled the associated variability of their chemical structures (46, 48, 50, 51). The *cwps* gene cluster is responsible for the biosynthesis of the two components that typically constitute the CWPS: (i) a peptidoglycan-embedded rhamnan (whose biosynthetic machinery is encoded by the 5' portion of the *cwps* cluster) and (ii) a surface-exposed polysaccharide pellicle (PSP) or an oligosaccharide side chain (the biosynthesis of which is performed by enzymes that are encoded by the 3' portion of the *cwps* cluster). The genetic diversity of this gene cluster among lactococcal strains, particularly in the 3' region corresponding to distinct glycosyltransferase encoding-genes, is mainly responsible for the biochemical diversity observed in the PSP/oligosaccharide side chain of the CWPS between strains. Based on the genetic differences in the genetic composition of the *cwps* gene cluster (46, 48, 50, 51) Lactococcal strains are grouped into four *cwps* genotypes designated by type A through to D, with C types further subdivided into eleven subtypes, named C_1 through to C_{11} (Figure 1) (48, 50, 52). A recent comparative analysis encompassing over 400 lactococcal sequences (including an industrial strain collection) also

indicates the presence of several additional *cwps* types (A-H) and subtypes, indicating the highly diverse nature of this gene cluster among lactococcal strains (53). Furthermore, functional analysis of individual genes in several *cwps* clusters has elucidated the biochemical pathway by which the CWPS is constructed. Based on this knowledge it is now possible to link the gene content of a given lactococcal *cwps* cluster to certain structural features of its corresponding CWPS, the presence of an oligosaccharide or polymeric (PSP) side chain and the number of monosaccharides that make up the side chain (subunit), and the presence of chemical modifications of the component monosaccharides (50). Furthermore, since the CWPS acts as the host-encoded receptor for many lactococcal phages, mutations in *cwps* genes involved in the biosynthesis of a given CWPS have been shown to cause alterations in its structure and in turn affect the sensitivity of such a *cwps* mutant strain to phage infection (45, 54, 55).

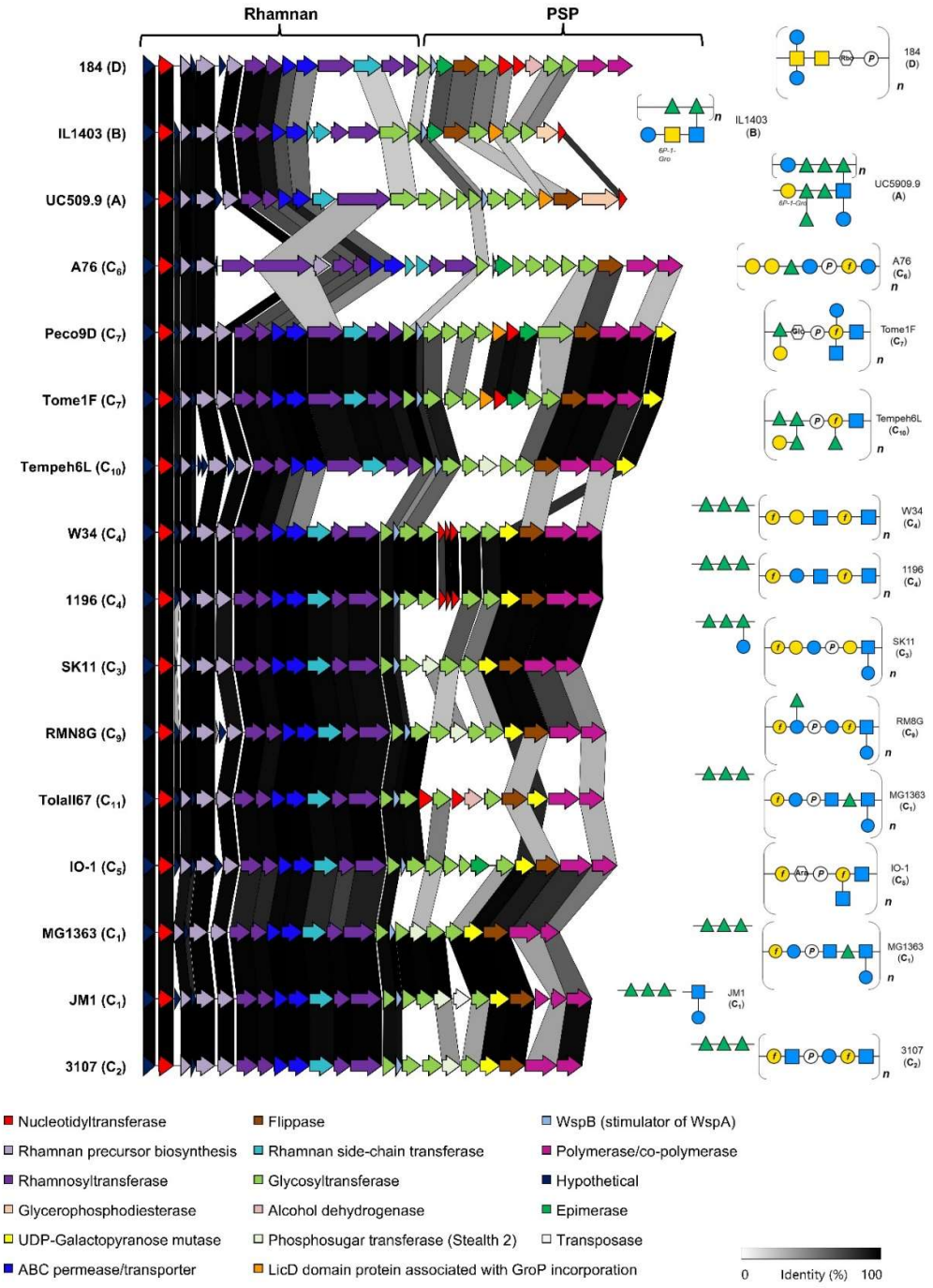


Figure 1. Schematic representation of the organisation of the gene cluster responsible for the biosynthesis of CWPS-types (A-D and C-subtypes C_1-C_7 & C_8-C_{10}) and their corresponding resolved chemical structures. Arrows represent open reading frames which are coloured according to their proposed function (indicated in the figure). Schematics depicting the biochemical structures of the Rhamnan (where known), and PSP/oligosaccharide side chains are shown on the right-hand side of the figure. Symbols Nomenclature for Glycans (SNFG) are used to represent

individual monosaccharides, linkages between monosaccharides can be found in Mahony *et al.* and Parlindungan *et al.* (50, 52).

LACTOCOCCAL BACTERIOPHAGES

All lactococcal phages belong the *Caudoviricetes* class (tailed phages) and possess double-stranded DNA (19). Lactococcal phages are classified into ten distinct taxonomic groups (Figure 2): *Skunavirus* (previously 936), P335, *Ceduovirus* (previously c2), *Whiteheadvirus* (previously 1358), *Questintvirus* (previously Q54), *Teubervirus* (previously P087), *Fremauxvirus* (previously 1706), *Audreyjarvisvirus* (previously 949), P034, and *Chopinvirus* (previously KSY1) (56, 57). Most lactococcal phages are siphoviruses possessing long contractile tails, however P034 and *Chopinvirus* members are podoviruses which have short, non-contractile tails (57). It is important to note that use of siphovirus (previously *Siphoviridae*) and podovirus (previously *Podoviridae*) as taxonomic names has recently been abolished, however they may still be useful as phage classifications based on morphological features (58). Members of the *Skunavirus* and *Ceduovirus* genera, and phages belonging to the P335 group are the most commonly encountered phages in industrial fermentation settings and have therefore received the most scientific attention. Both *Skunavirus* and *Ceduovirus* members are considered to be the most problematic due to their virulent lifestyle; P335 members, which can follow a virulent or temperate lifestyle, are commonly identified as prophages present on lactococcal chromosomes, yet are compared to the other two phage groups least frequently encountered as the cause of fermentation problems (5, 59).

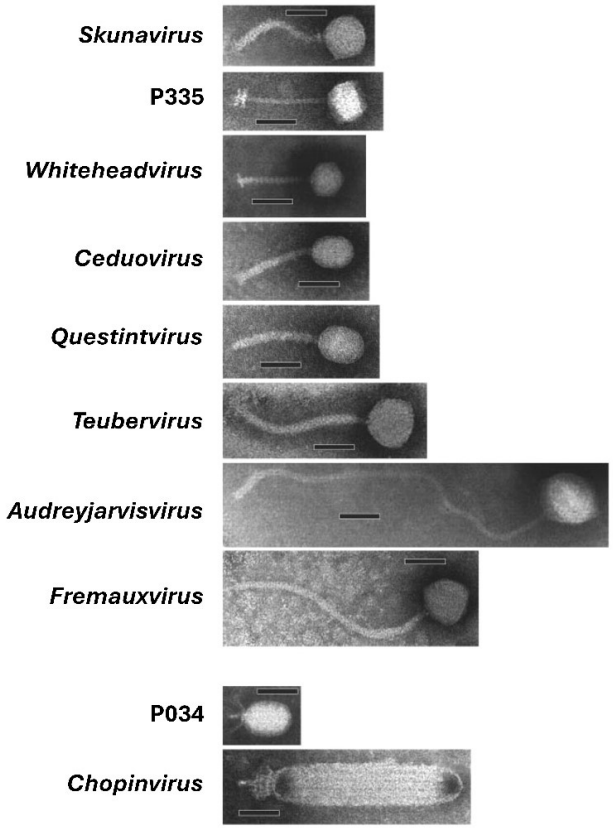


Figure 2. Ten taxonomic groups of lactococcal phages. Electron micrographs originally published in a review by Deveau, *et al* (58).

The overall genomic architecture of skunaviruses and ceduoviruses are highly conserved within their respective phage species (Figure 3) (60, 61). *Ceduovirus* genomes are typically ~22 kb in size and consist of an early- and late-expressed genetic region. Ceduoviruses fall into one two subgroups (c2- or bIL67-type phages) based on structural genes believed to be involved in host-range determination of these phages (60, 62). Ceduoviruses are less well understood compared to skunaviruses and P335 phages, in relation to gene function and phage-host interactions.

Skunavirus genomes are larger than those of ceduoviruses, and range in size from ~26-32 kB, comprising of genetic modules corresponding to tail morphogenesis, DNA packaging, and lysis (late-expressed region) and DNA replication (early- and middle-expressed regions) (51). Twenty-nine genes are reported to comprise the core genome of skunaviruses (51, 63). The majority (twenty genes) of these core genes are present in the late-expressed genetic module, while the remaining (nine) genes are encoded across the middle- and early- expressed modules (51). Although the genetic architecture of *Skunavirus* genomes is conserved, there are certain genetic regions that have been shown to exhibit a high degree of genetic diversity (64). For example, within the late-expressed region (tail morphogenesis) the neck passage structure (*nps*) and tail genes (particularly the tape measure protein and RBP) are hotspots for genetic diversity which has been correlated with higher recombination or mutation rates (64).

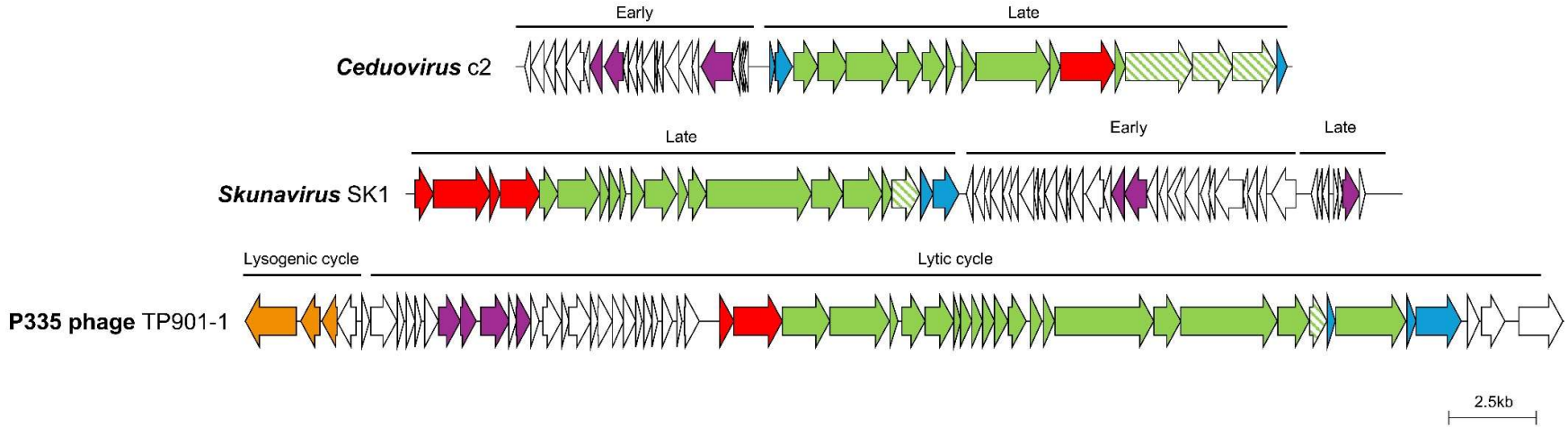


Figure 3. Schematic representation of the genomic architecture of most commonly isolated lactococcal phages: *Ceduovirus, Skunavirus,* and P335 phages. Each arrow represents an open reading frame and are coloured based on predicted function: purple for replication, blue for lysis, green for structural, red for packaging, and orange for lysogeny. The genes known or believed to be responsible for host recognition are green stripped.

P335 phage genomes range in size from ~32-54 Kbps and in contrast to skunaviruses, are very heterogenous which is exemplified by the lack of a defined core genome, i.e. no single gene is conserved among all known members of P335 phages (65, 66). The higher genetic diversity exhibited by P335 phages is likely as a result of their temperate ancestry which may have facilitated recombination between both temperate and virulent phages (57). Lactococcal genomes typically harbour one or more prophage associated regions (67).

Lactococcal phage-host interactions

To understand how phages influence the overall microbial community structure in fermented foods, it is important to consider the basis of phage-host interactions occurring in these bacterial populations. As mentioned above, phage infection commences with the recognition of, and adsorption to, a receptor on the host cell surface, often involving an initial reversible binding followed by irreversible binding and associated commitment to infection (68). Binding is facilitated by a phage-encoded anti-receptor, which typically comprises one or more receptor binding protein(s) (RBP) located at the distal end of the phage tail (often supported by auxiliary carbohydrate binding proteins) and the host-encoded, cell surface-exposed receptor (69, 70). The host binding machinery of lactococcal phages, located at the distal part of the phage tail, is encoded within the tail morphogenesis genetic module.

Considerable research has been undertaken to understand phage-LAB host interactions, which has rendered them a paradigm for a diverse range of microorganisms, particularly Gram-positive bacteria that are infected by tailed

phages. Despite the genetic diversity exhibited by siphophages, even of those infecting the same species, the genome architecture and functional synteny of the module responsible for tail morphogenesis is well conserved, incorporating genes encoding the following functions (in a 5' to 3' direction): the tail tape measure protein (TMP), the distal tail protein (Dit), and the tail-associated lysin (Tal). This region is typically followed by genes encoding the baseplate proteins, including the RBP, with auxiliary binding proteins in many cases (Figure 4). The RBP-encoding genes of many LAB phages were initially identified through comparative genome analysis (71–73). For example, the genomic region encoding the tail morphogenesis region of *Skunavirus* sk1 was first predicted by Chandry *et al.* due to size and isoelectric similarities in the genes and corresponding proteins encoded within this region to that of the known lambda phage tail region (73). Since then, molecular interactions between lactococcal phages and their cognate host have remained a focus of intense research scrutiny (60, 74–76).

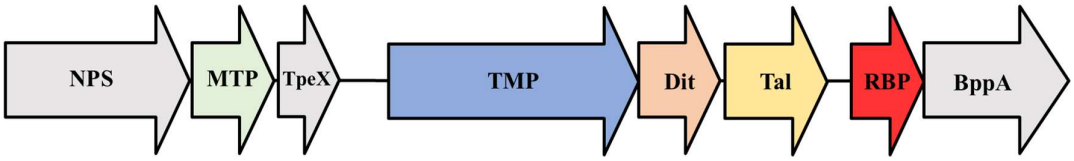


Figure 4. Schematic representation of genes commonly present in the tail morphogenesis region of lambda-like *Siphoviridae* recognising saccharidic receptors. The late-expressed genes that are commonly shared by these phages and make up the tail include the: major tail protein (MTP; green), tape measure protein (TMP; blue), distal tail protein (Dit; orange), tail-associated lysin (Tal; yellow), and RBP (red). Additional carbohydrate binding modules (CBMs) found in auxiliary binding proteins (grey) have been identified and characterised in certain phages, such as the: neck passage structure (NPS) and major tail extension protein (TpeX). Certain P335 phages also have an additional CBM (not found in other lactococcal phages) located in an auxiliary binding protein known as the accessory baseplate protein (BppA) (77, 78). It is noteworthy that in some cases (e.g. certain P335 lactococcal phages), the NPS-encoding gene is located downstream of the RBP-encoding gene.

As already mentioned above, most lactococcal phages recognise a saccharidic receptor known as the CWPS (47, 48, 50) (Table 2). Dupont and colleagues identified the role of the *cwps* gene cluster in lactococcal phage adsorption by means of random insertional mutagenesis of two *Lactococcus* strains (*L. lactis* IL1403 and Wg2) (55). Through this approach, bacteriophage insensitive mutant (BIM) derivatives exhibiting reduced phage adsorption capabilities were isolated. Of the 10 distinct groups of lactococcal phages, *Skunavirus*, P335, *Whiteheadvirus*, *Audreyjarvisvirus*, *Teubervirus*, and *Fremauxvirus* phages have been demonstrated to utilize saccharidic host receptors (Table 2) (47–49, 55, 77–81). Mediated by their RBP lactococcal phages specifically bind to the PSP/oligosaccharide side chain component of the CWPS, which is located on the host cell surface.

Table 2. Defined and putative phage receptors of lactococcal phages.

Phage Group	Representative phage(s)	Receptor type	Host receptor	Reference
Skunavirus	bIL170	Saccharidic	CWPS	(56)
P335	Tuc2009		CWPS and/or EPS	(78, 83)
Whiteheadvirus	1358		CWPS	(50)
Audreyjarvisvirus	949		CWPS	(48)
Teubervirus	P087		CWPS	(81)
Fremauxvirus	CHPC971		CWPS	(79)
Ceduovirus	c2	Proteinaceous	PIP	(84, 85)
	bIL67		YjaE	(86)
Questintvirus	Q54	Unknown		
P034	P034	Unknown		
Chopinvirus	KSY1	Unknown		

Ceduoviruses are the only lactococcal phage members known to recognise and bind irreversibly to a proteinaceous receptor (84). Relative to other lactococcal phages, there is limited information regarding the exact phage-encoded protein(s) responsible for host binding among members of the *Ceduovirus* group. While overall there is limited sequence divergence among

Ceduoviruses, a cluster of three genes displays reduced sequence similarity among members of this group. This cluster, which contains three late-expressed genes: *I14, I15,* and *I16* in phage c2 and its equivalents ORF34, ORF35, and ORF36 in bIL67, has been suggested to be responsible for host recognition in *Ceduovirus* phages c2 and bIL67 (72). This three-gene region is proposed to encode structural proteins responsible for binding to the host Pip or YjaE, however, the exact function of each gene in this process is yet to be elucidated (60, 72). The non-LAB *Bacillus subtilis*-infecting phage SPP1 also uses a proteinaceous receptor for infection and has been thoroughly studied. SPP1, like *Ceduovirus* phages, first binds reversibly to a saccharidic receptor and then irreversibly to a cell surface-located proteinaceous receptor, YueB, which bears similarity to the lactococcal PIP (86, 87). The distal tail complex of SPP1 is well described and is structurally similar to the tail of lactococcal phages. Due to these similarities, SPP1 has become a model for tailed phages that employ proteinaceous receptors (84).

Although much research has been dedicated in recent years to define phage-host interactions between *Siphoviridae* members and their Gram-positive hosts, in particular LAB, additional insights are required to fully elucidate these interactions at the molecular level. Also, due to the conserved nature of the tail structure of many Gram-positive host-infecting *Siphoviridae* phages, knowledge gained from understanding phage-host interactions in one group of bacteria (such as LAB) may be superimposed to better understand phage-host interactions involving other bacterial groups. As skunaviruses are one of the most problematic and isolated phages in mesophilic fermentation

systems the current body of knowledge regarding *Skunavirus* phage-host interactions will be examined in the ensuing section.

SKUNAVIRUS HOST RECOGNITION

The RBP of Skunaviruses was first identified in lactococcal phages sk1 and bIL170 using an *in-silico* approach, and functionally assigned following the isolation of recombinant phages encoding chimeric RBPs (71). The 3-dimensional structure of the RBP of lactococcal phage p2 was shown to represent a homotrimer of three domains: the shoulders, neck, and head (88). The first 135 N-terminal residues of an RBP typically correspond to both the shoulder and neck domains, and the C-terminal end (remaining residues) encodes the head domain which encompasses the receptor binding activity (89). Structurally, it was determined that there are five different “RBP-head” groups: p2, p2-like, bIL70, Tuc2009-like, and T4 gp12 group (Figure 5) (70). Although, p2 and p2-like RBP-head domains share significant structural similarity, all p2 group members possess four essential amino acid residues (Trp144, His232, Asp234, and Arg256) within the sugar binding pocket previously identified in the phage p2 RBP (88, 90). Conversely, not all of these binding site residues are conserved among p2-like RBP-head group members thereby resulting in different binding specificities than the p2 group members (70). The RBP head structure of *Skunavirus* bIL70 was previously determined via X-ray crystallography (91) and the head domain of bIL170 RBP-head group members (including phage bIL170) was found to share a similar yet distinct folding to phage p2 RBP head, has a distinct fold pattern (70). The remaining

two RBP head groups, Tuc2009-like and T4 gp12 differ considerably to that of the phage p2 RBP head. Tuc2009-like and T4 gp12 group RBP-head domains share significant structural similarity with those of the P335 phage Tuc2009 and the RBP (protein gp12) of *E. coli* phage T4, respectively.

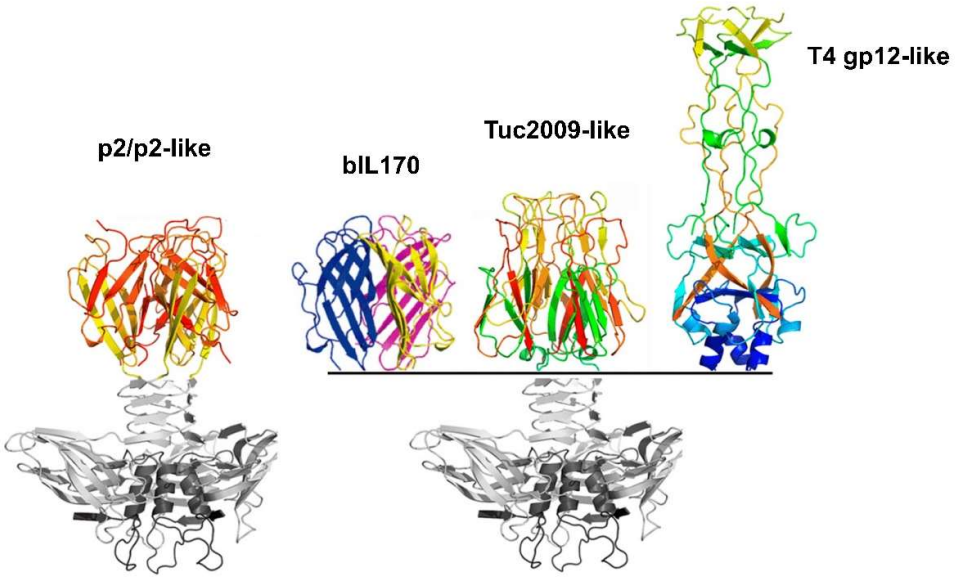


Figure 5. Structural representations of the four types of *Skunavirus* RBP head domain groups. Structures originally published by Hayes, *et al*. (71). The shoulder and neck domains are coloured grey, and the head domains are in various colours.

In addition, comparative analysis of C-terminal amino acid sequences from various RBPs facilitated a phylogenetic grouping revealing five RBP phylogroups that correlate with the specific *cwps* genotype of their corresponding host(s) as depicted in Figure 6 (46, 51, 70). Group I RBPs possess p2 head domains and recognise hosts with A C-type *cwps*, group II RBPs contain bIL170 heads and bind to B-type *cwps*, group III RBPs possess T4 gp12 RBP heads can infect both B and C-type strains, group IV RBPs possess p2-like head domains though the *cwps* type of their cognate host is

currently unknown, while group V RBPs have a Tuc2009-like head domains and can infect A-type *cwps* strains. Nonetheless, it is evident that a thorough refinement and expansion of the current classification system is needed as there are currently 251 publicly available RBPs in the GenBank database (as of September 2024) compared to the 91 sequences used in the comparative analysis described above (Figure 6). The need for an updated RBP classification system was also highlighted in a recent review by Romero, *et al.* following analysis of over 200 *Skunavirus* RBPs, many of which originating from an (unpublished) industrial phage collection (53).

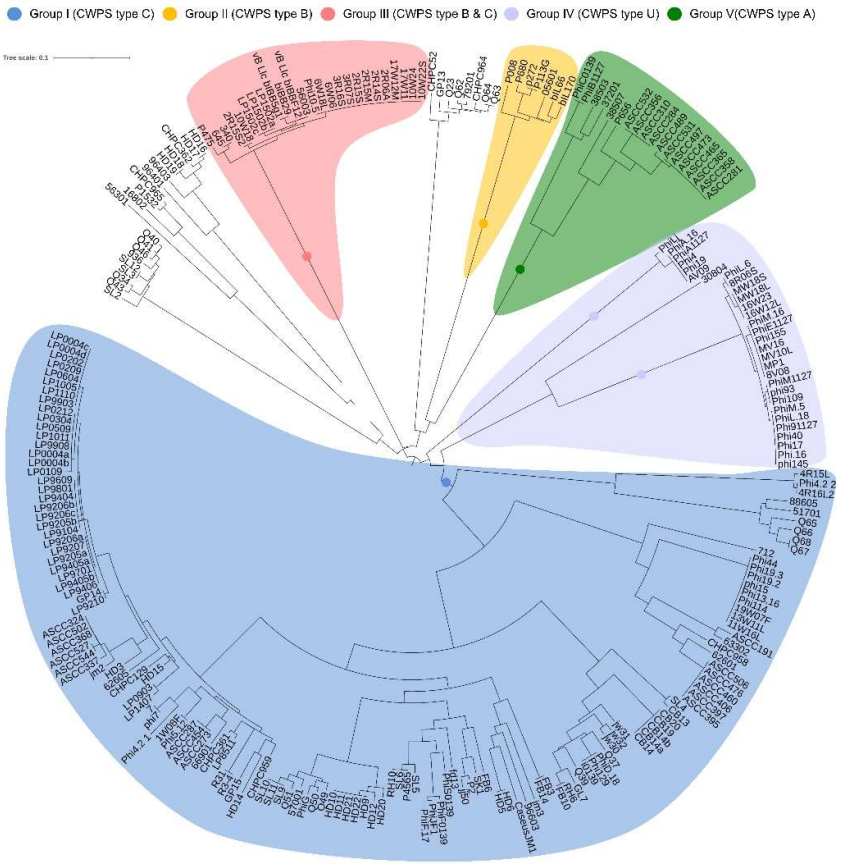


Figure 6. Phylogenetic tree of the C-terminal region of all 251 currently publicly available *Skunavirus* RBPs in the GenBank database.

In addition to the receptor binding ability of the RBP C-terminal or head domain, a lactococcal *Skunavirus* virion may possess auxiliary CBM decorations associated with the Dit, major tail protein (MTP), and neck passage structure (NPS) that may all contribute to host cell binding. These CBMs facilitate specific host binding and may indeed exhibit the same host specificity of its corresponding RBP albeit, in some cases, with an apparently reduced affinity relative to its RBP (70). In certain Skunaviruses (and P335 phages), the NPS forms a collar-whisker complex attached to the phage portal and contains a CBM involved in, but not essential to, host binding (92). In addition, a TpeX has been identified in certain Skunaviruses and results in the presence of additional decorations along the tail. Through fluorescence binding assays, the CBMs of the NPS and TpeX have been determined to exhibit the same host specificity as the RBP (though with inferior affinity when compared to that of the RBP) and bind favourably towards the ends of the cell where cell division occurs (70). Most lactococcal phages encode a Dit. In skunaviruses and P335 phages, it was determined that these Dits can either be classified as ‘classical’ or 'evolved’ (93–95). Evolved Dits are longer and contain an apparent domain insertion of an additional CBM which some cases has been found to exhibit the same host specificity as the RBP (94). More recently, a study found that skunaviruses encoding evolved Dits were still capable of infecting lactococcal strains following transformation with a plasmid encoding genes required for the biosynthesis of surface-associated exopolysaccharides (EPS), while infection by phages encoding classical Dits was inhibited (96). This thereby indicates that certain auxiliary CBMs, such as those encoded within evolved Dits, may not only complement RBP binding to

a host, but may also aid in expanding or enhancing it. Lactococcal phages tend to possess a variety of CBMs (in addition to the RBP) in somewhat conserved combinations, and very rarely contain none or all possible auxiliary CBMs (70).

ROLE OF (VIRULENT) PHAGES IN MODULATING MICROBIAL COMPOSITION OF COMPLEX FOOD SYSTEMS

In industrial settings virulent phages represent a double-edged sword, depending on the environment and their host. For example, phages have been used as a non-chemical biocontrol tool to eradicate contaminating pathogenic species such as *Listeria monocytogenes* from many food products (97). However, in the overall context of fermented food production, the presence of virulent phages that infect starter cultures is undesirable as they cause delayed or even failed fermentations with significant associated economic losses. In dairy fermentations, the impact of virulent phages on the fermentation process differs depending on the starter culture format, i.e. defined or undefined starter cultures, as well as the scale of fermentation, i.e. industrial scale versus artisanal production systems. Undefined starter culture systems are typically considered more phage-robust than defined starter cultures due to the culture complexity and strain diversity. For example, Erkus and colleagues demonstrated that phage-sensitive strains are replaced by phage-insensitive strains of the same lineage, allowing for continued fermentation when using the complex/undefined Gouda cheese starter culture Ur (98). The composition of the original Ur starter culture was dissected using

culture-dependent and independent methods, allowing for community monitoring using genetic lineage-specific qPCR. The Ur culture was shown to maintain the relative composition of the different lineages, despite phage (originating from the original starter culture) pressure on individual strains. This was determined to be due to heterogeneity of the culture and, more specifically, variations in phage sensitivity of strains within and between lineages (98).

Phage attack may nonetheless have a detrimental effect on both DSS and undefined starter cultures, although the severity of the impact and means to mitigate phage infection highly vary. In DSS cultures, the specific microbial composition, and phenotypic characteristics of each strain of the DSS culture is known. If one or more of the strains in such a DSS culture is partly or completely eliminated by virulent a phage, such strains have to be replaced by phage-unrelated strains or phage-insensitive derivatives (possessing the same technological characteristics). This process ensures successful fermentation and product consistency/quality. Conversely, in undefined starter cultures, if milk acidification continues as normal despite phage attack, the organoleptic properties and thus quality of the final product may be negatively impacted (99). Phages infecting NSLABs present in the starter culture do not tend to impact on acidification because these strains are typically utilised in fermentation specifically for the organoleptic properties they impart. In many cases, these inconsistencies are not observed until grading of the product occurs at which point flavour, aroma, eye formation and surface ripening are evaluated as appropriate to the specific product. Such product inconsistencies

are difficult to identify during production process and can be costly to food producers through product down-grading.

In addition to the negative impacts of phage predation on microbial communities during microbial fermentation processes, lytic (as well as lysogenic) phage also play an important role in the composition and evolution of starter cultures by driving the genetic diversity of bacterial strains. For example, simple blends composed of representative strains from different genetic lineages (with varying phage sensitivity profiles) were created from the undefined culture Ur (100). The relative abundance of the genetic lineages was monitored across sequential rounds of propagation in the absence or presence of phage pressure. Using this more defined version of a complex starter culture, the genetic lineages did not remain stable during sequential rounds of phage attack. However, phage-resistant variants eventually arose from phage-sensitive genetic lineages after which the cultures stabilized to the same relative composition as control blends without the presence of phage (100). This study demonstrates how phages are key contributors driving the diversity among bacterial strains.

Traditional and artisanal cheeses are produced employing starter cultures that typically consist of autochthonous bacteria already present in the fermentation materials (such as raw milk) or environment (e.g. fermentation vats). These cheeses are mostly produced using traditional production methods with fewer chemical and physical hurdles for phages to overcome, thereby allowing phage populations to emerge that are different when compared to those of modern fermentation facilities. For example, in Sicilian artisanal cheese facilities it was shown that 16 of the 18 phages isolated from

associated cheese whey and rennet samples belong to the *Audreyjarvisvirus* and *Teubervirus* lactococcal group phages, which are rarely isolated from whey samples obtained from modern, large-scale cheese factories (5). These phages isolated from artisanal fermentation systems were found to be much more heat sensitive when compared to other lactococcal phage groups more commonly isolated from large scale fermentations (such as Skunaviruses) and appear to thrive in traditional fermentation environments due to the lack of pasteurization (5). While phages are a driving force of bacterial evolution, they are also continuously adapting and evolving in response to their hosts when the latter acquire resistance. Phages may mediate transfer of genetic material via transduction (transfer of bacterial genetic material that has been packaged into the capsid of the phage) although this typically occurs at low frequencies. Temperate phages also have the potential to contribute to the evolution of a given bacterial species through an alternate form of transduction, whereby neighbouring genetic material to an induced prophage may be accidentally excised as well and transferred from one strain to another.

LACTOCOCCAL PHAGE DEFENSE ARSENAL

Primary defense: limiting phage adsorption

Since phages must recognise and bind to a specific host receptor, modification and/or mutation of the receptor typically results in a phage resistant phenotype (101). However, mutations to *cwps* gene cluster may negatively impact the growth of bacterial strains applied in the dairy fermentation setting. In addition to CWPS, some lactococcal strains can also

produce cell-associated exopolysaccharides (EPS). Lactococcal *eps* loci are typically plasmid-encoded (however they can be chromosomally associated in some cases (102)), and are highly diverse, and range in size from ~7 to 13 kb (103–106). EPS production usually imparts desired organoleptic properties, particularly texture, on the fermentation end-product. EPS biosynthesis has been associated with reduced phage sensitivity due to the EPS masking the host receptor of certain lactococcal phages, thereby reducing phage adsorption to the cell (107–109). Conversely, EPS has been found to be the phage receptor for certain P335 phages (82).

Prophage-encoded systems

Many phage defense systems in lactococcal strains such as restriction-modification (RM) and abortive infection (Abi)-like systems have been found to be plasmid-encoded. In contrast, superinfection exclusion systems (Sie) are typically prophage-encoded. Sie systems are membrane-associated proteins which prevent DNA injection (but not adsorption) by specific skunaviruses and the first lactococcal Sie system was identified in the temperate P335 phage Tuc2009 (110). Prophage sequences also facilitate superinfection immunity (Sii), where secondary or super-infection by a homologous temperate phage is blocked through expression of a repressor protein encoded by a resident prophage (111). In addition to Sie and Sii systems, (apparently defective) prophages have been found to encode orphan methyltransferases and occasionally Abi-like systems, thereby highlighting the positive role prophages may play within dairy fermentation cultures, by acting as reservoirs for genetic diversity and adding to a strain's repertoire of phage defense systems.

Plasmid-encoded systems

Abi-like systems are quite abundant in lactococcal strains and are typically plasmid-encoded (but may also be chromosomally or prophage-encoded (112)). Abi-like systems not only inhibit a crucial step of the phage cycle within the host cell (such as DNA replication, transcription, or translation) resulting in the death of the infected cell prior to completion of the phage lytic cycle thereby protecting the surrounding cells from infection. There has been a renewed interest recently in identifying and characterizing Abi systems in lactococcal strains since their initial discovery and characterisation (113, 114). There are currently 34 identified lactococcal Abi-like systems (11, 43, 115, 116). Most characterised Abi systems are single-gene systems, while some are two-component systems or three-component systems (115). The antiphage activity of Abi-like systems varies greatly in terms of activity against specific phages, the phage group(s) they target and the stage of phage replication they trigger/affect (43, 117). Most described Abi-like systems have been found to interfere with either replication or transcription. However, two recently described systems (Rhea and Kamadhenu) are predicted to interfere with DNA packaging or phage maturation rather than replication or transcription (118). Recently developed bioinformatic tools PADLOC and DefenseFinder that identify phage defense systems in bacterial genomes will likely aid efforts to identify novel Abi (and other phage defense) systems (119, 120).

Another abundant and typically plasmid-encoded phage defense system found in lactococcal strains are RM systems. These systems typically consist of a restriction endonuclease and a modification enzyme (i.e.

methyltransferase) (121). These enzymes will recognise and cleave foreign invading DNA, typically phage DNA, (restriction endonuclease) and protect the bacterial genome from the encoded system (modification enzymes). There are four Types of RM systems (I-IV) based on genetic structure, co-factor requirements, recognition sequence, and cleavage site (122). Prophage encoded RM proteins have also been found to protect a host against invading phage (123). Type I systems have three distinct structural genes encoding three subunits for restriction (R), modification (M), and specificity (S) (124). Type II RM systems have two components: a sequence specific endonuclease and a methyltransferase. Type III RM systems also have two distinct genes encoding subunits of a restriction complex that has both restriction and modification activity. Type IV RM systems have either one or two genes encoding proteins specific for restricting methylated DNA. Type I systems are the most common RM system in lactococcal plasmids (10).

It is important to note that phages have developed a number of counter defense measures against host-encoded phage defense systems. First, as a result of bacteria modifying their host receptors to invade phage infection, phages may modify or alter their RBPs (typically through mutations or genetic exchange). This highlights the constant coevolution of the *cwps* gene clusters and *rbp* genotypes exhibited through the ever increasing genetic and subsequent structural diversity of both host CWPSs and phage RBPs. Additionally, skunaviruses have adapted to RM systems by reducing the number of restriction sites present in a genome and by adding methyltransferase genes to their own genomes as protection (125, 126). Phage escape mutants are commonly isolated following interactions with Abi

systems by acquiring point mutation(s) or recombination with integrated prophage sequences (118). A schematic representation of bacteria encoded phage defense and phage counter defense mechanisms is shown in Figure 7.

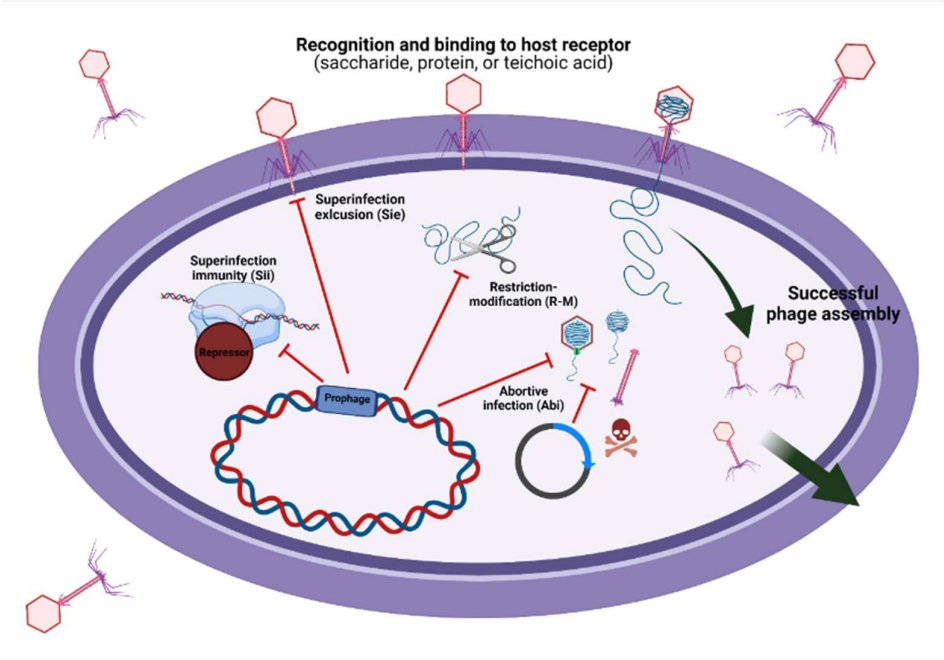


Figure 7. Schematic representation of commonly occurring phage-host interactions. After a phage recognises and binds to a specific host receptor, a number of plasmid-, chromosomal-, and/or prophage-derived anti-phage systems may impede successful phage proliferation, such as: Sii, Sie, RM, or Abi systems. Created with BioRender.com.

CURRENT APPROACHES TO CHARACTERISING COMPLEX FOOD SYSTEMS

Approaches to analysing undefined starter cultures

Undefined starter cultures are typically more phage robust than their defined counterparts. However, industry implemented remedies to delayed or failed fermentations as a result of phage attacks differ between the two culture systems. As mentioned above, phage-sensitive strains in DSS cultures can be

replaced by phage-unrelated strains or phage-insensitive derivates of the strain. Conversely starter culture rotation strategies, where cultures capable of producing a fermentation end-product with the same organoleptic properties (but intentionally a different phage sensitivity profile) are rotated. Starter culture rotation strategies may be more effective and informed with the knowledge of microbial diversity and interactions occurring in these complex fermentation systems. Therefore, employing effective monitoring and tracking systems of both the bacterial and phage populations within these settings is key.

Beyond the species or subspecies level of the major microbial components employed, very little is known regarding the microbial diversity and composition of undefined starter cultures (23). For successful starter rotation strategies, it is essential that all starter cultures in rotation impart the same organoleptic properties and quality of the final product. Therefore, it would be beneficial to define the functional and metabolic characteristics of the strains encompassing a given starter culture. Traditionally culture-based analyses have been used to evaluate the functional characteristics and associated roles of bacterial components of undefined starter cultures (27). This is a very useful and essential strategy to facilitate in-depth genomic and technological characterisation of individual strains present in undefined systems. However, culture-dependent methods are often limited by cultivation conditions and may not accurately reflect the exact composition of the culture. For example, a major limitation of culture-based methods is inability to isolate viable but non-culturable cells and low-abundance strains or species depending on the intensity of isolation efforts.

In recent years, there has been a surge in the application of culture-independent methods such as metagenomics and targeted sequencing approaches when studying fermented food systems (127, 128), particularly when attempting to characterise undefined starter cultures (24, 98, 128–132). Although metagenomic analysis of a given fermentation system yields an abundance of data and provides insights into the microbial diversity of the associated starter culture, ultimately metagenomics only provides a "snapshot" of the diversity at a particular moment in time. Currently, approaches to actively track and monitor the composition of starter cultures in a dynamic manner are established at species level. For example, the development of qPCR assays specific to certain species, subspecies, and/or genetic lineages has been a widely applied and useful strategy to both track and characterise targeted microbes in food systems (98, 133–139). However, most of these qPCR-based assays typically monitor only species/subspecies-level diversity shifts (133, 139–141). This is not particularly useful for undefined starter cultures that are dominated by a diverse array of strains belonging to the same species. Several studies have combined initial culture-based methods to define strain lineages present in an undefined starter culture, followed by either targeted amplicon sequencing (24) or specific qPCR assays (23, 98) to monitor and track these pre-determined lineages. Those approaches go beyond the species level to monitor microbial populations; however, these assays are culture-specific and are not broadly applicable (24, 98). Efforts to design widely-applicable culture-independent assays, which are able to go beyond the species/subspecies-level, would allow for the

characterisation and monitoring of strain-level microbial diversity in undefined starter cultures.

Approaches to assess phage diversity

In the modern food fermentation industry, owing to increasing consumer health and nutritional awareness of fermented foods, the establishment and implementation of a reliable and traceable manufacture system has been emphasized to achieve consistent, high-quality products. The crucial role of bacterial and/or fungal microbial communities in food fermentation processes has culminated in the generation of significant data pertaining to the microbiota of foods and food production environment and has been enhanced by recent developments in genome sequencing and meta-omics tools. In contrast, fewer studies pertaining to the food production environments and their phageome, which represents the overall phage community of a given sample or environment, have been published until recently (142–146).

To date, the presence of phages in fermentation facilities has typically been assessed through culture-dependent methods. These methods entail culture-dependent phage screening (e.g. through double-layer agar methods) or indirectly through milk acidification assays (147, 148). Other methods of assessing phage diversity in samples include transmission electron microscopy (TEM) (149) and epifluorescence microscopy (EFM) (150, 151), although this is a more typical approach to assess complex environmental ecosystems such as aquatic environments. As mentioned above, culture-dependent approaches are very informative, but ultimately rely on the ability to propagate phages present in a given sample. In addition, it is widely

understood that cultivation-based phage screening methods tend to identify the most dominant and culturable phages in a sample. This is particularly the case with skunaviruses due to their narrow host range (152, 153). Therefore, it is likely that the isolated phages represent a skewed portion of the overall phage population.

When metagenomic approaches became implemented for microbial compositional analysis of various environmental niches, phage-associated contigs began to be identified in metagenomic assemblies (127). However, many of these virome assemblies are considered to represent so-called viral dark matter as it does not correspond to any known viral sequence. However, to be able to evaluate the true phage diversity in a metagenome, more targeted extraction methods for viral particles and subsequent phageome assemblies were needed. In recent years, targeted nucleic acid extraction methods from enriched viral particle populations present in various environmental samples, such as dairy fermentation-associated niches, have been developed (146, 154, 155). This recent wave of virome sequencing and subsequent phageome analysis has and will continue to increase our understanding of phage prevalence, abundance, and dynamics throughout the fermentation process. For example, virome sequencing of kimchi (156) and cheese surfaces (143) has highlighted the viral diversity and its correlation with bacterial diversity. Notably, *Jung et al.* (2018) revealed that viral communities in kimchi have much more clear association with geographical origins than bacterial communities, facilitating an in-depth understanding of the fermented food ecosystem.

Another challenge associated with virome sequencing is extracting sufficient quantity and quality viral DNA from a sample in order to yield good quality viral metagenome assemblies. This challenge may be remedied through enrichment approaches however approaches such as filtration, ultracentrifugation, positive or negative selective virome capture, may cause incorrect information regarding the actual composition of the viral population in a sample (157). For example, ultracentrifugation through cesium chloride (CsCl) gradients, may result in a loss of phages that are outside the selected density gradient (due to differences in size and density) or that have reduced viral stability in the presence of CsCl (158). And positive selection approaches of viromes (such as those using probes in hybridization) are biased towards known virus sequences and therefore exclude uncharacterized viral sequences that may be present in a sample (157). To overcome these challenges, qPCR systems to identify and quantify a range of phage species of the sample pre- and post-enrichment could be applied to track the changing population landscape to complement sequencing strategies. This represents an opportunity to expand and deepen current understanding of the role, diversity and functionality of phages in food systems. To reduce the viral dark matter, not only an enrichment of viral sequences is required, but also the combinations with biological and molecular methods need to be improved.

Beyond the challenge of viral nucleic acid extractions and yields, accurate assembly of the phageome remains a subsequent challenge. Most currently available metagenome assembly pipelines are designed and optimised toward assembling bacterial genomes. This can result in the failed assembly of low coverage phage genomes or issues when attempting to

assemble phage genomes containing genomic repeats, typically leading to fragmented or incomplete assemblies (159, 160). Due to the modular and mosaic nature of most phage genomes (including lactococcal phage genomes) it is challenging to assemble complete and accurate phage genomes. Recently, a software tool called Phables has been developed, aimed at resolving complete phage genomes from fragmented viral assemblies (161). Viromes are initially assembled using typical assemblers, such as Spades (162), resulting in generation of a corresponding virome assembly graph. Phage-like unitigs (a continuous high-confidence sequence in an assembly graph without branches) from the assemblies are identified based on the presence of phage protein families commonly used when annotating viral sequences (also known as Prokaryotic Virus Remote Homologous Groups or PHROGs) (163). By modelling graphs of the phage components, using generated reads mapping data and flow decomposition, genomic paths and reads coverage values are obtained. These genomic paths are considered to be resolved, high-quality ‘phage-like contigs’. However, Phables often identifies certain bacterial plasmids (encoding proteins homologous to phage proteins) as phage-like components. Therefore, downstream filtering of bacterial contaminants from ‘phage-like contigs’ is typically needed if virome samples are mixed or contaminated with bacterial nucleic acids. Overall, as viral particle extraction methods, virome sequencing methods, and phageome assembly methods continue to improve so too will the quality of phageome analyses in a variety of niches including dairy fermentation facilities.

CONCLUSION AND INTRODUCTION TO THESIS

Phages have maintained a prominent role in modulating the microbial composition of fermented foods. The 'kill-the-winner' model of phage dynamics allows for the stabilisation of complex bacterial communities (164). This hypothesis states that an increase in a particular bacterial strain within a microbial community will coincide with an increase of phages that can infect that strain, thereby reducing its abundance and preventing a single strain from dominating the community. Phages, therefore, play an essential role in maintaining the diversity and stability of the complex microbial communities necessary for the production of fermented foods. Culture-based methods, and more recently, molecular and genomics-based methods have been instrumental in defining lactococcal phage-host interactions. There is a diverse range of globally applied fermentation practices that utilise LAB, as well as other microbes. The microbes used during fermentation, whether autochthonous or introduced through starter cultures, vary depending on the geography, environmental conditions, climate, fermentation practices, and raw materials used. In addition, the demand for fermented foods continues to increase, demonstrating the importance of generating deeper insights into the microbial interactions in these communities. With increasing demand for consistent, high-quality products there is increasing pressure to employ robust and reliable fermentation practices. In particular, there is increasing interest in plant-based dairy alternatives, and in order to gain a better understanding of the phage-host interactions occurring in these environments, phageome and metagenome analysis can be employed. Furthermore, there is an opportunity to expand current knowledge pertaining to the microbiota in fermented meat

and vegetable products and to determine their contributions to metabolite production as well as product safety and quality (165, 166).

With the expanding knowledge regarding metagenome and phageome information of fermented food systems, we are only now beginning to uncover the true complexity of these microbial communities. By combining both phageome and metagenomic analyses it will be possible to gain a better understanding of the ever-evolving phage-host interactions occurring in fermentation environments. Through the combination of cultivation-dependent and -independent approaches it is possible to achieve an in depth, holistic understanding of the genetic and functional diversity of microbial communities that are present in undefined starter cultures. Analysis of CBMs in phageome data, particularly those found in phage RBPs, can be used to predict potential hosts of the phages present. A link between *Skunavirus* RBPs and host CWPS has already been defined (and is in need of an expansion and refinement as more phages have been isolated and sequenced) (51, 53, 70, 81). With continued phage-host interaction studies of LAB this predictive strategy presents a paradigm for the microbial interactions of a range of Gram-positive bacteria and their infecting tailed phages.

Constant monitoring of the microbial community is essential to overcome the negative impacts of phages in fermentation and ensure the production of consistent, high-quality products. To date phageome and metagenomic sequencing in fermentation have provided just a snapshot of the diversity in specific communities. These tools need to be expanded and used to monitor shifts in populations over time and with changing environmental conditions. Metagenomics and phageomics should be used as a tool to

understand the phages (both virulent and temperate) and strains that are present in a specific fermentation factory. For instance, it may be possible to monitor the microbial community by assessing the presence of specific genes, such as those encoding phage-host receptors or strain/genetic-lineage specific genes. This can then be linked to the phageome, where problematic phages associated with a specific factory can also be monitored and detected. By utilising culturomics, metagenomics, and phageomics (and incorporating transcriptomics, proteomics and metabolomics) in combination, factory-specific detection and enumeration strategies can be developed and utilised. This will allow for a better understanding of the phage-host interactions occurring in these complex microbial communities and for more reliable and robust fermentation practices.

REFERENCES

1. Tamang JP, Watanabe K, Holzapfel WH. 2016. Diversity of microorganisms in global fermented foods and beverages. *Front Microbiol* **7**:377.
2. Fox PF, McSweeney PLH. 2017. Chapter 1 - Cheese: An Overview, p. 5-21. *In* McSweeney PLH, Fox PF, Cotter PD, Everett DW (ed), Cheese, Fourth edition. Academic Press, San Diego.
3. Galimberti A, Bruno A, Agostinetto G, Casiraghi M, Guzzetti L, Labra M. 2021. Fermented food products in the era of globalization: tradition meets biotechnology innovations. *Curr Opin Biotechnol* **70**:36-41.
4. McGrath S, Fitzgerald GF, Sinderen D van. 2004. Starter Cultures: Bacteriophages, p. 163-190. *In* Fox PF, McSweeney PLH, Cogan T, Guinee T (ed), Cheese: Chemistry, Physics and Microbiology, Third Edition. Academic Press, San Diego.
5. Mahony J, Moscarelli A, Kelleher P, Lugli GA, Ventura M, Settanni L, van Sinderen D. 2017. Phage biodiversity in artisanal cheese wheys reflects the complexity of the fermentation process. Viruses **9**:45.
6. Daly C, Fitzgerald GF, Davis R. 1996. Biotechnology of lactic acid bacteria with special reference to bacteriophage resistance. *Antonie Van Leeuwenhoek* **70**:99-110.
7. Johnson ME. 2013. Mesophilic and Thermophilic Cultures Used in Traditional Cheesemaking. *Microbiol Spectr* **1**:10.1128/microbiolspec.cm-0004-2012.

8. Mahony J, Bottacini F, van Sinderen D. 2023. Towards the diversification of lactococcal starter and non-starter species in mesophilic dairy culture systems. *Microb Biotechnol*. **16**:1745-1754.
9. Nomura M, Kobayashi M, Narita T, Kimoto-Nira H, Okamoto T. 2006. Phenotypic and molecular characterization of *Lactococcus lactis* from milk and plants. *J Appl Microbiol* **101**:396-405.
10. Kelleher P, Mahony J, Bottacini F, Lugli GA, Ventura M, van Sinderen D. 2019. The *Lactococcus lactis* pan-Plasmidome. *Front Microbiol* **10**:1-20.
11. Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D. 2014. The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. *FEMS Microbiol Rev* **38**:1066-1088.
12. Wels M, Siezen R, van Hijum S, Kelly WJ, Bachmann H. 2019. Comparative genome analysis of *Lactococcus lactis* indicates niche adaptation and resolves genotype/phenotype disparity. *Front Microbiol* **10**:4.
13. Kelleher P, Bottacini F, Mahony J, Kilcawley KN, van Sinderen D. 2017. Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation. *BMC Genomics* **18**:267.
14. Fusieger A, Carla M, Martins F, De Freitas R, Luís &, Nero A, Fernandes De Carvalho A. 2020. Technological properties of *Lactococcus lactis* subsp. *lactis* bv. *diacetylactis* obtained from dairy and non-dairy niches. *Braz J of Microbiol* **51**:313-321.

15. Torres Manno M, Zuljan F, Alarcón S, Esteban L, Blancato V, Espariz M, Magni C. 2018. Genetic and phenotypic features defining industrial relevant *Lactococcus lactis*, *L. cremoris* and *L. lactis* biovar. *diacetylactis* strains. *J Biotechnol* **282**:25-31.
16. Price CE, Zeyniyev A, Kuipers OP, Kok J. 2012. From meadows to milk to mucosa - adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. *FEMS Microbiol Rev* **36**:949-971.
17. Hols P, Hancy F, Fontaine L, Grossiord B, Prozzi D, Leblond-Bourget N, Decaris B, Bolotin A, Delorme C, Dusko Ehrlich S, Guédon E, Monnet V, Renault P, Kleerebezem M. 2005. New insights in the molecular biology and physiology of *Streptococcus thermophilus* revealed by comparative genomics. *FEMS Microbiol Rev* **29**:435-463.
18. Naureen Z, Dautaj A, Anpilogov K, Camilleri G, Dhuli K, Tanzi B, Maltese PE, Cristofoli F, Antoni L De, Beccari T, Dundar M, Bertelli M. 2020. Bacteriophages presence in nature and their role in the natural selection of bacterial populations. *Acta Biomed* **9**: e2020024.
19. Pujato SA, Quiberoni A, Mercanti DJ. 2019. Bacteriophages on dairy foods. *J Appl Microbiol* **126**:14-30.
20. Randazzo CL, Torriani S, Akkermans ADL, De Vos WM, Vaughan EE. 2002. Diversity, dynamics, and activity of bacterial communities during production of an artisanal sicilian cheese as evaluated by 16S rRNA analysis. *Appl Environ Microbiol* **68**:1882-1892.

21. Randazzo CL, Vaughan EE, Caggia C. 2006. Artisanal and experimental Pecorino Siciliano cheese: Microbial dynamics during manufacture assessed by culturing and PCR-DGGE analyses. *Int J Food Microbiol* **109**:1-8.

22. White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* **90**:e02152-23.

23. Smid EJ, Erkus O, Spus M, Wolkers-Rooijackers JCM, Alexeeva S, Kleerebezem M. 2014. Functional implications of the microbial community structure of undefined mesophilic starter cultures. *Microb Cell Fact* **13**:1-9.

24. Frantzen CA, Kleppen HP, Holo H. 2018. *Lactococcus lactis* Diversity in Undefined Mixed Dairy Starter Cultures as Revealed by Comparative Genome Analyses and Targeted Amplicon Sequencing of epsD. *Appl Environ Microbiol* **84**:e02199-17.

25. Timmons P, Hurley M, Drinan F, Daly C, Cogan TM. 1988. Development and use of a defined strain starter system for Cheddar cheese. *Int J Dairy Techno*l **41**:49-53.

26. Farkye NY. 2014. CHEESE | Microbiology of Cheesemaking and Maturation, p. 395-401. *In* Batt CA, Tortorello M Lou (ed), Encyclopedia of Food Microbiology, Second Edition. Academic Press, Oxford.

27. Bissonnette F, Labrie S, Deveau H, Lamoureux M, Moineau S. 2000. Characterization of mesophilic mixed starter cultures used for the manufacture of aged cheddar cheese. *J Dairy Sci* **83**:620-627.

28. Parente E, Guidone A, Matera A, De Filippis F, Mauriello G, Ricciardi A. 2016. Microbial community dynamics in thermophilic undefined milk starter cultures. *Int J Food Microbiol* **217**:59-67.

29. Lavelle K, Martinez I, Neve H, Lugli GA, Franz CMAP, Ventura M, Dal Bello F, van Sinderen D, Mahony J. 2018. Biodiversity of *Streptococcus thermophilus* phages in global dairy fermentations. *Viruses* **10**:577.

30. Meslier V, Loux V, Renault P. 2012. Genome sequence of *Lactococcus raffinolactis* strain 4877, isolated from natural dairy starter culture. *J Bacteriol* **194**:6364.

31. Tidona F, Meucci A, Povolo M, Pelizzola V, Zago M, Contarini G, Carminati D, Giraffa G. 2018. Applicability of *Lactococcus hircilactis* and *Lactococcus laudensis* as dairy cultures. *Int J Food Microbiol* **271**:1-7.

32. Schleifer KH, Kilpper-Bälz R. 1987. Molecular and chemotaxonomic approaches to the classification of Streptococci, Enterococci and Lactococci: a review. *Syst Appl Microbiol* **10**:1-19.

33. Li TT, Tian WL, Gu CT. 2021. Elevation of *Lactococcus lactis* subsp. *cremoris* to the species level as *Lactococcus cremoris* sp. nov. and transfer of *Lactococcus lactis* subsp. *tructae* to *Lactococcus cremoris* as *Lactococcus cremoris* subsp. *tructae* comb. nov. *Int J Syst Evol Microbiol* **71**: 004727.

34. Kimoto H, Nomura M, Suzuki I. 1999. Growth energetics of *Lactococcus lactis* subsp. *lactis* biovar *diacetylactis* in cometabolism of citrate and glucose. *Int Dairy J* **9**:857-863.

35. Kelly WJ, Ward LJH, Leahy SC. 2010. Chromosomal diversity in *Lactococcus lactis* and the origin of dairy starter cultures. *Genome Biol Evol* **2**:729-744.

36. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee J-H, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W, Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B, Mills D. 2006. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci USA* **103**:15611-15616.

37. Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M-C. 2001. Analysis of six prophages in *Lactococcus lactis* lL1403: different genetic structure of temperate and virulent phage populations Nucleic Acids Res. **29**:644-651.

38. Goh YJ, Goin C, O'Flaherty S, Altermann E, Hutkins R. 2011. Specialized adaptation of a lactic acid bacterium to the milk environment: the comparative genomics of *Streptococcus thermophilus* LMD-9. *Microb Cell Fact* **10**:S22.

39. van Mastrigt O, Di Stefano E, Hartono S, Abee T, Smid EJ. 2018. Large plasmidome of dairy *Lactococcus lactis* subsp. *lactis* biovar *diacetylactis* FM03P encodes technological functions and appears highly unstable. *BMC Genomics* **19**:620.

40. David B, Radziejwoski A, Toussaint F, Fontaine L, de Frahan MH, Patout C, van Dillen S, Boyaval P, Horvath P, Fremaux C, Hols P. 2017. Natural DNA transformation is functional in *Lactococcus lactis* subsp. *cremoris* KW2. *Appl Environ Microbiol* **83**:e01074-17.

41. Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM. 2005. Complete sequences of four plasmids of *Lactococcus lactis* subsp. *cremoris* SK11 reveal extensive adaptation to the dairy environment. *Appl Environ Microbiol* **71**:8371-8382.

42. Mills S, McAuliffe OE, Coffey A, Fitzgerald GF, Ross RP. 2006. Plasmids of lactococci—genetic accessories or genetic necessities? *FEMS Microbiol Rev* **30**:243-273.

43. Chopin MC, Chopin A, Bidnenko E. 2005. Phage abortive infection in lactococci: Variations on a theme. *Curr Opin Microbiol* **8**:473-479.

44. Labrie SJ, Samson JE, Moineau S. 2010. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**:317-327.

45. Chapot-Chartier MP, Vinogradov E, Sadovskaya I, Andre G, Mistou MY, Trieu-Cuot P, Furlan S, Bidnenko E, Courtin P, Péchoux C, Hols P, Dufre YF, Kulakauskas S. 2010. Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. *J Biol Chem* **285**:10464-10471.

46. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C. 2013. Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol* **79**:4385-4392.

47. Mahony J, Randazzo W, Neve H, Settanni L, van Sinderen D. 2015. Lactococcal 949 group phages recognize a carbohydrate receptor on the host cell surface. *Appl Environ Microbiol* **81**:3299-3305.

48. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, van Sinderen D. 2014. Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. *mBio* **5**:e00880-14.

49. Farenc C, Spinelli S, Vinogradov E, Tremblay D, Blangy S, Sadovskaya I, Moineau S, Cambillau C. 2014. Molecular insights on the recognition of a *Lactococcus lactis* cell wall pellicle by the phage 1358 receptor binding protein. *J Virol* **88**:7005-7015.

50. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, Chapot-Chartier M-P, Cambillau C, Holo H, van Sinderen D. 2020. The CWPS Rubik's cube: Linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. *Mol Microbiol* **114**:582-596.

51. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D. 2016. Comparative genomics and

functional analysis of the 936 group of lactococcal *Siphoviridae* phages. *Sci Rep* **6**:21345.

52. Parlindungan E, Sadovskaya I, Vinogradov E, Lugli GA, Ventura M, van Sinderen D, Mahony J. 2024. Novel cell wall polysaccharide genotypes and structures of lactococcal strains isolated from milk and fermented foods. *Int J Food Microbiol* **424**:110840.
53. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. 2020. Dairy lactococcal and streptococcal phage-host interactions: an industrial perspective in an evolving phage landscape. *FEMS Microbiol Rev* **44**:909-932.
54. Giesbers CAP, Fagan J, Parlindungan E, Palussière S, Courtin P, Lugli GA, Ventura M, Kulakauskas S, Chapot-Chartier MP, Mahony J, van Sinderen D. 2023. Reduced synthesis of phospho-polysaccharide in *Lactococcus* as a strategy to evade phage infection. *Int J Food Microbiol* **407**:110415.
55. Dupont K, Janzen T, Vogensen FK, Josephsen J, Stuer-Lauridsen B. 2004. Identification of *Lactococcus lactis* genes required for bacteriophage adsorption. *Appl Environ Microbiol* **70**:5825-5832.
56. Mahony J, van Sinderen D. 2014. Current taxonomy of phages infecting lactic acid bacteria. *Front Microbiol* **5**:7.
57. Deveau H, Labrie SJ, Chopin MC, Moineau S. 2006. Biodiversity and classification of lactococcal phages. *Appl Environ Microbiol* **72**:4338-4346.
58. Turner D, Shkoporov AN, Lood C, Millard AD, Dutilh BE, Alfenas-Zerbini P, van Zyl LJ, Aziz RK, Oksanen HM, Poranen MM, Kropinski

AM, Barylski J, Brister JR, Chanisvili N, Edwards RA, Enault F, Gillis A, Knezevic P, Krupovic M, Kurtböke I, Kushkina A, Lavigne R, Lehman S, Lobocka M, Moraru C, Moreno Switt A, Morozova V, Nakavuma J, Reyes Muñoz A, Rūmnieks J, Sarkar BL, Sullivan MB, Uchiyama J, Wittmann J, Yigang T, Adriaenssens EM. 2023. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch Virol* **168**:74.

59. Mahony J, Murphy J, van Sinderen D. 2012. Lactococcal 936-type phages and dairy fermentation problems: From detection to evolution and prevention. *Front Microbiol* **3**:1-9.
60. Chmielewska-Jeznach M, Bardowski JK, Szczepankowska AK. 2020. *Lactococcus Ceduovirus* phages isolated from industrial dairy plants—from physiological to genomic analyses. *Viruses* **12**:280.
61. Mahony J, Deveau H, McGrath S, Ventura M, Canchaya C, Moineau S, Fitzgerald GF, van Sinderen D. 2006. Sequence and comparative genomic analysis of lactococcal bacteriophages jj50, 712 and P008: evolutionary insights into the 936 phage species. *FEMS Microbiol Lett* **261**:253-261.
62. Chmielewska-Jeznach M, Bardowski JK, Szczepankowska AK. 2018. Molecular, physiological and phylogenetic traits of *Lactococcus* 936-type phages from distinct dairy environments. *Sci Rep* **8**:12540.

63. Frantzen CA, Holo H. 2019. Unprecedented diversity of lactococcal group 936 bacteriophages revealed by amplicon sequencing of the portal protein gene. *Viruses* **11**:443.

64. Castro-Nallar E, Chen H, Gladman S, Moore SC, Seemann T, Powell IB, Hillier A, Crandall KA, Chandr PS. 2012. Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage. *Genome Biol Evol* **4**:382-393.

65. J LS, Jytte J, Horst N, K VF, Sylvain M. 2008. Morphology, genome sequence, and structural proteome of type phage P335 from *Lactococcus lactis*. *Appl Environ Microbiol* **74**:4636-4644.

66. Kelly WJ, Altermann E, Lambie SC, Leahy SC. 2013. Interaction between the genomes of *Lactococcus lactis* and phages of the P335 species. *Front Microbiol* **4**:1-9.

67. Kelleher P, Mahony J, Schweinlin K, Neve H, Franz CM, van Sinderen D. 2018. Assessing the functionality and genetic diversity of lactococcal prophages. *Int J Food Microbiol* **272**:29-40.

68. Vinga I, Baptista C, Auzat I, Petipas I, Lurz R, Tavares P, Santos MA, São-José C. 2012. Role of bacteriophage SPP1 tail spike protein gp21 on host cell receptor binding and trigger of phage DNA ejection. *Mol Microbiol* **83**:289-303.

69. Dieterle ME, Fina Martin J, Durán R, Nemirovsky SI, Sanchez Rivas C, Bowman C, Russell D, Hatfull GF, Cambillau C, Piuri M. 2016. Characterization of prophages containing “evolved” Dit/Tal modules in the genome of *Lactobacillus casei* BL23. *Appl Microbiol Biotechnol* **100**:9201-9215.

70. Hayes S, Vincentelli R, Ramond L, Ramnd R, Nauta A, Sinderen D van, Cambillau C. 2019. Ubiquitous carbohydrate binding modules decorate 936 lactococcal siphophage virions. *Viruses* **11**:631.

71. Dupont K, Vogensen FK, Neve H, Bresciani J, Josephsen J. 2004. Identification of the receptor-binding protein in 936-species lactococcal bacteriophages. *Appl Environ Microbiol* **70**:5818-5824.

72. Millen AM, Romero DA. 2016. Genetic determinants of lactococcal *C2viruses* for host infection and their role in phage evolution. *J Gen Virol* **97**:1998-2007.

73. Chandry PS, Moore SC, Boyce JD, Davidson BE, Hillier AJ. 1997. Analysis of the DNA sequence, gene expression, origin of replication and modular structure of the *Lactococcus lactis* lytic bacteriophage sk1. *Mol Microbiol* **26**:49-64.

74. McDonnell B, Hanemaaijer L, Bottacini F, Kelleher P, Lavelle K, Sadovskaya I, Vinogradov E, Ver Loren van Themaat E, Kouwen T, Mahony J, van Sinderen D. 2020. A cell wall-associated polysaccharide is required for bacteriophage adsorption to the *Streptococcus thermophilus* cell surface. *Mol Microbiol* **114**:31-45.

75. Mahony J, Cambillau C, van Sinderen D. 2017. Host recognition by lactic acid bacterial phages. *FEMS Microbiol Rev* **41**:S16-S26.

76. Collins B, Bebeacua C, Mahony J, Blangy S, Douillard FP, Veesler D, Cambillau C, van Sinderen D. 2013. Structure and functional analysis of the host recognition device of lactococcal phage Tuc2009. *J Virol* **87**:8429-8440.

77. Legrand P, Collins B, Blangy S, Murphy J, Spinelli S, Gutierrez C, Richet N, Kellenberger C, Desmyter A, Mahony J. 2016. The atomic structure of the phage Tuc2009 baseplate tripod suggests that host recognition involves two different carbohydrate binding modules. *mBio* **7**:e01781-15.

78. Marcelli B, de Jong A, Karsens H, Janzen T, Kok J, Kuipers OP. 2019. A specific sugar moiety in the *Lactococcus lactis* cell wall pellicle is required for infection by CHPC971, a member of the rare 1706 phage species. *Appl Environ Microbiol* **85**:e01224-19.

79. Vegge CS, Vogensen FK, McGrath S, Neve H, van Sinderen D, Brøndsted L. 2006. Identification of the lower baseplate protein as the antireceptor of the temperate lactococcal bacteriophages TP901-1 and Tuc2009. *J Bacteriol* **188**:55-63.

80. Villion M, Chopin M-C, Deveau H, Ehrlich SD, Moineau S, Chopin A. 2009. P087, a lactococcal phage with a morphogenesis module similar to an *Enterococcus faecalis* prophage. *Virology* **388**:49-56.

81. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C, Vogensen FK, van Sinderen D. 2013. Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol* **79**:4385-4392.

82. Millen AM, Romero DA, Horvath P, Magill D, Simdon L. 2022. Host-encoded, cell surface-associated exopolysaccharide required for

adsorption and infection by lactococcal P335 phage subtypes. *Front Microbiol* **13**:971166.

83. Monteville MR, Ardestani B, Geller BL. 1994. Lactococcal bacteriophages require a host cell wall carbohydrate and a plasma membrane protein for adsorption and ejection of DNA. *Appl Environ Microbiol* **60**:3204-3211.
84. Valyasevi R, Sandine WE, Geller BL. 1991. A membrane protein is required for bacteriophage c2 infection of *Lactococcus lactis* subsp. *lactis* C2. *J Bacteriol* **173**:6095-6100.
85. Derkx PMF, Janzen T, Sørensen KI, Christensen JE, Stuer-Lauridsen B, Johansen E. 2014. The art of strain improvement of industrial lactic acid bacteria without the use of recombinant DNA technology. *Microb Cell Fact* **13**:S5.
86. Plisson C, White HE, Auzat I, Zafarani A, São-José C, Lhuillier S, Tavares P, Orlova E V. 2007. Structure of bacteriophage SPP1 tail reveals trigger for DNA ejection. EMBO J **26**:3720-3728.
87. Baptista C, Santos MA, São-José C. 2008. Phage SPP1 reversible adsorption to *Bacillus subtilis* cell wall teichoic acids accelerates virus recognition of membrane receptor YueB. *J Bacteriol* **190**:4989-4996.
88. Spinelli S, Desmyter A, Verrips CT, De Haard HJW, Moineau S, Cambillau C. 2006. Lactococcal bacteriophage p2 receptor-binding protein structure suggests a common ancestor gene with bacterial and mammalian viruses. *Nat Struct Mol Biol* **13**:85-89.

89. Siponen M, Spinelli S, Blangy S, Moineau S, Cambillau C, Campanacci V. 2009. Crystal structure of a chimeric receptor binding protein constructed from two lactococcal phages. *J Bacteriol* **191**:3220-3225.

90. Tremblay DM, Tegoni M, Spinelli S, Campanacci V, Blangy S, Huyghe C, Desmyter A, Labrie S, Moineau S, Cambillau C. 2006. Receptor-binding protein of *Lactococcus lactis* phages: identification and characterization of the saccharide receptor-binding site. *J Bacteriol* **188**:2400-2410.

91. Ricagno S, Campanacci V, Blangy S, Spinelli S, Tremblay D, Moineau S, Tegoni M, Cambillau C. 2006. Crystal structure of the receptor-binding protein head domain from *Lactococcus lactis* phage bIL170. *J Virol* **80**:9331-9335.

92. Vegge CS, Neve H, Brøndsted L, Heller KJ, Vogensen FK. 2006. Analysis of the collar-whisker structure of temperate lactococcal bacteriophage TP901-1. *Appl Environ Microbiol* **72**:6815-6818.

93. Mahony J, Oliveira J, Collins B, Hanemaaijer L, Lugli GA, Neve H, Ventura M, Kouwen TR, Cambillau C, van Sinderen D. 2017. Genetic and functional characterisation of the lactococcal P335 phage-host interactions. *BMC Genomics* **18**:146.

94. Hayes S, Vincentelli R, Mahony J, Nauta A, Ramond L, Lugli GA, Ventura M, van Sinderen D, Cambillau C. 2018. Functional carbohydrate binding modules identified in evolved Dits from siphophages infecting various Gram-positive bacteria. *Mol Microbiol* **110**:777-795.

95. Goulet A, Mahony J, Cambillau C, van Sinderen D. 2022. Exploring structural diversity among adhesion devices encoded by Lactococcal P335 phages with AlphaFold2. *Microorganisms* **10**:2278.

96. Millen AM, Magill D, Romero D, Simdon L. 2023. Evolved distal tail protein of skunaviruses facilitates adsorption to exopolysaccharide-encoding lactococci. *Microbiome Res Rep* **2**:26.

97. Carlton RM, Noordman WH, Biswas B, de Meester ED, Loessner MJ. 2005. Bacteriophage P100 for control of *Listeria monocytogenes* in foods: genome sequence, bioinformatic analyses, oral toxicity study, and application. *Regul Toxicol Pharmacol* **43**:301-312.

98. Erkus O, de Jager VCL, Spus M, van Alen-Boerrigter IJ, van Rijswijck IMH, Hazelwood L, Janssen PWM, van Hijum SAFT, Kleerebezem M, Smid EJ. 2013. Multifactorial diversity sustains microbial community stability. *ISME* J **7**:2126-2136.

99. de Melo AG, Rousseau GM, Tremblay DM, Labrie SJ, Moineau S. 2020. DNA tandem repeats contribute to the genetic diversity of *Brevibacterium aurantiacum* phages. *Environ Microbiol* **22**:3413-3428.

100. Spus M, Li M, Alexeeva S, Wolkers-Rooijackers JCM, Zwietering MH, Abee T, Smid EJ. 2015. Strain diversity and phage resistance in complex dairy starter cultures. *J Dairy Sci* **98**:5173-5182.

101. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. 2018. Cell wall glycans mediate recognition of the dairy bacterium *Streptococcus thermophilus* by bacteriophages. *Appl Environ Microbiol* **84**:e01847-18.

102. Poulsen VK, Derkx P, Oregaard G. 2019. High-throughput screening for texturing *Lactococcus* strains. *FEMS Microbiol Lett* **366**:fnz001.
103. Van Kranenburg R, Marugg JD, van Swam II, Willem NJ, de Vos WM. 1997. Molecular characterization of the plasmid-encoded *eps* gene cluster essential for exopolysaccharide biosynthesis in *Lactococcus lactis*. *Mol Microbiol* **24**:387-397.
104. van Kranenburg R, Kleerebezem M, de Vos WM. 2000. Nucleotide sequence analysis of the lactococcal EPS plasmid pNZ4000. *Plasmid* **43**:130-136.
105. R VE, M NJ. 1986. Involvement of a plasmid in production of ropiness (mucoidness) in milk cultures by *Streptococcus cremoris* MS. *Appl Environ Microbiol* **51**:677-682.
106. Kleerebezem M, van Kranenburg R, Tuinier R, Boels IC, Zoon P, Looijesteijn E, Hugenholtz J, de Vos WM. 1999. Exopolysaccharides produced by *Lactococcus lactis*: from genetic engineering to improved rheological properties? *Antonie Van Leeuwenhoek* **76**:357-365.
107. Looijesteijn PJ, Trapet L, de Vries E, Abee T, Hugenholtz J. 2001. Physiological function of exopolysaccharides produced by *Lactococcus lactis*. *Int J Food Microbiol* **64**:71-80.
108. Lucey M, Daly C, Fitzgerald GF. 1992. Cell surface characteristics of *Lactococcus lactis* harbouring pCI528, a 46 kb plasmid encoding inhibition of bacteriophage adsorption. *J Gen Microbiol* **138**:2137-2143.

109. Forde A, Fitzgerald GF. 1999. Analysis of exopolysaccharide (EPS) production mediated by the bacteriophage adsorption blocking plasmid, pCI658, isolated from *Lactococcus lactis* ssp. *cremoris* HO2. *Int Dairy J* **9**:465-472.

110. McGrath S, Fitzgerald GF, van Sinderen D. 2002. Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. *Mol Microbiol* **43**:509-520.

111. Durmaz E, Madsen SM, Israelsen H, Klaenhammer TR. 2002. *Lactococcus lactis* lytic bacteriophages of the P335 group are inhibited by overexpression of a truncated CI repressor. *J Bacteriol* **184**:6532-6543.

112. Prévots F, Daloyau M, Bonin O, Dumont X, Tolou S. 1996. Cloning and sequencing of the novel abortive infection gene AbiH of *Lactococcus lactis* ssp. *lactis* biovar. *diacetylactis* S94. *FEMS Microbiol Lett* **142**:295-299.

113. Hill C, Miller LA, Klaenhammer TR. 1990. Nucleotide sequence and distribution of the pTR2030 resistance determinant (*hsp*) which aborts bacteriophage infection in lactococci. *Appl Environ Microbiol* **56**:2255-2258.

114. Klaenhammer TR. 1987. Plasmid-directed mechanisms for bacteriophage defense in lactic streptococci. *FEMS Microbiol Lett* **46**:313-325.

115. Grafakou A, Mosterd C, Beck MH, Kelleher P, McDonnell B, de Waal PP, van Rijswijck IMH, van Peij NNME, Cambillau C, Mahony J, van

Sinderen D. 2024. Discovery of antiphage systems in the lactococcal plasmidome. *Nucleic Acids Res* **52**:9760-9776.

116. Philippe C, Cornuault JK, De Melo AG, Morin-Pelchat R, Jolicoeur AP, Moineau S. 2023. The never-ending battle between lactic acid bacteria and their phages. *FEMS Microbiol Rev* **47**:fuad035.
117. Durmaz E, Klaenhammer TR. 2007. Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. *J Bacteriol* **189**:1417-1425.
118. Grafakou A, Mosterd C, de Waal PP, van Rijswijck IMH, van Peij NNME, Mahony J, van Sinderen D. 2024. Functional and practical insights into three lactococcal antiphage systems. *Appl Environ Microbiol* **90**:e01120-24
119. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, Jackson SA. 2022. PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res* **50**:W541-W550.
120. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, Bernheim A. 2022. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**:2561.
121. Boyer HW. 1971. DNA Restriction and Modification Mechanisms in Bacteria. *Annu Rev Microbiol* **25**:153-176.
122. Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG, Murray NE. 2014. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res* **42**:3-19.

123. Patel PH, Maxwell KL. 2023. Prophages provide a rich source of antiphage defense systems. *Curr Opin Microbiol* **73**:102321.

124. O'Driscoll J, Glynn F, Cahalane O, O'Connell-Motherway M, Fitzgerald GF, van Sinderen D. 2004. Lactococcal plasmid pNP40 encodes a novel, temperature-sensitive restriction-modification system. *Appl Environ Microbiol* **70**:5546-5556.

125. Murphy J, Klumpp J, Mahony J, O'Connell-Motherway M, Nauta A, van Sinderen D. 2014. Methyltransferases acquired by lactococcal 936-type phage provide protection against restriction endonuclease activity. *BMC Genomics* **15**:831.

126. Murphy J, Mahony J, Ainsworth S, Nauta A, van Sinderen D. 2013. Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl Environ Microbiol* **79**:7547-55.

127. Jung JY, Lee SH, Kim JM, Park MS, Bae JW, Hahn Y, Madsen EL, Jeon CO. 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. *Appl Environ Microbiol* **77**:2264-2274.

128. Ferrocino I, Bellio A, Giordano M, Macori G, Romano A, Rantsiou K, Decastelli L, Cocolin L. 2018. Shotgun metagenomics and volatilome profile of the microbiota of fermented sausages. *Appl Environ Microbiol* **84**:e02120-17.

129. Mayo B, Rachid CTCC, Alegría Á, Leite AMO, Peixoto RS, Delgado S. 2014. Impact of next generation sequencing techniques in food microbiology. *Curr Genomics* **15**:293-309.

130. Somerville V, Berthoud H, Schmidt RS, Bachmann HP, Meng YH, Fuchsmann P, von Ah U, Engel P. 2022. Functional strain redundancy and persistent phage infection in Swiss hard cheese starter cultures. *ISME J* **16**:388-399.

131. Zotta T, Ricciardi A, Condelli N, Parente E. 2022. Metataxonomic and metagenomic approaches for the study of undefined strain starters for cheese manufacture. *Crit Rev Food Sci Nutr* **62**:3898-3912.

132. Porcellato D, de Paula Dias Moreira L, Franklin FV, da Silva Duarte V, Øyaas J, Svalestad F, Skeie S. 2024. Impact of microbial milk quality on mesophilic starter culture metabolism: a metaproteomics approach. *Int Dairy J* **159**:106070.

133. Achilleos C, Berthier F. 2013. Quantitative PCR for the specific quantification of *Lactococcus lactis* and *Lactobacillus paracasei* and its interest for *Lactococcus lactis* in cheese samples. *Food Microbiol* **36**:286-295.

134. Furet JP, Quénée P, Tailliez P. 2004. Molecular quantification of lactic acid bacteria in fermented milk products using real-time quantitative PCR. *Int J Food Microbiol* **97**:197-207.

135. Yu J, Wang HM, Zha MS, Qing YT, Bai N, Ren Y, Xi XX, Liu WJ, Menghe BLG, Zhang HP. 2015. Molecular identification and quantification of lactic acid bacteria in traditional fermented dairy foods of Russia. *J Dairy Sci* **98**:5143-5154.

136. Friedrich U, Lenke J. 2006. Improved enumeration of lactic acid bacteria in mesophilic dairy starter cultures by using multiplex

quantitative real-time PCR and flow cytometry-fluorescence in situ hybridization. *Appl Environ Microbiol* **72**:4163-4171.

137. Agrimonti C, Bottari B, Sardaro MLS, Marmiroli N. 2019. Application of real-time PCR (qPCR) for characterization of microbial populations and type of milk in dairy food products. *Crit Rev Food Sci Nutr* **59**:423-442.
138. Bottari B, Agrimonti C, Gatti M, Neviani E, Marmiroli N. 2013. Development of a multiplex real time PCR to detect thermophilic lactic acid bacteria in natural whey starters. *Int J Food Microbiol* **160**:290-297.
139. Fan X, Li X, Zhang T, Xu J, Shi Z, Wu Z, Wu J, Pan D, Du L. 2021. A Novel qPCR method for the detection of lactic acid bacteria in fermented milk. **Foods** 10:3066.
140. Jung MY, Chang Y -H., Kim W. 2010. A real-time PCR assay for detection and quantification of *Lactococcus garvieae*. *J Appl Microbiol* **108**:1694-1701.
141. Ruggirello M, Dolci P, Cocolin L. 2014. Detection and viability of *Lactococcus lactis* throughout cheese ripening. *PLoS One* **9**:e114280.
142. Townsend EM, Kelly L, Muscatt G, Box JD, Hargraves N, Lilley D, Jameson E. 2021. The human gut phageome: origins and roles in the human gut microbiome. *Front Cell Infect Microbiol* **11**:1-27.
143. Paillet T, Lamy-Besnier Q, Figueroa C, Petit M-A, Dugat-Bony E. 2024. Dynamics of the viral community on the cheese surface during

maturation and persistence across production years. *mSystems* **9**: e0020124.

144. Colombo S, Arioli S, Gargari G, Neri E, Della Scala G, Mora D. 2018. Characterization of airborne viromes in cheese production plants. *J Appl Microbiol* **125**:1444-1454.
145. Muhammed MK, Olsen ML, Kot W, Neve H, Castro-Mejía JL, Janzen T, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, Vogensen FK. 2018. Investigation of the bacteriophage community in induced lysates of undefined mesophilic mixed-strain DL-cultures using classical and metagenomic approaches. *Int J Food Microbiol* **272**:61-72.
146. Muhammed MK, Kot W, Neve H, Mahony J, Castro-Mejía JL, Krych L, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, van Sinderen D, Vogensen FK. 2017. Metagenomic analysis of dairy bacteriophages: extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. *Appl Environ Microbiol* **83**:e00888-17.
147. Lillehaug D. 1997. An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. *J Appl Microbiol* **83**:85-90.
148. Marcó MB, Moineau S, Quiberoni A. 2012. Bacteriophages and dairy fermentations. *Bacteriophage* **2**:149-158.
149. Ács N, Gambino M, Brøndsted L. 2020. Bacteriophage enumeration and detection methods. *Front Microbiol* **11**:594868.

150. Patel A, Noble RT, Steele JA, Schwalbach MS, Hewson I, Fuhrman JA. 2007. Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat Protoc* **2**:269-276.
151. Borsheim KY, Bratbak G, Heldal M. 1990. Enumeration and biomass estimation of planktonic bacteria and viruses by transmission electron microscopy. *Appl Environ Microbiol* **56**:352-356.
152. Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2**:63-77.
153. Rohwer F, Youle M. 2011. Consider something viral in your research. *Nat Rev Microbiol* **9**:308-309.
154. Göller PC, Haro-Moreno JM, Rodriguez-Valera F, Loessner MJ, Gómez-Sanz E. 2020. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil. *Microbiome* **8**:1-16.
155. Dugat-Bony E, Lossouarn J, De Paepe M, Sarthou A-S, Fedala Y, Petit M-A, Chaillou S. 2020. Viral metagenomic analysis of the cheese surface: a comparative study of rapid procedures for extracting viral particles. *Food Microbiol* **85**:103278.
156. Jung MJ, Kim MS, Yun JH, Lee JY, Kim PS, Lee HW, Ha JH, Roh SW, Bae JW. 2018. Viral community predicts the geographical origin of fermented vegetable foods more precisely than bacterial community. *Food Microbiol* **76**:319-327.

157. Kumar A, Murthy S, Kapoor A. 2017. Evolution of selective-sequencing approaches for virus discovery and virome analysis. *Virus Res* **239**:172-179.

158. Kleiner M, Hooper L V., Duerkop BA. 2015. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**:7.

159. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. 2019. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**:12.

160. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: An in silico evaluation of metagenome—enabled estimates of viral community composition and diversity. *PeerJ* **5**:e3817.

161. Mallawaarachchi V, Roach MJ, Decewicz P, Papudeshi B, Giles SK, Grigson SR, Bouras G, Hesse RD, Inglis LK, Hutton ALK, Dinsdale EA, Edwards RA. 2023. Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* **39**:btad586.

162. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**:455-477.

163. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit MA, Enault F. 2021. PHROG: families of

prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* **3**:lqab067.

164. Thingstad TF. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**:1320-1328.
165. Wen R, Li X-A, Han G, Chen Q, Kong B. 2021. Fungal community succession and volatile compound dynamics in Harbin dry sausage during fermentation. *Food Microbiol* **99**:103764.
166. Świder O, Wójcicki M, Bujak M, Juszczuk-Kubiak E, Szczepańska M, Roszko MŁ. 2021. Time evolution of microbial composition and metabolic profile for biogenic amines and free amino acids in a model cucumber fermentation system brined with 0.5% to 5.0% sodium chloride. *Molecules* **26**:5796.

Chapter 2. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures

This chapter has been published in:

White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* 90:e02152-23.

Experiments, analyses, and first draft preparations were performed by the candidate, except for the initial taxonomic classification of the metagenomes which was performed by Macrogen Europe.

ABSTRACT

The dairy fermentation industry relies on the activity of lactic acid bacteria in robust starter cultures to accomplish milk acidification. Maintenance of the composition of these starter cultures, whether defined or undefined, is essential to ensure consistent and high-quality fermentation end products. To date, limited information exists regarding the microbial composition of undefined starter culture systems. Here, we describe a culture-based analysis combined with a metagenomics approach to evaluate the composition of two undefined mesophilic starter cultures. In addition, we describe a qPCR-based genotype detection assay, which is capable of discerning nine distinct lactococcal genotypes to characterise these undefined starter cultures, and which can be applied to monitor compositional changes in an undefined starter culture during a fermentation.

INTRODUCTION

Lactic acid bacteria (LAB) are extensively used for the manufacture of a wide range of fermented food products. Fermentation can either occur spontaneously, utilising the endogenous microbiota of the fermentation environment or substrate, or through small- or large-scale practices which incorporate deliberately added starter cultures (1). In both situations, it is essential that the composition of the culture remains stable to ensure consistent and high-quality fermentation end products. Starter cultures are typically comprised of several bacterial strains to ensure the consistent manufacture of a final product with desired organoleptic properties. Mesophilic dairy fermentations primarily rely on the activity of two LAB species, i.e., *Lactococcus lactis* and *Lactococcus cremoris*, which until recently were classified as subspecies of *L. lactis* (2). The starter culture composition can be designated as either defined or undefined, where in the former case the starter culture consists of a small number of known strains (two to six strains) being present at an established relative abundance, while in the latter case it comprises an unknown, but presumably large number of strains present at an undetermined relative abundance (3). There are limitations and benefits associated with both defined and undefined starter culture systems, particularly with respect to bacteriophage susceptibility and robustness (4). Phage attacks can have a significant negative impact on starter cultures resulting in a delayed or even failed fermentation. Undefined starter cultures are typically considered to be more phage-robust when compared to defined starter cultures due to the superior culture diversity of the former (5). However, since the specific composition and relevant technical characteristics of the

strains applied in defined starter cultures are established, strains can be more readily replaced with phage-unrelated/insensitive strains without changing the (bio)technological integrity of the culture. Conversely, due to the unknown composition of undefined starter cultures, it is often difficult to mitigate the effects of phage infections or even to determine the impact of phages on the fermentation and ripening process until the final product is graded for its sensory attributes (6).

Lactococcal phages impose serious economic losses on the dairy industry and are therefore among the best studied phage-host systems. Most notably, the phage-host interactions of skunaviruses (formerly termed the 936 group phages) and (to a lesser extent) ceduoviruses (formerly termed the c2 group phages) have been extensively investigated since they are the most problematic virulent lactococcal phages (7–10). Phage infection typically commences with reversible binding of a phage's tail tip-associated anti-receptor to a receptor present on the host cell surface (11). The phage receptor binding protein (RBP) may recognise and bind to saccharidic, teichoic acid, or proteinaceous receptors (12). The RBPs of skunaviruses recognise and bind irreversibly to a saccharidic receptor on lactococcal cell surfaces, i.e. **c**ell **w**all **p**oly**s**accharide (CWPS) structures (13). The *cwps* gene cluster is responsible for the biosynthesis of a given lactococcal CWPS, which is composed of two elements: the peptidoglycan-embedded rhamnan and surface-exposed polysaccharide pellicle (PSP). There are currently four defined *cwps* genotypes (A-D) and several C-subtypes (C_1-C_8) recognized among lactococcal strains (14, 15). A recent review, of over 400 lactococcal strains, suggests that additional *cwps* genotypes are likely to exist, indicating

the continually evolving genetic composition of this gene cluster among lactococci (16). Recently, RBP phylogeny of skunaviruses has been shown to correspond with lactococcal *cwps* genotypes, which in turn is linked to the biochemical diversity of the associated surface-exposed PSP (i.e. the phage receptor) and thus to host range, demonstrating the industrial relevance of this gene cluster (14–18).

Starter culture rotation is a common strategy to counteract the effects of phage attacks during various dairy fermentations, by rotating starter cultures that are capable of generating products with reproducible organoleptic properties but have different phage sensitivity profiles. However, in order to implement an effective culture rotation strategy, it is important to understand both the microbial diversity of the starter cultures, as well as the diversity of phages present in the fermentation environment (19). In the case of undefined mixed starter cultures, typically very little is known about the microbial diversity and complex community dynamics, therefore this lack of knowledge may make it difficult to choose which starter culture to rotate in after disruptions due to phages (19, 20). In the current study, a combination of approaches was used to characterise two undefined mesophilic mixed starter cultures, including culture-based and metagenome analysis combined with the establishment of a qPCR assay (based on lactococcal *cwps* gene clusters) capable of discerning nine industrially relevant lactococcal genotypes. The metagenome, culture-based, and qPCR-based analyses were successfully used to not only discern the diversity of lactococci within starter cultures, but also to monitor changes in the strain composition of an undefined culture in the presence of phages.

MATERIALS AND METHODS

Lactococcal strains and bacteriophages

Bacterial strains and phages used in this study are listed in Supplementary Table S1. Bacterial strains were grown overnight at 30 °C in M17 broth (Oxoid, Hampshire, United Kingdom) supplemented with 0.5 % lactose (Sigma Aldrich, Gillingham, UK) (LM17). Phages were propagated (initially from single plaques, then using 100 µL of phage lysates after initial plaque propagation) in the same medium supplemented with 10 mM $CaCl_2$ (Sigma Aldrich) and added at the point of host strain inoculation (100 µL of fresh overnight culture). The mixture was incubated overnight at 30 °C and resultant phage lysates were then filtered (0.45 µM) and stored at 4 °C.

Metagenome analysis of undefined starter cultures

(i): Metagenome DNA extraction

The microbial diversity of two undefined industrial starter cultures, M1 and M2, commonly used in the production of semi hard cheeses, fermented milk (e.g., viili, kefir, and twaróg), and fresh cheeses (e.g., quark) was investigated. The cultures were activated by first diluting 10 g of lyophilised commercial starter culture in 207 g of sterilized (115 °C for 20 minutes) 10 % (w/v) reconstituted skimmed milk (RSM) powder. Following stomaching, in a Stomacher® 400 Circulator (Seward, UK), for 2 minutes at 230 rpm, 20 g of the first dilution was added to another 207 g of sterilized 10 % RSM. The second dilution was incubated overnight at 22 °C. The following day, 100 ml of pasteurized (90 °C for 20 minutes) 10 % RSM was inoculated with the

overnight culture (at 3 %) and incubated overnight again at the same temperature to fully activate the culture.

Metagenomic DNA of the activated starter cultures was extracted following an adapted protocol described previously by Erkus, et al., described as follows: 2.2 % trisodium citrate (Sigma Aldrich) solution (heated to 45 °C) was added to 2 g of activated starter culture to reach a total volume of 25 ml (20). The resulting solution was homogenised in a Stomacher® 400 Circulator (Seward) for 5 minutes at 230 rpm. The homogenised solution was centrifuged at 13,800 × *g* for 10 minutes at room temperature. The supernatant and fat layers were removed, and the cell pellet was resuspended in 1 ml of 2 % sodium citrate solution (45 °C) for washing. The resuspended solution was centrifuged at 10,800 × *g* for 5 minutes. The supernatant and remaining fat layer were removed and this washing step was repeated 2-3 more times (until full removal of the fat layer). The washed cell pellet was resuspended in 1 ml of lysis buffer [20 mM Tris-HCl buffer (pH=8), 2 mM Ethylenediaminetetraacetic acid (EDTA; Sigma Aldrich), 2 % polyethylene glycol (Sigma Aldrich)] with the addition of 50 µg/ml of lysozyme (Sigma Aldrich) and 100 U of mutanolysin (Sigma Aldrich). This cell suspension was incubated at 37 °C for 3 hrs. 250 µg/ml of proteinase K (Merck, NJ, USA) was then added and the solution was incubated at 56 °C for 1 hr. 1 ml of 96 % ethanol was added to the lysed solution and mixed gently. The precipitated DNA was further washed and purified using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol and eluted in a final volume of 300 µl. DNA concentrations were measured

using the Qubit dsDNA BR assay kit with Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA; Supplementary Table S2).

(ii): Illumina shotgun metagenome sequencing

Metagenomic DNA preparations from starters M1 and M2, following activation, were sequenced by Macrogen Europe (Amsterdam, The Netherlands) using the NovaSeq 6000 Illumina platform (Illumina, U.S.A.), generating paired-end reads with an average read length of 151 bp. The library was constructed using the TruSeq DNA PCR-Free kit and according to the TruSeq DNA PCR-Free Sample Preparation Guide, Part #15036187 Rev. D (Illumina, U.S.A.). Quality control (QC) was performed prior to library construction. The raw metagenomic reads of M1 and M2 were deposited in GenBank under Biosample accession numbers, SAMN37487907 (BioProject no. PRJNA1019656) and SAMN37493201 (BioProject no. PRJNA1019710), respectively.

(iii): Bioinformatic analysis of metagenomic sequences

Fastq files of paired-end reads were used as the input for genome assemblies using MEGAHIT (version v1.2.9). Assembled contigs were first aligned against the non-redundant (nr) database maintained by NCBI (https://ncbi.nlm.nih.gov), which includes most sequences available in RefSeq and GenBank, using BLASTn (version 2.11.0+; https://blast.ncbi.nlm.nih.gov/Blast.cgi) in order to taxonomically classify each contig using BLAST-based last common ancestor (LCA) taxonomy estimation (21). Reads were aligned against their corresponding assemblies to quantify the number of reads that mapped to each contig using BWA (version 0.7.17-

r1188). The relative abundance of the genus and/or species present in the starter cultures was then calculated by dividing the total number of reads mapped to all metagenome assembled contigs by the number of reads mapping to all contigs classified as a particular species or genus.

The relative abundance of *L. cremoris* and *L. lactis* was also determined through reads mapping of the genes *rpoB (llmg_1982* and *EFV54_RS09530*), *recA (llmg_0374* and *EFV54_RS01915*), and *pheS (llmg_2196* and *EFV54_RS10275*), corresponding to the species-specific representative strains *L. cremoris* MG1363 and *L. lactis* IL1403 using the widely applied 97 % clustering threshold (19). These genes were previously used to elevate *L. cremoris* from a subspecies of *L. lactis* to species level (2). The relative abundance of genotype-specific genes within each *cwps* cluster was also determined through metagenomic reads mapping using CoverM with stringent identity (95 %) and alignment cutoffs (75 %) to limit low quality mapping (22, 23). The following arguments were used: coverm genome --methods rpkm --min-read-percent-identity 95 --min-read-aligned-percent 75. Filtered raw metagenomic reads were mapped against the following conserved ORFs representing CWPS biosynthesis gene clusters A-D and C-subtypes C_1 to C_6: *UC509_0206* (*L. cremoris* UC509.9, representative of CWPS type A), *LLKf_0205* (*L. lactis* KF147, representative of CWPS type B), *llmg_0220* (*L. cremoris* MG1363,CWPS type C_1), *L3107_0192* (*L. cremoris* 3107, CWPS type C_2), *LACR_0215* (*L. cremoris* SK11, CWPS type C_3), *w34_004* (*L. cremoris* W34, CWPS type C_4), *lilo_0174* (*L. lactis* IO-1, CWPS type C_5), *llh_1295* (*L. cremoris* A76, CWPS type C_6), and *LL184_0294* (*L. lactis* 184,

CWPS type D). GenBank accession numbers for all strains listed above are presented in Supplementary Table S1.

(iv): Strain isolation and CWPS typing

Serial dilutions (prepared in Ringers' solution) of the undefined starter cultures M1 and M2 were directly (i.e., without activation, after initial dilution in RSM) plated on LM17 agar plates and incubated at 30 °C for 48 hrs. After incubation, 181 isolates from M1 and 183 from M2 were picked randomly and incubated in LM17 broth-containing polystyrene 96-well plates (Sarstedt, Nümbrecht, Germany) at 30 °C for a further 48 hours. Individual colonies were then plated on LM17 agar and after 24 hours at 30 °C, colonies were typed using colony PCR and previously described *cwps* genotyping (CWPS types A, B and C, including C-subtypes C_1 through to C_5) multiplex primers (15, 18). Additionally, the following primer pairs specific to C_6-type strains were designed: C6-fw 5'- ATATTGGTCAAGGCAAAGCT-3' and C6-rv 5'- TCAAACCAGTTCGTAAATCCT-3'. Primers were based on *llh_1295* (*L. cremoris* A76, CWPS type C_6) with a PCR product of 517 bp. The same PCR conditions were used as previously described (15). 16S rRNA gene amplification was performed on any isolates that did not yield a CWPS-type, as described previously (24). Sanger sequencing of 16S rRNA PCR products was performed by GENEWIZ (Azenta Life Science, Germany).

Phage screening and isolation

(i): Phage screening and phage isolation from whey

A collection of 92 isolates representing the four dominant *cwps* types (23 isolates of each of the C_1-C_4 types; Supplementary Table S3) isolated from two undefined starter cultures (M1 and M2; without activation) were selected for use in phage screening of whey samples. Three whey samples (T1, BD, and BA6) derived from factories that had utilized either one of these two starter cultures and where the potential presence of phages had been reported, were tested for the presence of phages (information on whey samples is presented in Supplementary Table S1). In a 96-well plate (Sarstedt), 20 µl overnight culture was added to a well containing 150 µL LM17 broth supplemented with 10 mM $CaCl_2$ (Sigma Aldrich) and 20 µl of a given whey sample diluted 1:20 in SM buffer [50 mM Tris-HCl, 100 mM NaCl (Sigma Aldrich), 10 mM $MgSO_4$ (Sigma Aldrich), 10 mM $CaCl_2$ (Sigma Aldrich)]. A second plate using 20 µl SM buffer in place of whey was prepared as a positive control for growth. Plates were incubated at 30 °C and the OD_{600nm} was measured at three time points (0 hr, 3 hr, and 6 hr), incubated overnight, and visually inspected for possible phage-induced lysis (by comparison to the uninfected control). Wells in which lysis of the cultures was observed (either through OD measurements during the first 6 hr or complete lysis following overnight incubation), the potential phage was diluted in SM buffer (1:10), filtered (0.45 µM), and spotted on its corresponding host to validate the presence of phages through double layer plaque assays (25). Phages were purified with one round of plaque purification and typed using a previously published 936 (now *Skunavirus*), c2 (now *Ceduovirus*), and P335 PCR-based typing system (26).

(ii): Host range analysis

Two representative phages were chosen for further studies, i.e., a *Skunavirus* (BA6-A5) and a *Ceduovirus* (T1-H4). The host range of two of the isolated phages was determined using the same 92 lactococcal strains used in the original whey-based phage screening. Host-range analysis was performed in liquid culture, where high titre lysates (i.e. phage lysates with titres of ~10^7-10^9 PFU/ml) of the two phages were co-incubated with a given lactococcal strain as described above. To confirm the results of the liquid host range studies, plaque assays were performed using the double-agar method as previously described (25).

Development and validation of a cwps locus-based lactococcal strain qPCR assay

(i): Primer design

Primers were designed to target the CWPS biosynthesis gene cluster genotype/subtypes A, B, C_1-C_6, and D and specifically using previously defined reference strains of each of the *cwps* genotypes mentioned (details of the strains, ORFs, and primers designed are available in Supplementary Table S1, Table S4, and Supplementary Figure S1) (14). For more information on the design and optimisation of the primer pairs see Supplementary Text. All GenBank accession numbers associated with these genomes are listed in Supplementary Table S1. As representative C_7 and C_8 strains are not available in our strain collection, they were not included in the design of this qPCR assay. To design the primers, the target ORFs (listed above) were submitted

to Primer3Plus (https://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) with optimal design settings set to select a primer pair with a melting temperature between 50-60 °C (with 55 °C being optimal) and a product size between 150-250 bp (with 200 bp being optimal) (27). The specificity of each primer set was confirmed using the BLASTn (https://blast.ncbi.nlm.nih.gov/Blast.cgi) search tool. A control primer set was included based on the conserved *groES* gene present in all lactococcal strains (28). The g*roES* sequences of nine representative strains (one of each *cwps* genotype) were retrieved from NCBI and aligned using ClustalW of the Megalign program (Version 5.01, Lasergene software package, DNASTAR Inc., Madison, WI, USA), to identify a conserved region for optimal primer design. The melting temperature of the *groES* primer set was determined using the OligoCalc tool, which was also used to check for primer self-complementarity (29). Similarly, the specificity of the primer set was confirmed using the BLASTn search tool. Primer sequences, expected amplicon sizes, and melting temperatures are presented in Supplementary Table S4. Primers were synthesized by Eurofins Genomics (Ebersberg, Germany).

(ii): Bacterial enumeration and DNA extraction for standard curves

Representative lactococcal strains (Supplementary Table S5) of each *cwps* genotype (A-D) were grown at 30 °C overnight in LM17 broth. Following overnight incubation, cells were enumerated on LM17 agar plates once colonies had grown at 30 °C under aerobic conditions to a sufficient size to be counted (24 – 48 hours). Enumerations were performed in triplicate using the spot plating technique. Concurrently, 2 mL of each of these overnight cultures was removed and centrifuged at 14,000 × *g* for 10 min. The supernatant was

removed, and the pellets were stored at -20 °C until required for DNA extractions, which was performed using the PureLink™ Genomic DNA Mini Kit (Invitrogen, UK), according to the manufacturer's protocol. All DNA samples were stored at -20 °C until required. DNA concentrations were measured using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, M.A., U.S.A.) or a Qubit dsDNA BR assay kit with Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA).

(iii): qPCR protocol

qPCR assays were performed using LightCycler® 480 SYBR Green I Master Mix (Roche Diagnostics, Mannheim, Germany), at a 1X final concentration, on a LightCycler® 480 Instrument II (Roche Diagnostics). Reactions were carried out in a reaction volume of 15 µl containing 5 µl of template DNA (between 50-100 ng per reaction). All primers were used at a final concentration of 625 nM. Thermal cycling conditions incorporated an initial denaturing step at 95 °C for 10 min followed by 45 cycles at 95 °C for 10 s, 55 °C for 15 s and 72 °C for 20 s. A melting-curve analysis was performed at the end of each qPCR assay, with one cycle at 95 °C for 1 min, 40 °C for 1 min, 65 °C for 1 s, and then gradually raised to 97 °C while continuously measuring the fluorescence. All standards and samples were analysed in triplicate.

(iv): Standard curves

DNA was obtained from representative lactococcal strains of each *cwps* genotype tested (as described above; the strains used are listed in Supplementary Table S5) was ten-fold serial diluted (10^0 to 10^{-7}) and used in

the qPCR assay. The cycle threshold (C_T) values obtained from the dilutions of DNA were plotted against the CFU/ml of the strains, as determined by viable plate counts of corresponding serial dilutions of each strain. Technical triplicate assays were performed for both the viable plate counts and qPCR analyses. The correlation coefficient (R^2) and primer efficiency were calculated for each primer set to determine the reliability of the assay (30). Primer efficiencies were calculated according to the equation $E = (10^{\left(-\frac{1}{b}\right)} - 1) \times 100\ \%$, where *b* is the slope of the standard curve. The detection limit for each primer set was defined as the lowest CFU/ml where the linearity of the standard curve was maintained. Melting curves were also created for each primer pair. The primer efficiency, R^2, and limit of detection for each primer pair are presented in Supplementary Table S5.

(v): Primer specificity

The specificity of each primer set was tested against DNA of strains (29 in total) harbouring target and non-target *cwps* gene clusters alongside melting analysis of the qPCR products obtained (30). 50 ng of DNA was used in each reaction, with ND denoting where a C_T value could not be determined at all or due to late C_T calls that have a higher uncertainty (within the last five cycles; Supplementary Table S6). All primer specificity analyses were performed as technical triplicates.

(vi): Quantification of cells present in undefined starter cultures M1 and M2 using qPCR

For absolute quantification of cells by qPCR, DNA was extracted from the M1 and M2 undefined starter cultures by firstly resuspending 5 g of the

lyophilised starter cultures (without activation) in 95 g of 10 % sterilized RSM and stomaching at 230 rpm for 5 mins. Total DNA was extracted and purified from 2.5 mL of the homogenized cultures as described above. 75 ng of genomic DNA from both M1 and M2 was subjected to qPCR with each *cwps* primer pair (Supplementary Table S4). Absolute quantification of cells in the samples was performed by interpolation on a standard regression curve of C_T values generated from genomic DNA extracted from reference strains of each *cwps* type (30). Determination of the slope and y-intercept used in the calculations is described in Supplementary Table S5. The qPCR results are expressed as Log(CFU/ml) for strains harbouring each *cwps* gene cluster tested (30-33). The subsequent relative abundance of the *cwps* types present in the starter cultures was then calculated by dividing the total CFU/mL of all *cwps* types by the CFU/mL corresponding to each *cwps* type.

Compositional impact of phage attacks on M1

(i): Protocol

10 g of lyophilised M1 starter culture was resuspended in 90 g of 10 % sterile RSM (prepared as described above). 20 ml of the first dilution was then transferred to 80 ml of 10 % sterilised RSM as the final dilution. To reflect typical fermentation conditions, 20 ml of pasteurized milk (as above) at pH 6.5 was inoculated with 4 % of the diluted lyophilised M1. Test cultures were additionally inoculated with 1 % of a whey or individual phage (at titres of ~10^7-10^9 PFU/ml). Three whey samples were tested (T1, BD, and BA6). The individual phages (isolated from phage screening of wheys, described in

supplementary text, and chosen at random to represent a *Skunavirus* and *Ceduovirus*) tested were BA6-A5 and T1-H4. The samples were incubated in a water bath at 32 °C for 6 hrs and were analysed in triplicate. After incubation, the pH of the whey-treated samples was measured and compared to an uninfected M1 control to first determine of the wheys and/or phages had any impact of the acidification of the culture.

(ii): Metagenomic DNA extraction of M1 culture following phage exposure

Following pH testing of the samples, 2 ml of each sample was added to 23 ml of 2.2 % trisodium citrate (45 °C) and homogenised in a Stomacher® 400 Circulator (Seward) at 230 bpm for 5 mins (20). The solution was centrifuged at 13,800 × *g* for 10 minutes at room temperature. The supernatant was removed, and the cell pellets stored at -20 °C prior to DNA extraction. The DNA extraction was performed using PureLink™ Genomic DNA Mini Kit (Invitrogen, UK) according to manufacturer's instructions, after an initial DNase treatment (20 units ml^{-1} DNase I recombinant, Sigma Aldrich for 30 minutes at room temperature, followed by DNase inactivation at 75 °C for 10 minutes). The absolute quantification and relative abundance of cells corresponding to the *cwps* genotypes in each sample was calculated as described above (See Quantification of cells present in undefined starter cultures M1 and M2 using qPCR method). Standard deviations and ordinary one-way ANOVA tests were performed, followed by Dunnett's multiple comparisons test, with a single pooled variance.

Data availability

The raw metagenomic reads of starter cultures M1 and M2 have been deposited in GenBank under BioProject accession numbers PRJNA1019656 and PRJNA1019710, respectively.

RESULTS

Microbial composition of two undefined mesophilic starter cultures

Metagenomic assessment of M1 and M2

Genomic DNA of both cultures was extracted following M1/M2 culture activation in pasteurized milk and sequenced (Supplementary Table S7). The assembled contigs from the sequenced metagenomes of M1 and M2 were taxonomically classified using BLASTn-based LCA taxonomy estimation. Then the relative abundance of the species present in these cultures was determined by mapping the raw metagenomic reads against these now taxonomically classified contigs (Figure 1 and Supplementary Table S8). Based on this analysis, both cultures were determined to be primarily composed of *L. cremoris* and/or *L. lactis* strains (>94 %). Due to the high sequence homology between the two species, it was not possible to accurately assign the reads to *L. lactis* or *L. cremoris* through the taxonomic classification of contigs alone. Therefore, the relative proportion of *L. lactis/L. cremoris* in these starter cultures was determined through reads mapping of three genes (*recA, pheS,* and *rpoB)* that have previously been applied in multi-locus sequence typing and differentiation of these two species, although here the 16S rRNA gene was excluded (Supplementary Figure S3) (2). In both starter cultures, *L. cremoris* was found to represent >99 % of the *L. cremoris/L. lactis* population. The next most abundant organisms in the cultures were *Leuconostoc garlicum* (~1.0 %) in M1 and *Lactococcus raffinolactis* (1.6 %) in M2 (Figure 1 and Supplementary Table S8). To further assess the diversity of the *L. cremoris* component of the M1/M2 cultures at the strain level, the raw

metagenomic reads were mapped against the genes specific to nine *cwps* gene cluster responsible for CWPS biosynthesis (A, B, C_1-C_6, and D; see Materials and Methods) (Figure 2 A and D). The *L. cremoris* constituent in M1 was assigned to C_2 (70 %), C_1 (27 %), and C_3 (3 %) *cwps* genotypes (Figure2A), while those present in M2 were shown to belong to the C_2 (64 %), C_3 (30 %), C_4 (5 %), C_1 (1 %), and C_5 (1 %) genotypes (Figure 2D).

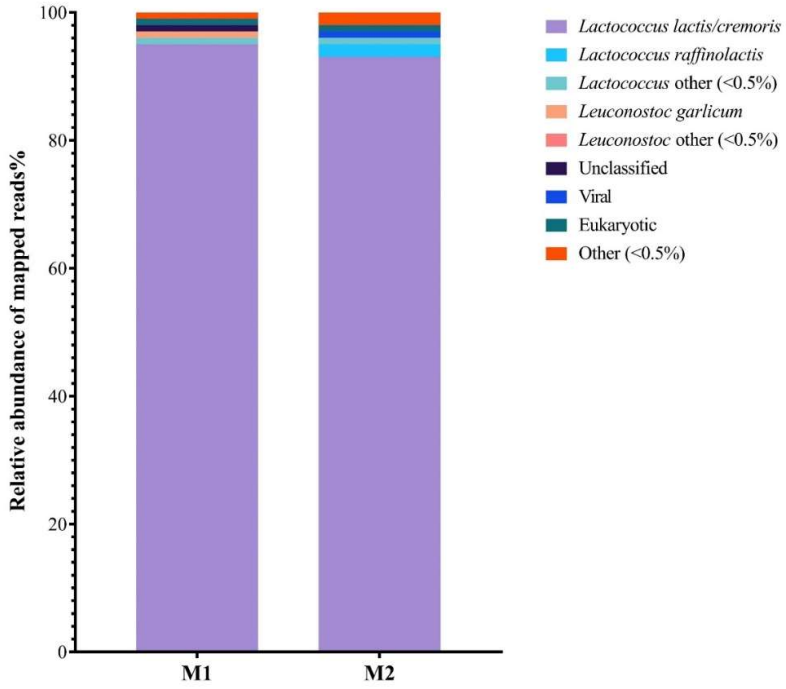


Figure 1. Taxonomic composition of M1 and M2 based on reads mapping of taxonomically classified contigs. Columns indicate the relative abundance of reads corresponding to contigs taxonomically classified as either bacterial (at the genus or species level), eukaryotic, or viral.

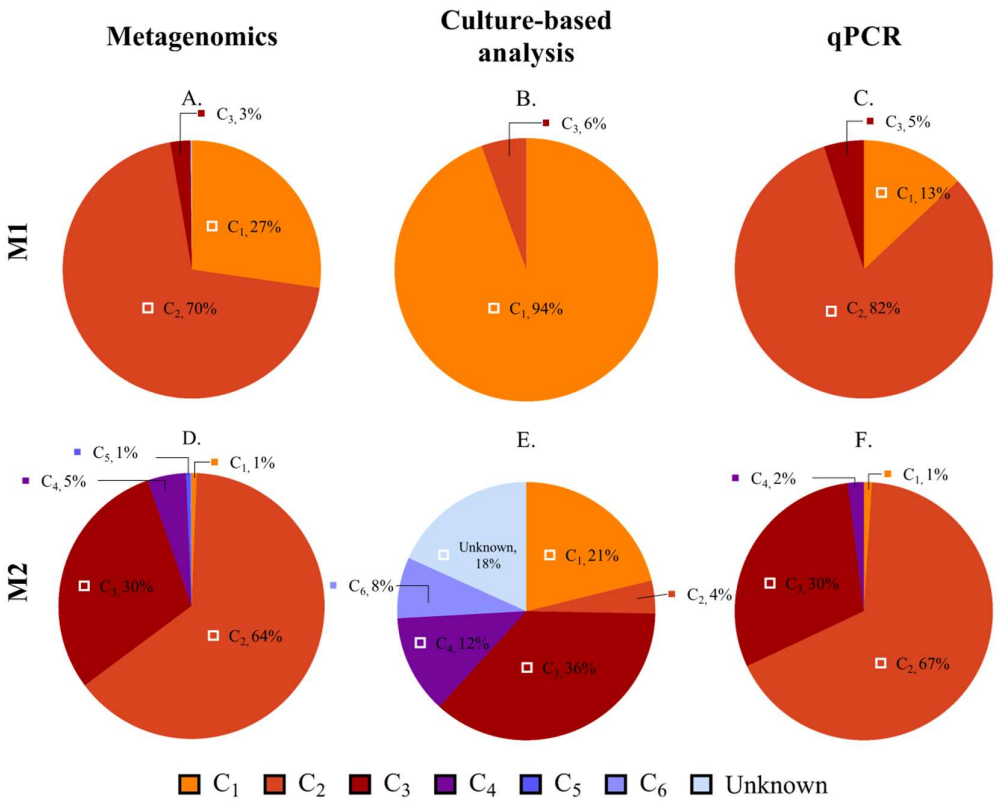


Figure 2. Comparison of culture-dependent and -independent methods of assessing the composition of undefined starter cultures. Comparison of metagenomics (A and D), culture-based analysis (B and E), and qPCR (C and F) as methods to characterise mesophilic undefined starter cultures M1 (A-C) and M2 (D-F) based on the *cwps* gene cluster. CWPS-types of isolates selected from M1 and M2 were typed using previously described multiplex PCR CWPS-typing assays (B and E). Isolates that were unable to be typed by any CWPS type primers (A-C, and C_1-C_5-subtypes) but yielded a product from the lactococcal control primers are indicated as unknown.

Culture-based analysis of M1 and M2

To investigate the composition of the M1 and M2 cultures through cultivation-based methods, isolates were selected from the original starter cultures (i.e., non-activated culture) following incubation on LM17 plates at 30 °C (N.B. there is no notable difference in composition between the activated and non-activated culture, Supplementary Figure S4). A total of 364 isolates

(181 isolates from M1 and 183 isolates from M2) were collected and typed by multiplex PCR (15, 18)) to determine the *cwps* genotype of the individual isolates. Of the 181 isolates retrieved from M1, 171 were C_1-type and ten were C_3-type strains (Figure 2B). Of the 183 isolates selected from M2, 36 were C_1-type, seven were C_2-type, 62 were C_3-type, 21 were C_4-type, and 13 C_6-type strains (Figure 2E). We were unable to type the remaining 44 isolates using the primers available in the multiplex typing assays, however 31 of these 44 isolates were determined to be *L. lactis/L. cremoris* strains (based on amplification of the lactococcal control, *rmlB*). The *cwps* gene clusters of these isolates are potentially novel and will be investigated in future studies. The species of the remaining 13 isolates were determined to be *Lactococcus laudensis* strains through 16S rRNA gene sequencing.

qPCR assay development

Set-up and validation of qPCR method

To develop a methodology to monitor the composition of mesophilic starter cultures and their typically dominant *L. lactis/cremoris* component, we developed a qPCR assay with primer pairs capable of differentiating nine industrially relevant strain genotypes based on distinct *cwps* gene clusters (A, B, C_1-C_6, and D) across two species (*L. lactis* and *L. cremoris*; Supplementary Table S4). Information on the validation and specificity of these primer pairs can be found in Supplementary Text. These primers can distinguish between nine distinct genotypes when assessed against a test panel of 29 lactococcal strains (Supplementary Table S6). In addition, the *groES* gene was shown to

be a highly reliable target for quantifying the total number of lactococcal cells present in a sample.

Application of a qPCR method for M1/M2 cwps-based lactococcal strain monitoring

The qPCR approach to quantify various lactococcal *cwps* genotypes was employed to assess the lactococcal-based undefined starter cultures M1 and M2 without activation. C_T values obtained from the individual *cwps* primer pairs were determined to be reliable if they were above the individual threshold of the primers (determined by the limit of detection and primer specificity analyses in complex medium LM17) and produced the expected melting curves. Based on the obtained results, both M1 and M2 were established to be composed of C-type strains only (C_T values obtained from the A, B, and D-type primer pairs exceeded the designated threshold). Based on the qPCR assay M1 was shown to be primarily composed of C_2 (82 %), C_1 (13 %), and C_3 (5 %) type strains (Figure 2C), whereas M2 contains C_2 (67 %), C_3 (30 %), and C_4 (2 %) type strains (Figure 2F). In addition, the total number of cells based on the *cwps* genotypes was quantified for M1 and M2 to be 8.65 ± 0.15 and 8.90 ± 0.11 log CFU/ml, respectively (Figure 3). These values were comparable to the total cell counts based on the *groES* primers 8.60 ± 0.15 log CFU/ml in M1 and 8.62 ± 0.12 log CFU/ml in M2. Therefore, the total CFU/ml of the combined *cwps* type strains quantified (based on the described *cwps* primers) was comparable to the CFU/ml quantified from primers targeting the house-keeping gene *groES,* indicating that the vast majority of lactococcal strains were typed and quantified (Figure 3). Total counts calculated using both the *cwps* and *groES* primers were also similar to the cell

counts obtained when the cultures were plated on LM17: 8.78 ± 0.10 log CFU/ml in the case of starter M1 and 7.18 ± 0.06 log CFU/ml in the case of starter M2. A notable limitation of the current methodology outlined here is that standard curves were based on DNA extracted from strains grown in complex medium LM17 and serial dilutions of the cultures plated on the same medium, whereas the starter cultures were grown in milk. It is recommended that standard curves should be prepared using the same matrix as the samples, where possible, which in this case would be milk (33). However, since some strains used in generating the standard curves were capable of growing well (or in some cases, at all) in milk, this was not possible in this case.

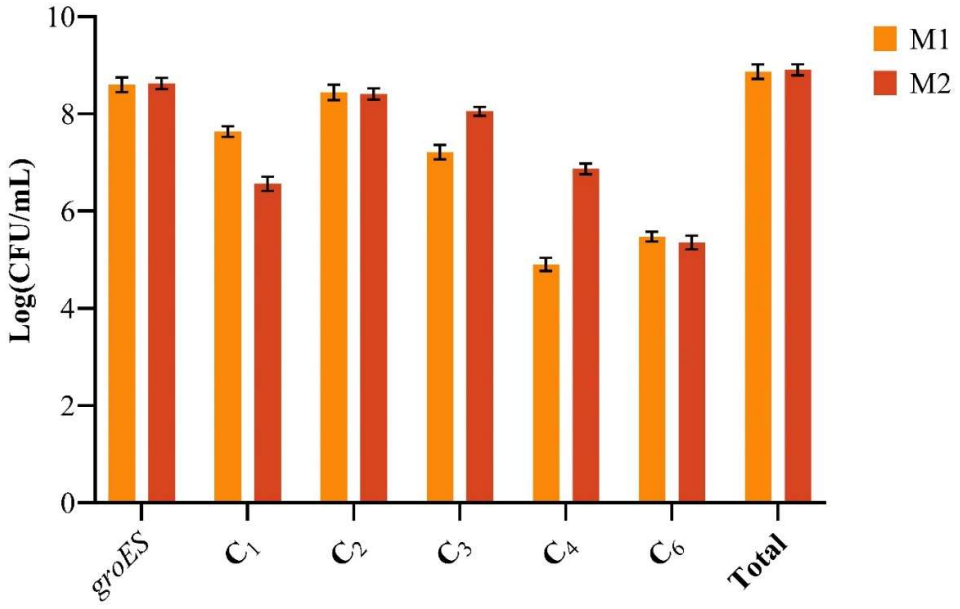


Figure 3. Absolute quantification of lactococcal strains of M1 and M2 using a cwps-based qPCR assay. Quantification of lactococcal strains grouped based on their cwps-types expressed as the estimated log CFU/mL (based on standard curves of each primer set against a representative strain).

Compositional changes caused by phage attack can be detected by the qPCR assay

The applicability of the qPCR approach to monitor culture compositional changes (of starter culture M1) following phage attack was evaluated using (a) industry-derived phage-containing whey samples and (b) individual phages that are known to infect at least one strain originating from M1. To this end, the composition of M1 following incubation with phage(s) from either whey samples originating from factories where fermentation (milk acidification) had slowed (T1, BA6, BD) or individual phages isolated from phage screening of whey samples (BA6-A5 and T1-H4), results of phage screening of whey can be found in Supplementary Text and Supplementary Figure S5) was quantified after 6 hrs of fermentation using the herein established qPCR assay and compared to the uninfected M1 control, to determine what (if any) compositional changes occurred as a result of phage attack. To determine if a significant change in composition occurred, ordinary one-way ANOVA tests followed by Dunnett's multiple comparisons tests, with a single pooled variance, were applied to determine if there was a significant difference between the cell counts observed for each primer pair (Figure 4).

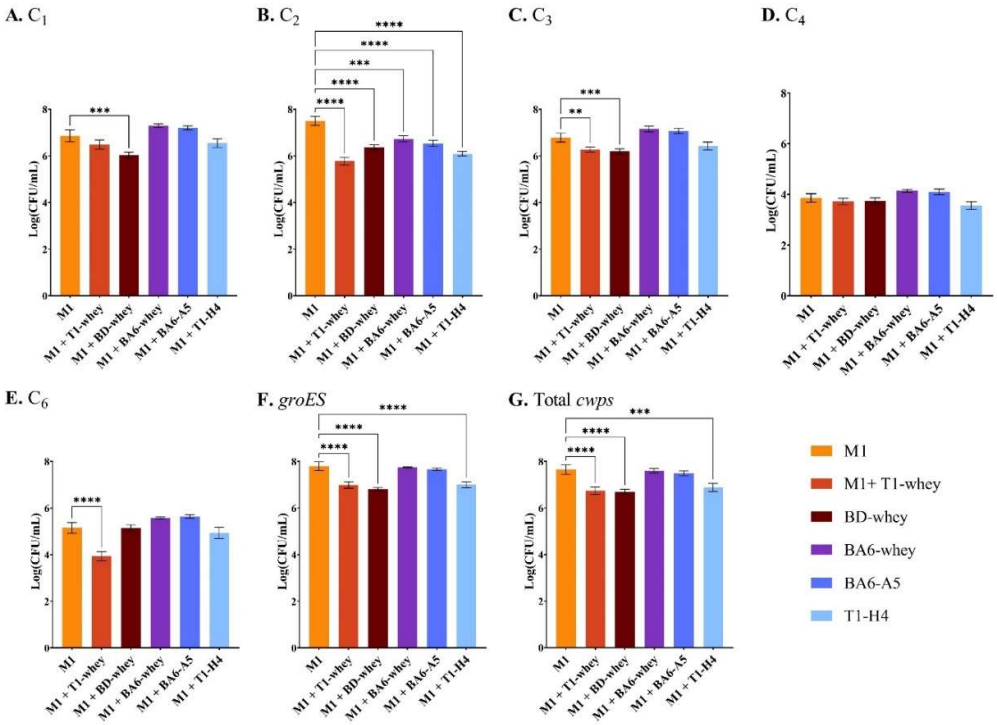


Figure 4. Absolute quantification of mesophilic undefined starter culture M1 following phage attack compared to an uninfected control using a *cwps*-based qPCR assay. Absolute abundance of *cwps* types present in M1 after 6 hr of normal fermentation conditions under phage pressure from individual phages (BA6-A5 and T1-H4) or phages present in whey samples (T1, BD, and BA6) compared to an uninfected M1 control. C_T values obtained from the primers corresponding to *cwps* types A, B, D, and C_5 strains were below the detection level and therefore excluded from this graph. A significant change (p-value < 0.01) in composition in a sample compared to the uninfected M1 control in a sample is indicated by star(s). Two stars indicates a p-value < 0.01, three stars a p-value < 0.001, and four star a p-value <0.0001.

All three whey samples tested, i.e., T1, BA6, BD inhibited acidification to some degree, with pH unit differences ranging from 0.44 to 1.07 compared to the uninfected control (Supplementary Figure S6). In these factories, the undefined starter culture(s) typically contribute(s) to the acidification of milk ultimately leading to the production of kefir (T1), a Polish curd cheese known as Twaróg (BA6), and a semi-hard cheese (BD). After 6 hr of fermentation in pasteurized milk, M1 reached a pH of 4.63 ± 0.06, whereas M1 starter to which

the phage-containing whey samples T1, BD, and BA6 had been added only reached pH values of 5.70 ± 0.27, 5.47 ± 0.12, and 5.07 ± 0.06, respectively. This indicates the presence of phages in these whey samples that clearly negatively impacted on the fermentation/acidification ability of M1. In addition to whey samples containing unknown phage levels and composition, two individual phages were tested. Phages BA6-A5 and T1-H4 were shown to significantly inhibit the fermentation ability of M1 (p-value < 0.01), with associated pH values of 5.10 ± 0.00 and 5.00 ± 0.00 after the 6-hr incubation period, compared to the uninfected M1 control reaching a pH of 4.63 ± 0.06 (Supplementary Figure S6).

Microbial DNA from the uninfected M1 starter culture and all whey/phage infected cultures was extracted and used as a template in the qPCR assay with all *cwps* and *groES* control primer pairs and used to quantify the CFU/ml of strains corresponding to each *cwps* type present in the samples (Figure 4). After 6 hr of fermentation in pasteurized milk, M1 was established to contain C_1 (6.86 ± 0.26 log CFU/ml), C_2 (7.50 ± 0.19 log CFU/ml), C_3 (6.78 ± 0.19 log CFU/ml), C_4 (3.86 ± 0.17 log CFU/ml), and C_6 (5.15 ± 0.23 log CFU/ml) -type strains, and with all other assessed genotypes being below the detection limit. Exposure of the M1 starter culture to phages present in whey sample T1 resulted in a significant decrease in strains with the *cwps* types C_2 (p-value < 0.0001) and C_6 (p-value < 0.0001), with final cell counts after phage infection of 5.78 ± 0.17 and 3.93 ± 0.20 log CFU/ml, respectively. Phages in whey BD resulted in significant decreases of M1-associated starter strains with *cwps* types C_1 (p-value < 0.001, 6.03 ± 0.12 log CFU/ml) and C_2 (p-value < 0.0001, 6.36 ± 0.11 log CFU/ml). Phages in whey BA6 resulted in significant

decreases (p-value < 0.001) of C_2 *cwps* type strains (6.73 ± 0.13 log CFU/ml). Infection by *Skunavirus* BA6-A5 and *Ceduovirus* T1-H4 resulted in a significant (p-value < 0.0001) decrease of C_2 type strains (6.53 ± 0.13 log CFU/ml and 6.09 ± 0.10 log CFU/ml, respectively). These changes in the absolute quantities of each *cwps* type resulted in shifts in the relative abundance of all *cwps* types present in M1 (Figure 5).

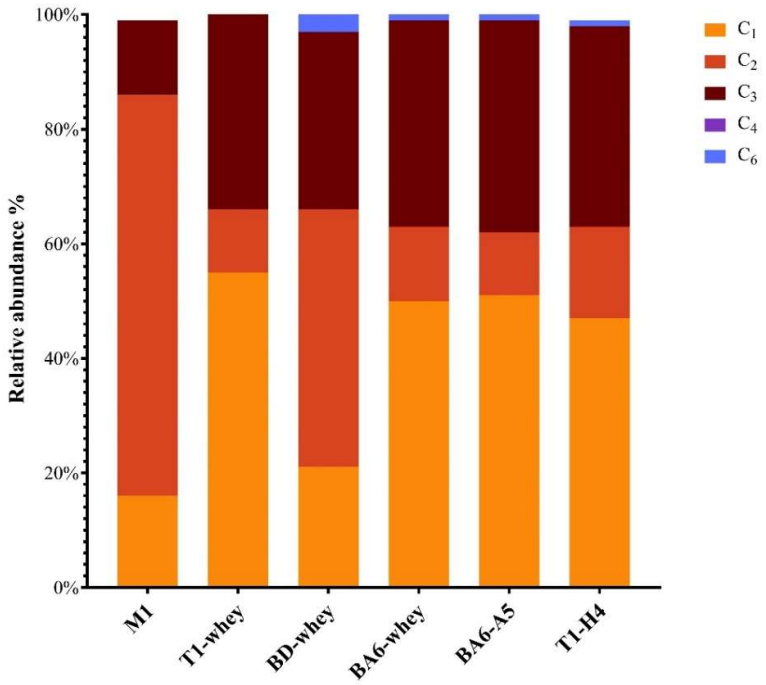


Figure 5. Relative abundance of *cwps* types present in mesophilic undefined starter culture M1 following phage attack compared to an uninfected control using *cwps* based qPCR assay. The relative abundance of *cwps* type strains present in the samples demonstrating the overall compositional changes in M1 as a result of phage infection. C_T values obtained from the primers corresponding to *cwps* types A, B, D, and C_5 strains were below the detection level and therefore excluded from this graph.

DISCUSSION

L. lactis and *L. cremoris* strains are mesophilic LAB widely used in dairy fermentation. To ensure consistent product quality, it is essential that the starter cultures are robust and perform reliably. Consequently, both culture-dependent and -independent approaches have been applied to explore the strain composition of a number of undefined starter cultures used in the production of Swiss hard cheeses (34), Gouda (20), and continental cheeses (19). In the present study, the microbial composition of two undefined industrial starter cultures, named M1 and M2, commonly used in the production of semi hard cheeses, fermented milk (e.g., viili, kefir, and twaróg) and fresh cheeses (e.g., quark) was evaluated using both culture-dependent and -independent methods. These microbial explorations led to the establishment of a qPCR assay (based on lactococcal *cwps* gene clusters) capable of discerning nine industrially relevant lactococcal genotypes, with the potential to be widely applicable, and thereby allowing for a better understanding of *L. lactis/L. cremoris* diversity in mesophilic starter cultures and ultimately aid in the implementation of effective starter culture rotation strategies.

Culture-dependent methods have historically been the most frequently used method to study complex microbial communities (35), although in recent years culture-independent methods such as metagenomics has become a more widespread application in the study of fermented foods (36). The benefits of metagenomics in fermented foods are particularly apparent in the study of traditional/artisanal fermented foods, undefined starter cultures, or spontaneous fermentation (19, 20, 34, 37–39). In these cases, the microbiomes are largely uncharacterised, rendering quality control and

product consistency a challenge. Metagenomics have allowed these complex communities to be characterised without the limitations associated with purely culture-dependent characterisation studies. In the current study, the culture-based profiles of M1 and M2 appeared to differ greatly from the compositions determined through metagenomic reads mapping of *cwps* types. For example, C_2 strains were determined in this manner to be the most dominant type in the M1 culture, whereas no C_2 strains were identified from the 181 isolates obtained from starter culture M1. Instead, the vast majority (94%) were C_1-type strains, and the remaining were C_3-type (6%). In M2, a similar trend was observed where C_2-type strains were expected to make up 70% of the culture (based on metagenomics), yet just seven out of the 170 *L. cremoris/L. lactis* isolates typed were C_2-type strains. These differences are likely to be due to cultivation conditions (LM17 instead of RSM and plating at 30 °C) or the presence of viable but non-culturable (VBNC) cells (30). On the other hand, culture-based analysis of M2 led to the isolation of several *L. laudensis* isolates as well as lactococcal isolates with a *cwps* gene cluster that could not be assigned to any of the currently known CWPS types, neither of which were identified by metagenomic analysis. In addition, the composition of a sample can only be reported in terms of relative abundance rather than absolute quantifications in purely metagenomic-based analyses (40). Furthermore, in order to evaluate the role and functionality of both dominant and minor components of these starter cultures it is necessary to isolate and characterise individual strains through culture-based methods. Therefore, these differences highlight the importance of a combined approach when evaluating the composition of undefined starter cultures.

In recent years, qPCR assays have been used extensively to track and characterise the microbiota of many complex cultures and food products (30, 41–44). This method is faster and more sensitive than many culture-dependent methods which can often be time consuming and limited by the cultivation conditions (medium composition, cultivation temperature, etc.). In addition, as shown above, culture-based analysis may not accurately reflect the true composition of the culture, as it does not take into account VBNC cells and is limited by cultivation conditions. Furthermore, where metagenomics and cultivation-based approaches typically provide a one-dimensional or 'snapshot' view of a culture under specific conditions, qPCR assays can be used to actively monitor the composition of a culture as a function of time and under varying conditions. However, these assays are often limited to defining the species/subspecies level abundance of cultures and provide little information regarding strain-level diversity or are specific to only a certain culture (20, 30, 41, 42). The majority of currently described LAB-based qPCR assays are limited to quantifying and monitoring certain species (*L. lactis, Streptococcus thermophilus, Lacticaseibacillus paracasei, Lactococcus garvieae,* etc.) during fermentation (30–32, 41, 45). However, a qPCR assay established by Erkus *et al*. went beyond the species level to monitor the population dynamics of seven genetic lineages of *L. lactis/cremoris* and *L. mesenteroides* strains in an undefined Gouda cheese starter culture (20). One limitation of this assay is that it was found to be quite culture-specific: in a subsequent study on undefined mesophilic mixed (DL) starter cultures used in the production of similar continental cheeses, 19 of their 127 *L. cremoris* isolates were found to contain gene markers specific to just one of the genetic

lineages established by Erkus (19). This indicates that the qPCR assay established by Erkus is highly culture-specific as it is based on genetic lineages present in starter culture Ur and may not be generally applied to other mesophilic starter culture systems. The qPCR-based assay described in the present study utilises sequence diversity within the variable *cwps* gene cluster which appears to be present in all known *L. lactis* and *L. cremoris* strains and therefore may be more widely applicable to mesophilic starter cultures in general. The *cwps* gene cluster is an ideal target as CWPS is a technologically relevant attribute of these species as it acts as a receptor for many lactococcal phages. Therefore, this qPCR assay can be useful both as a classification and risk evaluation tool from the perspective of defining what are likely to be phage-unrelated cultures, which is essential when implementing effective starter culture rotation strategies. The applicability of this assay was indeed demonstrated through the characterisation and quantification of *cwps* type lactococcal strains present in two undefined mesophilic starter cultures (M1 and M2). Overall, the cell numbers observed in both culture-based and qPCR-based analyses were comparable (within one log) for both cultures. It is possible that the subtle differences in cell counts are attributed to culturing conditions, the presence of VBNC cells, or dead cells. While the possible inclusion of dead cells is a limitation of both metagenomic and qPCR-based analysis, since the cells were pelleted and treated with DNase prior to DNA extraction, it is likely that DNA from lysed cells is eliminated before the DNA was extracted (30). Future improvements of these assays could include a live/dead evaluation step. Alternatively, an advantage of culture-independent analyses is that VBNC cells are still likely to be enumerated through both

qPCR and metagenomics, but not in culture-based assays (33). Also of note, the composition of the two starter cultures as determined through the herein developed qPCR assay is quite similar to that determined by metagenomics, in contrast to the findings of the culture-based analysis. This makes sense as both are culture-independent methods that are not biased by cultivation conditions and capable of enumerating both culturable and non-culturable cells.

It was also demonstrated that this qPCR assay may be used to monitor compositional changes of a mesophilic culture following phage attack. The compositional changes observed in these phage attacks correspond well to phage screening results from the whey samples and host ranges of the individual phages (Supplementary Text). For example, a significant drop in C_1 and C_2-type strains was observed in M1 following phage attack from whey BD. Phage screening identified several C_1, C_2, C_3, and C_4- type isolates that were sensitive to phage(s) present in the BD whey sample (all of which were typed as Skunaviruses). Since C_3 and C_4-type strains represent minor components of the M1 starter culture it seems logical that the C_1- and C_2-type strains were most impacted during the phage attacks. In addition, infection from phage(s) present in the T1 and BA6-wheys (as well as from individual phages BA6-A5 and T1-H4) decreases in the absolute quantities of certain cwps-type strains (C_2-types) following phage attack resulted in the increase in the relative abundance of other *cwps* type (C_1 and C_3-types). The PSP component of a host's CWPS is a major factor in determining if a phage can bind to a host cell and potentially infect and lyse a strain, therefore awareness of the CWPS type

or subtype of strains within a starter culture provides improved predictions of the risk and potential impact of phage infection (15, 18).

CONCLUSION

By investigating the composition of two undefined mesophilic starter cultures through metagenomics and culture-based analysis we demonstrated the importance of employing both culture-dependent and -independent methods when defining complex systems. In addition, a qPCR assay was established that is capable of discerning nine *cwps* genotypes of *L. lactis* and *L. cremoris* strains. Both metagenomics and qPCR-based methods were shown to be quite comparable, whereas the compositions of the two cultures as determined by culture-based methods clearly differed from culture-independent methods which can be attributed to differences in culturing conditions and the likely presence of VBNC cells. The added benefit and usefulness of the established qPCR assay to not only characterise, but actively monitor mesophilic undefined starter cultures was demonstrated by assessing culture composition following various phage attacks (using whey samples and individual phages). The results confirmed the robustness and flexibility of the undefined mesophilic cultures. The developed qPCR assay is expected to be widely applicable to mesophilic starter cultures, allowing for the quantification of lactococcal cells and specific genotypes as well as the identification of phage-susceptible components of a given culture, which made aid in efficient starter culture rotation practise. The *cwps* gene clusters of lactococcal strains are constantly evolving, likely because of continuous pressure from phage predation. Therefore, this qPCR assay is expected to be expanded or further fine-tuned for future purposes. Furthermore, as increasing sequence information becomes available for *L. raffinolactis, L. laudensis,* and *Leuconostoc mesenteroides/pseudomesenteroides* strains, it may be possible

to develop tailored and species-specific qPCR assays to evaluate (strain) diversity of these (and other) LAB species in mesophilic starter culture systems.

REFERENCES

1. Oliveira J, Mahony J, Hanemaaijer L, Kouwen TRHM, van Sinderen D. 2018. Biodiversity of bacteriophages infecting *Lactococcus lactis* starter cultures. J Dairy Sci **101**:96–105.
2. Li TT, Tian WL, Gu CT. 2021. Elevation of *Lactococcus lactis* subsp. *cremoris* to the species level as *Lactococcus cremoris* sp. nov. and transfer of *Lactococcus lactis* subsp. *tructae* to *Lactococcus cremoris* as *Lactococcus cremoris* subsp. *tructae* comb. nov. *Int J Syst Evol Microbiol* **71**:004727.
3. Daly C, Fitzgerald GF, Davis R. 1996. Biotechnology of lactic acid bacteria with special reference to bacteriophage resistance. *Antonie Van Leeuwenhoek* **70**:99–110.
4. McGrath S, Fitzgerald GF, van Sinderen D. 2007. Bacteriophages in dairy products: pros and cons. *Biotechnol J* **2**:450–455.
5. Smid EJ, Erkus O, Spus M, Wolkers-Rooijackers JCM, Alexeeva S, Kleerebezem M. 2014. Functional implications of the microbial community structure of undefined mesophilic starter cultures. *Microb Cell Fact* **13**:1–9.
6. de Melo AG, Rousseau GM, Tremblay DM, Labrie SJ, Moineau S. 2020. DNA tandem repeats contribute to the genetic diversity of *Brevibacterium aurantiacum* phages. *Environ Microbiol* **22**:3413–3428.
7. Mahony J, Murphy J, Van Sinderen D. 2012. Lactococcal 936-type phages and dairy fermentation problems: from detection to evolution and prevention. *Front Microbiol* **3**:1–9.

8. Chmielewska-Jeznach M, Bardowski JK, Szczepankowska AK. 2020. *Lactococcus Ceduovirus* phages isolated from industrial dairy plants—from physiological to genomic analyses. *Viruses* **12**:280.
9. Szczepańska AK, Hejnowicz MS, Kołakowski P, Bardowski J. 2007. Biodiversity of *Lactococcus lactis* bacteriophages in Polish dairy environment. *Acta Biochim Pol* **54**:151–158.
10. Murphy J, Royer B, Mahony J, Hoyles L, Heller K, Neve H, Bonestroo M, Nauta A, van Sinderen D. 2013. Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type cheese-producing plants. *J Dairy Sci* **96**:4945–4957.
11. Vinga I, Baptista C, Auzat I, Petipas I, Lurz R, Tavares P, Santos MA, São-José C. 2012. Role of bacteriophage SPP1 tail spike protein gp21 on host cell receptor binding and trigger of phage DNA ejection. *Mol Microbiol* **83**:289–303.
12. Dunne M, Hupfeld M, Klumpp J, Loessner MJ. 2018. Molecular basis of bacterial host interactions by gram-positive targeting bacteriophages. *Viruses* **10**:397.
13. Hayes S, Vincentelli R, Ramond L, Ramnd R, Nauta A, Sinderen D van, Cambillau C. 2019. Ubiquitous carbohydrate binding modules decorate 936 lactococcal siphophage virions. *Viruses* **11**:631.
14. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, Chapot-Chartier M-P, Cambillau C, Holo H, van Sinderen D. 2020. The CWPS Rubik's cube: linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. *Mol Microbiol* **114**:582–596.

15. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, Van Sinderen D. 2014. Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. *mBio* **5**:e00880-14.
16. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. 2020. Dairy lactococcal and streptococcal phage–host interactions: An industrial perspective in an evolving phage landscape. *FEMS Microbiol Rev* **44**:909–932.
17. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, Van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci Rep* **6**:21345.
18. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C, Vogensen FK, Van Sinderen D. 2013. Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol* **79**:4385–4392.
19. Frantzen CA, Kleppen HP, Holo H. 2018. *Lactococcus lactis* diversity in undefined mixed dairy starter cultures as revealed by comparative genome analyses and targeted amplicon sequencing of epsD. *Appl Environ Microbiol* **84**:e02199-17.
20. Erkus O, De Jager VCL, Spus M, Van Alen-Boerrigter IJ, Van Rijswijck IMH, Hazelwood L, Janssen PWM, Van Hijum SAFT, Kleerebezem M,

Smid EJ. 2013. Multifactorial diversity sustains microbial community stability. *ISME J* **7**:2126–2136.

21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–410.
22. Li Z, Pan D, Wei G, Pi W, Zhang C, Wang JH, Peng Y, Zhang L, Wang Y, Hubert CRJ, Dong X. 2021. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J* **15**:2366–2378.
23. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M. 2021. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat Commun* **12**:2009.
24. Parlindungan E, Lugli GA, Ventura M, van Sinderen D, Mahony J. 2021. Lactic acid bacteria diversity and characterization of probiotic candidates in fermented meats. *Foods* **10**:1519.
25. Lillehaug D. 1997. An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. *J Appl Microbiol* **83**:85–90.
26. Labrie S, Moineau S. 2000. Multiplex PCR for Detection and Identification of Lactococcal Bacteriophages. *Appl Environ Microbiol* **66**:987–994.

27. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**:W71-4.
28. Mahony J, McGrath S, Fitzgerald GF, Van Sinderen D. 2008. Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. *Appl Environ Microbiol* **74**:6206–6215.
29. Kibbe WA. 2007. OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res* **35**:W43–W46.
30. Achilleos C, Berthier F. 2013. Quantitative PCR for the specific quantification of *Lactococcus lactis* and *Lactobacillus paracasei* and its interest for *Lactococcus lactis* in cheese samples. *Food Microbiol* **36**:286–295.
31. Jung MY, Chang Y -H., Kim W. 2010. A real-time PCR assay for detection and quantification of *Lactococcus garvieae*. *J Appl Microbiol* **108**:1694–1701.
32. Ruggirello M, Dolci P, Cocolin L. 2014. Detection and Viability of Lactococcus lactis throughout Cheese Ripening. *PLoS One* **9**:e114280.
33. Postollec F, Falentin H, Pavan S, Combrisson J, Sohier D. 2011. Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol* **28**:848–861.
34. Somerville V, Berthoud H, Schmidt RS, Bachmann H-P, Meng YH, Fuchsmann P, von Ah U, Engel P. 2022. Functional strain redundancy

and persistent phage infection in Swiss hard cheese starter cultures. *ISME J* **16**:388–399.

35. Bissonnette F, Labrie S, Deveau H, Lamoureux M, Moineau S. 2000. Characterization of mesophilic mixed starter cultures used for the manufacture of aged cheddar cheese. *J Dairy Sci* **83**:620–627.
36. Mayo B, Rachid CTCC, Alegría Á, Leite AMO, Peixoto RS, Delgado S. 2014. Impact of Next Generation Sequencing Techniques in Food Microbiology. *Curr Genomics* **15**:293–309.
37. Jung JY, Lee SH, Kim JM, Park MS, Bae JW, Hahn Y, Madsen EL, Jeon CO. 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. *Appl Environ Microbiol* **77**:2264–2274.
38. Ferrocino I, Bellio A, Giordano M, Macori G, Romano A, Rantsiou K, Decastelli L, Cocolin L. 2018. Shotgun metagenomics and volatilome profile of the microbiota of fermented sausages. *Appl Environ Microbiol* **84**:e02120-17.
39. Zotta T, Ricciardi A, Condelli N, Parente E. 2022. Metataxonomic and metagenomic approaches for the study of undefined strain starters for cheese manufacture. *Crit Rev Food Sci Nutr* **62**:3898–3912.
40. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**:833–844.
41. Furet JP, Quénée P, Tailliez P. 2004. Molecular quantification of lactic acid bacteria in fermented milk products using real-time quantitative PCR. *Int J Food Microbiol* **97**:197–207.

42. Yu J, Wang HM, Zha MS, Qing YT, Bai N, Ren Y, Xi XX, Liu WJ, Menghe BLG, Zhang HP. 2015. Molecular identification and quantification of lactic acid bacteria in traditional fermented dairy foods of Russia. *J Dairy Sci* **98**:5143–5154.
43. Friedrich U, Lenke J. 2006. Improved enumeration of lactic acid bacteria in mesophilic dairy starter cultures by using multiplex quantitative real-time PCR and flow cytometry-fluorescence in situ hybridization. *Appl Environ Microbiol* **72**:4163–4171.
44. Bottari B, Agrimonti C, Gatti M, Neviani E, Marmiroli N. 2013. Development of a multiplex real time PCR to detect thermophilic lactic acid bacteria in natural whey starters. *Int J Food Microbiol* **160**:290–297.
45. Fan X, Li X, Zhang T, Xu J, Shi Z, Wu Z, Wu J, Pan D, Du L. 2021. A novel qPCR method for the detection of lactic acid bacteria in fermented milk. *Foods* **10**:3066.

SUPPLEMENTARY MATERIALS

Supplementary Text.

Validation and specificity of primers for qPCR assays

A qPCR system was established in order to characterise and quantitatively monitor specific strains/components of mesophilic undefined starter cultures based on the lactococcal *cwps* gene clusters, thereby capable of discerning nine distinct lactococcal genotypes. Primers were designed against a gene target within the region of the *cwps* gene cluster responsible for the biosynthesis of the PSP, that were unique to a particular *cwps* (sub)type while being conserved among members of said *cwps* (sub)type (the genes upon which the qPCR primers are based are indicated in Supplementary Figure S1). In most cases, the primers were designed against the same ORFs used in the previously described *cwps* genotyping multiplex primers, where applicable. In the case of CWPS type A where a suitable primer set could not be made from this ORF (*UC509_0206*) an alternative ORF (*UC509_0205*) which is still shared among A-type strains and divergent from other *cwps* types was selected. In the case of CWPS types D and C_6, which had not been included in the multiplex PCR system, a gene which is shared among strains of the same type and different from other *cwps* types was used. To ensure primer sets were well designed and reaction conditions were optimal, the amplification efficiency of each primer was calculated and determined to be within the accepted range of 90 to 110 % (1, 2). Primer efficiency curves were created for each primer pair (specific to each of the nine *cwps* genotypes and a *groES*-targeting control) against a single strain representing its particular target *cwps* type (Supplementary Figure S2). R^2 values were found to be within

an acceptable range of 0.9892 to 0.9999 for each primer set, demonstrating linearity of the data. Primer efficiencies ranged from 94.3 to 105.4 %. A limit of detection, using a pure culture, was determined for each primer set and ranged from 8.7 to 37.8 CFU/reaction (RXN).

The specificity of each primer set for their target *cwps* gene cluster was assessed for 29 lactococcal strains of varying *cwps* genotypes (six A-type, four B-type, four C_1-type, three C_2-type, two C_3 type, five C_4 type, one C_5, three C_6, and one D-type strains) using 50 ng of target or non-target DNA in each reaction (Table 5). Each primer pair displayed varying degrees of specificity to their target *cwps*-type strains, but overall, the primers were demonstrated to be highly specific. Typically, the target strains had lower C_T values ranging from 10.92 to 14.53. While the C_T values resulting from non-target strains varied more depending on the different primer pairs employed, the corresponding C_T values typically exceeded 20.88. The lowest C_T value obtained for a non-target strain for each primer pair was used as a cut-off, any C_T value below this cut-off was considered unreliable. In addition, using *groES* primers (used as a positive control and to enumerate total cell numbers) all 31 lactococcal strains yielded C_T values ranging between 11.03 and 15.88.

Determining the composition of starter cultures when activated vs not activated

In order to establish if there is a notable difference between the composition of starter cultures M1 and M2 when activated or not activated, genomic DNA of both cultures were extracted following culture activation or without activation (as previously described). The metagenomic DNA

preparations was then sequenced by GenProbio srl (Parma, Italy) using a NextSeq platform (Illumina, UK). According to the manufacturer's instructions, DNA library preparation was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA, USA). One ng input DNA from each sample was used for library preparation. The isolated DNA underwent fragmentation, adapter ligation, and amplification. The ready-to-go libraries were pooled equimolarly, denatured, and diluted to a sequencing concentration of 2 pM. Sequencing was performed on NextSeq 550 instrument (Illumina, San Diego, CA, USA), using the 2x150 bp High Output sequencing kit and spike-in of 1 % PhiX control library. The raw metagenomic reads of all four metagenomes were deposited in GenBank under the following Biosample accession numbers: M1_activated (SAMN37487908), M1_not-activated (SAMN37487909), M2_activated (SAMN37493202), and M2_not-activated (SAMN37493203). All M1 and M2-related sequences can be found under BioProject no. PRJNA1019656 and BioProject no. PRJNA1019710, respectively.

Short reads were filtered through the fastq-mcf script (https://github.com/ExpressionAnalysis/ea-utils). Polished contigs were then processed by MEGAnnotator (3) for the prediction of protein-encoding open reading frames (ORFs) using Prodigal (4). Predicted ORFs were functionally annotated by means of BLASTp using the software Diamond (5) performed against the NCBI nr database coupled with hidden Markov model profile (HMM) searches using InterProScan (6) performed against the manually curated Pfam-A database. Furthermore, tRNA genes were identified using tRNAscan-SE version 2.0 (7), while rRNA genes were detected using Barrnap

(https://github.com/tseemann/barrnap). The relative abundance of genotype-specific genes within each *cwps* clusters was determined through both qPCR and metagenomic reads mapping, as previously described.

Phage screening and phage-host range determination

Three whey samples from factories which were expected to contain phages because these samples corresponded to reduced acidification events during the fermentation process, were screened against a panel of 92 isolates originating from M1 and M2 (Supplementary Table S8) and representing the four most dominant *cwps* types present in these two undefined cultures (C_1-C_4). Overall, 22 of the 92 isolates were infected by phages (Supplementary Figure S5). Whey sample T1 was shown to elicit phage activity against a single C_1 isolate (M2-originated) and three C_2 M1-originated isolates. Whey sample BA6 was shown to exhibit phage activity against three C_2 isolates (M1-originated), a C_1 isolate (M1-originated), and a C_4 isolate (M2-originated), and whey sample BD exhibited the broadest infective activity with five C_1, three C_2, two C_3, and seven C_4 isolates being hit by apparent phages present in this whey. Phage presence was confirmed in double agar plaque assays. All but one of the identified phages were determined by multiplex PCR to represent members of *Skunaviruses*. The remaining phage, isolated from the T1 whey sample against a C_2-type isolate, was determined to be a *Ceduovirus*. Two phages, BA6-A5 (*Skunavirus*) and T1-H4 (*Ceduovirus*), were chosen for further host range analysis against the panel of 92 isolates originating from M1 and M2, as representatives of the two phage species found in the whey samples. BA6-A5 was capable of infecting two C_2-type isolates: MA5 and MB5. In addition to the C_2-type isolate that T1-H4 was isolated on, it was also able to infect eight C_1-type isolates (MA2, MA3, MB2, MB3, MC1, MC2, MF2, MG2).

Supplementary Table S1. Bacterial strains, phages, and wheys used in this study.

Lactococcal strains	CWPS type	Genbank Accession number	Source/Reference
L. lactis 275	A	CP015897	UCC
L. cremoris UC509.9*	A	CP003157	UCC
L. cremoris 158	A	CP015894	UCC
L. lactis UL8	A	CP015908	UCC
L. lactis C10	A	CP015898	UCC
L. cremoris WM1	A	CP032500	UCC
L. lactis IL1403*	B	AE005176	(Bolotin et al., 2001)
L. lactis KF147	B	CP001834	(Siezen et al., 2010)
L. lactis Bu260	B	WJUP00000000	Marcelli & De Jong, 2019)
L. lactis UC77	B	CP015906	UCC
L. cremoris NZ9000	C_1	CP002094	(Linares, Kok, & Poolman, 2010)
L. cremoris JM1	C_1	CP015899	UCC
L. cremoris H12	C_1	Unpublished	This study
L. cremoris MG1363*	C_1	AM406671	(Linares et al., 2010)
L. cremoris FA-H10	C_1	Unpublished	This study
L. cremoris 3107*	C_2	CP031538	(Ostergaard Breum, Neve, Heller, & Vogensen, 2007)
L. cremoris JM2	C_2	CP015900	UCC
L. cremoris JM4	C_2	CP015909	UCC
L. cremoris H2	C_2	Unpublished	Host of BA6-A5, this study
L. cremoris KW2	C_2	CP004884	(Kelly, Altermann, Lambie, & Leahy, 2013)
L. cremoris JM3	C_3	CP015901	UCC
L. cremoris C34	C_3	Unpublished	This study
L. cremoris SK11*	C_3	CP000425	(Vos et al., 1989)
L. cremoris W34*	C_4	CP032430	UCC
L. cremoris 1196	C_4	CP032148	UCC
L. cremoris FA-B7	C_4	Unpublished	This study
L. cremoris FN-H3	C_4	Unpublished	This study
L. cremoris FA-D5	C_4	Unpublished	This study
L. lactis IO-1*	C_5	AP012281	(Kato et al., 2012)
L. cremoris A76*	C_6	CP003132	(Bolotin, Quinquis, Ehrlich, & Sorokin, 2012)
L. cremoris NG-A6	C_6	Unpublished	This study
L. cremoris NG-F5	C_6	Unpublished	This study
L. lactis 184	D	CP015895	UCC

* Indicates strains on which qPCR primers were designed.

Whey samples	Year of isolation	Country of origin	Product
T1	2019	Spain	Kefir
BD	2021	Denmark	Semi-hard cheese
BA6	2020	Poland	Twaróg

Lactococcal phages	Isolated host CWPS type	Phage type	Source/reference
BA6-A5	C_2	*Skunavirus* (936)	This study, isolated from wheys from factories employing M1 starter cultures in starter culture rotation systems.
T1-H4	C_2	*Ceduovirus* (c2)	

Supplementary Table S2. DNA concentration of samples used in metagenomic sequencing.

Sample	Concentration (ng/µl)
M1	3.44
M2	5.82
M1 (activated)	15.4
M1 (not activated)	40.2
M2 (activated)	16.7
M2 (not activated)	23.9

Supplementary Table S3. List of *L. lactis or L. cremoris* (accurate species identification could not be determined without whole-genome sequencing) isolates used in phage screening.

Isolate	Origin	CWPS-type
MA1	M1	C_1
MB1	M1	
MC1	M1	
MD1	M1	
ME1	M1	
MF1	M1	
MG1	M1	
MH1	M1	
MA2	M1	
MB2	M1	
MC2	M1	
MD2	M1	
ME2	M1	
MF2	M1	
MG2	M1	
MH2	M1	
MA3	M1	
MB3	M1	
MC3	M1	
MD3	M2	
ME3	M2	
MF3	M2	
MG3	M2	
MA4	M2	C_2
MB4	M2	
MC4	M2	
MD4	M2	
ME4	M2	
MF4	M2	
MG4	M2	
MH4	M1	
MA5	M1	
MB5	M1	
MC5	M1	
MD5	M1	
ME5	M1	
MF5	M1	
MG5	M1	
MH5	M1	
MA6	M1	
MB6	M1	
MC6	M1	
MD6	M1	
ME6	M1	
MF6	M1	
MG6	M1	

Isolate	Origin	CWPS-type
MA7	M1	C_3
MB7	M1	
MC7	M1	
MD7	M2	
ME7	M2	
MF7	M2	
MG7	M2	
MH7	M2	
MA8	M2	
MB8	M2	
MC8	M2	
MD8	M2	
ME8	M2	
MF8	M2	
MG8	M2	
MH8	M2	
MA9	M2	
MB9	M2	
MC9	M2	
MD9	M2	
ME9	M2	
MF9	M2	
MG9	M2	
MA10	M1	C_4
MB10	M2	
MC10	M2	
MD10	M2	
ME10	M2	
MF10	M2	
MG10	M2	
MH10	M2	
MA11	M2	
MB11	M2	
MC11	M2	
MD11	M2	
ME11	M2	
MF11	M2	
MG11	M2	
MH11	M2	
MA12	M2	
MB12	M2	
MC12	M2	
MD12	M2	
ME12	M2	
MF12	M1	
MG12	M1	

Supplementary Table S4. Oligonucleotides used in this study.

Target CWPS type/subtype	Target ORFs	Primer	Sequence (5' to 3')	GC %	Primer length (bp)	Amplicon size (bp)	Amplicon T_m (°C)
A	*UC509_0205*	cwpsA-F	AATTTCCCCAACAGCAAGTG	45	20	182	82
		cwpsA-R	TCCAAAATAGCCAATGGTCA	45	20		
B	*LLKf_0205*	cwpsB-F	ACGGCCGTATGAAACTCATC	50	20	157	84
		cwpsB-R	TGTGCCCTCTTCCTGCTTAG	55	20		
C_1	*llmg_0220*	cwpsC1-F	AGACTGGGATGCATGGGAAA	50	20	178	82.5
		cwpsC1-R	TGCCATTGTTTCTCCCCAAT	45	20		
C_2	*L3107_0192*	cwpsC2-F	CAATATGGCCTTGTTGCTGA	45	20	154	79
		cwpsC2-R	CCCATTCAATCGCAATTTTT	35	20		
C_3	*LACR_0215*	cwpsC3-F	TCCATTCTTGGGAGGCATAG	50	20	185	81
		cwpsC3-R	GAAAAGGATACCGATTTTTCCA	36	22		
C_4	*w34_004*	cwpsC4-F	GTTGCGGGTAAACTTCCAAA	45	20	205	84
		cwpsC4-R	GGACTAACG GCAGACCTTCA	55	20		
C_5	*lilo_0174*	cwpsC5-F	GACATTGGTATGACCGGAAAAG	45	22	226	80.5
		cwpsC5-R	CTGCTAAGTTATCGAACGC	47	19		
C_6	*llh_1295*	cwpsC6-F	TCGGAATTTACGGATACAACG	43	21	213	82.5
		cwpsC6-R	ATAACCCGCACTGAGCATTC	50	20		
D	*LL184_0294*	cwpsD-F	GCGATTTAGCGATGGAAAAG	45	20	184	82
		cwpsD-R	GAAAGCCAAGCTGCTGATTC	50	20		
groES		groES-F	TACGTGTAAAAGAAGAAGAGG	38	21	143	86
		groES-R	TACGAGTGGAGAAATCAGTG	45	20		

Supplementary Table S5. Primer efficiencies associated with representative strains of target CWPS type/subtype.

Primers	A	B	C_1	C_2	C_3	C_4	C_5	C_6	D	groES
Representative strain	*L. lactis* 275	*L. lactis* KF147	*L. cremoris* JM1	*L. cremoris* JM2	*L. cremoris* SK11	*L. cremoris* 1196	*L. lactis* IO-1	*L. cremoris* A76	*L. lactis* 184	*L. cremoris* JM2
Slope	-3.199	-3.229	-3.370	-3.406	-3.375	-3.468	-3.220	-3.427	-3.303	-3.361
Y-intercept	33.482	36.078	31.545	32.827	33.23	32.356	36.062	32.666	35.543	33.219
R2	0.998	0.9952	0.9999	0.9960	0.9986	0.9993	0.9892	0.9997	0.9972	0.9999
Efficiency (%)	105.4	104.0	98.0	96.6	97.8	94.3	104.5	95.8	100.8	98.4
Limit of detection (CFU/rxn)	37.8	11.3	0.3	5.8	7.8	4.2	12.4	4.2	2.4	5.8

Supplementary Table S6. Primer specificity.

		Primers									
Target CWPS type/subtype	**Strain***	**A**	**B**	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**D**	**groES**
A	**275**	**13.13 ± 0.09**	32.95 ± 0.23	31.15 ± 0.18	31.16 ± 0.33	34.43 ± 1.52	32.76 ± 0.25	39.09 ± 1.58	32.98 ± 0.78	36.82 ± 1.14	**13.14 ± 0.07**
	UC509.9	**14.53 ± 0.04**	30.56 ± 0.16	35.55 ± 1.09	34.84 ± 1.17	35.39 ± 0.76	34.59 ± 0.27	ND	37.37 ± 0.35	38.87 ± 1.96	**14.80 ± 0.01**
	158	**11.52 ± 0.02**	30.69 ± 0.22	29.08 ± 0.09	30.33 ± 0.15	31.67 ± 0.09	32.10 ± 0.08	ND	29.54 ± 0.10	34.69 ± 0.80	**11.20 ± 0.06**
	UL8	**11.77 ± 0.03**	31.05 ± 0.24	30.50 ± 0.14	29.90 ± 0.39	35.50 ± 1.00	29.66 ± 0.09	ND	31.18 ± 0.52	36.52 ± 0.72	**12.35 ± 0.21**
	C10	**12.16 ± 0.03**	32.71 ± 1.00	34.65 ± 1.23	29.17 ± 0.23	38.36 ± 1.93	33.58 ± 0.43	ND	39.56 ± 0.41	37.17 ± 0.91	**12.71 ± 0.06**
	WM1	**11.74 ± 0.01**	31.09 ± 0.34	30.26 ± 0.37	32.15 ± 0.43	34.88 ± 0.49	31.39 ± 0.34	ND	32.79 ± 1.34	33.93 ± 0.95	**11.93 ± 0.04**
B	**IL1403**	26.81 ± 0.09	**11.70 ± 0.08**	24.66 ± 7.82	26.31 ± 0.10	38.26 ± 2.47	33.68 ± 0.32	ND	38.73 ± 2.19	36.42 ± 0.92	**11.03 ± 0.03**
	KF147	31.11 ± 0.17	**12.43 ± 0.16**	35.98 ± 0.78	30.93 ± 0.21	38.38 ± 2.81	31.29 ± 0.19	29.58 ± 0.65	37.27 ± 1.92	35.55 ± 0.29	**12.64 ± 0.05**
	Bu260	29.25 ± 0.08	**12.14 ± 0.01**	20.88 ± 0.15	31.30 ± 0.35	31.12 ± 0.28	30.83 ± 0.40	ND	31.37 ± 0.42	33.80 ± 0.26	**11.28 ± 0.05**
	UC77	26.46 ± 0.14	**11.73 ± 0.04**	24.39 ± 6.87	27.08 ± 0.07	32.62 ± 0.70	31.02 ± 0.42	ND	29.92 ± 0.16	33.80 ± 0.26	**12.24 ± 0.05**
C_1	NZ9000	29.08 ± 0.20	27.26 ± 0.31	**12.08 ± 0.01**	30.41 ± 0.36	30.48 ± 0.10	30.80 ± 0.43	ND	28.55 ± 0.10	31.91 ± 0.16	**11.37 ± 0.06**
	JM1	29.04 ± 0.12	29.59 ± 0.33	**11.94 ± 0.06**	34.02 ± 0.59	35.97 ± 3.49	32.28 ± 0.38	ND	30.52 ± 0.05	33.60 ± 0.69	**12.69 ± 0.07**
	H12	31.11 ± 0.29	33.28 ± 0.46	**11.68 ± 0.06**	27.35 ± 0.18	33.88 ± 0.67	33.49 ± 0.70	36.40 ± 3.15	37.49 ± 2.18	35.71 ± 0.60	**12.12 ± 0.01**
	MG163	30.43 ± 0.28	30.82 ± 0.17	**13.64 ± 0.02**	30.86 ± 0.28	31.26 ± 0.16	33.34 ± 0.17	UC	32.50 ± 0.28	36.12 ± 1.57	**15.35 ± 0.38**
C_2	JM2	28.16 ± 0.10	31.74 ± 0.37	30.27 ± 0.43	**12.66 ± 0.03**	31.07 ± 0.49	30.65 ± 0.43	ND	27.77 ± 0.10	36.22 ± 1.07	**12.87 ± 0.02**
	H2	31.10 ± 0.53	32.89 ± 0.30	28.30 ± 0.17	**12.06 ± 0.02**	33.75 ± 1.57	38.17 ± 1.77	35.07 ± 1.07	37.95 ± 2.90	35.53 ± 0.94	**12.24 ± 0.02**
	KW2	30.51 ± 0.24	33.74 ± 0.67	35.55 ± 0.36	**10.92 ± 0.01**	33.17 ± 0.41	35.56 ± 1.19	ND	35.74 ± 0.33	29.19 ± 0.21	**11.53 ± 0.01**
C_3	C34	31.63 ± 0.44	32.42 ± 0.19	30.77 ± 0.08	29.65 ± 0.11	**11.72 ± 0.14**	34.96 ± 0.33	36.05 ± 0.72	37.89 ± 2.00	35.26 ± 0.50	**11.90 ± 0.02**
	SK11	30.31 ± 0.20	29.67 ± 0.19	31.93 ± 0.15	32.38 ± 0.73	**12.48 ± 0.03**	33.07 ± 0.82	37.43 ± 2.25	32.89 ± 0.24	35.18 ± 1.08	**14.13 ± 0.08**
C_4	W34	30.59 ± 0.78	27.08 ± 0.08	33.04 ± 0.53	31.33 ± 0.25	31.07 ± 0.30	**11.91 ± 0.03**	26.80 ± 0.02	34.98 ± 0.16	35.86 ± 0.26	**11.91 ± 0.02**
	1196	31.00 ± 0.10	33.74 ± 0.22	32.07 ± 0.44	30.37 ± 0.13	32.69 ± 0.47	**12.10 ± 0.10**	26.67 ± 0.10	38.67 ± 2.30	36.39 ± 0.93	**29.94 ± 0.20**
	FA-B7	29.14 ± 0.38	30.40 ± 0.06	31.43 ± 0.33	31.70 ± 0.29	32.66 ± 0.77	**13.01 ± 0.34**	ND	28.29 ± 0.01	35.10 ± 0.70	**12.96 ± 0.04**
	FN-H3	27.92 ± 0.05	32.25 ± 0.28	31.96 ± 0.21	31.07 ± 0.13	35.10 ± 4.29	**13.62 ± 0.02**	ND	22.91 ± 0.03	35.07 ± 0.84	**13.77 ± 0.03**
	FA-D5	28.25 ± 0.21	31.81 ± 0.18	31.06 ± 0.51	32.79 ± 0.51	32.42 ± 0.12	**13.05 ± 0.02**	ND	23.37 ± 0.10	33.93 ± 0.27	**13.56 ± 0.02**
C_5	**IO-1**	30.51 ± 0.04	32.04 ± 0.14	34.66 ± 5.46	31.23 ± 0.08	34.55 ± 1.82	31.69 ± 0.15	**11.59 ± 0.01**	37.93 ± 3.59	35.94 ± 0.77	**11.97 ± 0.01**
C_6	A76	31.49 ± 0.05	32.86 ± 0.93	38.49 ± 2.62	34.12 ± 0.98	37.17 ± 2.49	32.86 ± 0.15	27.69 ± 0.16	**11.52 ± 0.06**	35.90 ± 0.41	**13.76 ± 0.06**
	NG-A6	30.57 ± 0.15	33.02 ± 0.32	31.81 ± 0.18	32.54 ± 0.40	34.12 ± 0.46	30.19 ± 0.16	ND	**13.58 ± 0.01**	34.70 ± 0.81	**14.03 ± 0.01**
	NG-F5	30.32 ± 0.02	32.70 ± 0.48	32.89 ± 0.17	32.88 ± 0.32	ND	28.49 ± 0.02	ND	**12.03 ± 0.02**	34.72 ± 0.70	**13.35 ± 0.08**
D	**184**	30.45 ± 0.42	31.68 ± 0.17	31.09 ± 0.37	30.93 ± 0.09	34.58 ± 1.73	32.65 ± 0.38	ND	34.59 ± 0.68	**13.83 ± 0.02**	**13.80 ± 0.03**

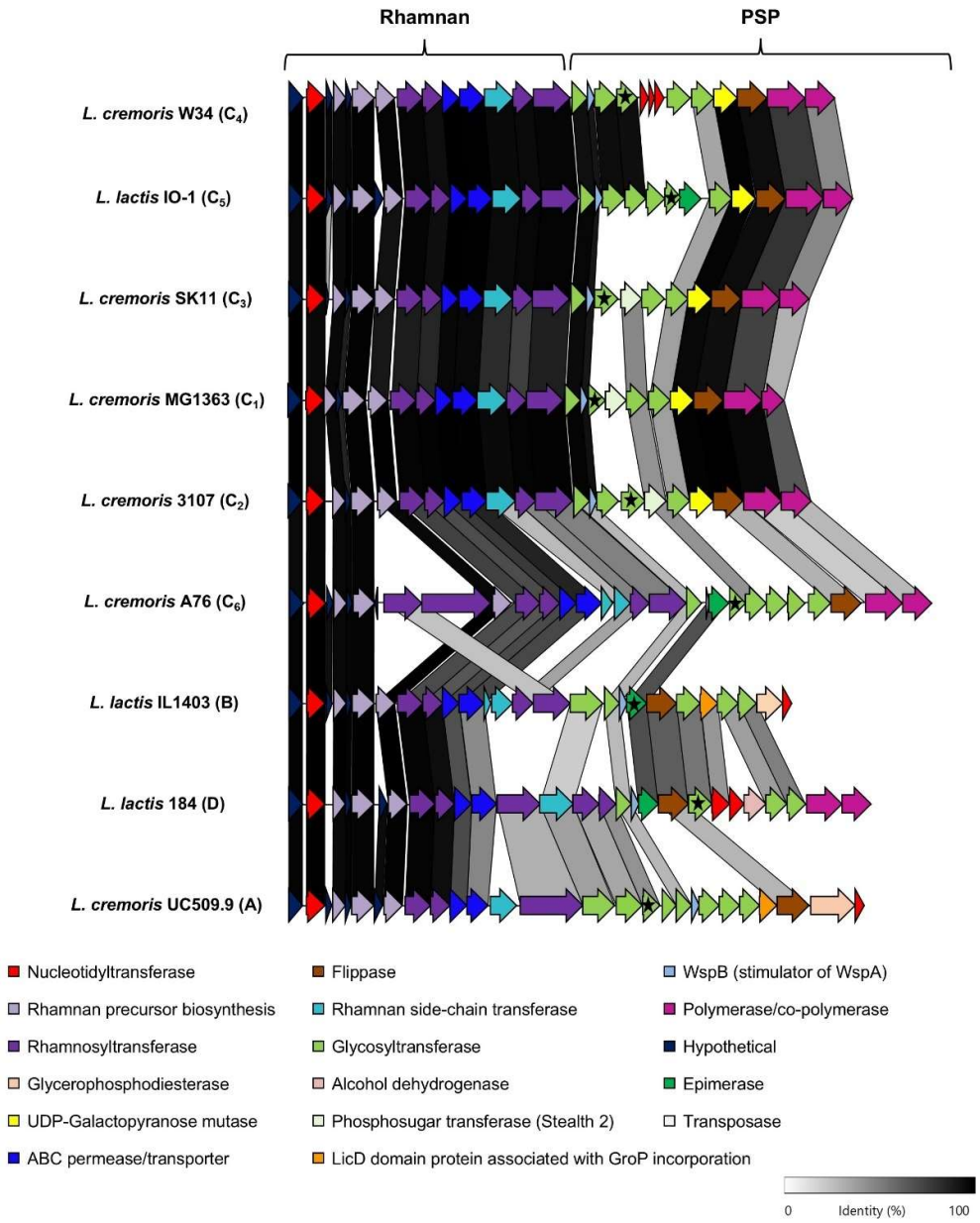
* *L. lactis* strains are indicated in bold face text.; ND: not detected or uncertain call, C_T call of high uncertainty (last five cycles)

Supplementary Table S7. Sequencing information of M1 and M2 metagenome.

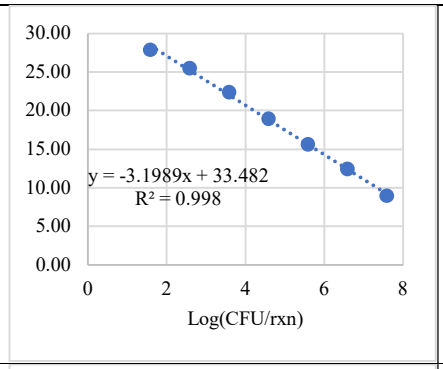
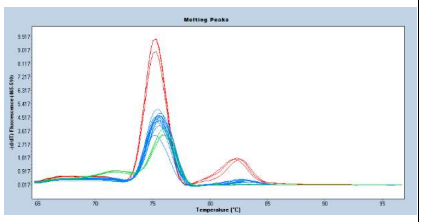
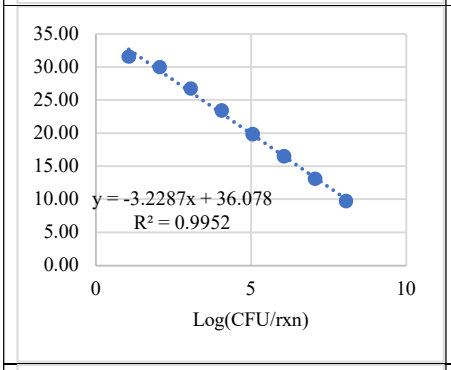
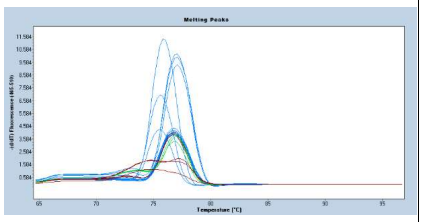
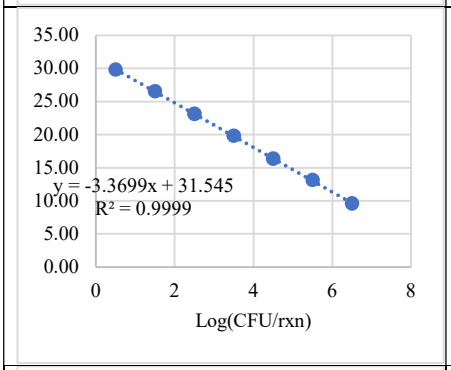
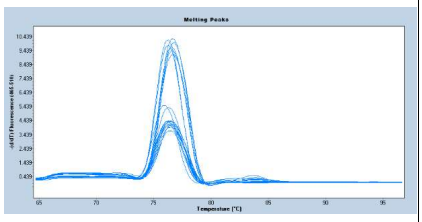
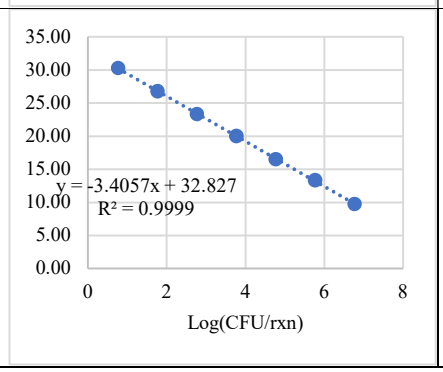
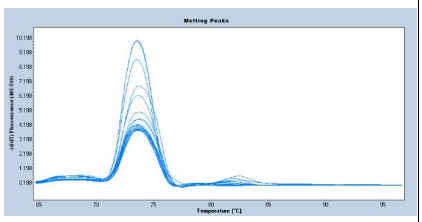
Sample	Total length (bp)	Number of contigs	Average contig length	Largest contig
M1	8 571 770	7008	1223	74 627
M2	8 463 832	6071	1394	86 782

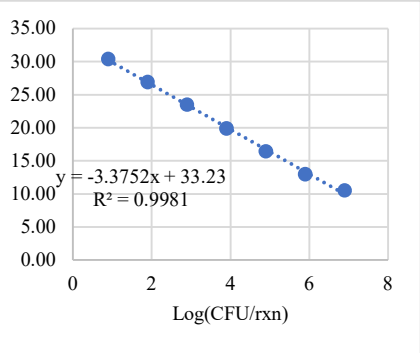
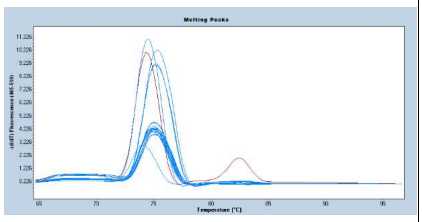
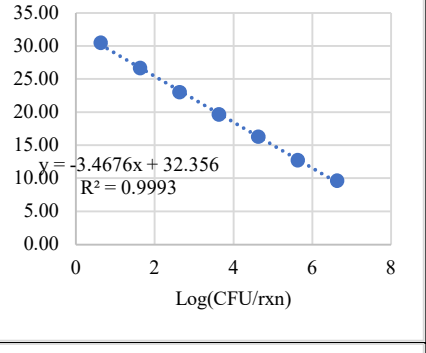
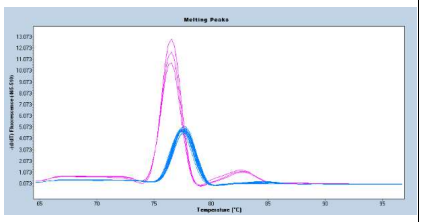
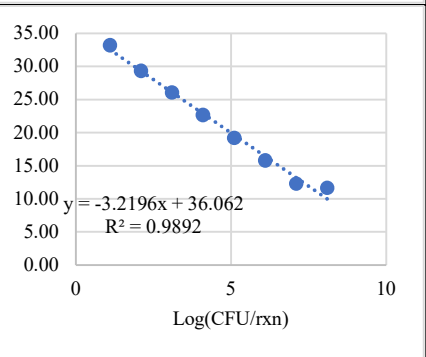
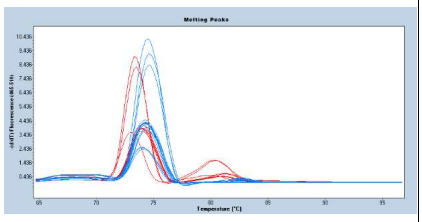
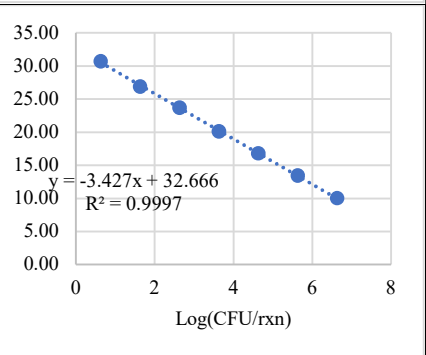
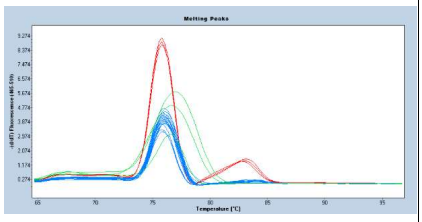
Supplementary Table S8. Breakdown of relative abundance of species

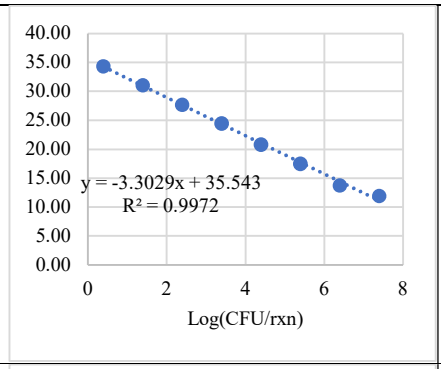
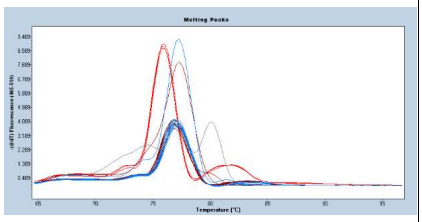
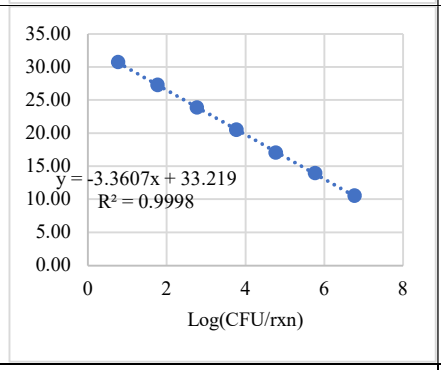
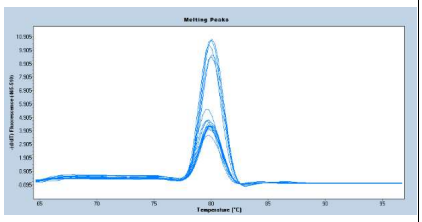
Taxonomy of contigs after LCA		M1	M2
***Lactococcus* sp.**	*Lactococcus lactis/cremoris*	96 %	94 %
	Lactococcus raffinolactis	0 %	2 %
	Lactococcus other (<0.5 %)	1 %	1 %
***Leuconostoc* sp.**	*Leuconostoc garlicum*	1 %	0 %
	Leuconostoc other (<0.5 %)	0 %	0 %
Unclassified		1 %	0 %
Viral		0 %	1 %
Eukaryotic		1 %	1 %
Other (<0.5 %)		1 %	2 %



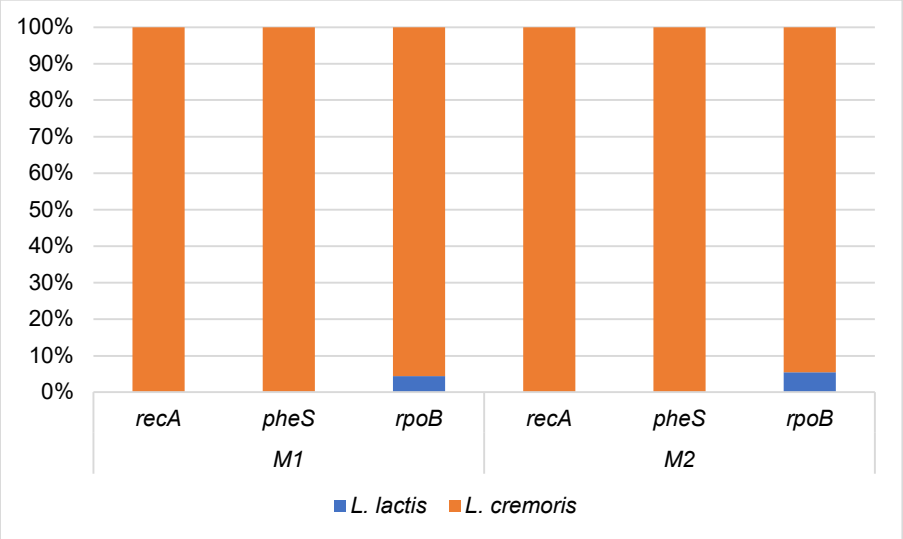
Supplementary Figure S1. Schematic representation and comparison of the CWPS biosynthesis gene cluster (A-D and C-subtypes C_1-C_6). Arrows represent open reading frames which are coloured according to their proposed function (indicated in the figure). The unique genes upon which the qPCR primers are targeting are indicated by a star.

A	CT→	30.00 25.00 20.00 15.00 10.00 5.00 0.00 0 2 4 6 8 y = -3.1989x + 33.482 R² = 0.998 Log(CFU/rxn)	
B		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 0 5 10 y = -3.2287x + 36.078 R² = 0.9952 Log(CFU/rxn)	
C_1		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 0 2 4 6 8 y = -3.3699x + 31.545 R² = 0.9999 Log(CFU/rxn)	
C_2		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 0 2 4 6 8 y = -3.4057x + 32.827 R² = 0.9999 Log(CFU/rxn)	

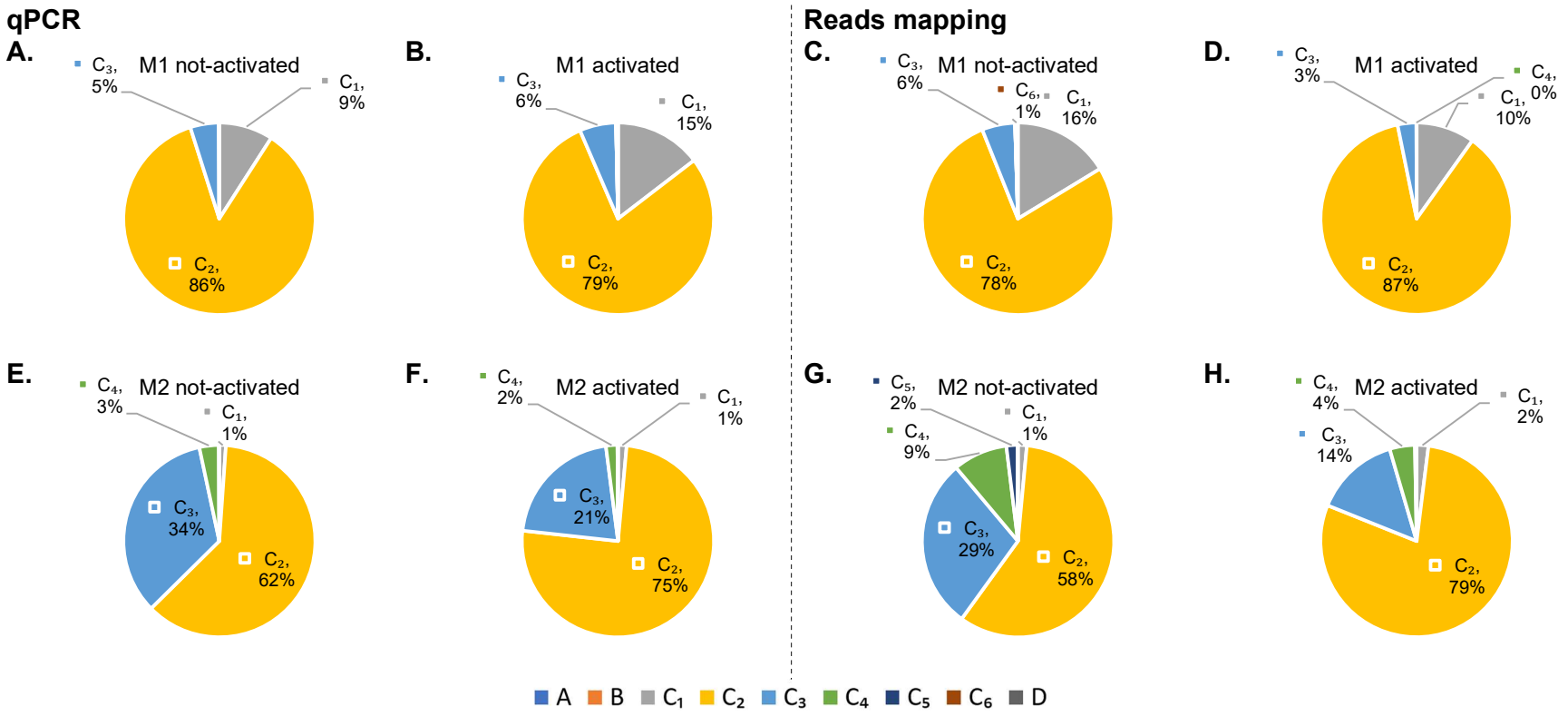
C_3		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 y = -3.3752x + 33.23 R² = 0.9981 0 2 4 6 8 Log(CFU/rxn)	Melting Peaks Temperature (°C)
C_4		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 y = -3.4676x + 32.356 R² = 0.9993 0 2 4 6 8 Log(CFU/rxn)	Melting Peaks Temperature (°C)
C_5		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 y = -3.2196x + 36.062 R² = 0.9892 0 5 10 Log(CFU/rxn)	Melting Peaks Temperature (°C)
C_6		35.00 30.00 25.00 20.00 15.00 10.00 5.00 0.00 y = -3.427x + 32.666 R² = 0.9997 0 2 4 6 8 Log(CFU/rxn)	Melting Peaks Temperature (°C)

D			
groES			
	Log(CFU/rxn) →		

Supplementary Figure S2. Standard curves/primer efficiency and melting curves of *cwps* qPCR primers. Standard curves, for each primer pair, of the threshold cycle (C_T) values plotted against the Log CFU/rxn (based on plating) generated from ten-fold serial dilutions of DNA extracted from representative *L. lactis/L. cremoris* strains (first column). The second column shows amplicon melting curves for each primer set, showing the specificity of the amplification.



Supplementary Figure S3. Proportion of *L. cremoris* vs *L. lactis* strains in starter cultures. The relative abundance of *L. cremoris* and *L. lactis* strains making up the M1 and M2 starter cultures was determined with reads mapping of the genes *recA, pheS*, and *rpoB* corresponding to a representative *L. cremoris* and *L. lactis* strain.

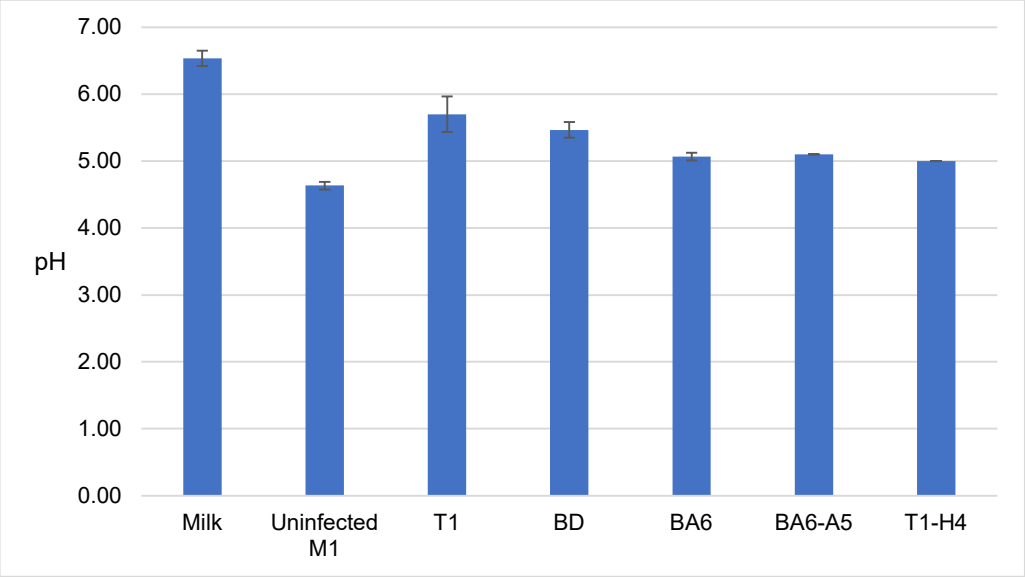


Supplementary Figure S4. Composition of activated vs not activated starter cultures. Relative abundance of *cwps*-types present in M1 (A-D) and M2 (E-H) before (A, C, E, and G) and following culture activation (B, D, F, and H) as determined by qPCR and metagenomics.

CWPS type	Isolate	M1 or M2 derived	T1	BA6	BD
C_1	MA1	M1			■
	MD1	M1			■
	ME1	M1		■	
	MH1	M1			■
	MH2	M1			■
	MC3	M1			■
	MF3	M2	■		
C_2	MH4	M1	*		■
	MA5	M1		■	■
	MB5	M1	■		
	MA6	M1	■		
	MF6	M1		■	
C_3	MC7	M1			■
	MD8	M2			■
C_4	MB10	M2			■
	ME10	M2			■
	MC10	M2		■	
	MC11	M2			■
	MD11	M2			■
	MG11	M2			■
	MD12	M2			■
	MG12	M1			■

*Phage typed as a *Ceduovirus*. All other phages are skunaviruses.

Supplementary Figure S5. Phage screening of whey samples. Phage(s) identified in wheys T1, BA6, and BD after phage screening against a panel of 92 isolates representing the four most dominant *cwps*-types present in the M1 and M2 starter cultures. Black squares indicate presence of phage(s) infecting a strain and white the lack of phage infection. An asterisk indicates the single phage that was typed as a *Ceduovirus*. All other phages are skunaviruses.



Supplementary Figure S6. pH of samples after phage attacks from whey samples (T1, BD, and BA6) and individual phages (BA6-A5 and T1-H4) compared to an uninfected M1 control and pasteurized milk. Samples were measured after 6 hrs of normal fermentation conditions following phage attack from individual phages (BA6-A5 and T1-H4) or phage(s) present in whey samples (T1, BD, and BA6) compared to an uninfected M1 control.

SUPPLEMENTARY MATERIAL REFERENCES

1. Bergsveinson J, Pittet V, Ziola B. 2012. RT-qPCR analysis of putative beer-spoilage gene expression during growth of *Lactobacillus brevis* BSO 464 and *Pediococcus claussenii* ATCC BAA-344T in beer. *Appl Microbiol Biotechnol* **96**:461–470.
2. Johansen P, Vindeløv J, Arneborg N, Brockmann E. 2014. Development of quantitative PCR and metagenomics-based approaches for strain quantification of a defined mixed-strain starter culture. *Syst Appl Microbiol* **37**:186–193.
3. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. 2016. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. *FEMS Microbiol Lett* **363**:fnw049.
4. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.
5. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**:366–368.
6. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240.
7. Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**:9077–9096.

Chapter 3. Profiling phage-host interactions between *Skunavirus* receptor binding proteins and lactococcal cell wall polysaccharide structures

This chapter is in preparation as a manuscript as:

White K, Eraclio G, McDonnell B, Lugli GA, Crowley T, Ventura M, Volontè F, Cambillau C, Nally K, Dal Bello F, Mahony J, van Sinderen D. Profiling phage-host interactions between *Skunavirus* receptor binding proteins and lactococcal cell wall polysaccharide structures.

Experiments, analyses, and first draft preparations were performed by the candidate, except for AlphaFold2 structural predictions and analysis which was performed by C.C. and fluorescence-activated cell sorting which was performed by T.C.

ABSTRACT

Dairy fermentations utilising mesophilic starter cultures rely on the activity of specific lactic acid bacteria (LAB) such as *Lactococcus lactis* and *Lactococcus cremoris* for the acidification of milk. This biotechnological process can be affected by bacteriophage infection of LAB starter strains resulting in failed or delayed fermentations. The majority of studied lactococcal phages commence infection with the binding of a tail-associated receptor binding protein (RBP) to a host cell surface-exposed cell wall polysaccharide (CWPS). In the present study, phage prevalence and diversity in whey samples originating from fermentations performed in various European countries employing undefined mesophilic starter cultures were investigated using phageome analysis. The range of *Skunavirus* RBP genotypes and associated RBP-CWPS binding abilities in the phageomes was evaluated, resulting in the refinement and expansion of the *Skunavirus* RBP grouping system and the identification of several heretofore unknown *Skunavirus* RBP (sub)groups. These findings vastly improve our knowledge on lactococcal *Skunavirus* RBP diversity and their specificity towards CWPS receptor structures, thereby improving the predictability of fermentation outcomes and robustness of starter culture rotations or blends.

INTRODUCTION

Milk fermentations involving mesophilic starter cultures are typically reliant on the activity of *Lactococcus cremoris* and/or *Lactococcus lactis* for the manufacture of a variety of fermented dairy products. The primary role of these microorganisms in fermentations is to produce lactic acid, which leads to a pH reduction and consequent coagulation of milk and curd formation. Bacterial dairy starter cultures are broadly classified as either defined (a small number of strains in a known combination) or undefined (an unknown number of strains in an undetermined relative abundance) (1). There are advantages and disadvantages associated with either of these starter culture approaches, particularly with respect to bacteriophage (or phage) susceptibility, which represents one of the most significant challenges in the fermentation industry. Phage infection can retard or even arrest fermentation, which can lead to substantial economic losses (2). Therefore, phages associated with dairy fermentations (particularly lactococcal phages) have been extensively studied in recent decades (3–8).

Lactococcal phages are classified into ten distinct taxonomic groups: *Skunavirus* (previously 936), P335, *Ceduovirus* (previously c2), *Whiteheadvirus* (previously 1358), *Questintvirus* (previously Q54), *Teubervirus* (previously P087), *Fremauxvirus* (previously 1706), *Audreyjarvisvirus* (previously 949), P034, and *Chopinvirus* (previously KSY1) (9). Skunaviruses, ceduoviruses, and P335 phages are, by a substantial margin, the most common in commercial, high intensity dairy fermentations (6, 10). Skunaviruses and ceduoviruses are both strictly virulent, whereas P335

phages may either be virulent or temperate in nature (11, 12). Despite efforts to limit and control phage proliferation in dairy fermentation factories, lactococcal phages remain a persistent issue (12–14). Phage populations in dairy fermentation environments, therefore, require constant monitoring and investigation. Historically, culture-based methods were applied to detect, enumerate and characterise phages (15). However, these methods typically identify the most dominant and propagatable phages present, thereby not comprehensively accounting for the overall phage population, in particular low abundance phages or phages that do not or poorly propagate under laboratory conditions. The recent surge in metagenomics (and phageomics, i.e. the study of all phage genetic material within a sample) platforms and analysis tools allow a broader, in-depth understanding of phage diversity, prevalence and abundance in a number of food fermentation processes (16–19). Traditional culture-based methods are biased towards the most dominant and propagatable portions of the viral fraction, whereas phageome analysis provides insights into the prevalence, diversity, abundance, and community dynamics occurring in these fermentation environments. Although virome analyses have been performed for various ecological niches, there is a lack of assembly methods that are optimised towards viral rather than bacterial genome assembly leading to an underrepresentation of the true viral population (20, 21). This is particularly problematic when resolving phage genomes, which tend to be highly mosaic compared to bacterial genomes and often harbour repeat regions. The software, Phables has recently been developed and is capable of identifying phage-like components from assembly graphs and more accurately and completely resolving phage genome

assemblies (22). Nevertheless, for functional characterisation of phages it is still necessary to identify the combination of a particular phage and its associated sensitive host suggesting that a culture-based combined with DNA-based analyses may be most suitable to provide a comprehensive view of the viral population (23).

Phage infection typically commences with the recognition and reversible binding of a phage-encoded receptor binding protein (RBP) to a specific receptor present on a host's cell surface. The RBP of skunaviruses is located at the distal end of the phage tail as part of a heteropolymeric protein complex (24). The RBP recognizes and binds to a specific cell wall polysaccharide (CWPS), (part of) which is exposed on the host's cell surface (25–28). A lactococcal CWPS is composed of two distinct parts: a peptidoglycan-embedded rhamnan and a surface-exposed polysaccharide pellicle (PSP), the latter believed to act as the primary phage receptor (29, 30). The PSP of lactococcal strains is chemically diverse and this is underpinned by the observed genetic variability within the 3' region of lactococcal *cwps* gene clusters, which is responsible for PSP biosynthesis. Currently, four *cwps*-types, termed *cwps* A through to *cwps* D, and 11 C-subtypes (designated C_1 through to C_{11}) are described although it is expected that the diversity of these gene clusters and their associated structures far exceeds those identified to date (25–27). The compositional and structural diversity of lactococcal PSPs partly underpin the typically narrow host range of skunaviruses (27), while it also explains the associated diversity in the *Skunavirus* RBPs (26, 28, 31).

All *Skunavirus* RBPs possess a conserved structure consisting of three modules: the so-called “head”, “neck”, and “shoulder” domains (32). The C-terminally located RBP head domain encompasses a carbohydrate binding activity which facilitates specific binding to the host’s saccharidic receptor, i.e. the PSP of the CWPS as mentioned above. *Skunavirus* RBPs have been classified into five groups (groups I - V) based on amino acid sequence similarity levels of their C-terminal (encoding the RBP head domain) sequences (28, 31). Recently, a phylogenetic analysis of over 200 *Skunavirus* RBP sequences from an industrial phage collection indicated the potential presence of additional RBP groups (3). Bacteria often incorporate mutations or alterations to host receptors to evade phage infection (33). Phages, in turn, may modify their RBPs, typically by acquiring mutations or through genetic exchange, highlighting the constant co-evolution of host-encoded receptors and phage-encoded RBPs. The need for continuous and comprehensive examination of binding specificity correlations between *Skunavirus* RBPs and host CWPSs is evident, therefore (34, 35).

In the current study, viromes (all virus-like particles and corresponding nucleic acids within a sample) from 27 whey samples collected from several dairy fermentation factories across Europe employing mesophilic, undefined starter cultures were extracted and sequenced. The corresponding phageomes (assembled phage-like genetic components) were analysed to investigate the diversity of RBP genotypes of skunaviruses in these factories facilitating an expanded and refined *Skunavirus* RBP classification system. The emergent *Skunavirus* RBP (sub)groups identified in the analysed whey phageomes were examined through structural predictions using AlphaFold2

and *in vitro* binding assays. A previously described **p**hage **R**BP-**a**ctivated **c**ell **s**orting (PhRACS) approach was used to identify specific RBP-host interactions (23). As a result of the co-evolutionary dynamics between bacteria and phages, the structural diversity observed for both lactococcal host cell wall receptors and *Skunavirus* RBPs (due to bacteria-phage co-evolution) appears to be a key contributor to the perpetual battle between bacteria and their infecting phages.

MATERIALS AND METHODS

Cultivation/propagation conditions and storage of lactococcal strains and bacteriophages

Bacterial strains used in this study are presented in Supplementary Table S1. Lactococcal strains were grown overnight at 30 °C in M17 broth (Millipore Sigma Aldrich, Gillingham, UK) supplemented with 0.5 % lactose (LM17). Phages were propagated (initially from single plaques, then using 100 µL of phage lysates after initial plaque propagation) in the same medium supplemented with 10 mM $CaCl_2$ and added at the point of host strain inoculation (100 µL of fresh overnight culture), after which the infected culture was incubated overnight at 30 °C. The resulting phage lysates were filtered (0.45 µM; Sarstedt, Nümbrecht, Germany) and stored at 4 °C.

Virome extraction and sequencing

Dairy-derived whey samples (Table 1) were acquired from factories where diminished milk acidification activity of starter cultures was observed. The obtained whey samples were stored at -20 °C prior to processing. Viral particles were isolated from whey samples using a previously described viral extraction method, with minor adjustments, as follows: samples were thawed on ice and 5 to 10 ml of whey (depending on originally received volume) was centrifuged at 300 × *g* for 5 mins at 4 °C to remove any large particulate debris present in the sample. The supernatant was treated with 1 M NaCl at 4 °C for at least 1 hr to release any bacterium-bound phages. The pH of the whey

supernatant was adjusted to ~4.6 ± 0.05 using 0.1 M HCl or 0.1 M NaOH. The whey supernatant was centrifuged at 28,000 × *g* for 15 mins (4 °C). The resulting supernatant was filtered using a double filtration step: first with a 0.45 µm filter, followed by a 0.2 µM filter (Sarstedt). The viral component of the whey filtrate was precipitated with 10 % (w/v) polyethylene glycol (PEG) 6000 overnight at 4 °C. Precipitated virions were collected by centrifugation at 15,000 × *g* for 15 minutes (4 °C) and the pellet was resuspended in 1 ml SM buffer (50 mM Tris-HCl, 100 mM NaCl, 10 mM $MgSO_4$, 10 mM $CaCl_2$) (23). This suspension was treated with DNase I (20 U/mL; Sigma Aldrich, Gillingham, UK) for 15 minutes at RT, to remove any contaminating DNA, followed by DNase inactivation at 75 °C for 10 minutes. Viral DNA isolation was then performed using the Norgen Biotek Phage DNA isolation kit (Norgen Biotek, Ontario, Canada) according to the manufacturer's instructions. Viral DNA was quantified using the Qubit dsDNA HS Assay Kit with Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) according to manufacturer's guidelines.

Table 1. Sample information, absolute, and filtered reads numbers for sequenced whey virome samples.

Sample code	Date	Country	Sequencing pair-end reads (bp)	Quality filtered reads (bp)
T1	06/2019	Spain	1133867	1037855
BH1	07/2021		2568678	2221469
BR	01/2022		6214791	6060432
BT	04/2022		3547612	2996672
AD1	12/2019	Austria	2551593	2077891
AF3	02/2020	France	2315752	1912768
AF9	02/2020		3409093	2891303
BQ*	11/2022		5063789	4927315
BA6	10/2020	Poland	3717233	3460088
BB4	10/2020		2614285	2521910
BZ	10/2022		6194746	6031224
AO	05/2020	Sweden	5660413	4315150
V1	06/2019	Denmark	1271906	1196962
V2	06/2019		1811806	1618823
AE3	02/2020		4217248	3945684
AI3	03/2020		2463249	2341785
AM2	03/2020		108056	33317
AN3	04/2020		1263488	1068294
AN4	04/2020		3292791	3131523
AV6	09/2020		3224986	2895954
BC9	02/2021		4802406	4506953
BD	03/2021		5123477	4697018
BE9	03/2021		5456904	5056351
BV	07/2022		5059960	4832490
CA	11/2022		7704661	7436758
CD	01/2023		6824593	6609464
BP	10/2021	Czechia	6013683	5434002

* Different factory in France

Sequencing libraries were prepared using the NexteraXT DNA Library Preparation kit (Illumina, California, USA), following the manufacturer's guidelines. One ng of DNA was used for library preparation undergoing fragmentation and amplification. Paired-end (300 cycles, 2 × 150 bp) sequencing was performed by GenProbio srl (Parma, Italy) on a NextSeq platform (Illumina, UK). Phageomes were assembled employing two different

assembly and annotation methods. For the first method, viral DNA was quality filtered to remove low-quality reads and improve final quality of the sequenced paired end reads prior to genome assembly. Using these filtered FASTQ files, phage contigs were assembled and annotated using the METAnnotatorX2 pipeline (36).

The second method employs a bioinformatics tool, Phables, which resolves phage genomes from viral metagenomic assemblies (22). Prior to running Phables, an assembly graph was generated from the FASTQ files using the open-source assembler SPAdes v3.15.4 (22, 37). The reads and subsequent SPAdes-generated assembly graph (in Graphical Fragment Assembly (GFA) format) were submitted to the Phables pipeline (https://github.com/Vini2/phables) to generate the most probable combinations of varying phage genome segments (using read mapping information, graph algorithms, and flow decomposition techniques) using default parameters (22). From the 'all_phage_like_edges.fasta' output assembly file (containing sequences from all phage-like components, both resolved and unresolved), open reading frames (ORFs) were predicted using Prodigal v2.6.3 prediction software (38). ORFs were annotated using BLASTP 2.6.0+ (39) against the non-redundant protein databases curated by NCBI (https://ncbi.nlm.nih.gov).

Bioinformatic analysis of RBP sequences

Skunavirus RBP sequences were identified from the phageome assemblies based on BLASTP analysis. RBP head domain protein homology

detection and initial structure prediction were determined using HHpred (Homology detection and structure production by HMM-HMM comparison) using default settings. Sequence alignments of RBP head domains (C-terminal ~110 amino acid residues of RBP) were performed using MultAlin (40) employing phageome-identified *Skunavirus* RBP sequences and previously sequenced *Skunavirus* RBPs retrieved from the NCBI viruses sequence data base (https://www.ncbi.nlm.nih.gov/labs/virus/). Sequence comparison of RBP head domains was performed by all-against-all, bi-directional BLAST alignment with an alignment (*E*-value) cut-off value of <0.0001 and greater than 50 % identity across at least 50 % of the amino acid sequence. ClustalW multiple alignments were conducted to compare amino acid sequences of *Skunavirus* RBP head domains, a phylogenetic tree of sequences was constructed using ClustalW (bootstrapped with 1,000 replicates) and visualised via ITOL (Interactive Tree of Life) (41).

Structural predictions of RBPs using Alphafold2

Structural predictions were performed with either a Colab notebook running AlphaFold v2.3 (42, 43) or HPC resources from GENCI-IDRIS running AlphaFold v2.3.1 (44). The pLDDT values (Supplementary Figure S1) of predicted structures were stored in the PDB files as B-factors and plotted with the Colab or the IDRIS resources. The final predicted protein or domain structures were submitted to the Dali server (44, 45) to identify the closest structural homologues in the protein database (PDB). Visual representations of the structures were prepared with ChimeraX (46).

Phageome analysis

To determine the relative abundance of phages present in the assessed whey samples, a phage database was created using publicly available and taxonomically classified lactococcal and streptococcal phage genomes retrieved from NCBI viruses (https://www.ncbi.nlm.nih.gov/labs/virus/). The phage database encompassed a total of 1,155 lactococcal and streptococcal phage genomes, grouped according to their genera (ten lactococcal and five dairy streptococcal genera). This database was used to quantify the number of reads (from the raw, FASTQ files quality filtered using NGS QC Toolkit v2.3.3 (47)) that align against each phage genus. CoverM v0.4.0 (https://github.com/wwood/CoverM) was used to determine the proportion of reads that align against each of the mapped phage groups. The relative abundance of phages was determined using the following arguments: coverm genome --methods relative_abundance --min-read-percent-identity 80 --min-read-aligned-percent 50. The relative abundance of *Skunavirus* RBP genes present in each sample was determined through reads mapping of a representative RBP-encoding nucleotide sequence of each *Skunavirus* RBP group/subgroup (refined and expanded *Skunavirus* RBP groups will be described below) CoverM v0.4.0 by calculating the RPKM (reads per kilobase of exon per million reads mapped) using the following arguments: coverm contig --methods rpkm --min-read-percent-identity 86 --min-read-aligned-percent 95 --min-covered-fraction 55.

Molecular cloning

Primers used in this study (Table 2) were synthesized by Eurofins Genomics (Germany). PCR amplifications were performed using Phusion High Fidelity DNA polymerase. Primer pairs were used to amplify RBP-encoding genes (list of oligonucleotides and targeted ORFs are presented in Table 2) using whey virome DNA as template. PCR products were purified using GenElute™ PCR Clean-Up Kit (Sigma-Aldrich, UK). Purified PCR products were cloned in the green fluorescent protein (GFP) fusion vector pHTP9 (GFP) using the NZYEasy Cloning Kit (NZYTech genes & enzymes, Portugal) according to manufacturer's instructions, and introduced into competent *E. coli* BL21 (DE3) by heat shock at 42 °C for 40 s and plated on LB agar supplemented with 50 µg/ml kanamycin (Sigma, UK) (48). The integrity of recombinant plasmid sequences was verified using Sanger sequencing (GeneWiz, Germany) using vector (pHTP9-F GAATGAAAAACGCGACCACATGGTG and pHTP-R GGTTATGCTAGTTATTGCTCAGCG) and/or insert-specific primers (49).

Table 2. Oligonucleotides used in this study.

Targeted RBP(s)	Targeted *orf*	Oligonucleotide	Sequence (5' -3')
AO-RBP2, AD1-RBP5	orf_{AO-2}, orf_{AD1-5}	AO2AD15F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAATTTCACG
		AO2AD15_R	**TCAGCGGAAGCTGAGG**TTATTTAATAAAGTAGCTTGCG
BB4-MG12	$orf_{BB4-MG12}$	C-type_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAACTTTAC
		B-MG12_R	**TCAGCGGAAGCTGAGG**TTATTTTATAAAGTAACTGG
V2-RBP2	orf_{V2-2}	C-type_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAACTTTAC
		V2-2_R	**TCAGCGGAAGCTGAGG**TTATTTGATGAAATACATGGC
BA6-RBP1	orf_{BA6-1}	BA61_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAACTTTACATTCTTTAG
		BA16_R	**TCAGCGGAAGCTGAGG**TTACAATAAAGGAGCTGATAAAGCG
BB4-RBP2	orf_{BB4-2}	BB4-RBPF*	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAAATTCACGTT
		BB4-RBPR*	**TCAGCGGAAGCTGAGG**CTAATTTATACGTTTCCAAATAT
BA6-RBP2	orf_{BA6-2}	BA62_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAAATTCACG
		BA62_R	**TCAGCGGAAGCTGAGG**CTAATTTATACGTTTCCAAATATAG
BB4-RBP1	orf_{BB4-1}	BB42_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAAATTCACGTT
		BB42_R	**TCAGCGGAAGCTGAGG**CTAATTTATACGTTTCCAAATAT
AD1-RBP2	orf_{AD1-2}	AD1-2_F	**TCAGCAAGGGCTGAGG** ATGACAATTAAAAGCTTCACATTTTTC
		AD1-2_R	**TCAGCGGAAGCTGAGG** TTATGCTGGGAATGAATCACCAGTG
AE3-RBP3, BB4-RBP4	orf_{AE3-3}, orf_{BB4-4}	AE3-3_F	**TCAGCAAGGGCTGAGG** ATGACAATTAAAAACTTCACATTTTTC
		AE3-3_R	**TCAGCGGAAGCTGAGG**TTATTTATGCTTGCTAGCTTGTC
BQ-RBP1	orf_{BQ-1}	BQ1_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAACTTTACATTCTTTAG
		BQ1_R	**TCAGCGGAAGCTGAGG**TCAATATTCTAATACGTGCTGTAG
AO-RBP4	orf_{AO-4}	AO4_F	**TCAGCAAGGGCTGAGG**ATGACGATTAAAAACTTTACATTTTTC
		AO4_R	**TCAGCGGAAGCTGAGG**CTATTTTTTAGTTGCTATAAATTC
CA_RBP5	orf_{CA-5}	CA5_F	**TCAGCAAGGGCTGAGG** ATGACAATTAAAAACTTTACATTTTTC
		CA5_F	**TCAGCGGAAGCTGAGG**TTATAAAGTTTTTGTCGCAAATAC
AO-RBP1	orf_{AO-1}	AO1_F	**TCAGCAAGGGCTGAGG**ATGACAATTAAAAATTTTACATTTTTCAGTCC
		AO1_R	**TCAGCGGAAGCTGAGG**CTAATTTCTATAAAGTTTTACAATC

*Primers first described in previous study (23).

Recombinant protein production

2 ml aliquots of *E. coli* BL21 (DE3) carrying recombinant plasmids of interest were inoculated in 200 mL of LB medium supplemented with 50 µg/mL of kanamycin and 10 % glycerol followed by incubation at 37 °C in a shaking incubator at 150 rpm until cultures reached an OD_{600nm} between ~0.3-0.7. Proteins were then induced with 0.05 mM isopropyl-β-d-1-thiogalactopyranoside (IPTG) and incubated overnight at 20 °C in a shaking incubator at 250 rpm. Cells were harvested by centrifugation at 4,600 × *g* for 30 mins and resuspended in 10-20 mL (depending on size of cell pellets) of lysis buffer (50 mM Tris-HCl, 300 mM NaCl, 10 % glycerol, 1 % Triton X-100 pH 7.5) and incubated shaking at 4 °C for at least an hour. The cells were disrupted by sonication in a MSE Soniprep (Sanyo, Japan) at maximum amplitude for five 30 s cycles followed by a 30 s rest, with a holding step on ice for at least 2 minutes following the third round of sonication. Proteins were separated from cell debris with centrifugation at 15,000 × *g* for 30 mins (4 °C). His-tagged proteins were purified with standard Ni-NTA agarose columns (Qiagen, Manchester, UK) in protein buffer (50 mM Tris-HCl, 300 mM NaCl, 50 mM $CaCl_2$, pH 7.5) with increasing concentrations of imidazole (50-250 mM). Purified proteins were analysed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). Protein concentrations were quantified using the Qubit Protein Assay Kit with Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) according to manufacturer's guidelines. Purified proteins were dialysed against protein buffer stored at 4 °C.

GFP-RBP-host binding assays and fluorescence microscopy

Binding activity assessment of the His-tagged GFP-RBP fusion proteins (^{His}GFP-RBP) was performed with cells of various lactococcal strains (Supplementary Table S1), which had been grown to mid-log phase (OD_{600} ~0.4-0.7), harvested by centrifugation at 10,000 × *g* for 5 mins and resuspended in an equal volume of SM buffer. 100 µL of the bacterial resuspension was mixed with 12.5 µg of each GFP-RBP fusion protein in protein storage buffer, added to a well in a 96-well round (U) bottom plates (Thermo Scientific, Waltham, M.A., U.S.A.) containing 100 µL of SM buffer, and incubated at 30 °C for 12.5 mins. Control wells containing only 12.5 µg of ^{His}GFP-RBP in SM buffer was included in each assay. Cells were centrifuged (2,000 × *g* for 5 mins) and washed three times with 200 µL of SM buffer. Prior to fluorescence quantification, the OD_{600} was measured and transferred to black polystyrene 96-well plates (Nunc; Thermo Fisher Scientific, Waltham, USA). Fluorescence was quantified with a Tecan Infinite 200 PRO series plate reader (Männedorf, Switzerland). GFP-fusion proteins were excited at a wavelength of 488 nm, with a z-stack of 20,000 µm and gain of 80. Arbitrary fluorescence units (FU) were determined with the following formula: $\frac{Sample\ FU - control\ well\ FU}{OD600\ of\ sample}$. A relative binding affinity score was then calculated as follows: $\frac{Average\ FU\ of\ a\ ^{His}GFP::RBP\ \text{to a strain}}{Average\ FU\ of\ a\ ^{His}GFP::RBP\ \text{to a optimal host strain}} \times 100\ \%$, where the optimal host strain is the strain to which a given ^{His}GFP-RBP consistently binds with the highest affinity.

For fluorescence microscopy, 6 µL of lactococcal cell suspensions labelled with fluorescent ^{His}GFP-RBP (as described above) were transferred

onto a glass slide and covered with a cover slip. Fluorescence imaging was performed using an Olympus BX53 fluorescence microscope (Olympus, Tokyo, Japan).

Isolating novel phage-host combinations through PhRACS

Sorting of ^{His}GFP-RBP$_{AO-4}$ and ^{His}GFP-RBP$_{BQ-1}$ was performed in the same manner as described above for the binding assays, using activated bulk starter culture M1 (using the so-called PhRACS method as previously described by Casey *et al.* (23)). M1 is an undefined mesophilic industrial starter culture commonly used in the production of semi-hard cheeses, fermented milk, and fresh cheeses. 10 g of lyophilized M1 was diluted in 207 g of sterilized (115 °C for 20 min) 10 % (wt/vol) reconstituted skimmed milk (RSM) powder. Following homogenization in a Stomacher 400 Circulator (Seward, UK) for 2 min at 230 rpm, this mixture was inoculated into LM17 (at 2 %) and incubated overnight at 22 °C. LM17 was inoculated overnight and incubated at 30 °C until it reached an OD_{600} ≈ 0.3-0.4. ^{His}GFP-RBP fusion protein-labelled bacteria were sorted with a BD FACSARIA Fusion cell sorter with 85 µM nozzle and a 488 nm laser with 530/30 filter for GFP excitation and detection. GFP-positive events were sorted as single-cells into 96-well plates containing 150 µL of LM17. 96-well plates were then incubated at 30 °C for 48 hrs. Isolates were subcultures into 96-well plates containing LM17, grown overnight, and stored at -20 °C in 25 % glycerol.

RESULTS

Phageome of dairy-derived whey samples

To gain deeper insights into phage-host interactions occurring in dairy fermentation factories employing mesophilic undefined starter cultures, the viral fraction of 27 whey samples collected from several factories across Europe utilising mesophilic undefined starter cultures (typically supplemented with additional strains/cultures) was extracted. Virion-extracted, double-stranded DNA was sequenced with a NextSeq Illumina shotgun metagenomic sequencing platform, generating an average sequencing output of 3,666,334 high quality reads per sample (Table 1). To assess phage prevalence and diversity of these whey samples, the sequencing reads were mapped against an in-house database of reference phage genomes of representatives of the ten lactococcal and five *S. thermophilus* currently described phage genera (encompassing 640 lactococcal and 515 dairy streptococcal phages from published phage genomes, respectively; Figure 1A). This mapping demonstrated that skunaviruses are the dominant phage group in 18 of the 27 analysed whey samples, followed by the P335 (three samples) and ceduoviruses (two samples). In two whey samples collected in France, AF3 and AF9, streptococcal phages belonging to the *Moineauvirus* and *Brussowvirus* genera were the most dominant. The presence of streptococcal phage sequence in these samples may emanate from phages from fermentations incorporating *S. thermophilus* containing starter cultures within the factory or may indicate that adjunct *S. thermophilus* strains were added to the mesophilic fermentation. The percentage of reads that did not align to any of the reference genomes varied greatly (1-100 %), mainly due to

contaminating bacterial DNA or (in fewer cases) the presence of non-lactococcal/streptococcal phage genomes (Supplementary Table S2). Reads were assembled using the METAnnotatorX2 and Phables assembly pipelines and select *Skunavirus* contigs (confirmed through BLASTN analysis) were investigated for further analysis as described below.

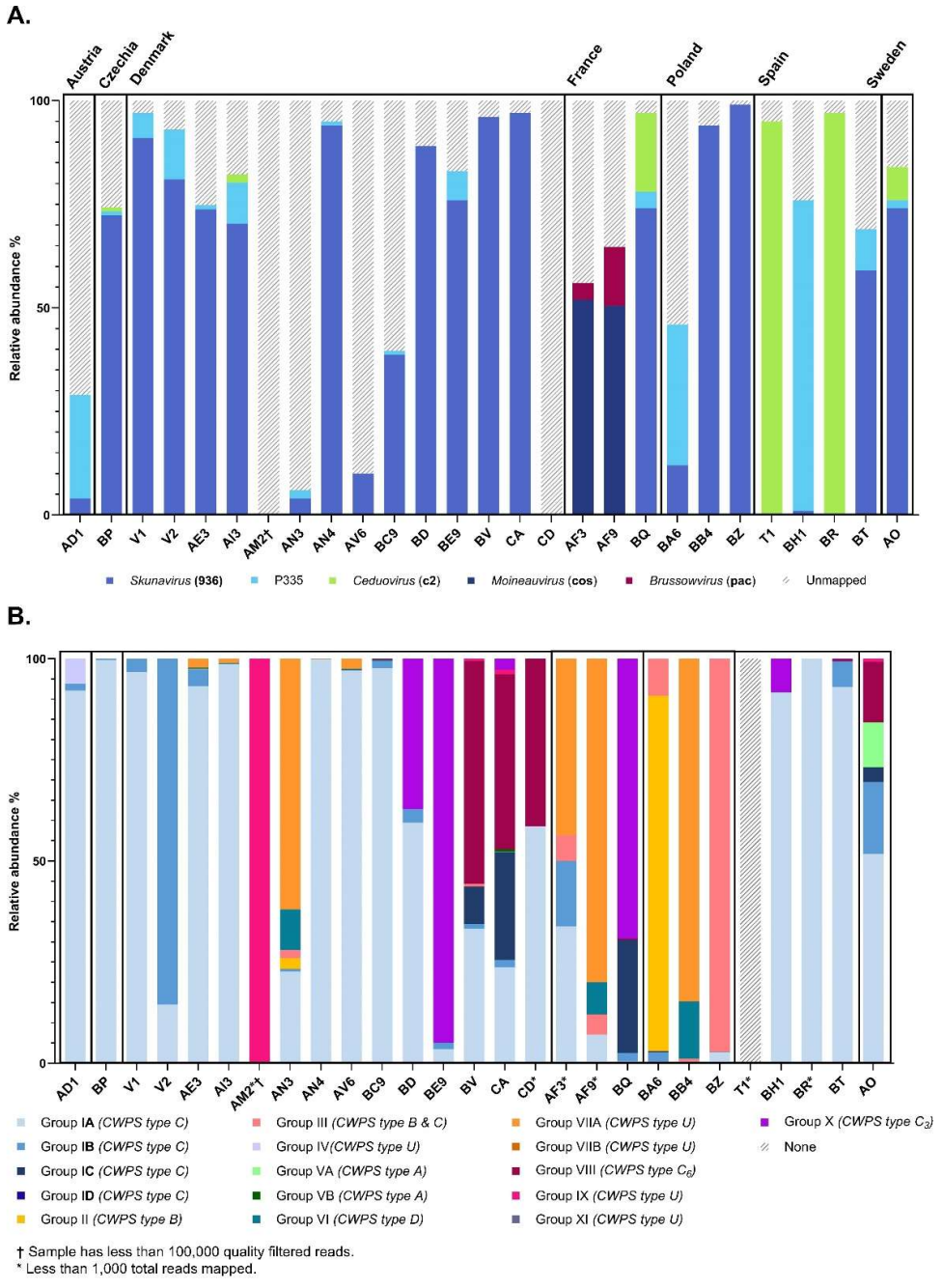


Figure 1. Phage diversity present in whey virome samples from dairy fermentation factories across Europe based on reads mapping. (A) Relative abundance of lactococcal and streptococcal phages. (B) Relative abundance of *Skunavirus* RBP groups. Reads that do not map to any of the phage database or RBP groups, either due to novel phage sequences (not present in any reference genome) or contaminating DNA (mainly bacterial) are depicted as diagonal grey lines.

Analysis of Skunavirus RBPs

Since members of the *Skunavirus* genus highly dominate the phageome in the majority of examined whey samples, it was decided to exclusively focus on these phages. As RBP-mediated adsorption of a *Skunavirus* to its host is the first step in the infection process, the obtained phageome sequences were scrutinized for predicted *Skunavirus* RBP-encoding genes. Using METAnnotatorX2- and Phables-assembled *Skunavirus* contigs, 81 *Skunavirus* RBP sequences (after extracting distinct RBP sequences within a given sample, i.e. RBPs with less than 95 % amino acid identity to any other RBP identified within the same sample) were identified using BLASTP. *Skunavirus* RBP sequences are known to exhibit considerable similarity within the first ~135 N-terminal amino acid residues, which correspond to the shoulder and neck domains of the RBP, while considerable variation is observed in the remaining ~110 C-terminal residues, representing the RBP head domain, which interacts with the host receptor represented by the CWPS. A previous study of 91 *Skunavirus* RBP sequences had identified five RBP groups (I-V) based on varying similarity levels that were found to correspond to phage host CWPS preference (26, 28). With access to 332 RBP sequences (251 publicly available RBPs and the 81 phageome-derived RBPs described herein), a refinement and expansion of the previously

defined *Skunavirus* RBP groups was established through comparative analysis of the C-terminal region (~110 residues of the C-terminal end of each RBP) of all (i.e. 332) currently available *Skunavirus* RBPs. A multiple amino acid sequence alignment of the RBP head domain of all 332 RBPs was performed (alignments were generated using MultAlin; Supplementary Figure S2 and S3), and a phylogenetic tree was generated to visualise RBP groups (Figure 2). Through this analysis, 11 distinct groups of RBPs were identified (as well as the presence of several subgroups). RBP groups were strictly defined here based on a particular group sharing a minimum of 50 % amino acid identity across 90 % of the RBP head domain with other group members. Amino acid identity within the head domain was typically considerably higher than the minimum threshold (between ~82-95 % identity). However for groups where there was an observed divergence in the amino acid sequence of the head domain among group members (but still sufficient identity to fall within a single group based on the defined parameters), subgroups were defined in which RBPs shared at least 65 % of amino acid residues within the head domain of all subgroup members (a cut-off that was supported by subsequent host binding assays and/or structural analysis). This analysis revealed the presence of 11 RBP groups (groups I through XI) and with subgroups of groups I (IA-ID), V (VA & VB), and VII (VIIA & VIIB), representing a significant expansion of the previously defined five RBP phylogenetic groups. One *Skunavirus*, 30804, does not group within an RBP group based on the phylogenetic analysis and may represent a future additional group. The relationship between the proposed refinement and expansion to the RBP groups is presented in Table 3 and will be elaborated upon below. Four of the

current 11 RBP groups (I-III and V) directly correspond to one of the previous groups (albeit with the addition of subgroups for groups I and V), while group IV has been refined to represent two groups (groups IV and VI based on the defined parameters), and the remaining five groups are completely novel.

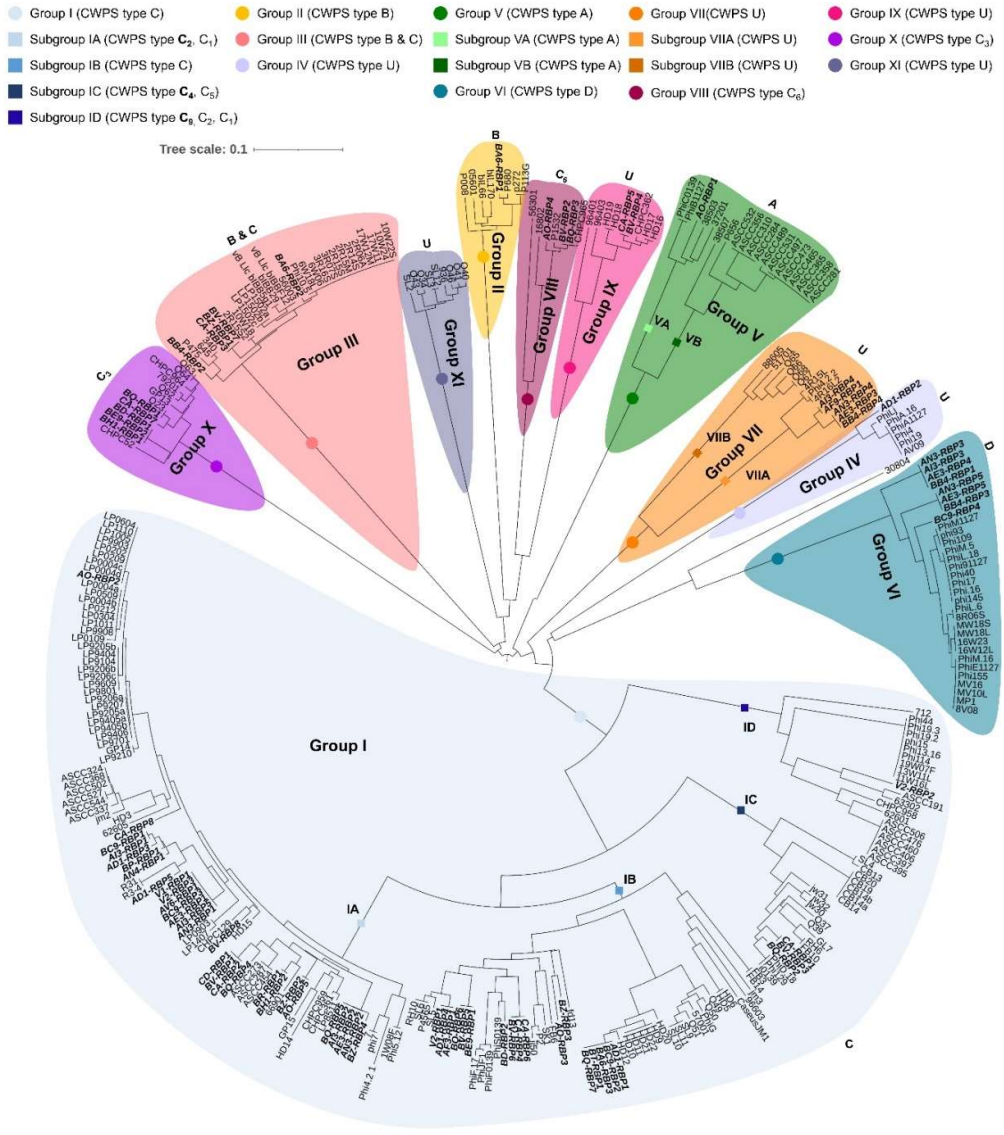


Figure 2. Phylogenetic tree of the C-terminal region of the RBP from whey phageome derived assemblies (bolded) alongside previously sequenced *Skunavirus* RBP sequences. Amino acid sequences of *Skunavirus* RBPs identified from whey phageomes (81 RBP sequences) and publicly available *Skunavirus* genomes, 332 RBPs in total, were aligned and grouped together based on sequence similarity and corresponded to host CWPS (where possible).

Table 3. RBP groups as determined through phylogenetic analysis. Correlation to the previously described RBP grouping system and preference of an RBP group to a particular CWPS(s) is described.

Group/subgroup	Relation to previous RBP grouping*	CWPS specificity
Subgroup IA	Group I	C (C_2, C_1)†
Subgroup IB		C*
Subgroup IC		C (C_4, C_5)†
Subgroup ID		C (C_9,C_2, C_1)†
Group II	Group II	B
Group III	Group III	B & C*
Group IV	Group IV	U
Subgroup VA	Group V	A*
Subgroup VB		
Group VI	Group IV	D
Subgroup VIIA	Novel	U
Subgroup VIIB	Novel	U
Group VIII	Novel	C_6
Group IX	Novel	U
Group X	Novel	U
Group XI	Novel	C_3

*Host specificity determined in previous studies (26, 28).

†Preference for particular C-subtypes determined through fluorescent binding assays, affinity to a particular subtype listed in descending order.

The 81 phageome-derived RBPs identified in the current study were distributed among ten of these 11 RBP groups, as presented in Supplementary Table S3, where RBPs are named according to their whey phageome from which they had been identified and numbered based on the number of distinct RBPs found within a sample. The majority (50 of 81) of phageome-identified RBP sequences are members of RBP group I, thus representing the most

abundant RBPs in the assessed phageomes (Figure 2). Group I RBPs all share the amino acid residues Trp 144, Arg 256, Asp 234, and His 232 (as determined from MultAlin generated alignments of the RBP head domains; Supplementary Figure S2) from the RBP head of phage p2, that have previously been determined to be located within its sugar binding crevice (31, 32, 34). A previous study had suggested that additional amino acid residues (other than the four conserved residues mentioned above) impact on specific RBP binding abilities, justifying the decision to introduce subgroups (which was supported by differences in host specificity of subgroups, which will be discussed below) (50). The remaining phageome-derived RBPs belong to (sub)groups II, IV, or VA (one RBP identified for each of these), group IX (two), group VIII (three), (sub)groups III, VIIB, and X (five each), and eight were found to belong to group VI. Reads mapping was performed to determine the relative abundance of each of the (re)defined *Skunavirus* RBP (sub)groups (Figure 1B and Supplementary Table S4). Of note, group I RBPs (specifically subgroup IA) were found to be the most prevalent in 14 of the 27 analysed whey samples and present at some abundance (>2 %) in all but five samples (AM2, T1, BQ, BA6, and BB4; Figure 1B). A notable trend among the Danish whey samples, is that novel RBP groups VIII and X began appearing in samples collected in 2021 and progressively became more dominant in samples collected in 2022 and 2023. Whereas group I RBPs were the most dominant in most (7/11) Danish samples collected between 2019 and 2021 (apart from AM2, AN3, and BE9). In addition, lactococcal RBPs are present in two whey samples (AF3 and AF9) collected in France (Figure 1B) despite streptococcal phages being the dominant phage genera in these specific samples (Figure 1A).

Structural predictions of novel RBPs

To determine the impact of the sequence diversity exhibited by the phageome-derived RBPs on the structural complexity and diversity of *Skunavirus* RBPs, six phageome-derived RBPs were selected as representatives of their groups (IV and VI-X) for structural analysis using AlphaFold2. This expands on the previous structural analysis of four RBP head domain groups (represented by groups I-III and V) (31). Through this approach, six novel *Skunavirus* RBP structure predictions were generated (Figure 3). All six RBP group representatives possess the conserved modular RBP structure observed in all skunaviruses, consisting of the head, neck, and shoulder domains (51). However, group VII (subgroup VIIB) representative AE3-RBP3 exhibits a distinct structure including a very long neck domain, similar to those of lactococcal P335 phages, LC3 and Q33 (Figure 3 C), although not every member of this group possesses the same extended neck domain (52). Of the six novel RBP structures, only the representative RBP head domain of group IV (AD1-RBP2) shares any residues (Asn 234 and Arg 255/256) located in the binding crevice of the RBPs of phages p2 (32, 34) and TP901-1 (53) (Supplementary Figure S4). The head domains of group IV, VI, and VII and group VIII and IX, respectively, appear to share similar folding characteristics to each other but have different amino acid residues in the binding crevice (Figure 3D & E). Additionally, the group X representative RBP sequence, BQ-RBP1, has a unique folding pattern within its head domain when compared to the other RBPs (Figure 3F).

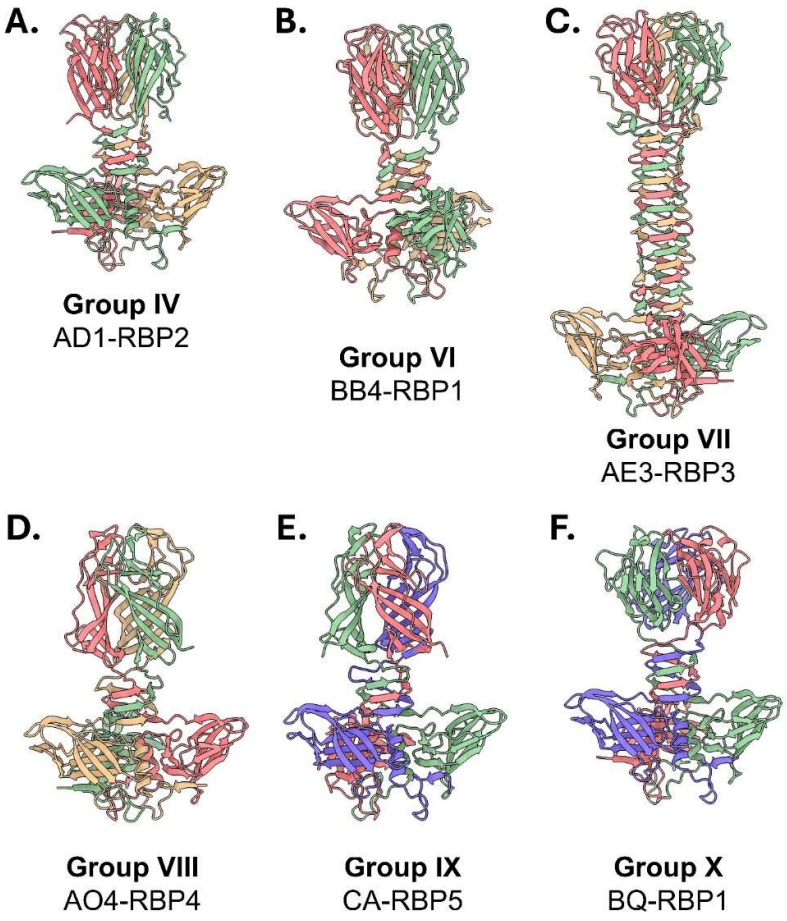


Figure 3. Structural analysis of RBP groups (IV and VI-X). Structures of representatives of the novel and re-defined group IV as determined with AlphaFold2: (A) AD1-RBP2 (group IV), (B) BB4-RBP1 (group VI), (C) AE3-RBP3 (group VII), (D) AO-RBP4 (group VIII), (E) CA-RBP5 (Group IX), and (F) BQ-RBP1 (Group X).

Fluorescent binding assays employing His-tagged GFP-fusion RBPs

Host recognition by skunaviruses occurs through highly specific, RBP-mediated binding to the CWPS of their corresponding host. The compositional and structural diversity of lactococcal CWPSs is therefore believed to at least partly account for the highly specific nature of lactococcal (*Skunavirus*) phage-host interactions. To experimentally explore the specificity of these

interactions among the identified RBPs, we selected 15 distinct RBP-encoding genes, representing RBP groups I-X (and more specifically subgroups IA, IC, ID, VA, and VIIB) and cloned these into the expression vector pHTP9 to create constructs for the production of RBP-GFP fusion proteins of ~54-64-kDa (See Materials and Methods; Table 1). In all cases but one ($RBP_{BB4\text{-}MG12}$) RBP sequences were selected from phageome sequence assemblies. These RBP-GFP fusion proteins were used in binding assays employing available lactococcal strains representative of 13 currently defined *cwps* gene clusters (CWPS-types A, B, C_1-C_7, C_9-C_{11}, and D). Where successful, the results obtained from these binding experiments are summarised in Figure 4 and will be discussed in further detail below.

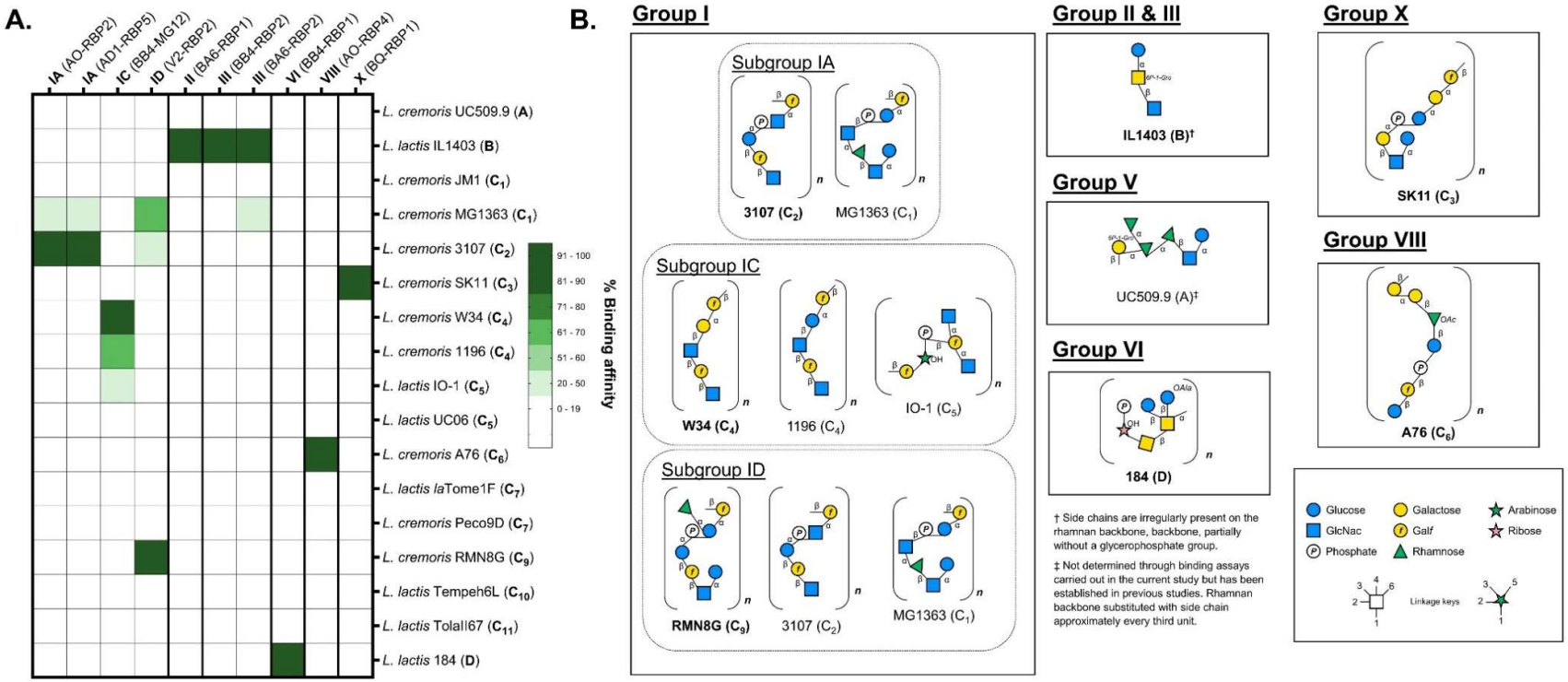


Figure 4. Binding patterns of *Skunavirus* phage RBPs to lactococcal strains of various CWPS-types. (A) It was possible to determine the binding specificity of ten of the initial 15 ^{His}GFP-RBP representing eight RBP (sub)groups. Percent of binding affinity is expressed as a ratio of the quantitative determination of binding (average FU) of a ^{His}GFP-RBP to lactococcal strains representing different CWPS-types divided by the highest binding affinity (average FU) of a ^{His}GFP-RBP to a particular lactococcal CWPS-type representative. (B) Symbolic representation of PSP component of the lactococcal strains ^{His}GFP-RBP were capable of binding to. The optimal host strain (the strain to which a given ^{His}GFP-RBP consistently binds with the highest affinity) is bolded. Full representation of PSPs can be found in Supplementary Table S5.

Binding specificity of group I, II, and III RBP representatives

There are currently 11 recognized C-subtype(s) *cwps* gene clusters, which are associated with the production of structurally diverse PSPs (25, 54). Therefore, here we sought to determine if the amino acid differences found among the group I RBP head domains (leading to the recognition of four subgroups) correspond to specific binding activity to particular C-subtype strain(s). Subgroup IA representatives, ^{His}GFP-RBP$_{AO-2}$ and ^{His}GFP-RBP$_{AD1-5}$ were both shown to be capable of binding to strain 3107 (C_2) with high affinity, at 1.50×10^5 and 1.81×10^5 FU, respectively, and to strain MG1363 (C_1) with significantly lower affinity, at 5.57×10^4 and 4.10×10^4 FU (p-value < 0.01; Figure 5 A&B). Subgroup IC representative, ^{His}GFP-RBP$_{BB4-MG12}$, predominantly binds to C_4-type strain W34 at 2.08×10^5 FU, and with decreased binding to C_4-type strain 1196 at 1.29×10^5 FU, and C_5-type strain IO-1 at 6.88×10^4 FU (Figure 5C). ^{His}GFP-RBP$_{V2-2}$ (subgroup ID) is capable of binding to both MG1363 (C_1) at 6.50×10^4 FU and 3107 at 4.13×10^4 FU, while it binds to strain RMN8G (C_9) with significantly higher affinity at 1.07×10^5 FU (p-value < 0.0001; Fig 5. D).

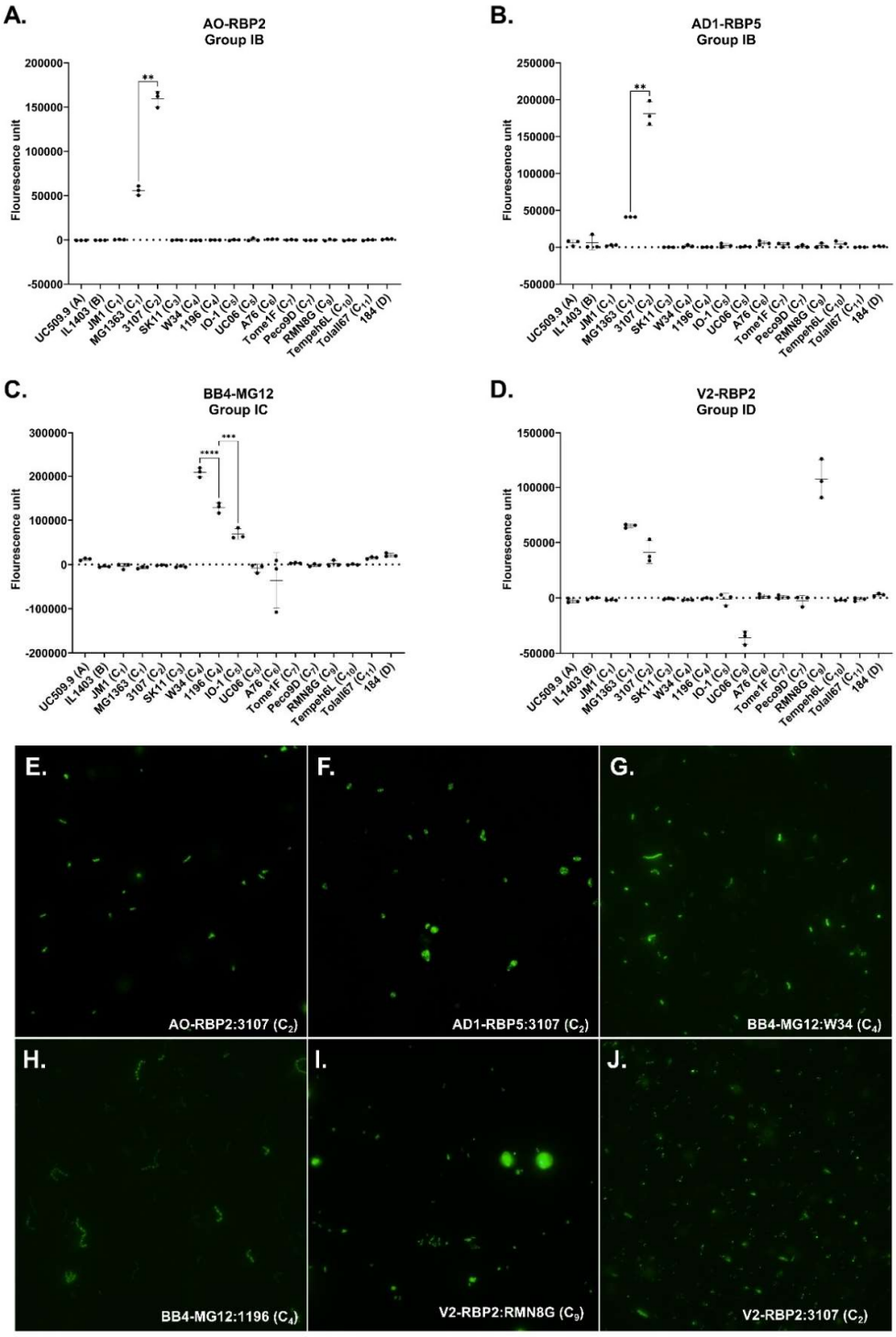


Figure 5. Quantitative determination and fluorescence imaging of group I ^{His}GFP-RBPs to various lactococcal CWPS-type strains. (A-D) Fluorescence-based quantification of binding of the indicated ^{His}GFP-RBPs (group I) to representative lactococcal strains encoding various *cwps* genotypes (A, B, C_1-C_7, C_9-C_{11}, and D). (E-J) Fluorescence imaging of ^{His}GFP-RBPs to indicated lactococcal strains.

Group II (^{His}GFP-RBP$_{BA6-1}$) and group III (^{His}GFP-RBP$_{BB4-2}$ and ^{His}GFP-RBP$_{BA6-2}$) representative RBPs specifically recognise and bind to the *cwps* B-type strain IL1403 at 1.92×10^5, 2.24×10^5, and 3.03×10^5 FU, respectively (Figure 6). ^{His}GFP-RBP$_{BA6-2}$ of group III was also found to bind to MG1363 (*cwps* C_1) at a significantly reduced affinity (7.97×10^4 FU). Group V (subgroup VA) representative ^{His}GFP-RBP$_{AO-1}$ did not exhibit detectable binding to any of the tested lactococcal strains, although phages within this group have previously been shown to infect *cwps* A-type strains (28, 31).

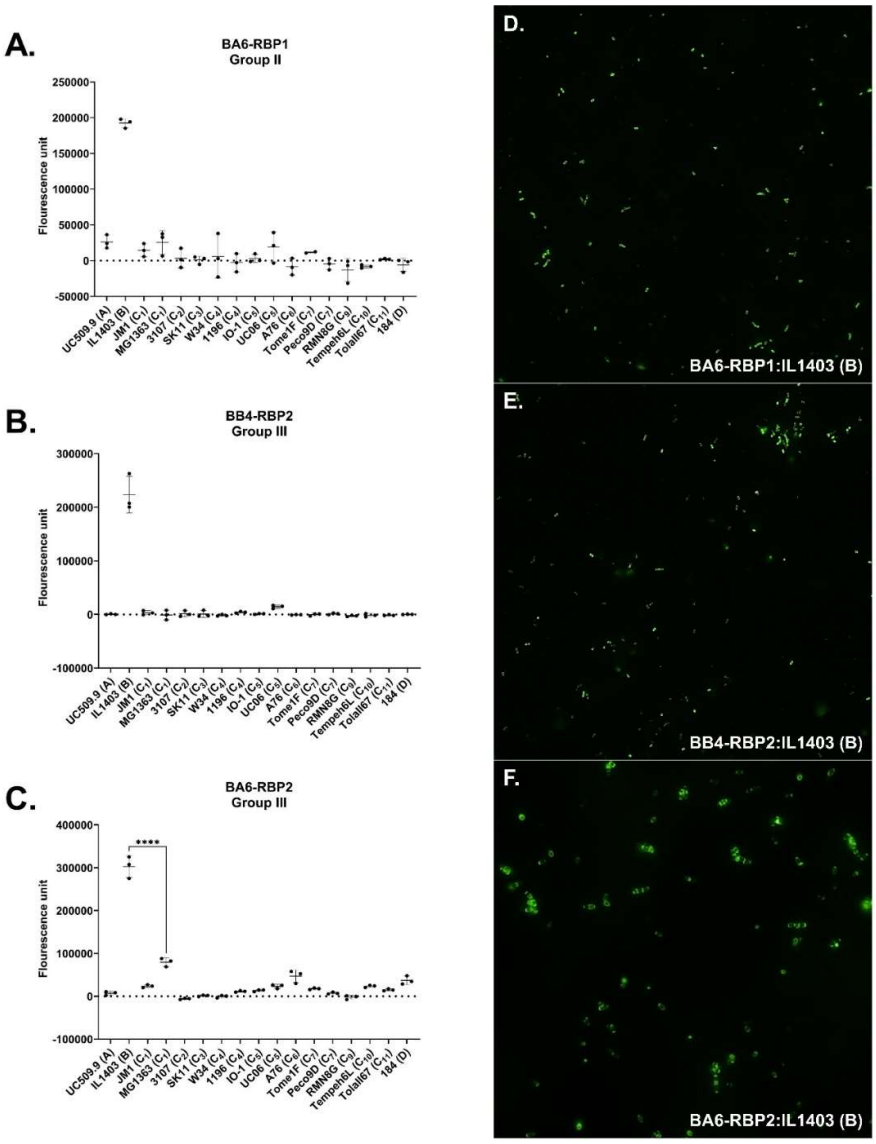


Figure 6. Quantitative determination and fluorescence imaging of group II and III ^{His}GFP-RBPs to various lactococcal CWPS-type strains. (A-C) Fluorescence-based quantification of binding of the indicated ^{His}GFP-RBPs (groups II and III) to representative lactococcal strains encoding various *cwps* genotypes (A, B, C_1-C_7, C_9-C_{11}, and D). (D-F) Fluorescence imaging of indicated ^{His}GFP-RBPs to *L. lactis* IL1403 (B-type).

Investigating binding capabilities of emergent RBP groups

Binding assays were conducted on (representatives of) previously unstudied/unknown RBP groups to evaluate the link to their host CWPS type preference of identified potential host strains (Figure 7). Group VI representative ^{His}GFP-RBP$_{BB4-1}$ binds specifically to D-type strain 184, at 1.85 × 10^5 FU (Figure 7A&D) (25). Group VIII representative ^{His}GFP-RBP$_{AO-4}$ binds specifically to C_6-subtype strain A76 at 5.34 × 10^4 FU (Figure 7B&E). Group X representative, ^{His}GFP-RBP$_{BQ-1}$, binds to SK11 (C_3) with high affinity at 1.59 × 10^5 FU (Fig 7C&F). Group IV representative RBP (^{His}GFP-RBP$_{AD1-2}$), group VII (subgroup VIIA) RBPs (^{His}GFP-RBP$_{AE3-3}$ and ^{His}GFP-RBP$_{BB4-4}$), and group IX RBP (^{His}GFP-RBP$_{CA-5}$) did not appear to bind to any of the CWPS-type strains with particularly high affinity (compared to the other RBPs studied and visually confirmed with fluorescence imaging) and therefore the specificity of these RBP groups could not be elucidated in the current study.

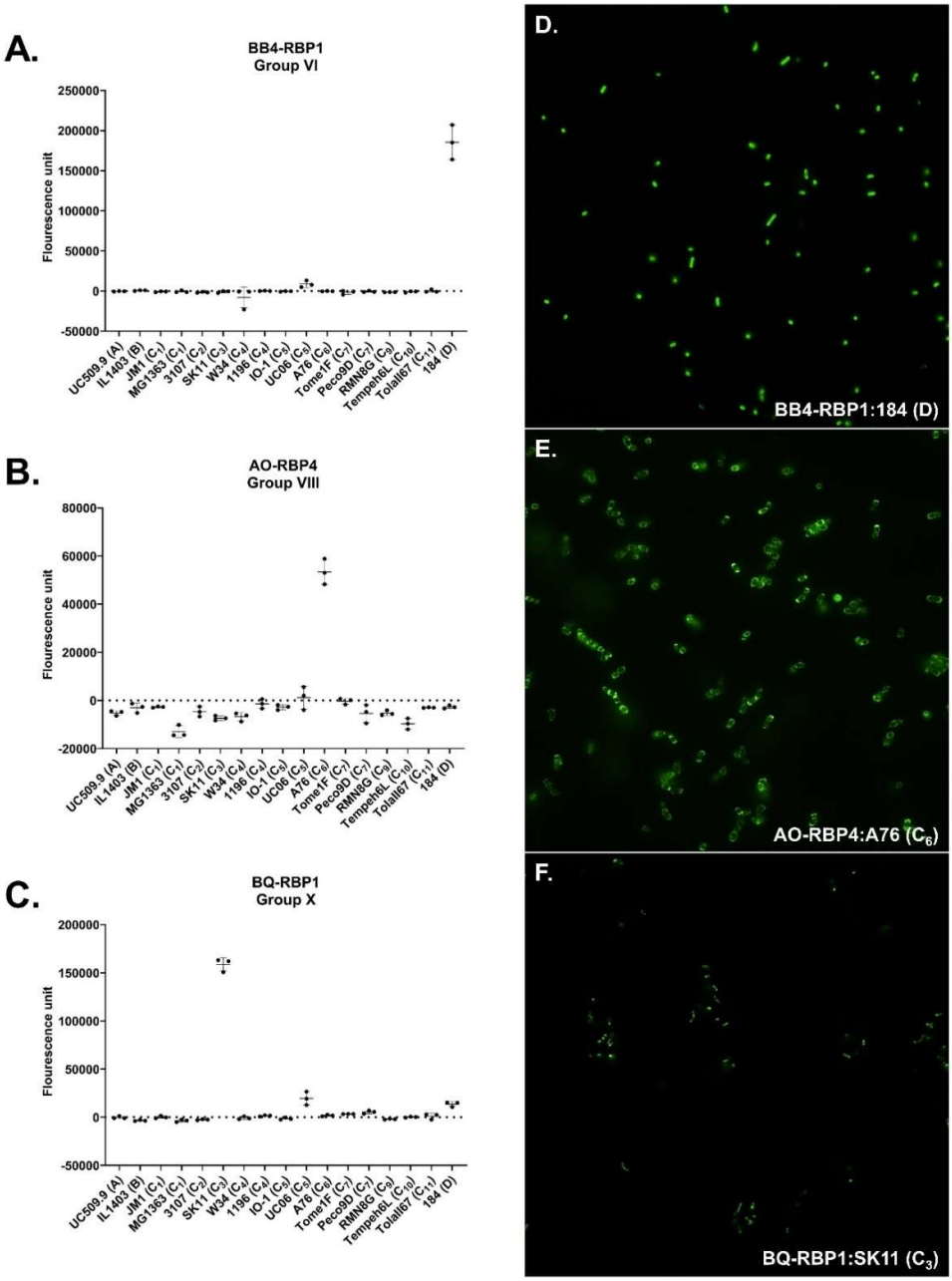


Figure 7. Quantitative determination and fluorescence imaging of novel p2-like ^{His}GFP-RBPs to various lactococcal CWPS-type strains. (A-C) Fluorescence-based quantification of binding of the indicated ^{His}GFP-RBPs (groups VI, VIII, and X) to representative lactococcal strains encoding various *cwps* genotypes (A, B, C_1-C_7, C_9-C_{11}, and D). (D-F) Fluorescence imaging of ^{His}GFP-RBPs to indicated lactococcal strains.

Phage-host retrieval using PhRACS

The so-called phage RBP-activated cell sorting or PhRACS method (23) was recently established as an effective method of retrieving low abundance phage-host combinations from complex samples. This is achieved through specific labelling and isolation of culturable bacterial hosts with fluorescently tagged phage RBPs using flow cytometry. Notably, this method had previously been applied to the successful isolation of a phage-host combination for BB4-RBP2 (group III), resulting in the isolation of a low abundance lactococcal strain (harbouring a B-type *cwps* cluster) and infecting phage (BB4-SC34) with a 100 % identical RBP to BB4-RBP2 (23). Therefore to complement the results of the binding assays, attempts were made to isolate individual phage-host combinations corresponding to the following RBPs representing groups II and IV through X (predominantly representing the novel or refined RBP groups) through the so-called PhRACS method: AD1-RBP2, BB4-RBP1, BA6-RBP1, AO-RBP1, AE3-RBP3, AO-RBP4, BQ-RBP1, and CA-RBP5 (23). To isolate novel phage-host combinations from dairy-derived whey samples, ^{His}GFP-RBP corresponding to the above RBP-encoding genes identified from phageome data representing several RBP groups were used to identify GFP-positive cells from an undefined mesophilic starter culture (M1) employed in dairy fermentation facilities. Of the RBPs attempted, only with ^{His}GFP-RBP$_{AO-4}$ and ^{His}GFP-RBP$_{BQ-1}$ (determined to bind specifically to C_6 and C_3-type strains, respectively, see previous paragraph) was it possible to successfully retrieve potential hosts through PhRACS. Of the total population of cells, 1.7 % and 0.7 % of cells were found to be GFP-labelled by ^{His}GFP-RBP$_{AO-4}$ and ^{His}GFP-RBP$_{BQ-1}$ (respectively) and were then

sorted into a 96-well plate through single-cell sorting (Supplementary Figure S5). From ^{His}GFP-RBP$_{AO-4}$ and ^{His}GFP-RBP$_{BQ-1}$ sorting, 52 and 83 culturable isolates were successfully retrieved following 24-72 hrs of incubation at 30 °C, respectively, and were genotyped with previously defined *cwps*-type specific multiplex PCR systems (26, 27). 42 of the 52 ^{His}GFP-RBP$_{AO-4}$ sorted isolates were typed as *cwps* subtype C_6 and the remaining ten were C_1-subtype strains. Of the 83 ^{His}GFP-RBP$_{BQ-1}$ sorted isolates, 75 were typed as *cwps* subtype C_3 and the remaining eight were C_1-subtype strains. Representative isolates of these subtypes were tested for binding against both ^{His}GFP-RBPs to identify potential hosts. ^{His}GFP-RBP$_{AO-4}$ was shown to bind to the C_6 isolates and ^{His}GFP-RBP$_{BQ-1}$ to the C_3 isolates (neither the C_1 isolates). These isolates were used in phage isolation from the original whey samples (AO and BQ) through spot assays, leading to the isolation of skunaviruses AO-4 and BQ-1. The attempts to use PhRACS to isolate the potential hosts of AD1-RBP2, BB4-RBP1, BA6-RBP1, AO-RBP1, AE3-RBP3, and CA-RBP5 were unsuccessful, probably as a result of suitable hosts either occurring in very low abundance (below the sensitivity of this approach) or not being present in the starter culture population used in the study.

DISCUSSION

In recent years, consumer expectations of consistent and high-quality fermentation products have led to an increase in both metagenome and virome studies of various food production systems and products (55–59). This has generated a wealth of knowledge on the microbiota and community dynamics occurring in various fermentation systems, facilitating more reliable methods of community monitoring. Studies of the viral metagenomics of food systems has revealed an abundance of, as yet, unclassified viral “dark matter” (i.e. not aligning to any reference viral sequences), in food systems (60). Consequently, more targeted virome extractions and phageome-focused bioinformatic analysis methods have been developed (16–19). In the current study, the viromes of 27 whey samples were successfully extracted/sequenced and the subsequent phageomes analysed using two different assembly methods to yield as many phage-like contigs as possible (METAnnotatorX2 and Phables). The majority of whey phageomes mostly consisted of lactococcal phage sequences, skunaviruses were the principal phage group present in almost all assessed samples (Figure 1A). However, both T1 and BR whey samples originated from a Spanish factory and were shown to predominantly contain ceduoviruses. Both skunaviruses and ceduoviruses are strictly lytic and represent the most problematic phages in commercial dairy fermentation facilities. *Skunavirus* sequences were also found to be the most abundant phage type in 18 of the 27 whey phageomes studied. This is in line with several other (culture-based) studies investigating the diversity of phages associated with dairy fermentation facilities using mesophilic starter cultures (10, 61, 62).

As adsorption is crucial in the initiation of the *Skunavirus* infection process, LAB strains have been found to mutate or alter their host receptors (33, 63). In turn, phages may modify or mutate their RBPs or even acquire novel genes that allow them to overcome the modifications to host receptors (50, 64–66). Such systems of modifying RBPs include diversity-generating retroelements (DGRs), first characterised in *Bordetella* phage BPP-1 (67) and identified in other lysogenic phages infecting various Bacteroidota, Pseudomonadota, Bacillota, Actinomycetota (68, 69), as well as homologous recombination (HR) events between closely related phages or prophages, which has been observed between LAB phages (66, 70). Another RBP diversification system identified in *Bifidobacterium* strains is the Rin shufflon system that generates RBP variants between a constant N-terminus and variable C-terminus RBP regions utilising a tyrosine-family recombinase (65). *Skunavirus* escape-mutants that can overcome phage-resistant derivatives of the lactococcal strain 3107 (with reduced synthesis of PSP), it appears that mutations to key amino acid residues likely involved in saccharidic binding is an effective strategy used by skunaviruses to counteract host receptor modifications (50). Therefore, the diversity of skunaviruses in these factories was the primary focus of the investigation, particularly the diversity of RBPs, which are responsible for the initial binding and recognition of phages to a cognate host, genotypes present in the facilities.

Group I RBPs are structurally highly conserved (resemble phage p2 RBP) and share four amino acid residues (Trp 144, Arg 256, Asp 234, and His 232) which have previously been determined to be found within the binding crevice of phage p2 (31), while it has been suggested recently that additional

residues are involved in saccharidic binding (50). Detailed analysis of the head domain of group I RBPs present in the whey phageomes led to the identification of notable amino acid differences within the RBP head domains. Consequently, group I RBPs were divided into four subgroups. Representative RBPs from three of the four group I subgroups (A, C, and D) were used to generate ^{His}GFP-RBPs that were subsequently tested for their ability to bind to lactococcal strains representing 15 distinct CWPSs (A, B, C_1-C_7, C_9-C_{11}, and D). Here we have confirmed that group I RBPs bind specifically to C-type *cwps* strains, and that differences in key amino acid residues in the sugar binding domain are linked to a preference for certain C-subtypes. In the case of group IA this appears to be C-subtype C_2 (and comparatively low affinity to C_1), for group IC it is C-subtype C_4 (and comparatively low affinity to C_5), and for group ID it is C-subtype C_9 (with low affinity to C_1 and C_2-types). This aligns with previous host range studies of a number of skunaviruses (included in the RBP group analysis) possessing group I RBPs that have been found to infect hosts possessing various C-type *cwps* genotypes (26, 28).

Both the group II representative ^{His}GFP-RBP$_{BA6\text{-}1}$ and group III representatives (^{His}GFP-RBP$_{BB4\text{-}2}$ and ^{His}GFP-RBP$_{BA6\text{-}2}$) were found to bind to B-type strain IL1403, aligning with previous CWPS preferences of phages within these groups (26, 31). However, certain skunaviruses possessing group III RBPs have also been shown to have the ability to infect C-type lactococcal strains (based on host range studies) (26, 31). Binding affinity of the subgroup VA ^{His}GFP-RBP representative could not be confirmed through fluorescent binding assays, however previous studies have found that skunaviruses possessing similar RBPs are capable of infection lactococcal strains encoding

A-type *cwps* gene clusters (26, 31). As the mesophilic starter cultures used in these fermentation facilities are (or are derived from) starter cultures primarily composed of various *cwps* type C lactococcal strains (59), it is likely these phages infect strains that have been added to supplement the starter cultures in these fermentation facilities. Supplementation of additional lactococcal strains/culture was reported in the Polish factories where samples BA6 and BB4 were collected from (and group II and III RBPs were identified in) as well as the Swedish factory where AO whey was collected (and AO-RBP1, subgroup VA, was identified).

In addition, to confirming and expanding upon previously described links between RBP type and what *cwps*-type strains they have the capacity to infect we also investigated the specificity of several refined and novel RBP types possessing so-called “p2-like” head domains (31). Based on initial HHpred analysis, a number of RBP head domains (that within the current study formed seven distinct groups) were found to possess a “p2-like” head domain but upon sequence analysis had significantly different amino acid composition and were structurally quite distinct. Attempts were made to correlate these novel RBP groups to a particular pellicle genotype of potential host strains through fluorescent binding assays. This was successful for only three (groups VI, VIII, and X) of the seven RBP groups. The binding affinity of the remaining RBP groups (groups IV, VII, and IX) remains unknown as representative ^{His}GFP-RBPs of these groups did not appear to bind to any of the CWPS-type strains tested with particularly high affinity. This may be due to poorly functioning ^{His}GFP-RBPs or these RBPs recognise CWPSs that were not included in this study (either CWPS-types not yet identified or defined, or ones

not included in our strain collection, such as C_8-subtype strains). It is expected, as more lactococcal phages and strains are isolated the RBP grouping system (and links between groups and host CWPS) will need to be further polished and (likely) expanded.

Along with the phageomic analysis outlined above, the PhRACS approach was used here to exploit the specific phage receptor binding protein (RBP)-host cell surface receptor interactions facilitating the isolation of phage-host combinations from factory-derived whey samples and a mesophilic undefined starter culture. This led to isolation of a novel *Skunavirus* AO-4 (identified through Sanger sequencing of the RBP gene) capable of infecting C_6 strains, which has not yet been described (although phage genomes with similar RBP sequences have been deposited in GenBank, but the host *cwps* of these phages has not been described (71)).

It is important to note that the ^{His}GFP-RBP fusions were expressed only as a single RBP, whereas the complete baseplate of *Skunavirus* p2 was found to consist of a Dit (distal tail protein) hexameric ring, with an RBP trimer attached to each Dit monomer (24). Structurally however, all assessed phages were found to possess an attachment to the Dit similar to that of phages p2 (24) and 1358 (72) implying that their RBPs undergo a rotation to be activated and point to the receptor at the bacterial surface. In addition to the RBP there are also may be additional carbohydrate binding modules (CBMs) found in auxiliary binding proteins such as the neck passage structure (NPS), major tail extension protein (TpeX), and evolved Dit that have been found to contribute to host cell binding (31). Therefore, the addition of multiple RBP trimers and

auxiliary CBMs may impact the phage-host interaction during adsorption and this study describes only the most probable host of skunaviruses possessing a particular RBP. Despite these limitations, by refining and expanding the RBP groups it is possible through an analysis of RBP sequences within a population to better predict potential hosts of phages present. With a better understanding of the phage-host interactions occurring within these complex microbial communities it would then be possible to improve starter culture rotations strategies based on both the abundant and emerging phage populations within a fermentation environment. Importantly, this study further demonstrates the specific nature of the interactions of *Skunavirus* RBPs and their lactococcal host CWPS. And even possibly highlighting a reasonable demand for further investigations into the RBP-CWPS interactions of other gram positive ovococcoid bacteria.

REFERENCES

1. Daly C, Fitzgerald GF, Davis R. 1996. Biotechnology of lactic acid bacteria with special reference to bacteriophage resistance. *Antonie Van Leeuwenhoek* **70**:99-110.
2. Samson JE, Moineau S. 2013. Bacteriophages in food fermentations: new frontiers in a continuous arms race. *Annu Rev Food Sci Technol* **4**:347–368.
3. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. 2020. Dairy lactococcal and streptococcal phage–host interactions: An industrial perspective in an evolving phage landscape. *FEMS Microbiol Rev* **44**:909–932.
4. Spinelli S, Veesler D, Bebeacua C, Cambillau C. 2014. Structures and host-adhesion mechanisms of lactococcal siphophages. *Front Microbiol* **5**:3.
5. Mahony J, van Sinderen D. 2015. Novel strategies to prevent or exploit phages in fermentations, insights from phage–host interactions. *Curr Opin Biotechnol* **32**:8–13.
6. Deveau H, Labrie SJ, Chopin MC, Moineau S. 2006. Biodiversity and classification of lactococcal phages. *Appl Environ Microbiol* **72**:4338–4346.
7. Dupont K, Vogensen FK, Neve H, Bresciani J, Josephsen J. 2004. Identification of the receptor-binding protein in 936-species lactococcal bacteriophages. *Appl Environ Microbiol* **70**:5818–5824.
8. Muhammed MK, Olsen ML, Kot W, Neve H, Castro-Mejía JL, Janzen T, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, Vogensen FK.

2018. Investigation of the bacteriophage community in induced lysates of undefined mesophilic mixed-strain DL-cultures using classical and metagenomic approaches. *Int J Food Microbiol* **272**:61–72.

9. Mahony J, van Sinderen D. 2014. Current taxonomy of phages infecting lactic acid bacteria. *Front Microbiol* **5**:7.
10. Oliveira J, Mahony J, Hanemaaijer L, Kouwen TRHM, van Sinderen D. 2018. Biodiversity of bacteriophages infecting *Lactococcus lactis* starter cultures. *J Dairy Sci* **101**:96–105.
11. Mahony J, Oliveira J, Collins B, Hanemaaijer L, Lugli GA, Neve H, Ventura M, Kouwen TR, Cambillau C, van Sinderen D. 2017. Genetic and functional characterisation of the lactococcal P335 phage-host interactions. *BMC Genomics* **18**:146.
12. Mahony J, Murphy J, Van Sinderen D. 2012. Lactococcal 936-type phages and dairy fermentation problems: From detection to evolution and prevention. *Front Microbiol* **3**:1–9.
13. Marcó MB, Moineau S, Quiberoni A. 2012. Bacteriophages and dairy fermentations. *Bacteriophage* **2**:149–158.
14. Pujato SA, Quiberoni A, Mercanti DJ. 2019. Bacteriophages on dairy foods. *J Appl Microbiol* **126**:14–30.
15. Bissonnette F, Labrie S, Deveau H, Lamoureux M, Moineau S. 2000. Characterization of mesophilic mixed starter cultures used for the manufacture of aged cheddar cheese. *J Dairy Sci* **83**:620–627.
16. Dugat-Bony E, Lossouarn J, De Paepe M, Sarthou AS, Fedala Y, Petit MA, Chaillou S. 2020. Viral metagenomic analysis of the cheese

surface: a comparative study of rapid procedures for extracting viral particles. *Food Microbiol* **85**:103278.

17. Paillet T, Lamy-Besnier Q, Figueroa C, Petit M-A, Dugat-Bony E. 2024. Dynamics of the viral community on the cheese surface during maturation and persistence across production years. *mSystems* **9**: e0020124.
18. Colombo S, Arioli S, Gargari G, Neri E, Della Scala G, Mora D. 2018. Characterization of airborne viromes in cheese production plants. *J Appl Microbiol* **125**:1444–1454.
19. Muhammed MK, Kot W, Neve H, Mahony J, Castro-Mejía JL, Krych L, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, van Sinderen D, Vogensen FK. 2017. Metagenomic analysis of dairy bacteriophages: extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. *Appl Environ Microbiol* **83**:e00888-17.
20. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: An in silico evaluation of metagenome—enabled estimates of viral community composition and diversity. *PeerJ* **5**:e3817.
21. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. 2019. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**:12.
22. Mallawaarachchi V, Roach MJ, Decewicz P, Papudeshi B, Giles SK, Grigson SR, Bouras G, Hesse RD, Inglis LK, Hutton ALK, Dinsdale

EA, Edwards RA. 2023. Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* **39**:btad586.

23. Casey E, McDonnell B, White K, Stamou P, Crowley T, O'Neill I, Lavelle K, Hayes S, Lugli GA, Arboleya S, James K, Ventura M, Martinez I, Gueimonde M, Dal Bello F, Nally K, Mahony J, van Sinderen D. 2022. Needle in a Whey-Stack: PhRACS as a Discovery Tool for Unknown Phage-Host Combinations. *mBio* **13**:e03334-21.
24. Sciara G, Bebeacua C, Bron P, Tremblay D, Ortiz-Lombardia M, Lichière J, Van Heel M, Campanacci V, Moineau S, Cambillau C. 2010. Structure of lactococcal phage p2 baseplate and its mechanism of activation. *Proc Natl Acad Sci USA* **103**:15611-15616.
25. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, Chapot-Chartier M-P, Cambillau C, Holo H, van Sinderen D. 2020. The CWPS Rubik's cube: linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. *Mol Microbiol* **114**:582–596.
26. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C. 2013. Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol* **79**:4385–4392.
27. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, Van Sinderen D. 2014. Differences in lactococcal cell wall polysaccharide

structure are major determining factors in bacteriophage sensitivity. *mBio* **5**:e00880-14.

28. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. *Sci Rep* **6**:21345.
29. Sadovskaya I, Vinogradov E, Courtin P, Armalyte J, Meyrand M, Giaouris E, Palussière S, Furlan S, Péchoux C, Ainsworth S, Mahony J, van Sinderen D, Kulakauskas S, Guérardel Y, Chapot-Chartier M-P. 2017. another brick in the wall: a rhamnan polysaccharide trapped inside peptidoglycan of *Lactococcus lactis*. *mBio* **8**: e01303-17.
30. Chapot-Chartier MP, Vinogradov E, Sadovskaya I, Andre G, Mistou MY, Trieu-Cuot P, Furlan S, Bidnenko E, Courtin P, Péchoux C, Hols P, Dufre YF, Kulakauskas S. 2010. Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. *J Biol Chem* **285**:10464-10471.
31. Hayes S, Vincentelli R, Ramond L, Ramnd R, Nauta A, Sinderen D van, Cambillau C. 2019. Ubiquitous carbohydrate binding modules decorate 936 lactococcal siphophage virions. *Viruses* **11**:631.
32. Spinelli S, Desmyter A, Verrips CT, De Haard HJW, Moineau S, Cambillau C. 2006. Lactococcal bacteriophage p2 receptor-binding protein structure suggests a common ancestor gene with bacterial and mammalian viruses. *Nat Struct Mol Biol* **13**:85–89.
33. Szymczak P, Filipe SR, Covas G, Vogensen FK, Neves AR, Janzen T. 2018. Cell wall glycans mediate recognition of the dairy bacterium

Streptococcus thermophilus by bacteriophages. *Appl Environ Microbiol* **84**:e01847-18.

34. Tremblay DM, Tegoni M, Spinelli S, Campanacci V, Blangy S, Huyghe C, Desmyter A, Labrie S, Moineau S, Cambillau C. 2006. Receptor-binding protein of Lactococcus lactis phages: Identification and characterization of the saccharide receptor-binding site. *J Bacteriol* **188**:2400–2410.
35. Hyman P. 2013. Phage Receptor, p. 277–279. *In* Maloy, S, Hughes, K (ed), Brenner's Encyclopedia of Genetics, Second Edition. Academic Press, San Diego.
36. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura M. 2021. METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. *mSystems* **6**:e00583-21.
37. White K, Eraclio G, Lugli GA, Ventura M, Mahony J, Bello FD, van Sinderen D. 2024. A Metagenomics Approach to Enumerate Bacteriophages in a Food Niche, p. 185–199. *In* Tumban, E (ed), Bacteriophages: Methods and Protocols. Springer US, New York, NY.
38. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–410.

40. Corpet F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**:10881-10890.

41. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**:W242–W245.

42. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589.

43. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Applying and improving AlphaFold at CASP14. *Proteins* **89**:1711-1721.

44. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods* **19**:679–682.

45. Holm L. 2022. Dali server: structural unification of protein families. *Nucleic Acids Res* **50**:W210–W215.

46. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**:70–82.

47. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.

48. van Die IM, Bergmans HEN, Hoekstra WPM. 1983. Transformation in *Escherichia coli*: studies on the role of the heat shock in induction of competence. *J Gen Microbiol* **129**:663–670.

49. Lavelle K, Goulet A, McDonnell B, Spinelli S, van Sinderen D, Mahony J, Cambillau C. 2020. Revisiting the host adhesion determinants of *Streptococcus thermophilus* siphophages. *Microb Biotechnol* **13**:1765–1779.

50. Giesbers CAP, Fagan J, Parlindungan E, Palussière S, Courtin P, Lugli GA, Ventura M, Kulakauskas S, Chapot-Chartier MP, Mahony J, van Sinderen D. 2023. Reduced synthesis of phospho-polysaccharide in *Lactococcus* as a strategy to evade phage infection. *Int J Food Microbiol* **407**:110415.

51. Stefano R, Valérie C, Stéphanie B, Silvia S, Denise T, Sylvain M, Mariella T, Christian C. 2006. Crystal structure of the receptor-binding protein head domain from *Lactococcus lactis* Phage bIL170. *J Virol* **80**:9331–9335.

52. Goulet A, Mahony J, Cambillau C, van Sinderen D. 2022. Exploring structural diversity among adhesion devices encoded by Lactococcal P335 phages with AlphaFold2. *Microorganisms* **10**:2278.

53. Spinelli S, Campanacci V, Blangy S, Moineau S, Tegoni M, Cambillau C. 2006. Modular structure of the receptor binding proteins of Lactococcus lactis phages. The RBP structure of the temperate phage TP901-1. *J Biol Chem* **281**:14256-14262.

54. Parlindungan E, Sadovskaya I, Vinogradov E, Lugli GA, Ventura M, van Sinderen D, Mahony J. 2024. Novel cell wall polysaccharide genotypes and structures of lactococcal strains isolated from milk and fermented foods. *Int J Food Microbiol* **424**:110840.

55. Frantzen CA, Kleppen HP, Holo H. 2018. *Lactococcus lactis* Diversity in Undefined Mixed Dairy Starter Cultures as Revealed by Comparative Genome Analyses and Targeted Amplicon Sequencing of *epsD*. *Appl Environ Microbiol* **84**:e02199-17.

56. Erkus O, de Jager VCL, Spus M, van Alen-Boerrigter IJ, van Rijswijck IMH, Hazelwood L, Janssen PWM, van Hijum SAFT, Kleerebezem M, Smid EJ. 2013. Multifactorial diversity sustains microbial community stability. *ISME* J **7**:2126-2136.

57. Somerville V, Berthoud H, Schmidt RS, Bachmann HP, Meng YH, Fuchsmann P, von Ah U, Engel P. 2022. Functional strain redundancy and persistent phage infection in Swiss hard cheese starter cultures. *ISME J* **16**:388–399.

58. Jung JY, Lee SH, Kim JM, Park MS, Bae JW, Hahn Y, Madsen EL, Jeon CO. 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. *Appl Environ Microbiol* **77**:2264–2274.

59. White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* **90**:e02152-23.

60. Mahony J, van Sinderen D. 2022. Virome studies of food production systems: time for 'farm to fork' analyses. *Curr Opin Biotechnol* **73**:22–27.

61. Chmielewska-Jeznach M, Bardowski JK, Szczepankowska AK. 2018. Molecular, physiological and phylogenetic traits of *Lactococcus* 936-type phages from distinct dairy environments. *Sci Rep* **8**:12540.

62. Jolicoeur AP, Lemay ML, Beaubien E, Bélanger J, Bergeron C, Bourque-Leblanc F, Doré L, Dupuis MÈ, Fleury A, Garneau JE, Labrie SJ, Labrie S, Lacasse G, Lamontagne-Drolet M, Lessard-Hurtubise R, Martel B, Menasria R, Morin-Pelchat R, Pageau G, Samson JE, Rousseau GM, Tremblay DM, Duquenne M, Lamoureux M, Moineau S. 2023. Longitudinal study of *Lactococcus* phages in a Canadian cheese factory. *Appl Environ Microbiol* **89**:e0042123.

63. Dupont K, Janzen T, Vogensen FK, Josephsen J, Stuer-Lauridsen B. 2004. Identification of *Lactococcus lactis* genes required for bacteriophage adsorption. *Appl Environ Microbiol* **70**:5825–5832.

64. Philippe C, Cornuault JK, de Melo AG, Morin-Pelchat R, Jolicoeur AP, Moineau S. 2023. The never-ending battle between lactic acid bacteria and their phages. *FEMS Microbiol Rev* **47**:fuad035.

65. Mavrich TN, Casey E, Oliveira J, Bottacini F, James K, Franz CMAP, Lugli GA, Neve H, Ventura M, Hatfull GF, Mahony J, van Sinderen D. 2018. Characterization and induction of prophages in human gut-associated *Bifidobacterium* hosts. *Sci Rep* **8**:12772.

66. Bouchard JD, Moineau S. 2000. Homologous recombination between a lactococcal bacteriophage and the chromosome of its host strain. *Virology* **270**:65-75.

67. Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, Baker S, Davis P, Simmonds M, Churcher C, Mungall K, Quail MA, Preston A, Harvill ET, Maskell DJ, Eiserling FA, Parkhill J, Miller JF. 2004. Genomic and genetic analysis of *Bordetella* bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol* **186**:1503–1517.

68. Ventura M, Lee JH, Canchaya C, Zink R, Leahy S, Moreno-Munoz JA, O'Connell-Motherway M, Higgins D, Fitzgerald GF, O'Sullivan DJ, van Sinderen D. 2005. Prophage-like elements in bifidobacteria: Insights from genomics, transcription, integration, distribution, and phylogenetic analysis. *Appl Environ Microbiol* **71**:8692–8705.

69. Benler S, Cobián-Güemes AG, McNair K, Hung SH, Levi K, Edwards R, Rohwer F. 2018. A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* **6**:191.

70. Labrie SJ, Moineau S. 2007. Abortive infection mechanisms and prophage sequences significantly influence the genetic makeup of emerging lytic lactococcal phages. *J Bacteriol* **189**:1482-1487.

71. Marcelli B. 2020. PhD thesis. University of Groningen, Groningen, Netherlands. *Lactococcus lactis* bacteriophages: phage-host interaction and phage transduction.

72. Spinelli S, Bebeacua C, Orlov I, Tremblay D, Klaholz BP, Moineau S, Cambillau C. 2014. Cryo-Electron microscopy structure of lactococcal siphophage 1358 virion. *J Virol* **88**:8900–8910.

SUPPLEMENTARY MATERIALS

Supplementary Table S1. Bacterial strains used in this study.

Lactococcal strains	CWPS type	Genbank Accession number	Source/Reference
L. cremoris UC509.9	A	CP003157	UCC
L. lactis IL1403	B	AE005176	(1)
L. cremoris MG1363	C_1	AM406671	(2)
L. cremoris 3107	C_2	CP031538	(3)
L. cremoris SK11	C_3	CP000425	(4)
L. cremoris W34	C_4	CP032430	UCC
L. cremoris 1196	C_4	CP032148	UCC
L. lactis IO-1	C_5	AP012281	(5)
L. lactis UC06	C_5	CP015902	UCC
L. cremoris A76	C_6	CP003132	(6)
L. lactis Tome1F	C_7	NZ_JAHIBU000000000	UCC
L. cremoris Peco9D	C_7	NZ_JAHIBV000000000	UCC
L. cremoris RMN8G	C_9	NZ_JAHIBS000000000	UCC
L. lactis Tempeh6L	C_{10}	NZ_JAHIBT000000000	UCC
L. lactis TolaII67	C_{11}	JAHIBR000000000	UCC
L. lactis 184	D	CP015895	UCC

Supplementary Table S2. Whey virome reads taxonomic profiling.

Genera	T1	BH1	BR	BT	AD1	AF3	AF9	BQ	BA6	BB4	BZ	AO	BS	V1	V2	AE3	AI3	AM2	AN3	AN4	AV6	BC9	BD	BE9	BV	CA	CD	BP
Eukaryota-Fungi_Cantharellus	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	10%	0%
Eukaryota-Fungi_Echinosphaeria	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Eukaryota-Fungi_Geoglossum	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	6%	0%	0%	0%	0%	0%	27%	0%
Eukaryota-Protist_Rostrostelium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%
Prokaryote_Achromobacter	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
Prokaryote_Acinetobacter	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Agrobacterium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%
Prokaryote_Bacillus	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	2%	31%	34%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Bacteroides	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Prokaryote_Confluentibacter	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%
Prokaryote_Cutibacterium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Enterococcus	0%	0%	0%	0%	1%	1%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Faecalibacterium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Prokaryote_Flavobacterium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	4%	0%
Prokaryote_Herbaspirillum	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%
Prokaryote_Klebsiella	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Lactobacillus	0%	0%	0%	0%	1%	1%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Lactococcus	13%	32%	12%	6%	95%	1%	0%	3%	84%	1%	0%	4%	4%	4%	4%	12%	12%	34%	46%	3%	4%	1%	0%	2%	0%	0%	0%	5%
Prokaryote_Lelliottia	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	26%	0%	0%	0%	0%	0%	0%
Prokaryote_Microbacterium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
Prokaryote_Mycolicibacterium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Ochrobactrum	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Prokaryote_Pantoea	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
Prokaryote_Paraburkholderia	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%
Prokaryote_Pseudomonas	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	10%	0%
Prokaryote_Rahnella	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%
Prokaryote_Rhizobium	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%
Prokaryote_Sporosarcina	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%
Prokaryote_Stenotrophomonas	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	2%	0%	0%	16%	0%	0%	0%	0%	0%	0%
Prokaryote_Streptococcus	0%	0%	0%	0%	1%	14%	10%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%
Prokaryote_Xanthomonas	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
Virus_Demerecviridae	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	11%	0%
Virus_Myoviridae	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%	25%	0%
Virus_Podoviridae	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	9%	0%
Virus_Poxviridae	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Virus_Siphoviridae	86%	68%	87%	94%	2%	81%	89%	95%	13%	99%	100%	94%	96%	96%	96%	84%	86%	3%	10%	97%	76%	36%	100%	98%	100%	100%	0%	93%

Supplementary Table S3. Phageome-derived RBP classification.

RBP	Length	Head domain	RBP Group	Subgroup
AD1-RBP1	269	p2 group 1	I	B
AD1-RBP4	269	p2 group 1	I	B
AF3-RBP1	269	p2 group 1	I	B
AI3-RBP2	269	p2 group 1	I	B
AO-RBP3	269	p2 group 1	I	B
BA6-RBP3	269	p2 group 1	I	B
BC9-RBP2	269	p2 group 1	I	B
BD-RBP2	241	p2 group 1	I	B
BE9-RBP1	269	p2 group 1	I	B
BQ-RBP6	269	p2 group 1	I	B
BQ-RBP7	269	p2 group 1	I	B
BT-RBP1	269	p2 group 1	I	B
BV-RBP5	269	p2 group 1	I	B
BV-RBP6	269	p2 group 1	I	B
BZ-RBP3	269	p2 group 1	I	B
CA-RBP4	269	p2 group 1	I	B
CA-RBP6	269	p2 group 1	I	B
V2-RBP1	269	p2 group 1	I	B
V2-RBP3	269	p2 group 1	I	A
AD1-RBP3	269	p2 group 1	I	A
AD1-RBP5	269	p2 group 1	I	A
AE3-RBP1	269	p2 group 1	I	A
AE3-RBP2	269	p2 group 1	I	A
AI3-RBP1	269	p2 group 1	I	A
AI3-RBP5	269	p2 group 1	I	A
AN3-RBP1	269	p2 group 1	I	A
AN3-RBP2	269	p2 group 1	I	A
AN4-RBP1	269	p2 group 1	I	A
AO-RBP2	269	p2 group 1	I	A
AO-RBP5	241	p2 group 1	I	A
AV6-RBP1	269	p2 group 1	I	A
BC9-RBP1	269	p2 group 1	I	A
BC9-RBP3	269	p2 group 1	I	A
BH1-RBP2	269	p2 group 1	I	A
BP-RBP1	269	p2 group 1	I	A
BQ-RBP4	269	p2 group 1	I	A
BQ-RBP5	269	p2 group 1	I	A
BR-RBP1	241	p2 group 1	I	A
BV-RBP1	269	p2 group 1	I	A
BV-RBP8	269	p2 group 1	I	A
BZ-RBP2	269	p2 group 1	I	A
BZ-RBP4	269	p2 group 1	I	A
CA-RBP2	269	p2 group 1	I	A
CA-RBP8	269	p2 group 1	I	A
CD-RBP1	269	p2 group 1	I	A
V1-RBP1	269	p2 group 1	I	A
BV-RBP3	275	p2 group 1	I	C
BQ-RBP2	274	p2 group 1	I	C

CA-RBP1	275	p2 group 1	I	C
V2-RBP2	257	p2 group 1	I	D
BA6-RBP1	272	bIL70 group 3	II	
BA6-RBP2	274	Tuc2009-like group 4	III	
BB4-RBP2	274	Tuc2009-like group 4	III	
BV-RBP7	274	Tuc2009-like group 4	III	
BZ-RBP1	274	Tuc2009-like group 4	III	
CA-RBP3	274	Tuc2009-like group 4	III	
AD1-RBP2	273	p2-like group 2	IV	
AO-RBP1	275	Phage T4 gp12 group 5	V	A
BC9-RBP4	264	p2-like group 2	VI	
AE3-RBP4	272	p2-like group 2	VI	
AE3-RBP5	272	p2-like group 2	VI	
AI3-RBP3	273	p2-like group 2	VI	
AN3-RBP3	273	p2-like group 2	VI	
AN3-RBP5	273	p2-like group 2	VI	
BB4-RBP1	273	p2-like group 2	VI	
BB4-RBP3	273	p2-like group 2	VI	
AE3-RBP3	348	p2-like group 2	VII	A
AF9-RBP1	316	p2-like group 2	VII	A
AI3-RBP4	317	p2-like group 2	VII	A
AN3-RBP4	348	p2-like group 2	VII	A
BB4-RBP4	348	p2-like group 2	VII	A
AO-RBP4	267	p2-like group 2	VIII	
BQ-RBP3	267	p2-like group 2	VIII	
BV-RBP2	267	p2-like group 2	VIII	
BV-RBP4	269	p2-like group 2	IX	
CA-RBP5	269	p2-like group 2	IX	
BD-RBP1	285	p2-like group 2	X	
BE9-RBP2	285	p2-like group 2	X	
BH1-RBP1	285	p2-like group 2	X	
BQ-RBP1	285	p2-like group 2	X	
CA-RBP7	285	p2-like group 2	X	

Supplementary Table S4. Relative abundance of *Skunavirus* RBP groups present in whey virome samples. The total number of reads, in a given sample, that mapped against any of the RBP groups is indicated in red.

Whey Viromes:			AD1	BP	V1	V2	AE3	AI3	AM2*†	AN3	AN4	AV6	BC9	BD	BE9	BV	CA	CD*	AF3*	AF9*	BQ	BA6	BB4	BZ	T1*	BH1	BR*	BT	AO
(Total reads mapped)			6095	143019	37885	23440	120063	58212	54	2979	106815	11911	66043	225378	225432	179810	264538	63	70	385	240842	25955	138646	240708	0	1282	50	77519	128418
RBP groups	Group IA	*(CWPS type C)*	92.2	99.7	96.7	14.5	93.2	98.7	0.0	22.7	99.9	97.1	97.7	59.5	3.5	33.2	23.7	58.5	33.9	7.1	0.4	0.3	0.0	2.8	0.0	91.7	100.0	93.0	51.8
	Group IB		1.6	0.2	3.3	85.5	4.1	0.0	0.0	0.6	0.0	0.0	1.7	3.4	1.6	1.2	1.9	0.0	16.1	0.0	2.2	2.3	0.0	0.1	0.0	0.0	0.0	6.3	17.8
	Group IC		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.2	26.7	0.0	0.0	0.0	27.9	0.3	0.0	0.0	0.0	0.0	0.0	0.1	3.6
	Group ID		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Group II	*(CWPS type B)*	0.0	0.0	0.0	0.0	0.1	0.0	0.0	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	87.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Group III	*(CWPS type B & C)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.1	0.0	0.0	0.0	0.7	0.1	0.0	6.3	5.0	0.0	9.1	1.1	97.1	0.0	0.0	0.0	0.0	0.0
	Group IV	*(CWPS type U)*	6.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Group VA	*(CWPS type A)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.2
	Group VB	*(CWPS type A)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Group VI	*(CWPS type D)*	0.0	0.0	0.0	0.0	0.3	0.2	0.0	10.1	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	7.9	0.0	0.0	14.2	0.0	0.0	0.0	0.0	0.0	0.0
	Group VIIA	*(CWPS type U)*	0.0	0.0	0.0	0.0	2.2	1.0	0.0	62.0	0.1	2.5	0.0	0.0	0.0	0.0	0.0	0.0	43.6	80.0	0.0	0.1	84.7	0.0	0.0	0.0	0.0	0.0	0.0
	Group VIIB	*(CWPS type U)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Group VIII	*(CWPS type C_α)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	55.1	43.2	41.5	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.4	15.0
	Group IX	*(CWPS type U)*	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
	Group X	*(CWPS type C_β)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.2	94.9	0.0	2.7	0.0	0.0	0.0	69.1	0.0	0.0	0.0	0.0	8.3	0.0	0.1	0.0
	Group XI	*(CWPS type U)*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	None		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0

† Sample has less than 100,000 quality filtered reads.
* Less than 1,000 total reads mapped.

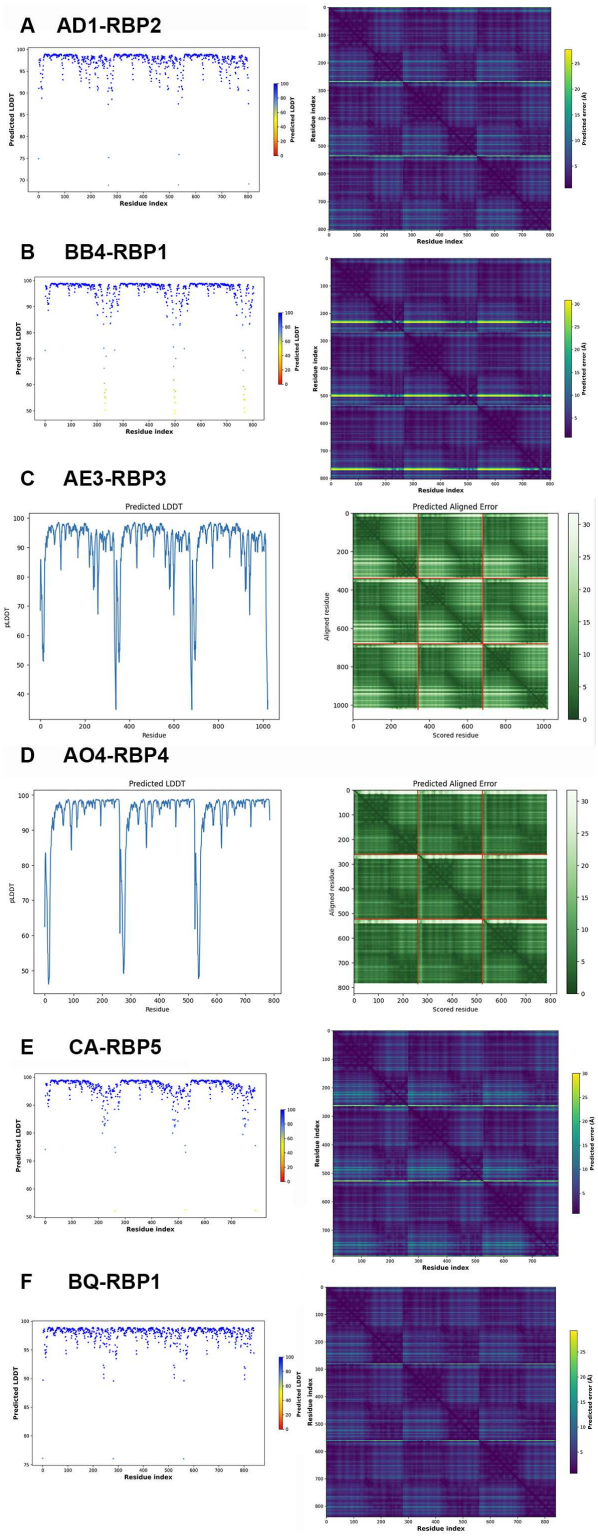
Supplementary Table S5. Full representation of PSP component of the lactococcal strains ^{His}GFP-RBP were capable of binding to.

^{His}GFP-RBP	RBP Cluster	Strain & CWPS	PSP structure
AO-RBP2, AD1-RBP5	IA	3107 (C_2)*	-2-β-Gal*f*-6-α-GlcNAc-*P*-6-α-Glc-3-β-Gal*f*-3-β-GlcNAc-
		MG1363 (C_1)*	-2-β-Gal*f*-6-a-Glc-*P*-6-β-GlcNAc-3-α-Rha-3-β-GlcNAc- |6 α-Glc
BB4-MG12	IC	W34 (C_4)*	-6-β-Gal*f*-6-α-Glc-6-β-GlcNAc-3-β-Gal*f*-3-β-GlcNAc-
		1196 (C_4)*	-6-β-Gal*f*-6-α-Gal-6-β-GlcNAc-3-β-Gal*f*-3-β-GlcNAc-
		IO-1 (C_5)*	-2-β-Gal*f*-5-Ara-ol-1-*P*-2-β-Gal*f*-3-α-GlcNAc- |3 α-GlcNAc
V2-RBP2	ID	RMN8G (C_9)*	α-Rha |2 -2-β-Gal*f*-6-α-Glc-1-*P*-6-α-Glc-3-β-Gal*f*-3-β-GlcNAc- |6 α-Glc
		3107 (C_2)*	-2-β-Gal*f*-6-α-GlcNAc-*P*-6-α-Glc-3-β-Gal*f*-3-β-GlcNAc-
		MG1363 (C_1)*	-2-β-Gal*f*-6-a-Glc-*P*-6-β-GlcNAc-3-α-Rha-3-β-GlcNAc- |6 α-Glc
BA6-RBP1	II	IL1403 (B)†	β-Glc-4-α-GalNAc(6*P*-1-Gro)-3-β-GlcNAc-
BB4-RBP2	III		
BA6-RBP2			
BB4-RBP1	VI	184 (D)*	β-Glc |3 -6-α-GalNAc-3-β-GalNAc-5-Ribitol-1-*P*- |4 β-Glc*6OAla*
AE3-RBP3, BB4-RBP4	VIIB	N.D.	
AD1-RBP2	IV	N.D.	
BQ-RBP1	X	SK11 (C_3)*	-6-β-Gal*f*-6-α-Gal-6-α-Glc-*P*-6-α-Gal-3-β-GlcNAc- |6 α-Glc
AO-RBP4	VIII	A76 (C_6)*	-6-β-Gal-2-α-Gal-3-β-RhaOAc-4-β-Glc6*P*-6-β-Gal*f*-6-β-Glc-
CA-RBP5	IX	N.D.	
AO-RBP1‡	V	UC509.9 (A)‡	β-Gal(6*P*-1-Gro)-2-α-Rha-2-α-Rha-3-β-GlcNAc- |3 |6 α-Rha α-Glc

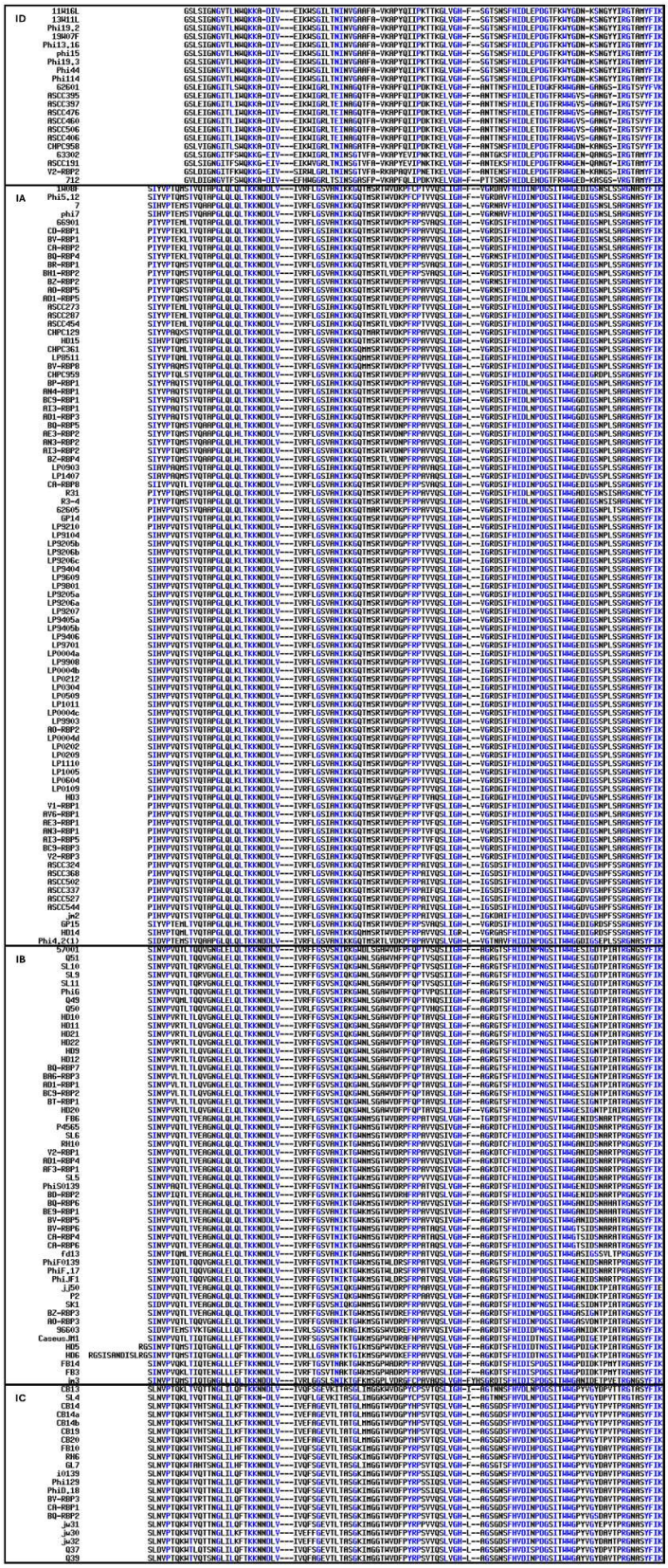
* Side chains are present as repeating units on the rhamnan backbone.

† Side chains are irregularly present on the rhamnan backbone, backbone, partially without a glycerophosphate group.

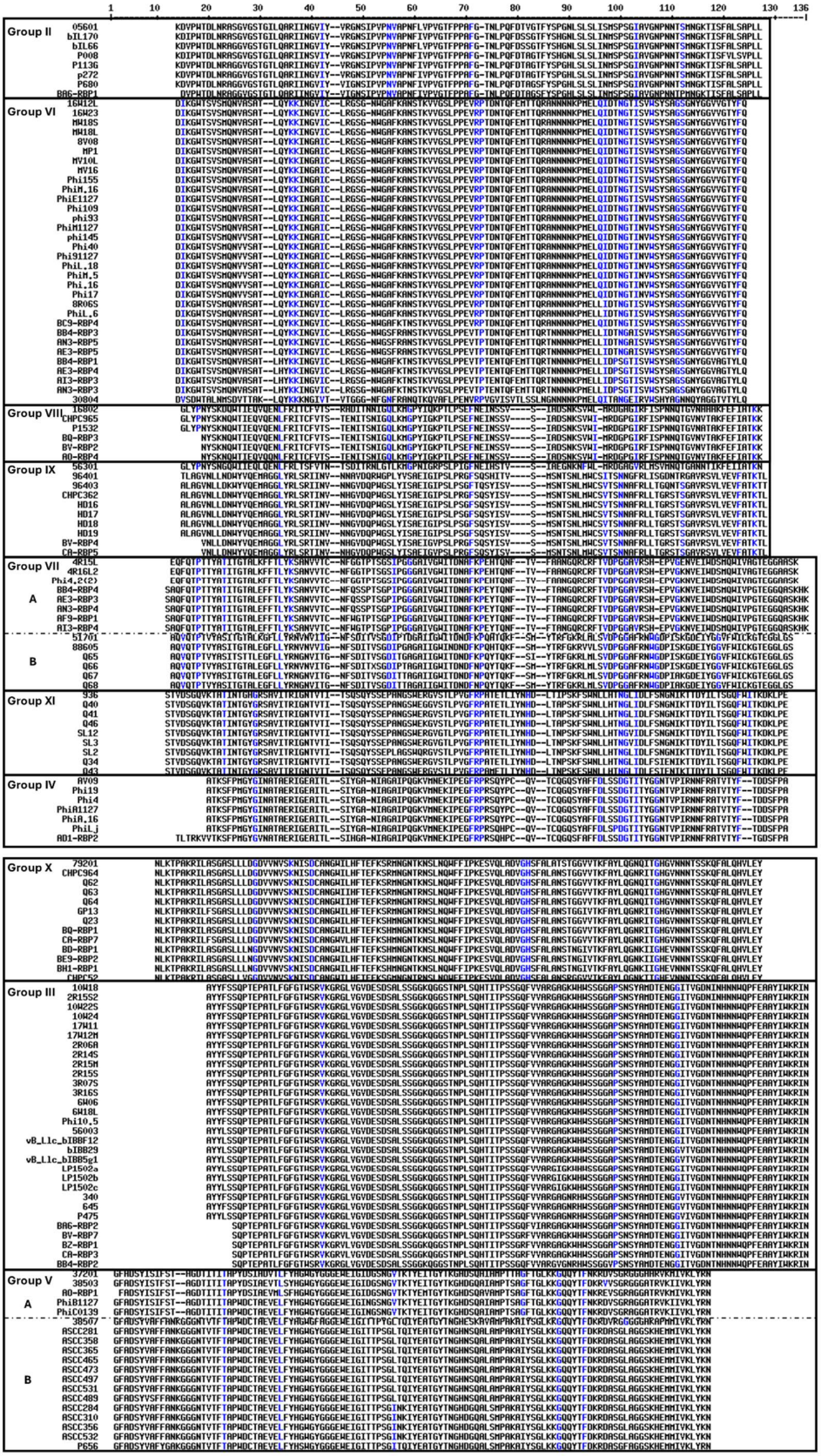
‡ Not determined through binding assays carried out in the current study but has been established in previous studies (7, 8). Rhamnan backbone substituted with side chain approximately every third unit.



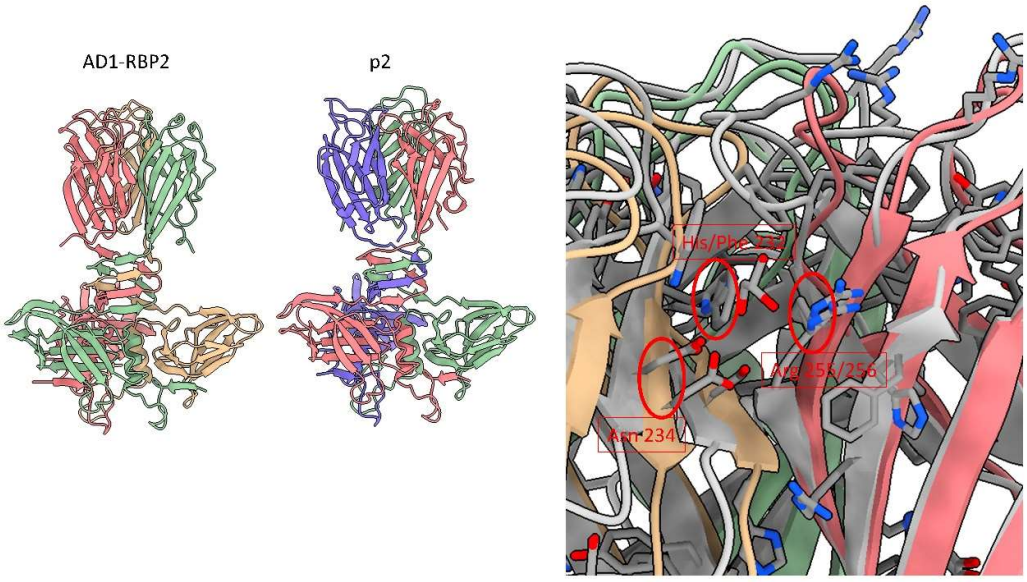
Supplementary Figure S1. pLDDT values and predicted aligned errors (PAE) plots of predicted structures. The pLDDT plots (left) and PAE plots (right) correspond to RBP trimers; the numbering is that of the main RBPs Figure.



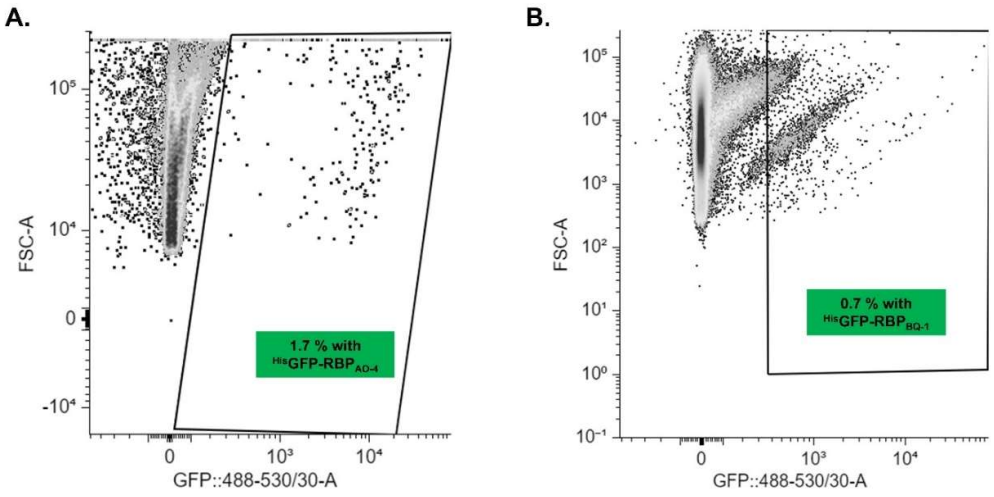
Supplementary Figure S2. Sequence alignment of group I RBP head domains.



Supplementary Figure S3. Sequence alignment of RBP groups II-XI head domains.



Supplementary Figure S4. Group IV representative AD1-RBP2 shares two residues that are involved in the binding of phages p2 and TP901-1.



Supplementary Figure S5. Dot analysis of undefined culture M1 labelled with (A) ^{His}GFP-RBP$_{AO-4}$ and (B) ^{His}GFP-RBP$_{BQ-1}$.

SUPPLEMENTARY MATERIALS REFERENCES

1. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A. 2001. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res* **11**:731–753.
2. Linares DM, Kok J, Poolman B. 2010. Genome sequences of *Lactococcus lactis* MG1363 (revised) and NZ9000 and comparative physiological studies. *J Bacteriol* **192**:5806–5812.
3. Østergaard Breum S, Neve H, Heller KJ, Vogensen FK. 2007. Temperate phages TP901-1 and φLC3, belonging to the P335 species, apparently use different pathways for DNA injection in *Lactococcus lactis* subsp. *cremoris* 3107. *FEMS Microbiol Lett* **276**:156–164.
4. de Vos WM, Vos P, de Haard H, Boerrigter I. 1989. Cloning and expression of the *Lactococcus lactis* subsp. *cremoris* SK11 gene encoding an extracellular serine proteinase. *Gene* **85**:169–176.
5. Kato H, Shiwa Y, Oshima K, Machii M, Araya-Kojima T, Zendo T, Shimizu-Kadota M, Hattori M, Sonomoto K, Yoshikawa H. 2012. Complete genome sequence of *Lactococcus lactis* IO-1, a lactic acid bacterium that utilizes xylose and produces high levels of L-lactic acid. *J Bacteriol* **194**:2102–2103.
6. Bolotin A, Quinquis B, Ehrlich SD, Sorokin A. 2012. Complete genome sequence of *Lactococcus lactis* subsp. *cremoris* A76. *J Bacteriol* **194**:1241–1242.

7. Hayes S, Vincentelli R, Ramond L, Ramnd R, Nauta A, Sinderen D van, Cambillau C. 2019. Ubiquitous carbohydrate binding modules decorate 936 lactococcal siphophage virions. *Viruses* **11**:631.
8. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, Van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci Rep* **6**:21345.

Chapter 4. Biodiversity of skunaviruses isolated from dairy fermentation factories employing mesophilic undefined starter cultures

This chapter is in preparation as a manuscript as:

White K, Eraclio G, McDonnell B, Lugli GA, Ventura M, Biere N, Hille F, Volontè F, Franz C, Cambillau C, Dal Bello F, Mahony J, van Sinderen D. Biodiversity of skunaviruses isolated from dairy fermentation factories employing mesophilic undefined starter cultures.

Experiments, analyses, and first draft preparations were performed by the candidate, except for AlphaFold2 structural predictions and analysis which was performed by C.C. and electron microscopy which was performed by N.B., F.H., and C.F.

ABSTRACT

Mesophilic starter cultures rely on the activity of lactic acid bacteria such as *Lactococcus lactis* and *Lactococcus cremoris* to acidify milk in dairy fermentations. Bacteriophage infection of starter bacteria remains one of the most significant threats to a successful fermentation, as they may cause costly disruptions and loss of production. The exclusively virulent skunaviruses represent one of the most problematic lactococcal phage genera in dairy fermentation facilities and have consequently been studied extensively in recent decades. In the present study, the diversity of lactococcal phages in whey samples originating from fermentations employing undefined mesophilic starter cultures was assessed by culture-dependent phage screening, targeted isolation based on predicted receptor binding protein specificity, and **ph**age **R**BP-**a**ctivated **c**ell **s**orting (PhRACS). The genomes of 26 distinct *Skunavirus* isolates were sequenced and compared to all publicly available *Skunavirus* genomes. Furthermore, phageome-derived *Skunavirus* contigs were analysed to determine the diversity of phages present in the dairy fermentation facilities. Analysis of auxiliary carbohydrate binding module (CBM)-containing proteins encoded by the *Skunavirus* isolates revealed the presence of novel CBMs. Finally, host range analysis revealed a clear correlation between specific *Skunavirus* receptor binding protein phylogroups and the cell wall polysaccharide-associated genotype of the corresponding host strain.

INTRODUCTION

Bacteriophages (or phages) are viruses that infect and replicate within bacteria, and that occupy practically every ecological niche on Earth (1). One niche where they interfere with human food production activity is the dairy fermentation environment. Dairy fermentations primarily rely on the activity of members of the lactic acid bacteria (LAB), such as particular species of *Lactococcus, Leuconostoc*, *Lactobacillus, Pediococcus, Enterococcus* and *Streptococcus*, to acidify milk or to contribute to flavour formation during the maturation and ripening processes (2, 3). For large-scale industrial fermentations, LAB typically form the basis of starter cultures which can possess a defined (small number of strains in a known composition) or undefined (unknown number of strains in an unknown combination) composition (4). Fermentations are also characterized based on their optimal growth temperature and can be classified as either thermophilic or mesophilic (5). Mesophilic starter cultures tend to incorporate various lactococcal strains, typically *Lactococcus lactis* and *Lactococcus cremoris* while they may also contain *Leuconostoc* spp. (6).

Extensive efforts focused on characterising the diversity of lactococcal phages have led to the identification of ten distinct taxonomic groups (7). In fermentation environments, the three most frequently encountered lactococcal phage groups are the exclusively virulent skunaviruses (previously termed the 936 group phages) and ceduoviruses (previously termed the c2 group phages), and the P335 phages whose members can either be virulent or temperate (8–10). Infection of starter cultures by virulent lactococcal phages causes slow or failed fermentations, resulting in substantial economic losses

(11). Despite the extensive research efforts to solve “the phage problem” over the past four decades, they remain a considerable challenge in the fermentation sector. The dominance of skunaviruses in global mesophilic dairy fermentations is reflected by the dominance of genomes of these lactococcal phages in public databases, i.e. there are currently 212 complete *Skunavirus* genomes published in GenBank (https://www.ncbi.nlm.nih.gov/labs/virus/; as of June 20, 2024) (12–18). However, studies investigating phage populations associated with artisanal cheese whey samples have found higher instances of so-called “rare” lactococcal phage groups such as *Audreyjarvisvirus* (previously the 949 group phages) and *Teubervirus* (previously termed the P087 group phages) (19).

The dominance of lactococcal skunaviruses in industry paired with the availability of extensive genomic data has facilitated thorough analysis of their genome architecture, functional assignments, host interactions and 3-D structure reconstructions of individual proteins and protein complexes (20–24). Skunaviruses, like other *Siphoviridae* phages, possess an isometric capsid and a long, non-contractile tail (25). These are double-stranded (ds)DNA-containing phages with genomes ranging in size from 26-32 kb, with a modular genetic architecture comprised of regions involved in DNA packaging, capsid and tail morphogenesis, and cell lysis (late-expressed genes) and replication (early and middle-expressed genes) (24, 26–31). The genome architecture of the *Skunavirus* tail morphogenesis module is well conserved, and typically incorporates genes encoding the major tail protein, tail tape measure protein (TMP), distal tail protein (Dit), tail-associated lysin (Tal), and receptor binding protein (RBP). However, there is notable sequence diversity within confined

regions of certain genes encoding the host recognition device within the morphogenesis module (32). Within the *Skunavirus* genome, three distinct regions have been identified that exhibit relatively high genetic diversity: the first (~2 kb) corresponds to the late-expressed region encompassing the neck passage structure protein-encoding gene; the second region (~2 kb) contains the late/early gene expression junction (encoding the TMP, RBP, holin, lysin, and early expressed genes of unknown function) with the highest level of sequence variation within the gene encoding the RBP, while the third region (~2.5 kb) encompasses the start of the early-expressed region and entire middle-expressed region (32). In addition, these regions of genetic diversity have been correlated with a higher recombination rate suggesting the important role of recombination (in addition to point mutations) in generating genetic novelty and diversity of, and associated adaptation by skunaviruses (32).

The *Skunavirus* host adsorption apparatus or baseplate, located at the distal end of the phage tail, consists of heteropolymeric assembly of Dit, Tal, and RBP subunits, with Dit forming a hexameric ring in which an RBP trimer attaches to each Dit monomer (23). In skunaviruses, the C-terminal “head” domain of the RBP is responsible for the specific recognition of and binding to a surface-exposed host receptor moiety thereby commencing phage infection (24). The RBP may be supported by auxiliary carbohydrate binding module (CBM)-containing proteins, located at the distal end of the phage tail or proximal end as so-called “whiskers” that can form a neck passage structure (23, 33). One example of these proteins includes the so-called “evolved” Dit proteins which possess an internal insertion encoding a CBM which has

previously been shown to promote host binding (34, 35). CBMs present in the RBP and auxiliary binding proteins facilitate binding to the lactococcal cell surface-exposed saccharidic receptor, also known as the **c**ell **w**all **p**oly**s**accharide (CWPS), which consists of a relatively conserved rhamnan backbone and a variable side chain, the latter also referred to as the polysaccharide pellicle or PSP (22, 36, 37). There is considerable sequence diversity within the *cwps* gene cluster, responsible for the biosynthesis of these two CWPS elements, and this is consistent with the compositional and structural diversity of the associated CWPSs of lactococcal strains (particularly that of the surface-exposed PSP) (38–40).

In the current study, we report the isolation, genome sequence analysis, and characterisation of 27 lactococcal phages (26 *Skunavirus* and one *Ceduovirus*) isolated from whey samples collected from various European dairy fermentation factories which employ mesophilic, undefined starter cultures. These phages were isolated through both culture-dependent and -independent techniques. The phylogeny of these novel *Skunavirus* isolates was investigated in relation to publicly available *Skunavirus* genomes. *Skunavirus* RBP head domains were shown to be correlated to the CWPS genotype of phage-sensitive hosts with binding resolution at the *cwps* types and C-subtypes established. Understanding both the diversity of *cwps* (sub)types of strains within a starter culture as well as the RBP diversity of the *Skunavirus* population present within dairy fermentation facilities, will assist in risk management of the potential impact of phage infection on the associated starter cultures. In addition, phageome-derived *Skunavirus* contigs obtained from the same samples were investigated to determine the presence of

additional (partial) phage genomes that may be underrepresented in culture-based screening approaches.

MATERIALS AND METHODS

Bacterial strains, bacteriophages and growth conditions

Bacterial strains used in this study are presented in Supplementary Table S1. Lactococcal strains were grown overnight at 30 °C in M17 broth (Millipore Sigma Aldrich, Gillingham, UK) supplemented with 0.5 % lactose (LM17). Phages were propagated on appropriate host strains (at time of inoculation with host) in LM17 broth supplemented with 10 mM $CaCl_2$ (Millipore Sigma Aldrich, Gillingham, UK) and incubated at 30 °C until lysis occurred. Phage lysates were filtered through a 0.45 µM filter (Sarstedt, Nümbrecht, Germany) and stored at 4 °C. Plaque assays were performed using the previously described double agar method for both phage isolation and host range analysis (41).

Phage screening and isolation

32 whey samples (Supplementary Table S2) collected (between 2019 and 2023) from dairy fermentation factories employing mesophilic undefined starter cultures (typically supplemented with additional strains/cultures) from seven countries across Europe (Spain, Austria, France, Poland, Sweden, Denmark, and Czech Republic) were found to have phage activity. The whey samples were screened for the presence of phages that infect strain(s) from a collection of 92 isolates (previously isolated from mesophilic undefined starter cultures M1 and M2; isolates listed in Supplementary Table S3) representing four *cwps* genotypes (23 isolated of each of the C_1-C_4 subtypes). Whey samples were also screened against representative strains (from an in-house

strain collection) of *cwps* B and D genotypes (Supplementary Table S1). Whey samples were filtered using 0.45 µm filters prior to phage screening to remove residual bacterial debris and stored at -20 °C until use. In a 96-well plate (Sarstedt), 20 µl of a fresh overnight culture was added to a well containing 150 µL LM17 broth supplemented with 10 mM $CaCl_2$ and 20 µl of a given filtered whey sample diluted 1:20 in SM buffer [50 mM Tris-HCl, 100 mM NaCl, 10 mM $MgSO_4$, 10 mM $CaCl_2$]. A second plate using 20 µl SM buffer in place of filtered whey was prepared as a positive control for growth. Plates were incubated at 30 °C and were visually examined for lysis (by comparison to the uninfected control), initially after 6 hrs and following overnight incubation for the presence of presumed phage infection.

Phage isolates representing a spread of geographic sampling locations and host *cwps* genotypes were selected for further analysis. Lysates of potential phages (in which culture lysis was observed either during the first 6 hr or complete lysis following overnight incubation) were diluted in SM buffer (1:10), filtered (0.45 µm), and spotted on its corresponding host to validate the presence of phages through double layer plaque assays (41). Phages were propagated from a single plaque against an appropriate lactococcal host. A single plaque was added to 10 mL of LM17 (supplemented with 10 mM $CaCl_2$) concurrently inoculated with 100 µL of the appropriate host (from a fresh overnight culture) and incubated at 30 °C until lysis had occurred. Phage lysates were filtered through a 0.45 µm filter and stored at 4 °C. The selected phages were genotyped using a previously described multiplex PCR system capable of identifying *Skunavirus*, *Ceduovirus*, and P335 phages (42).

Phage DNA extraction, sequencing, and annotation

Genomic DNA of selected phages (Table 1) was extracted from 2 mL of fresh phage lysate (>10^8 PFU/mL) using the Norgen Biotek Phage DNA isolation kit (Norgen Biotek, Ontario, Canada) according to the manufacturer's instructions. Phage genome sequences were determined by GenProbio srl (Parma, Italy) using a MiSeq platform (Illumina, San Diego, CA, USA) according to the supplier's protocol. Phage genome libraries were prepared using an Illumina Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA). Libraries were quantified using a fluorometric Qubit quantification system (Life Technologies, USA), and loaded on a 2200 TapeStation instrument (Agilent Technologies, USA). Sequencing was performed using the Illumina MiSeq platform with a 600-cycle flow cell version 3 (Illumina Inc., San Diego, CA, USA).

Fastq files of Illumina paired end reads (250 bp) obtained from phage sequencing efforts were used as input for genome assembly through the MEGAnnotator2 pipeline (43). The SPAdes program v3.14 was used for *de novo* assembly of phage sequences with the pipeline option "--careful" and a list of k-mer sizes of 21,33,55,77,99,127 (44). Contigs were then employed by MEGAnnotator2 for the prediction of protein-encoding open reading frames (ORFs) using Prodigal (45). Predicted ORFs were functionally annotated by means of DIAMOND against the NCBI RefSeq database and INTERPRO (46) against the PFAM database. Furthermore, tRNA genes were identified using tRNAScan-SE version 2.0 (47).

Table 1. Characteristics of the novel lactococcal phages isolated from whey samples as part of this study.

Phage	Host CWPS type/species	Year of collection (whey)	Phage genus	Genome size (bp)	GC%	No. ORFs	Best Hit (% coverage, % nt ID)
T1-C2	C_2 (*L. cremoris*)	2019	*Ceduovirus*	21771	36.33	36	BIM BV-114 (98, 93.03)
V1-MA6	C_2 (*L. cremoris*)	2019	*Skunavirus*	30346	34.85	57	16802 (78, 93.18)
T1-MH4	C_2 (*L. cremoris*)	2019		29002	35.26	52	66901 (87, 92.19)
V2-MF8	C_3 (*L. cremoris*)	2019		29823	35.25	50	P680 (77, 95.55)
V2-MB6	C_2 (*L. cremoris*)	2019		30278	34.66	51	37301 (76, 93.92)
AD1-ME6	C_2 (*L. cremoris*)	2019		29486	34.84	54	37201 (78, 94.02)
T1-MF3	C_1 (*L. cremoris*)	2019		29716	34.90	51	P4565 (76, 94.78)
AN4-MF5	C_2 (*L. cremoris*)	2020		30334	34.58	52	38503 (74, 90.44)
AN3-MA5	C_2 (*L. cremoris*)	2020		29514	34.65	52	FB3 (83, 94.49)
AV6-MC5	C_2 (*L. cremoris*)	2020		31100	34.42	56	PhiG (80, 95.08)
BA6-ME1	C_1 (*L. cremoris*)	2020		29849	34.47	55	p272 (84, 94.96)
BA6-B*	B (*L. lactis*)	2020		29963	35.41	57	p272 (83, 93.75)
BB4-MB5	C_2 (*L. cremoris*)	2020		29444	34.88	51	Phi4 (81, 94.45)
BB4-SC34†	B (*L. lactis*)	2020		31874	35.27	57	P475 (84, 95.08)
BB4-MG12	C_4 (*L. cremoris*)	2020		29565	35.13	53	Phi29 (85, 95.12)
AM2-MF12	C_4 (*L. cremoris*)	2020		29999	35.13	57	62601 (80, 91.88)
BA6-MF12	C_4 (*L. cremoris*)	2020		29899	34.59	58	62601 (81, 92.30)
AO-4†	C_6 (*L. cremoris*)	2020		30371	35.31	54	16802 (88, 94.36)
AO-MH4	C_2 (*L. cremoris*)	2020		27071	35.00	48	66901 (89, 93.26)
AO-MA1	C_1 (*L. cremoris*)	2020		32986	34.80	55	fd13 (81, 93.31)
BB4-D*	D (*L. lactis*)	2020		29715	34.84	53	38507 (75, 93.07)
BD-MH5	C_2 (*L. cremoris*)	2021		30798	34.71	55	fd13 (82, 93.94)
BD-MC7	C_3 (*L. cremoris*)	2021		30717	35.28	51	645 (72, 95.16)
BQ-1	C_3 (*L. cremoris*)	2022		28122	35.44	52	79201 (84, 95.62)
BR-MC7	C_3 (*L. cremoris*)	2022		30717	35.28	51	645 (72, 95.16)
BZ-MH2	C_1 (*L. cremoris*)	2022		31036	35.17	54	fd13 (94, 96.84)
CA-MA6	C_2 (*L. cremoris*)	2023		30324	34.61	54	CHPC361 (80, 95.79)

* Isolated from lab strains, this study
† Isolated by PhRACS, previous studies ((60) and Chapter 3).

Bioinformatic analysis

Protein sequence and structure homology analysis was performed using HHpred (Homology detection and structure prediction by HMM-HMM comparison) using standard mode (48). Sequence alignments were performed with Multalin (http://multalin.toulouse.inra.fr/multalin/) (49). *Skunavirus* RBPs and auxiliary host binding proteins that contain CBMs (i.e. NPS, major tail protein extension (TpeX), and Dit) were identified from the *Skunavirus* genomes based on BLASTp analysis. The CBMs within these proteins were determined using HHpred (Homology detection and structure production by HMM-HMM comparison) using default settings. Sequence alignments of the auxiliary host binding proteins were performed using Multalin using sequences identified in this study and previously sequenced phages retrieved from NCBI viruses (https://www.ncbi.nlm.nih.gov/labs/virus/). Auxiliary host binding proteins found to contain novel CBM sequences (i.e. CBMs were considered novel if there was less than 55 % (amino acid) sequence identity with previously characterised auxiliary binding proteins) were selected for structural modelling with AlphaFold2 (50, 51).

Comparison of individual encoded phage proteins was performed by all-against-all, bi-directional BLAST alignment with an alignment (*E*-value) cut-off value of 0.0001 and greater than 50 % identity across at least 50 % of the amino acid sequence (52). Multiple alignment of the nucleotide sequences (or protein sequence) of isolated skunaviruses genomes or proteins in this study and previously sequenced *Skunavirus* members were performed using ClustalW software. The alignment was employed to generate both rooted and unrooted phylogenetic trees using the iTOL v6 software (https://itol.embl.de/).

The Markov Clustering (MCL) algorithm was employed via the mclblastline pipeline v12-0678 as previously described using a cut-off value of 50 % identity across at least 50 % of the amino acid sequence (53). A binary matrix containing the presence/absence of each MCL family was computed and used as an input of the MeV suite (v 4.9.0) (54). The produced MCL matrix was analysed using a two-way hierarchical clustering (HCL) to group phage genomes according to their similarity at the protein level. CAGECAT version 1.0 web platform (https://cagecat.bioinformatics.nl/tools/clinker) was used to generate alignment of phage genomes (55). Putative orphan methyltransferases (MTases) were analysed using REBASE (http://rebase.neb.com/rebase/rebase.html) (56). Novel auxiliary host binding proteins were validated through structural predictions using AlphaFold2.

The presence or absence of closely related *Skunavirus* genomes in 27 whey viromes (previously sequenced, see Chapter 3 of this thesis) was determined through reads mapping. CoverM v0.4.0 (https://github.com/wwood/CoverM; “contig” mode) was employed to determine the covered fraction of a corresponding *Skunavirus* genome- or phageome-derived contig. The reads mapped to the contigs with identity >95 % and coverage >95 %, and only the contigs/genomes where >90 % of their length was covered by reads were reported. Hadamard scores, which represent a quantitative evaluation of the average nucleotide identity (ANI) and coverage between DNA sequences, were calculated using the ANIm method from pyANI (https://github.com/widdowquinn/pyani) (57) in order to assess overlaps between isolated/sequenced phages and phageome-derived contigs.

Transmission electron microscopy (TEM)

Phage lysates of isolates T1-MH4 and V2-MF8 were purified on a discontinuous 3M & 5M caesium chloride (CsCl) gradient and were dialysed against SM buffer as previously described (58). Following cesium chloride purification, the phage samples were fixed to a carbon film on a grid for 20 min, washed twice with distilled water and stained with 2 % uranyl acetate. The images were captured using a Talos L120C transmission electron microscope (ThermoFisher, Dreieich, Germany) and a CETA 16M camera (ThermoFisher) at an accelerating voltage of 80 kV. The measurements of the phages were performed using Thermo Scientific Velox software.

Host range

To investigate a link between the RBP phylogroups and corresponding host *cwps* genotypes, the host range of the 26 skunaviruses was determined using a panel of 45 strains whose speciation and *cwps* genotype had previously been determined (Supplementary Table S1). The host range strain panel was made up of a selection of 26 strains used in the original 92 isolate screening panel, while the remaining strains were sourced from our in-house strain collection (the *cwps* type and species of each strain is listed in Supplementary Table S1). Host range analysis was performed using spot assays, where 10 µL of each phage lysate (diluted 1:10 to achieve titres of between 10^8 and 10^9 PFU/mL) was dropped on LM17 double-layer agar plates (supplemented with 10 mM $CaCl_2$) containing 200 µL of the selected lactococcal strains (from fresh overnight cultures). After incubation overnight

at 30 °C, plates were examined for the presence of a zone of lysis and individual plaques were counted. The efficiency of plaquing (EOP) was calculated by dividing the titre of the phage (PFU/ml) on the tested strain by the phage titre on its respective propagating strain.

RESULTS

Prevalence of lactococcal phages in whey samples

Phage screening against dominant strains present in associated starter cultures

To assess the presence and diversity of lactococcal phages in fermentation facilities employing mesophilic undefined starter cultures, phage screening was conducted on 32 whey samples (Supplementary Table S2) collected from factories where the presence of phage(s) had been suspected due to slowed fermentation (milk acidification) issues. These 32 whey samples were collected from factories in factories in Spain (four samples), Austria (one sample), France (four samples), Poland (three samples), Sweden (one sample), Denmark (18 samples), and the Czech Republic (one sample) between 2019 and 2023. Although the exact undefined starter culture(s) employed in each factory at the time of sample collection is not known, two mesophilic starter cultures (M1 and/or M2) used for the production of various semi-hard cheeses, fermented milk (e.g., viili, kefir, and twaróg) and fresh cheeses (e.g., quark) may have been employed in some of these facilities as part of their starter culture rotation system in these factories. In a previous study M1 and M2 were determined to be primarily composed of lactococcal strains [see Chapter 2 of this thesis; (59)] and the *cwps* genotypic composition of these starter cultures was determined through culture-based analysis, metagenomics, and qPCR. A panel of 92 lactococcal isolates [as previously described in Chapter 2 of this thesis; (59)], representing the four most dominant *cwps* types (23 isolates of each of the C_1-C_4 subtypes) present in

these undefined cultures and a panel of B- and D-type lactococcal strains (from in-house strain collections) was used to screen the whey samples for phages.

Through this analysis, it was established that 29 of the 32 whey samples assessed (representing 90.6 % of samples) exhibited presumptive phage activity against at least one lactococcal strain in the panel based on lysis observed in wells inoculated with isolates compared to uninfected controls (Figure 1). This phage screening yielded a total of 268 (of 2,944 combinations tested) potential phages as based on the observed well lysis: 16 against C_1 isolates, 94 against C_2 isolates, 53 against C_3 isolates, and 105 against C_4 isolates. Of the 92 bacterial isolates employed in the screening panel, 63 (68 %) were infected by at least one presumptive phage. The three whey samples causing lysis of the largest number of isolates (Figure 1) were samples BQ (43 of 92 isolates), BV (36 of 92 isolates), and CA (31 of 92 isolates); all of which were isolated in 2022 from factories in Spain (BQ) or Denmark (BV and CA).

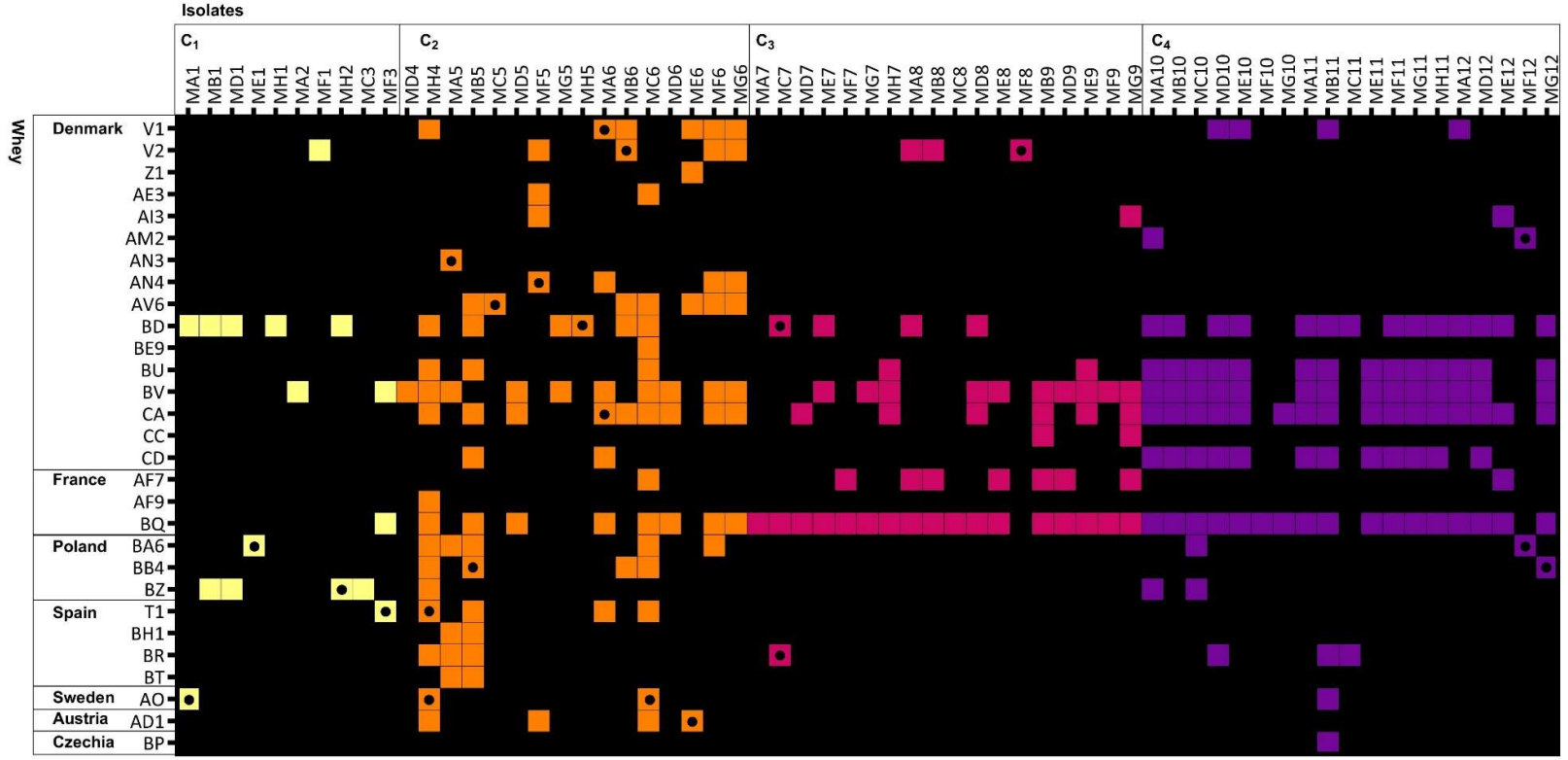


Figure 1. Phage screening of whey samples against a panel of lactococcal strains. Presumptive phage activity is indicated by coloured squares or absence of activity by black squares. The 22 phage-positive whey sample-host combinations selected for further analysis are indicated by a black circle within a coloured square. Colouring of the squares is indicative of the *cwps* genotype of the isolate the phage activity was against; yellow for C_1, orange for C_2, pink for C_3, and purple for C_4 isolates.

Of these 268 presumptive phage-positive whey sample-host combinations, 22 were selected as a representation of the assessed geographic locations and host *cwps* genotypes for phage screening, purification, and subsequent genome sequencing. Following confirmation of phage presence through plaque assays, individual plaque isolates were propagated on their corresponding sensitive lactococcal host to a high titre (at least 10^7 pfu/mL). The phages were initially characterised through multiplex PCR typing capable of identifying *Skunavirus*, *Ceduovirus*, and P335 phages, revealing that 22 of the 23 selected phages are *Skunavirus* members, with the remaining phage typed as a *Ceduovirus*.

Previous analysis of the diversity of RBP genotypes present in the phageome of these whey samples (see Chapter 3 of this thesis), indicated the presence of phages capable of infecting strains possessing a wide variety of *cwps* genotypes (such as B and D-types), although such strains had not been isolated from the starter cultures [see Chapter 2 of this thesis; (59)]. Therefore, to isolate such phages, targeted phage detection was conducted using whey samples BA6 and BB4 against strains possessing B and D *cwps* types from our in-house strain collection. BA6 was used in screening against four different B-type lactococcal strains, allowing the isolation of a phage capable of infecting IL1403 (phage BA6-B). Two D-type strains, *L. lactis* 184 and 18-16, were used for the targeted detection of a D-type infecting phage leading to the isolation of a phage capable of infecting *L. lactis* 18-16 from whey BB4 (phage BB4-D).

The so-called **ph**age **R**BP-**a**ctivated **c**ell **s**orting (PhRACS) approach was used in previous studies for the isolation of three low abundance (60) and

novel phage-host combinations (see Chapter 3 of this thesis). The genomes of the associated three skunaviruses, i.e. BB4-SC34 (B-type), AO-4 (C_6), and BQ-1 (C_3), were sequenced in addition to the 23 skunaviruses isolated above yielding a total of 26 sequenced *Skunavirus* genomes (for detail regarding the one isolated *Ceduovirus,* see Supplementary Text and Supplementary Figure S1).

Comparative genomic analysis and phylogeny of sequenced Skunaviruses

The genomes of the 26 selected *Skunavirus* members isolated in this study ranged in length from 27.1 to 33.0 Kbp (30.1 Kbp on average) and possessed between 48 and 58 ORFs (53.4 on average) with a consistent genome architecture. The genome sequences were first compared to publicly available genomes (available in NCBI as of June 2024) through BLASTn analysis to determine their relatedness/novelty (Table 1).

A phylogenetic tree of the nucleotide sequence of 212 publicly available *Skunavirus* genomes (those indicated as a complete nucleotide sequence on submission) combined with the 26 *Skunavirus* genomes from this study was generated (Figure 2). With the inclusion of the dairy streptococcal *Moineauvirus* SW8 as an outgroup, five major branches of skunaviruses can be discerned (Figure 2). The majority (14 of 26) of the skunaviruses isolated in this study cluster within one of these five major branches (which also includes previously described skunaviruses PhiG, PhiJF1, PhiF0139, and PhiF.17 isolated from dairy factories in the Netherlands (15, 20) with the

remaining 12 genomes being spread amongst the remaining four branches, thus reflecting the genetic diversity of skunaviruses isolated from the various dairy fermentation facilities. In addition, the skunaviruses isolated from the whey samples appear to cluster independent of the geographical origin of the phages (Figure 2; Supplementary Figure 2). Overall, the 26 novel *Skunavirus* genomes represent a diverse collection of skunaviruses when compared to the nucleotide sequence of other publicly available *Skunavirus* genomes (Figure 2). To further investigate the genome content and diversity of skunaviruses, a combined MCL and HCL analysis of the novel and publicly available skunaviruses was conducted allowing the identification of 24 protein families common among all sequenced skunaviruses, thus representing the core genome (Supplementary Figure S2; Supplementary Table S4). This is a decrease from a previous MCL analysis of 90 *Skunavirus* genomes which reported a total of 29 core genes, which is likely due to the significant (now at 238) increase in *Skunavirus* genomes used in the analysis (20). The five genes that appear to now be missing from the core genome include genes encoding the predicted tail terminator, distal tail protein, two hypothetical proteins from the early expressed region, and middle expressed protein 2 (Supplementary Table S4). This may be due to the cut-off (greater than 50 % identity across at least 50 % of the amino acid sequence) used in the MCL analysis, functional replacement of genes, or poor/incomplete genome assemblies. To determine if incomplete genome assemblies may be the result of the apparent loss of five genes from the core genome of skunaviruses, a second MCL and HCL analysis was generated using only high quality and complete genome assemblies (determined by the indication of a complete assembly during

genome submission and, in some cases, manual confirmation). This secondary analysis of 121 skunaviruses revealed that the apparently correct number of core *Skunavirus* genes may be 27 (rather than 24) with a hypothetical protein from the early expressed region and middle expressed protein 2 no longer present in the core genome (Supplementary Figure S3).

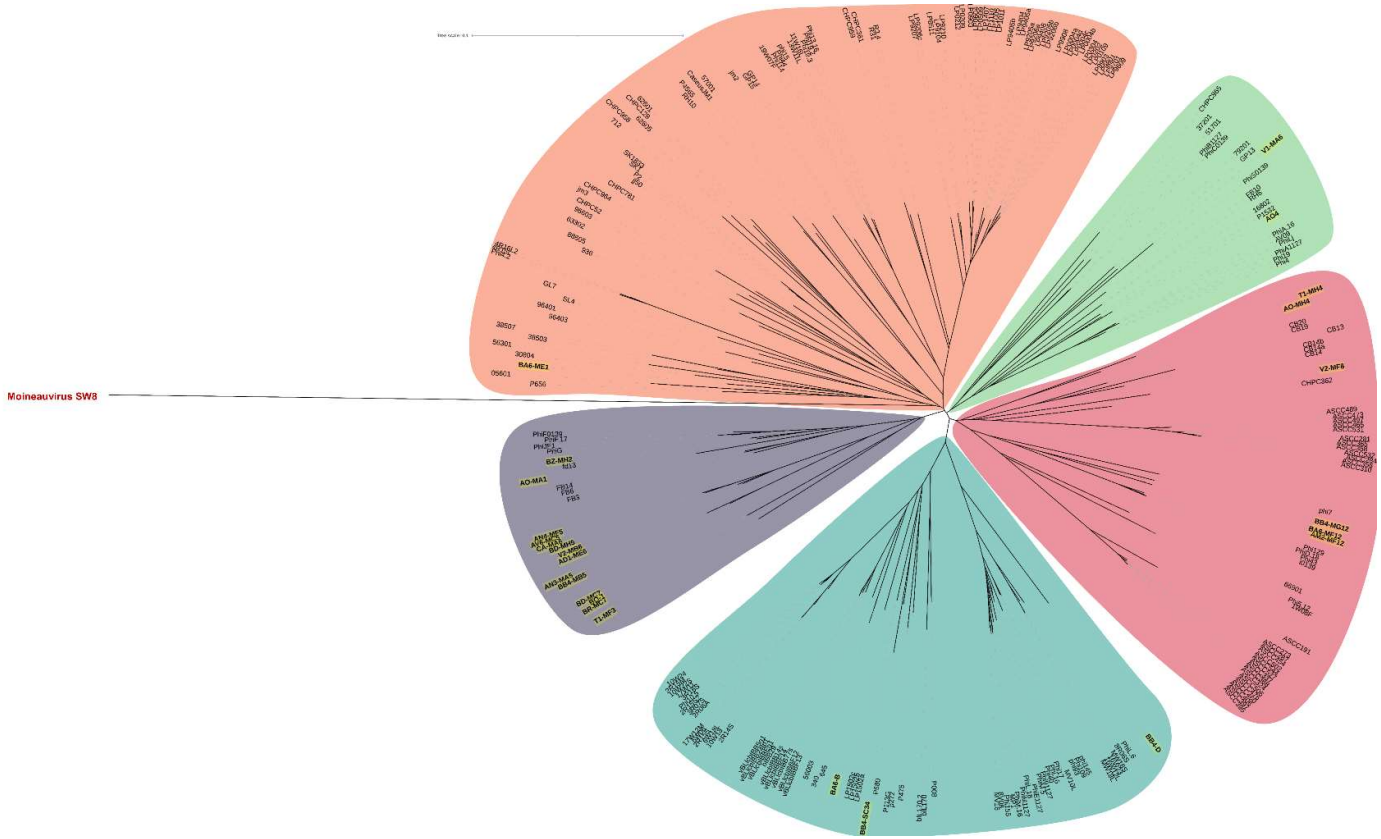


Figure 2. Phylogenetic tree of 238 *Skunavirus* genomes. Phage phylogeny and maximum-likelihood tree was constructed based on multiple alignments of publicly available whole genome sequences and the 26 skunaviruses sequenced within this study (names are highlighted in yellow) using ClustalW and iTOL. The identified, visually apparent five major branches of skunaviruses are highlighted in distinct colours. The dairy streptococcal *Moineauvirus* SW8 was included as an outgroup.

16 *Skunavirus* genomes, believed to be representative of the five genetic clusters mentioned above, were used to create a schematic overview of the organisation and sequence relatedness of the whey-isolated *Skunavirus* genomes (Figure 3). The genetic elements involved in DNA packaging, host recognition and binding, virion morphogenesis, and lysis are well conserved among these *Skunavirus* genomes. However, there is considerable genetic diversity within the regions that specify the host binding machinery (RBPs and auxiliary binding proteins) in these phages. Certain core genetic elements (with known functions) found in the early- and middle-expressed regions such as the genes encoding single-stranded DNA binding (SSB) protein, recombinase Sak (sensitivity to AbiK) protein, and a RuvC-like Holliday endonuclease are highly conserved among the genomes, which is consistent with previous studies (20). In addition, all but five of the *Skunavirus* genomes analysed in this study encode one or more putative orphan methyltransferase(s) (MTases) within the replication region and six possess additional methyltransferase(s) within the DNA packaging module. A previous study found that MTases encoded within the packaging module in skunaviruses (Phi145 and Phi93) provided protection against host-encoded restriction endonucleases (61). However, there is limited similarity between the six MTases encoded within the DNA packaging module identified in the analysed phage genomes and those in the previous study that were determined to provide protection against R-M systems of a particular specificity. Therefore, the REBASE webtool (http://rebase.neb.com/rebase/rebase.html) was used to predict the type and specificity of all putative MTases (Supplementary Table S5). The closely

related skunaviruses BD-MC7, BR-MC7, and BQ-1 all share two genes encoding identical MTases within the DNA packaging module and neither shared significant identity (<32 %) with any known MTase in the database (Supplementary Table S5). The three remaining skunaviruses (AO-MA1, T1-MF3, and BB4-MG12) each encode one distinct MTase (sharing <67.87 % amino acid identity to MTases encoded in the skunaviruses sequenced within this study) in their DNA packaging module. Phages AO-MA1 and BB4-MG12 each encode an MTase (AO-MA1_02 and BB4-MG12_02; Supplementary Table S5) which shares significant sequence identity (97 %) with Type II MTases M.PspTAD18ORF83P (CTSAG specificity) and M.Lph93Dam (GATC specificity), respectively. The remaining MTase (T1-MF3_02) most closely matched to M.PspTAD18ORF83P as well , though with lower (69 %) sequence identity.

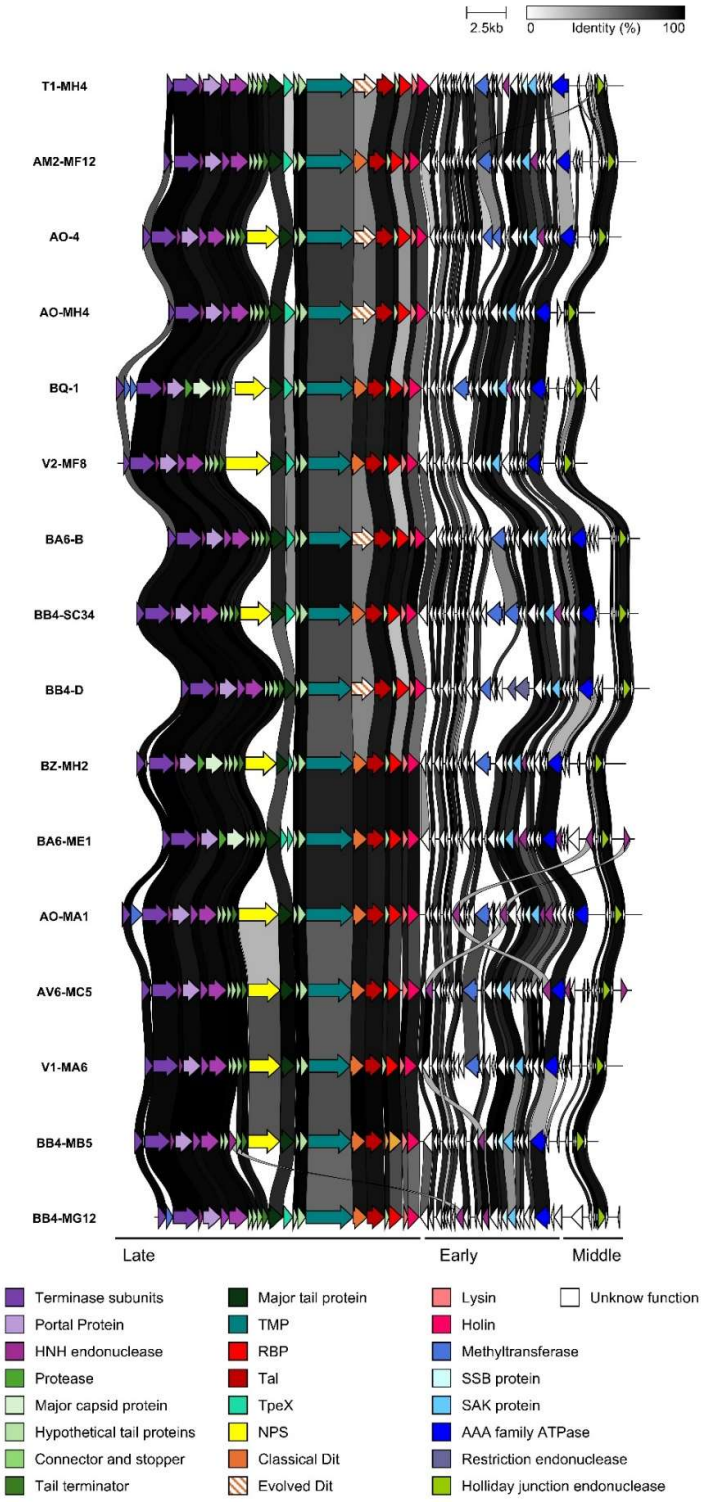


Figure 3. Schematic representation of the genetic architecture of 16 newly sequenced skunaviruses. Arrows represent predicted genes which are colour-coded based on the presumed functions of the encoded proteins indicated in the panel at the bottom of the figure.

Identification and characterisation of Skunavirus carbohydrate binding modules

Skunavirus binding to the surface-exposed saccharidic host receptor is facilitated by the phage-encoded RBP, and this function is often supplemented by auxiliary binding proteins encompassing CBMs. These auxiliary CBMs may be associated with the distal tail protein (Dit), major tail protein (TpeX), and/or neck passage structure (NPS), and have been found to complement and aid in host cell binding (33, 34). *Skunavirus* RBPs are comprised of head (which incorporates a carbohydrate binding site that facilitates receptor binding), neck, and shoulder domains. These RBPs have been categorised into distinct phylogenetic groups, which are based on amino acid sequence similarity within the C-terminal region of the RBP (which encompasses the head domain), and which correspond to the CWPS of the strain to which the phage can bind, thus being an important host range determinant [(20, 22), see Chapter 3 of this thesis]. To categorise the *Skunavirus* RBPs identified in this study, the C-terminal ~110 residues of the newly identified RBPs were aligned against the head domains of all available *Skunavirus* RBP sequences (a total of 358 RBP sequences; Figure 4). The majority of the isolated skunaviruses (18 of 26) encode group I RBPs (11 subgroup IA, four subgroup IB, and three subgroup IC RBPs). Group I RBPs have previously been found to recognise and bind CWPS type C strains (20, 22). A recent study (see Chapter 3 of this thesis) expanded on the binding specificity of the group I RBPs and introduced four distinct group I RBP subgroups. Subgroups IA RBPs were found to have a binding preference for the CWPS of C_2-subtype (and C_1-subtype to a lesser extent) strains and members of subgroup IC RBPs exhibit a binding preference

for the CWPS of C_4-subtype (and C_5-subtype to a lesser extent) strains (see Chapter 3 of this thesis). Two RBPs were found to belong to group III (which have a binding preference for B and C-type strains), four to group X (preference for C_3-subtype strains), one to group VI RBP (preference for D-type strains), and one to group VIII RBP (preference for C_6 subtype strains).

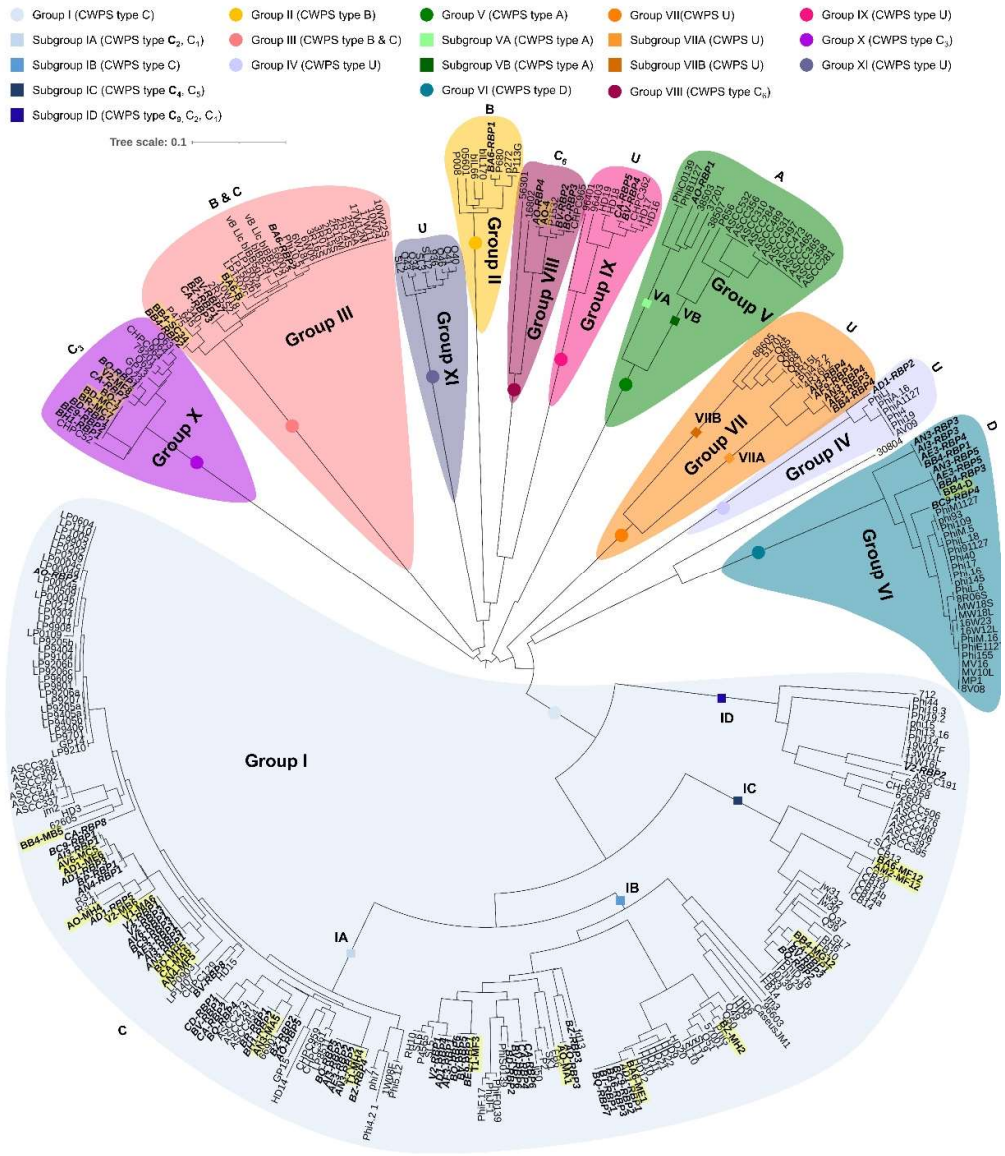


Figure 4. Phylogenetic tree of the C-terminal domain of the 358 *Skunavirus* RBP sequences. RBPs were aligned and grouped together based on sequence similarity and correspond to host *cwps* (where possible). RBPs corresponding to phages isolated within this study are bolded and highlighted. RBP sequences from previously

analysed whey phageome derived assemblies (see Chapter 3 of this thesis) are italicised and bolded.

The presence and diversity of auxiliary CBMs present in the phage genomes sequenced in this study were evaluated (Table 2). Lactococcal Dits can be classified as classical or evolved, where evolved Dits are larger and typically contain an additional CBM. Evolved Dit-associated CBMs have been found to possess the same host specificity as their corresponding RBP and were previously found to belong to one of four classes (34). Of the 26 skunaviruses, five were found to possess evolved Dits (Table 2). Three phages (BA6-B, BB4-D, and AO-4) possess Dits that are similar to one of the previously described evolved Dit classes (Table 2; Supplementary Figure S4). However, two phages (T1-MH4 and AO-MH4) both possessed the same CBM insertion in their respective Dit proteins, though this CBM shares limited amino acid sequence identity (<70 %) with other lactococcal Dit proteins and we therefore propose this to represent a novel class (Class 5 Dit) of evolved Dit proteins (Supplementary Figure S4). Structure predictions of the T1-MH4 Dit using AlphaFold2 revealed that it is closely related to an uncharacterised protein encoded in the genome of a non-motile facultative anaerobic *Leptotrichia* spp. (Figure 5D). Transmission electron microscopy images of T1-MH4 did not yield any discernible feature associated with the distal tail region of this phage (Figure 5C).

Table 2. Types of auxiliary carbohydrate domains observed in the genomes of 26 sequenced skunaviruses.

Phage	Host CWPS Type/species	RBP Group	NPS		TpeX		Dit	
			Group	Size (aa)	Subgroup	Size (aa)	Type	Size (aa)
BA6-ME1	C_1 (*L. cremoris*)	IB			***VI***	224	Classical	298
T1-MF3	C_1 (*L. cremoris*)	IB	B	672			Classical	298
BZ-MH2	C_1 (*L. cremoris*)	IB	A	656			Classical	298
AO-MA1	C_1 (*L. cremoris*)	IB	***I***	804			Classical	298
V2-MB6	C_2 (*L. cremoris*)	IA	A	676			Classical	298
AN3-MA5	C_2 (*L. cremoris*)	IA	B	666			Classical	298
AN4-MF5	C_2 (*L. cremoris*)	IA	A	676			Classical	298
V1-MA6	C_2 (*L. cremoris*)	IA	B	672			Classical	298
T1-MH4	C_2 (*L. cremoris*)	IA			***V***	208	***Class 5***	498
AV6-MC5	C_2 (*L. cremoris*)	IA	B	676			Classical	298
BB4-MB5	C_2 (*L. cremoris*)	IA	B	666			Classical	298
AD1-ME6	C_2 (*L. cremoris*)	IA	A	676			Classical	298
BD-MH5	C_2 (*L. cremoris*)	IA	A	676			Classical	298
AO-MH4	C_2 (*L. cremoris*)	IA			***V***	227	***Class 5***	498
CA-MA6	C_2 (*L. cremoris*)	IA	A	676			Classical	298
BB4-MG12	C_4 (*L. cremoris*)	IC			I	193	Classical	298
AM2-MF12	C_4 (*L. cremoris*)	IC			I	206	Classical	298
BA6-MF12	C_4 (*L. cremoris*)	IC			I	206	Classical	298
BA6-B	B (*L. lactis*)	III			I	206	Class 3	470
BB4-SC34	B (*L. lactis*)	III	B	641	I	204	Classical	298
BB4-D	D (*L. lactis*)	VI					Class 2	486
AO-4	C_6 (*L. cremoris*)	VIII	B	672			Class 3	474
V2-MF8	C_3 (*L. cremoris*)	X	***H***	924	I	187	Classical	298
BQ-1	C_3 (*L. cremoris*)	X	***J***	658			Classical	298
BD-MC7	C_3 (*L. cremoris*)	X	***J***	658	I	175	Classical	298
BR-MC7	C_3 (*L. cremoris*)	X	***J***	658	I	175	Classical	298

Novel CBM groups, subgroups, and types identified in this study are in bold-face text and italicised.

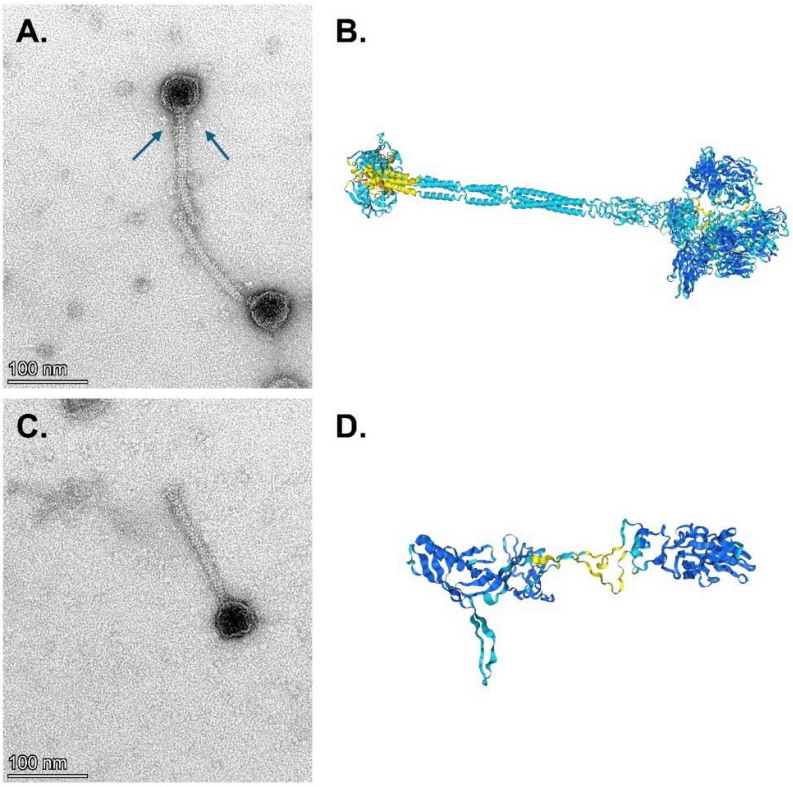


Figure 5. Electron micrographs of select skunaviruses and their corresponding auxiliary binding proteins encoding novel carbohydrate binding modules. Micrograph of *Skunavirus* V2-MF8 (A) with arrows marking the whiskers with small bobble-like structures that alongside collar structures make up the neck passage structure of the phage. (B) Structural prediction of the V2-MF8 NPS protein. (C) Micrograph of *Skunavirus* T1-MH4 and (D) predicted ribbon structure of its corresponding distal tail protein encoding a novel CBM (not visible in the EM).

Many skunaviruses have been found to encode neck passage structures (NPS), which form a collar-whisker complex attached to the phage portal and which (may) contain CBM(s) that aid in host binding (33, 62). Previous studies have indicated the presence of seven (group A-G) NPS groups corresponding to distinct neck passage structures and/or limited amino acid sequence identity (33). 13 (of 18 possessing an NPS) skunaviruses were found to possess Group A or B NPS structures (Table 2). The NPSs of the remaining five phages share limited amino acid identity with other *Skunavirus*

NPS proteins and were, therefore, considered novel NPSs. These novel NPS groups will be referred to as Groups H, I, and J (Supplementary Table S6). Structural prediction of *Skunavirus* V2-MF8 novel NPS yielded a unique structure which was found to encode three different CBMs for each NPS monomer at the C-terminal end (Figure 5B). *Skunavirus* V2-MF8 was therefore assessed by electron microscopy for its morphological characteristics (Figure 5A). Collar structures and whiskers with small, bobble-like structures were clearly visible resembling the whisker-like appendages with small terminal knots observed in certain skunaviruses (such as Phi15 and PhiG, for example (15)). The observed whiskers were 23.04 ± 4.44 nm (*n*= 19) in length with the associated whisker balls being about 9.474 ± 1.188 nm (*n*= 17) wide (Supplementary Table S7).

Finally, 11 skunaviruses encode a major tail extension (TpeX) protein which result in the presence of additional decorations along the tail, and which have also been found to exhibit the same host specificity as their corresponding RBP, with more favourable binding at the cell poles where division occurs (but with significantly reduced affinity) (20, 33). There are currently four defined TpeX subgroups that all correspond to the Tuc2009 BppA CBM but vary in amino acid composition and/or length (33). 11 of the skunaviruses characterized in this study encode a TpeX, most of which (8/11) are members of subgroup I. The remaining three TpeX, incorporating a Tuc2009 BppA domain, form two novel subgroups. Two TpeX proteins (encoded by phages T1-MH4 and AO-MH4) cluster closely together and form a novel subgroup V, while the remaining TpeX (from phage BA6-ME1) forms

subgroup VI (relation to previous TpeX subgroups is summarised in Supplementary Table S8).

Linking Skunavirus RBPs to host range

To further understand the link between the host range of skunaviruses and their corresponding RBP, host range analysis of the 26 isolated skunaviruses was undertaken against a panel of 45 lactococcal strains (26 coming from the original 92 isolate screening panel, the remaining from in-house strain collections; strains listed in Supplementary Table S1) possessing various *cwps* genotypes (A, B, C_1-C_6, and D-types) and correlated to their defined RBP groups (Figure 6; Supplementary Table S9). A strong correlation was observed between the *Skunavirus* RBP (sub)group and the *cwps*-type of the lactococcal strain they infect, being consistent with previous studies [see Chapter 3 of this thesis, (22, 33)]. For example, all skunaviruses with group I RBPs were found to infect C-type strains; with subgroup IA infecting C_2-type strains, subgroup IB mainly infecting C_1-type strains, and IC infecting C_4-type strains (see Chapter 3 of this thesis). However, a *Skunavirus* with subgroup a IB RBP was also capable of infecting three C_2-type strains (Figure 6). Therefore, this host range analysis supports and supplements the most recent study on RBP specificity as determined through fluorescent binding assays (see Chapter 3 of this thesis), as the binding specificity of subgroup IB RBPs was not determined through fluorescent binding assays.

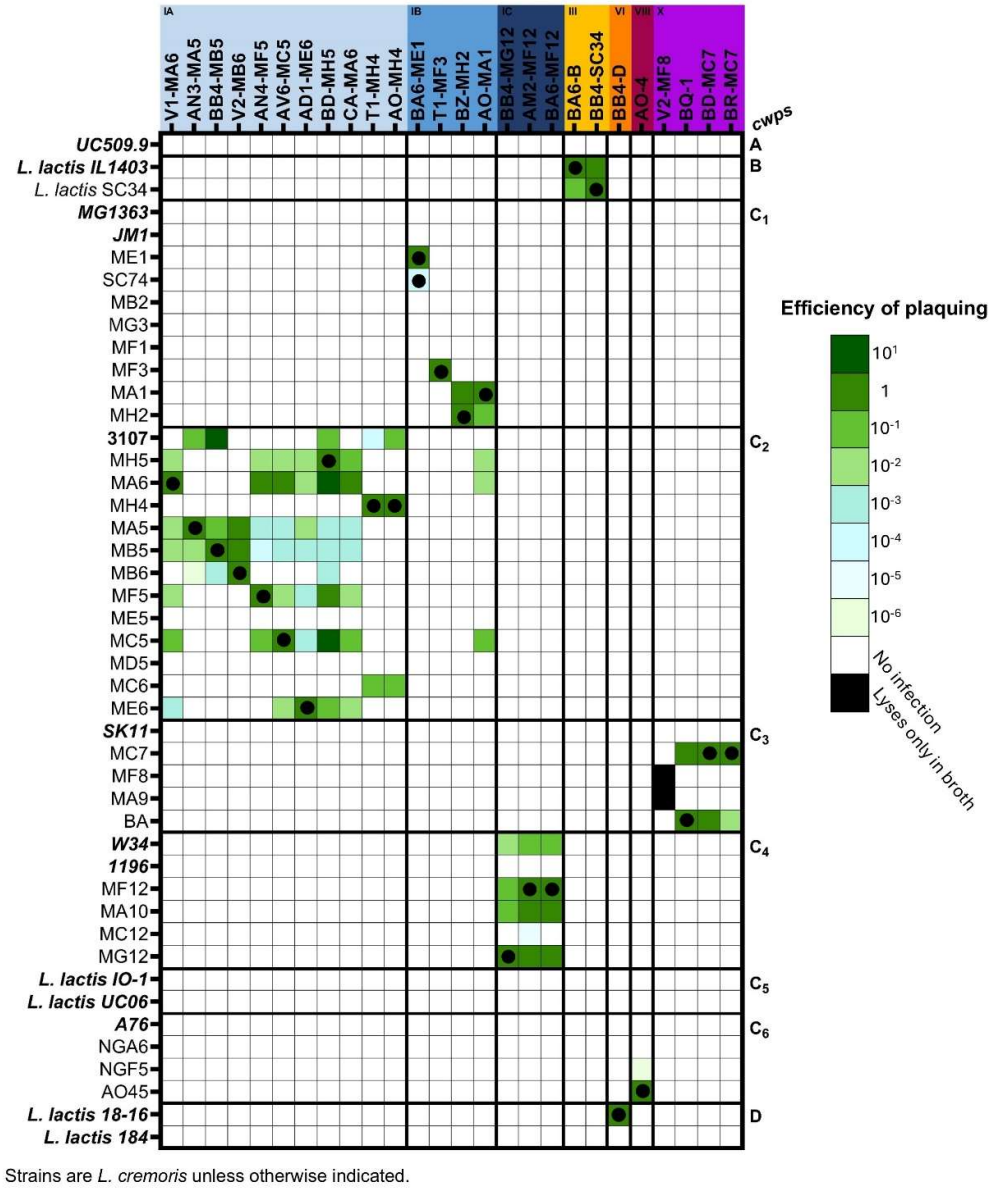


Figure 6. Host range of the 26 skunavirus isolated within this study. Host range was determined by spot testing against a panel of 45 lactococcal strains (strains belong to the species *L. cremoris*, unless otherwise indicated in the figure). *Skunavirus* names are coloured according to their RBP group. Squares are colour-coded based on the calculated EOP of a phage against a particular strain (indicated in the figure legend) relative to its propagating host (indicated by a black circle within a square). Detailed EOP information summarised in this figure is presented in Supplementary Table S9. White squares indicate no infection. Black squares indicate infection by a phage determined by lysis in broth that does not produce visible plaques.

Combinatorial approach to investigate phage diversity

Initial analysis of phage diversity present in these fermentation facilities was conducted through virome sequencing and subsequent phageome analysis of 27 (of the 32) whey samples focused on investigating the diversity and binding specificity of *Skunavirus* RBPs present in the phageomes (see Chapter 3 of this thesis). Here, to complement the traditional approach of phage isolation and sequencing, complete (or nearly complete, i.e. >25.0 Kbp) *Skunavirus* genomes from the whey phageome assemblies were identified (Table 3) and aligned against the sequenced *Skunavirus* genomes (Figure 7). From the analysed 27 whey phageome assemblies, only 11 distinct (within a sample) complete *Skunavirus* genomes or near complete (i.e. contigs >25.0 Kbp) *Skunavirus* contigs could be identified. These contigs ranged in length from 25.7 to 32.0 Kbp and were isolated from 11 samples (AD1, BP, V2, AE3, AI3, AN3, AV6, BD, BA6, BB4, and BZ). To determine if any of these retrieved contigs closely resembles one of the isolated/sequenced phages, Hadamard scores (product of average nucleotide identity (ANI) and coverage) were generated for each isolated/sequenced *Skunavirus* against the 11 phageome-derived *Skunavirus* contigs (Figure 7A). In addition, a phylogenetic tree of the isolated/sequence skunaviruses and phageome-derived *Skunavirus* contigs was generated (Figure 7B). Validated by manual inspection, it was determined that two phageome-derived *Skunavirus* contigs, Con_AE3-1 and Con_AI3-1, are essentially identical to the isolated (and subsequently sequenced) skunaviruses V1-MA6 (Hadamard score 1.00) and CA-MA6 (Hadamard score 0.98), respectively. Additionally, two phageome-derived *Skunavirus* contigs, Con_AN3-1 and Con_BD-1, are nearly identical to the genomes of

skunaviruses V1-MA6 (Hadamard score 0.90) and BD-MC7 (Hadamard score 0.84), respectively. The absence of certain genes in these two phageome-derived contigs (presumably due to incomplete assemblies) is the likely cause of the reduced Hadamard scores. Nevertheless, the genes present in the two phageome-derived contigs are identical to those encoded in the respective genomes of the isolated/sequenced skunaviruses. The seven remaining phageome-derived *Skunavirus* contigs therefore correspond to skunaviruses that were not isolated within this study by means of cultivation-based techniques. This therefore demonstrates that phage diversity was not fully covered by the traditional (i.e. cultivation-based) isolation and sequencing method described herein. Moreover, the identification of 11 near-complete phageome-derived *Skunavirus* contigs in 27 whey phageome reveals the limitations of a “metagenomics-only” approach to investigate phage diversity. This may be due to a need to increase sequencing depth (an average sequencing output of 3.7 million high quality reads was generated per sample Supplementary Table S2), or due to difficulties in accurately assembling closely related and mosaic *Skunavirus* genomes.

Table 3. Characteristics of 11 phageome-derived contigs.

Phageome-derived contig	Year of collection (whey)	Genome size (bp)	GC%	Best Hit (% coverage, % nt ID)	No. ORFs			
					Late	Early	Middle	Total
Con_AD1-1	2019	32001	35.43	96401 (74, 94.07)	23	26	4	**53**
Con_BP-1	2021	30841	34.75	Phi4 (75, 94.83)	23	29	4	**56**
Con_V2-1	2019	30670	35.01	GP13 (48, 93.75)	22	30	4	**56**
Con_AE3-1	2020	30747	35.12	CHPC362 (77, 94.28)	21	31	4	**56**
Con_AI3-1	2020	30315	34.86	PhiG (82, 95.04)	22	27	6	**55**
Con_AN3-1	2020	27314	35.18	FB6 (87, 89.70)	18	30	4	**52**
Con_AV6-1	2020	26492	35.12	2R06A (72, 94.07)	20	27	4	**51**
Con_BD-1	2021	25726	35.34	CHPC964 (75, 91.49)	15	24	3	**42**
Con_BA6-1	2020	29666	35.25	p272 (84, 92.99)	21	29	5	**55**
Con_BB4-1	2020	30105	35.00	16802 (70, 91.55)	22	25	5	**52**
Con_BZ-1	2022	29052	35.31	vB_Llc_bIBBF13 (99, 96.19)	21	30	4	**55**

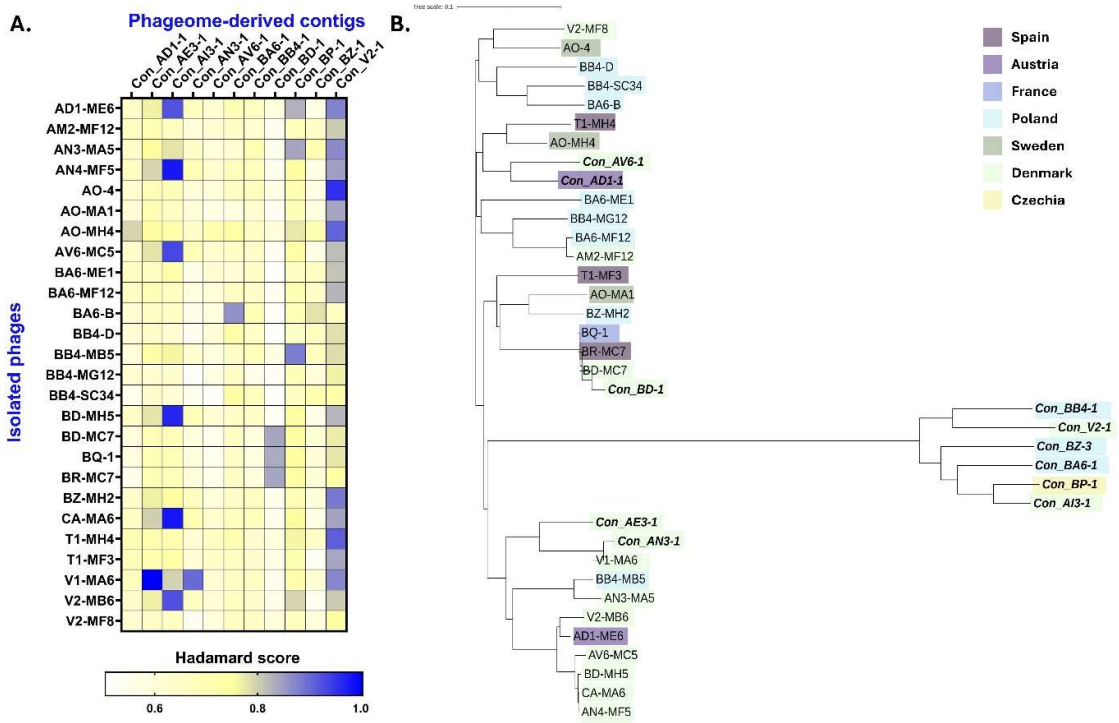


Figure 7. Comparative analysis of phageome-derived *Skunavirus* contigs corresponding to putatively complete or nearly complete *Skunavirus* genomes and the 26 isolated and sequenced skunaviruses. (A) Heat map indicating the Hadamard score (product of average nucleotide identity and coverage) of each isolated/sequenced *Skunavirus* against the 11 phageome-derived *Skunavirus* contigs was generated and indicated by the colour of the squares. (B) A phylogenetic tree was generated of the phageome-derived and isolated/sequenced *Skunavirus* contigs and genomes. The geographical origin of each genome/contig is indicated by a certain colour (outlined in the figure legend).

To garner insights into the phage diversity in the 27 samples, for which virome data was available, the isolated/sequence and phageome-derived *Skunavirus* genomes/contigs were mapped against the raw virome reads to determine which skunaviruses (or very closely related ones) may be present in which dairy fermentation facilities (Figure 8). Certain skunaviruses such as V1-MA6 and CA-MA6, for example, appear (with >99 % covered bases) to be persistent in the Danish dairy facilities being present in whey samples (eight

and four, respectively) collected between 2019 and 2021. Other skunaviruses, such as T1-MH4 and BB4-MB5 (or very closely related phages), may be associated with the undefined starter cultures as they are present in several whey samples across (five and four of the six countries, respectively) geographical locations. In certain cases, phages (for example BA6-MF12 and AN3-MA5) that were isolated from a particular whey sample during the phage screening process outlined in this study were not identified in the same corresponding whey viromes (but were detected in other virome samples). In addition, two of the isolated phages (BB4-MG12 and BZ-MH2) were not identified in any whey virome. This may be due to the phages being in lower abundance or due to limitations associated with virome extractions and sequencing.

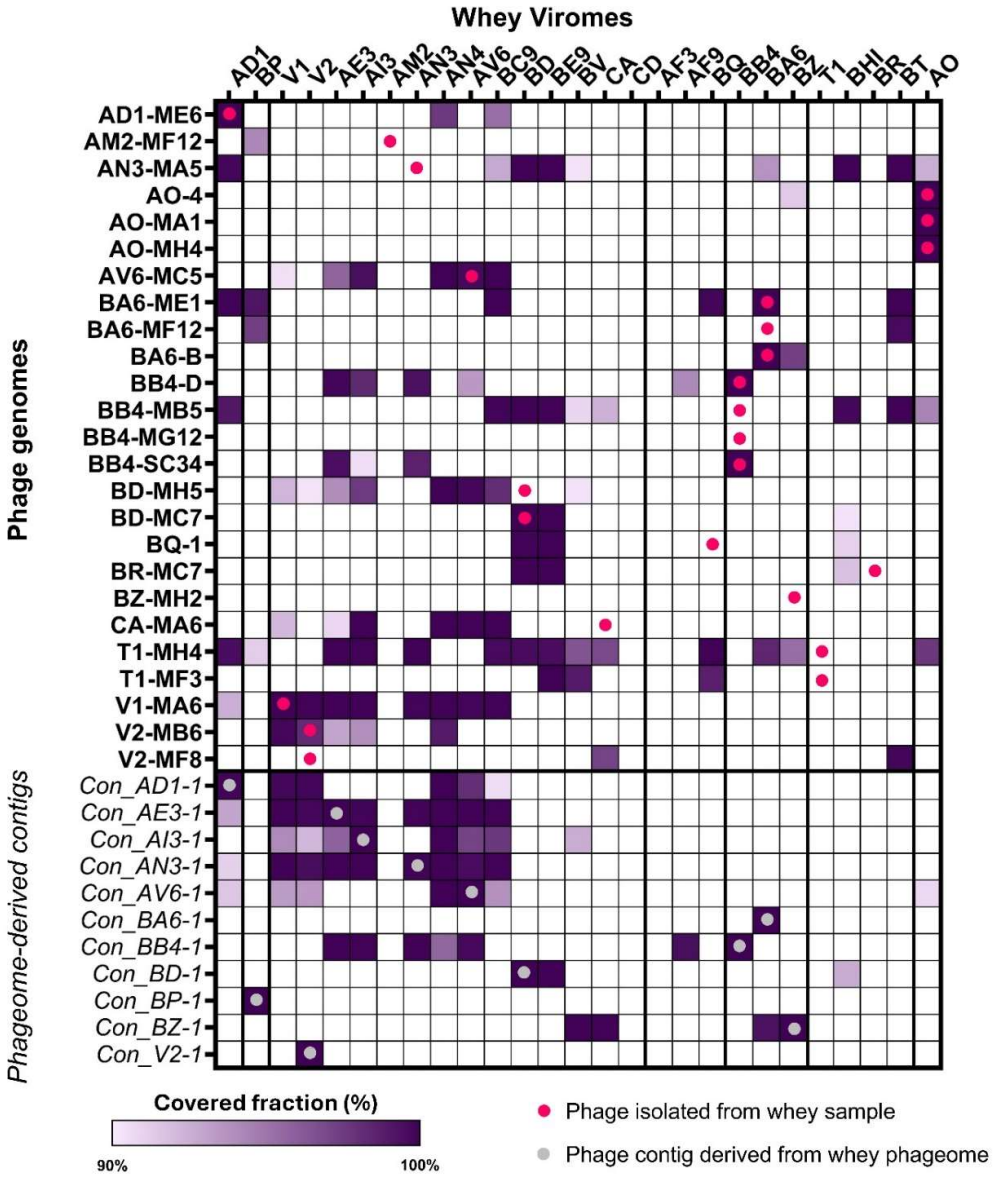


Figure 8. Presence or absence of a *Skunavirus* genome/contig in the virome reads of 27 previously sequenced whey viromes. The covered fraction of a particular (or closely related) genome/contig was determined through reads mapping of phageome-derived *Skunavirus* contigs and isolated/sequenced *Skunavirus* against raw reads of 27 whey viromes. The minimum covered fraction depicted in the figure is 90 %, squares are coloured according to the covered fraction (between 90-100 %).

DISCUSSION

Phage infection remains one of the main challenges in dairy fermentation systems, typically resulting in slowed fermentations (11). In particular, skunaviruses remain the most endemic and problematic lactococcal phage group in facilities utilising mesophilic starter cultures (mainly composed of *L. lactis* and *L. cremoris* strains) (9). Several decades of research has underpinned the current depth of understanding of this group of virulent lactococcal phages in relation to their genetic, morphological, and molecular characteristics, and yet they persist (12–18). Therefore, consistent screening and monitoring of phage populations associated with starter cultures and/or fermentation facilities is essential to minimise the impact and reduce the risk of phage infection of starter cultures. Phage screening has traditionally been performed through culture-dependent methods (13, 15, 16, 21, 63). While this is a widely used and informative approach, culture-dependent methods of phage detection typically represent the dominant and culturable phages present in a sample. Metagenomics, and more recently viromics/phageomics, have been utilised to determine the phage prevalence and diversity in a number of ecological niches (64–66). The development of more targeted methods for viral particles and associated nucleic acids facilitate improved phageome analysis, particularly in dairy settings where contaminating protein and bacterial cells/debris has been a major challenge when extracting phage particles (67). Furthermore, there continue to be challenges associated with assembling high-quality phage genomes from virome data due to the modular and mosaic nature of most phage genomes (68, 69). Phables has recently been introduced as a software tool with an improved ability to resolve phage

genomes accurately from viral metagenome assemblies (70). While this and other similar tools have improved scientific endeavours to defining viral diversity based on sequence information, biological and functional information is lacking with a phageomics only approach. In the current study, the biodiversity of phages present in several dairy fermentation facilities across Europe employing mesophilic undefined starter cultures was assessed. 27 lactococcal phages (26 *Skunavirus* and one *Ceduovirus*) were isolated through traditional culture-dependent phage screening methods, targeted isolation based on predicted RBP specificity, and PhRACS. The genetic novelty, CBMs, and host range of the *Skunavirus* genomes was investigated. Throughout the study culture-dependent and -independent methods were combined to conduct a robust investigation of the phage diversity in dairy fermentation facilities.

MCL/HCL and phylogenetic analysis of the 26 *Skunavirus* genomes revealed that the skunaviruses cluster independently of geographic location and were spread across the visually apparent five major branches of skunaviruses (Figure 2, Supplementary Figure S2). However, it is important to note that although the isolated skunaviruses within the study clustered independently of the geographic location of their corresponding fermentation facility, 14 of these skunaviruses make up the majority of one of the five major branches (14 of 22 total members). This is possibly as a result of closely related strains being present in the mesophilic undefined starter cultures used in disparate locations. The phages isolated in the current study were shown to infect strains possessing C_1-C_4, C_6, B, and/or D *cwps* genotypes. The host range of these phages was consistent with the recently established predicted

binding specificity of their associated RBP (see Chapter 3 of this thesis). In addition, a number of novel CBMs were identified in the NPS, evolved Dit, and major tail extension proteins of several phages. Of note, two skunaviruses were found to encode *nps* genes that resulted in the appearance of collar structures and whiskers with small, bobble-like structures that resemble the same appendages observed in other skunaviruses (Figure 5 A and B) (15). The collar structures associated with the NPS also resemble the KIS (LL-K insertion sequence) elements identified in *Lactobacillus delbrueckii* phages (71). The KIS-element was first described in *L. delbrueckii* subsp. *lactis* phage LL-K as an insertion element encoding two putative open reading frames (72). Electron micrographs of *L. delbrueckii* phages Ld25A and Ld17 revealed that one of the genes encoded in KIS-element was responsible for presence of collar neck passage structures in both phages (71). A notable difference is that in the case of *L. delbrueckii* phages, the apparent NPS is encoded across two open reading frames whereas they are encoded by a single gene (*nps*) for the lactococcal phages. The diversity of *Skunavirus* RBPs continues to increase (see Chapter 3 of this thesis) highlighting the continuous adaptation and evolution of these phage components to diversifying host receptors (20, 22, 33). In the present study, genetic and subsequent structural diversity was observed in the CBMs of auxiliary binding proteins including Dit and TpeX. In congruence with previous studies, these genome regions (tail morphogenesis, RBP, and NPS) are considered diversity hotspots (32). *Skunavirus* RBPs are primarily linked to the host range of a phage but additional CBMs present on auxiliary binding proteins have been shown to contribute to the binding process (33, 34). The auxiliary binding protein TpeX is associated with the

presence of additional decorations along the phage tail while NPS proteins form a collar-whisker complex attached at the head-tail interface of the phage (33). CBMs present in both the NPS and TpeX have previously been found to bind favourably towards the polar ends of a host cell (where cell division occurs) maintaining the same host specificity as their corresponding RBP (although with reduced affinity) (33). However, in the current study, skunaviruses encoding subgroup IC, group III, and group X RBPs were found to encode the same subgroup I TpeX (Table 2). In addition, with the emergence of novel CBMs in the TpeX, NPS, and Dit it may be necessary to re-examine whether these CBMs always maintain the same host specificity as their RBP or if they may aid in expanding a phage's host range. Interestingly, it has recently been shown that evolved Dits in skunaviruses aid in adsorption to exopolysaccharide (EPS) receptors (35).

Based on the RBP diversity uncovered in the whey samples through phageome analysis (see Chapter 3 of this thesis), culture-independent analysis (through phageomics) was used to guide or facilitate the culture-based isolation of phage-host combinations, indicating that a combinational approach to phage screening will yield a more complete view of phage diversity present in a sample. This led to the isolation of phages infecting B-type, D-type, and C_6-type strains. This is the first description of skunaviruses capable of infecting C_6 and D-type lactococcal strains, respectively. Although *Skunavirus* genomes encoding similar RBPs have previously been published, the *cwps* genotype of the strains they were capable of infecting had not been described (73).

To complement the culture-based phage isolation and characterisation described here, 11 complete (or nearly complete) phageome-derived *Skunavirus* contigs generated from the previously described phageome assemblies (see Chapter 3 of this thesis) were compared to the 26 isolated/sequenced *Skunavirus* genomes. The mosaic nature of *Skunavirus* genomes renders accurate assemblies from phageome-derived sequence reads a significant challenge (70). The depth of sequencing used in the current study (average 3.7 M quality filtered reads per sample) is consistent with other fermentation viromes studied recently (64, 67). However, deeper virome sequencing potentially complemented by long read sequencing may improve the accuracy and completeness of assembled phage genomes from virome data. Nonetheless this analysis highlights both the limitations and benefits of applying culture-dependent or -independent analysis methods in isolation. Accurate and continuous monitoring of phage populations within dairy fermentation systems is likely one of the most important practices necessary to define the dominant and low abundance phages present in a sample, which in turn will be crucial to improve phage-related risk assessment and consequent predictive and planning capabilities in dairy fermentations.

CONCLUSION

By investigating the diversity of lactococcal phages present in whey samples originating from various fermentation factories employing mesophilic undefined starter cultures 26 *Skunavirus* and one *Ceduovirus* were isolated and sequenced. An extensive host range analysis employing strains corresponding to 11 different cell wall polysaccharide structures, has further resolved the specificity of a *Skunavirus* RBP to a corresponding host *cwps* genotype. We conclude that the combination of phageome analysis and conventional culture-based approaches is likely the most effective route to developing a holistic view of the phage diversity in fermentation samples or indeed those from a variety of ecological niches. Furthermore, it is likely that the benefits of a combinational approach will increase as nucleic acid extraction, sequencing, and assembly methods of virome/phageome data continue to improve. By establishing a deeper understanding of both current and potentially problematic phage populations present in dairy fermentation factories, more informed, robust, and efficient culture rotation strategies may be developed.

REFERENCES

1. Hendrix RW, Smith MCM, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc Natl Acad Sci USA* **96**:2192–2197.
2. Gobbetti M, De Angelis M, Di Cagno R, Mancini L, Fox PF. 2015. Pros and cons for using non-starter lactic acid bacteria (NSLAB) as secondary/adjunct starters for cheese ripening. *Trends Food Sci Technol* **45**:167–178.
3. Tamang JP, Watanabe K, Holzapfel WH. 2016. Diversity of microorganisms in global fermented foods and beverages. *Front Microbiol* **7**:377.
4. Daly C, Fitzgerald GF, Davis R. 1996. Biotechnology of lactic acid bacteria with special reference to bacteriophage resistance. *Antonie Van Leeuwenhoek* **70**:99–110.
5. Johnson ME. 2013. Mesophilic and Thermophilic Cultures Used in Traditional Cheesemaking. *Microbiol Spectr* **1**:10.1128/microbiolspec.cm-0004-2012.
6. Li TT, Tian WL, Gu CT. 2021. Elevation of *Lactococcus lactis* subsp. *cremoris* to the species level as *Lactococcus cremoris* sp. nov. and transfer of *Lactococcus lactis* subsp. *tructae* to *Lactococcus cremoris* as *Lactococcus cremoris* subsp. *tructae* comb. nov. *Int J Syst Evol Microbiol* **71**:004727.

7. Deveau H, Labrie SJ, Chopin MC, Moineau S. 2006. Biodiversity and classification of lactococcal phages. *Appl Environ Microbiol* **72**:4338–4346.
8. Mahony J, Oliveira J, Collins B, Hanemaaijer L, Lugli GA, Neve H, Ventura M, Kouwen TR, Cambillau C, van Sinderen D. 2017. Genetic and functional characterisation of the lactococcal P335 phage-host interactions. *BMC Genomics* **18**:146.
9. Mahony J, Murphy J, Van Sinderen D. 2012. Lactococcal 936-type phages and dairy fermentation problems: From detection to evolution and prevention. *Front Microbiol* **3**:1–9.
10. Millen AM, Romero DA. 2016. Genetic determinants of lactococcal *C2viruses* for host infection and their role in phage evolution. *J Gen Virol* **97**:1998-2007.
11. Samson JE, Moineau S. 2013. Bacteriophages in food fermentations: new frontiers in a continuous arms race. *Annu Rev Food Sci Technol* **4**:347–368.
12. Moineau S, Borkaev M, Holler BJ, Walker SA, Kondo JK, Vedamuthu ER, Vandenbergh PA. 1996. Isolation and characterization of Lactococcal bacteriophages from cultured buttermilk plants in the United States. *J Dairy Sci* **79**:2104–2111.
13. Jolicoeur AP, Lemay ML, Beaubien E, Bélanger J, Bergeron C, Bourque-Leblanc F, Doré L, Dupuis MÈ, Fleury A, Garneau JE, Labrie SJ, Labrie S, Lacasse G, Lamontagne-Drolet M, Lessard-Hurtubise R, Martel B, Menasria R, Morin-Pelchat R, Pageau G, Samson JE, Rousseau GM, Tremblay DM, Duquenne M, Lamoureux

M, Moineau S. 2023. Longitudinal study of *Lactococcus* phages in a Canadian cheese factory. *Appl Environ Microbiol* **89**:e0042123.

14. Miklič A, Rogelj I. 2003. Characterization of lactococcal bacteriophages isolated from Slovenian dairies. *Int J Food Sci Technol* **38**:305–311.
15. Murphy J, Royer B, Mahony J, Hoyles L, Heller K, Neve H, Bonestroo M, Nauta A, van Sinderen D. 2013. Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type cheese-producing plants. *J Dairy Sci* **96**:4945–4957.
16. Oliveira J, Mahony J, Hanemaaijer L, Kouwen TRHM, van Sinderen D. 2018. Biodiversity of bacteriophages infecting *Lactococcus lactis* starter cultures. *J Dairy Sci* **101**:96–105.
17. Szczepańska AK, Hejnowicz MS, Kołakowski P, Bardowski J. 2007. Biodiversity of *Lactococcus lactis* bacteriophages in Polish dairy environment. *Acta Biochim Pol* **54**:151–158.
18. Moineau S, Fortier J, Ackermann H-W, Pandian S. 1992. Characterization of lactococcal bacteriophages from Quebec cheese plants. *Can J Microbiol* **38**:875–882.
19. Mahony J, Moscarelli A, Kelleher P, Lugli GA, Ventura M, Settanni L, van Sinderen D. 2017. Phage biodiversity in artisanal cheese wheys reflects the complexity of the fermentation process. Viruses **9**:45.
20. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. *Sci Rep* **6**:21345.

21. Chmielewska-Jeznach M, Bardowski JK, Szczepankowska AK. 2018. Molecular, physiological and phylogenetic traits of *Lactococcus* 936-type phages from distinct dairy environments. *Sci Rep* **8**:12540.

22. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C, Vogensen FK, Van Sinderen D. 2013. Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol* **79**:4385–4392.

23. Sciara G, Bebeacua C, Bron P, Tremblay D, Ortiz-Lombardia M, Lichière J, Van Heel M, Campanacci V, Moineau S, Cambillau C. 2010. Structure of lactococcal phage p2 baseplate and its mechanism of activation. *Proc Natl Acad Sci USA* **107**:6852–6857.

24. Cecilia B, Denise T, Carine F, Marie-Pierre C-C, Irina S, Marin van H, David V, Sylvain M, Christian C. 2013. Structure, Adsorption to Host, and Infection Mechanism of Virulent Lactococcal Phage p2. J *Virol* **87**:12302–12312.

25. Casey CN, Morgan E, Daly C, Fitzgerald GF. 1993. Characterization and classification of virulent lactococcal bacteriophages isolated from a Cheddar cheese plant. *J of Appl Bacteriol* **74**:268–275.

26. Chandry PS, Moore SC, Boyce JD, Davidson BE, Hillier AJ. 1997. Analysis of the DNA sequence, gene expression, origin of replication and modular structure of the *Lactococcus lactis* lytic bacteriophage sk1. *Mol Microbiol* **26**:49–64.

27. Scaltriti E, Polverini E, Grolli S, Eufemi E, Moineau S, Cambillau C, Ramoni R. 2013. The DNA binding mechanism of a SSB protein from *Lactococcus lactis* siphophage p2. *Biochim Biophys Acta* **1834**:1070–1076.

28. Scaltriti E, Moineau S, Launay H, Masson J-Y, Rivetti C, Ramoni R, Campanacci V, Tegoni M, Cambillau C. 2010. Deciphering the function of lactococcal phage ul36 Sak domains. *J Struct Biol* **170**:462–469.

29. D BJ, Sylvain M. 2004. Lactococcal Phage Genes Involved in Sensitivity to AbiK and Their Relation to Single-Strand Annealing Proteins. *J Bacteriol* **186**:3649–3652.

30. Bidnenko E, Ehrlich SD, Chopin M-C. 1998. *Lactococcus lactis* phage operon coding for an endonuclease homologous to RuvC. *Mol Microbiol* **28**:823–834.

31. Jakob H, M RG, Karin H, Sylvain M. 2009. Identification and Characterization of the Phage Gene sav, Involved in Sensitivity to the Lactococcal Abortive Infection Mechanism AbiV. *Appl Environ Microbiol* **75**:2484–2494.

32. Castro-Nallar E, Chen H, Gladman S, Moore SC, Seemann T, Powell IB, Hillier A, Crandall KA, Chandr PS. 2012. Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage. *Genome Biol Evol* **4**:382–393.

33. Hayes S, Vincentelli R, Ramond L, Ramnd R, Nauta A, Sinderen D van, Cambillau C. 2019. Ubiquitous carbohydrate binding modules decorate 936 lactococcal siphophage virions. *Viruses* **11**:631.

34. Hayes S, Vincentelli R, Mahony J, Nauta A, Ramond L, Lugli GA, Ventura M, van Sinderen D, Cambillau C. 2018. Functional carbohydrate binding modules identified in evolved Dits from siphophages infecting various Gram-positive bacteria. *Mol Microbiol* **110**:777–795.

35. Millen AM, Magill D, Romero D, Simdon L. 2023. Evolved distal tail protein of skunaviruses facilitates adsorption to exopolysaccharide-encoding lactococci. *Microbiome Res Rep* **2**:26.

36. Dupont K, Vogensen FK, Neve H, Bresciani J, Josephsen J. 2004. Identification of the receptor-binding protein in 936-species lactococcal bacteriophages. *Appl Environ Microbiol* **70**:5818–5824.

37. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, Van Sinderen D. 2014. Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. *mBio* **5**:e00880-14.

38. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, Chapot-Chartier M-P, Cambillau C, Holo H, van Sinderen D. 2020. The CWPS Rubik's cube: linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. *Mol Microbiol* **114**:582–596.

39. Sadovskaya I, Vinogradov E, Courtin P, Armalyte J, Meyrand M, Giaouris E, Palussière S, Furlan S, Péchoux C, Ainsworth S, Mahony J, van Sinderen D, Kulakauskas S, Guérardel Y, Chapot-Chartier M-

P. 2017. Another brick in the wall: a rhamnan polysaccharide trapped inside peptidoglycan of *Lactococcus lactis*. *mBio* **8**: e01303-17.

40. Chapot-Chartier MP, Vinogradov E, Sadovskaya I, Andre G, Mistou MY, Trieu-Cuot P, Furlan S, Bidnenko E, Courtin P, Péchoux C, Hols P, Dufre YF, Kulakauskas S. 2010. Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. *J Biol Chem* **285**:10464-10471.
41. Lillehaug D. 1997. An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. *J Appl Microbiol* **83**:85–90.
42. Steve L, Sylvain M. 2000. Multiplex PCR for Detection and Identification of Lactococcal Bacteriophages. *Appl Environ Microbiol* **66**:987–994.
43. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura M. 2021. METAnnotatorX2: a comprehensive tool for deep and shallow metagenomic data set analyses. *mSystems* **6**:e00583-21.
44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**:455-477.
45. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.

46. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240.

47. Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**:9077–9096.

48. Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**:W244–W248.

49. Corpet F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**:10881–10890.

50. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589.

51. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu

K, Hassabis D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577**:706–710.

52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–410.
53. Enright AJ, Dongen S Van, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**:1575–1584.
54. Saeed Ai, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M. 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**:374–378.
55. Gilchrist CLM, Chooi Y-H. 2021. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**:2473–2475.
56. Roberts RJ, Vincze T, Posfai J, Macelis D. 2009. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **43**:D298-D299.
57. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. 2016. Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. *Anal Methods* **8**:12-24.
58. Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning: a laboratory manual, Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
59. White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A

multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* **90**:e02152-23.

60. Casey E, McDonnell B, White K, Stamou P, Crowley T, O'Neill I, Lavelle K, Hayes S, Lugli GA, Arboleya S, James K, Ventura M, Martinez I, Gueimonde M, Dal Bello F, Nally K, Mahony J, van Sinderen D. 2022. Needle in a Whey-Stack: PhRACS as a Discovery Tool for Unknown Phage-Host Combinations. *mBio* **13**:e03334-21.
61. Murphy J, Klumpp J, Mahony J, O'Connell-Motherway M, Nauta A, van Sinderen D. 2014. Methyltransferases acquired by lactococcal 936-type phage provide protection against restriction endonuclease activity. *BMC Genomics* **15**:831.
62. Vegge CS, Neve H, Brøndsted L, Heller KJ, Vogensen FK. 2006. Analysis of the collar-whisker structure of temperate lactococcal bacteriophage TP901-1. *Appl Environ Microbiol* **72**:6815–6818.
63. Bissonnette F, Labrie S, Deveau H, Lamoureux M, Moineau S. 2000. Characterization of mesophilic mixed starter cultures used for the manufacture of aged cheddar cheese. *J Dairy Sci* **83**:620–627.
64. Dugat-Bony E, Lossouarn J, De Paepe M, Sarthou AS, Fedala Y, Petit MA, Chaillou S. 2020. Viral metagenomic analysis of the cheese surface: A comparative study of rapid procedures for extracting viral particles. *Food Microbiol* **85**:103278.
65. Paillet T, Lamy-Besnier Q, Figueroa C, Petit M-A, Dugat-Bony E. 2024. Dynamics of the viral community on the cheese surface during maturation and persistence across production years. *mSystems* **9**: e0020124.

66. Colombo S, Arioli S, Gargari G, Neri E, Della Scala G, Mora D. 2018. Characterization of airborne viromes in cheese production plants. *J Appl Microbiol* **125**:1444–1454.

67. Muhammed MK, Kot W, Neve H, Mahony J, Castro-Mejía JL, Krych L, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, van Sinderen D, Vogensen FK. 2017. Metagenomic analysis of dairy bacteriophages: extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. *Appl Environ Microbiol* **83**:e00888-17.

68. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: An in silico evaluation of metagenome—enabled estimates of viral community composition and diversity. *PeerJ* **5**:e3817.

69. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. 2019. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**:12.

70. Mallawaarachchi V, Roach MJ, Decewicz P, Papudeshi B, Giles SK, Grigson SR, Bouras G, Hesse RD, Inglis LK, Hutton ALK, Dinsdale EA, Edwards RA. 2023. Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* **39**:btad586.

71. Casey E, Mahony J, O'Connell-Motherway M, Bottacini F, Cornelissen A, Neve H, Heller KJ, Noben JP, Dal Bello F, van Sinderen D. 2014. Molecular Characterization of three *Lactobacillus delbrueckii* subsp. *bulgaricus* phages. *Appl Environ Microbiol* **80**:5623–5635.

72. Forsman P, Alatossava T. 1994. Repeated sequences and the sites of genome rearrangements in bacteriophages of *Lactobacillus delbrueckii* subsp. *lactis*. *Arch Virol* **137**:43–54.

73. Marcelli B. 2020. PhD thesis. University of Groningen, Groningen, Netherlands. *Lactococcus lactis* bacteriophages: phage-host interaction and phage transduction.

SUPPLEMENTARY MATERIALS

Supplementary text.

Comparative genomic analysis of Ceduovirus T1-MH4

A single *Ceduovirus* (T1-C2) was isolated from whey sample T1 and its genome was sequenced in the context of this study. The genome comprises 21.8 Kbp, encompasses 36 genes and is most closely related to *Ceduovirus* BIM BV-114 (Table 1). T1-C2 can be characterized as a c2-type *Ceduovirus*, which have previously reported to recognise the PIP (phage infection protein) receptor (2). Comparative genomic analysis of T1-C2 to *Ceduovirus* c2 shows consistent genomic architecture and considerable amino identity among genes. The most notable sequence differences can be found in proteins corresponding to those responsible for host recognition and host range determination (I14, I15, and I16) and the tail tape measure protein (Supplementary Figure S1).

Supplementary Table S1. Bacterial strains used in this study.

Strains	Species	CWPS type	GenBank Accession	Source/References
UC509.9	*L. cremoris*	A	CP003157	UCC
IL1403	*L. lactis*	B	AE005176	(Bolotin et al., 2001)
SC34	*L. lactis*	B	Unpublished	(Casey et al., 2020)
ME1	*L. cremoris*	C_1	Unpublished	This study
SC74	*L. cremoris*	C_1	Unpublished	(Casey et al., 2020)
MB2	*L. cremoris*	C_1	Unpublished	This study
MG3	*L. cremoris*	C_1	Unpublished	This study
MF1	*L. cremoris*	C_1	Unpublished	This study
MF3	*L. cremoris*	C_1	Unpublished	This study
MA1	*L. cremoris*	C_1	Unpublished	This study
JM1	*L. cremoris*	C_1	CP015899	UCC
MG1363	*L. cremoris*	C_1	AM406671	(Linares et al., 2010)
MA6	*L. cremoris*	C_2	Unpublished	This study
MH4	*L. cremoris*	C_2	Unpublished	This study
MA5	*L. cremoris*	C_2	Unpublished	This study
MB5	*L. cremoris*	C_2	Unpublished	This study
MB6	*L. cremoris*	C_2	Unpublished	This study
MF5	*L. cremoris*	C_2	Unpublished	This study
ME5	*L. cremoris*	C_2	Unpublished	This study
MC5	*L. cremoris*	C_2	Unpublished	This study
MD5	*L. cremoris*	C_2	Unpublished	This study
MC6	*L. cremoris*	C_2	Unpublished	This study
ME6	*L. cremoris*	C_2	Unpublished	This study
3107	*L. cremoris*	C_2	CP031538	(Ostergaard Breum, Neve, Heller, & Vogensen, 2007)
MF8	*L. cremoris*	C_3	Unpublished	This study
MA9	*L. cremoris*	C_3	Unpublished	This study
BA	*L. cremoris*	C_3	Unpublished	(Chapter 3)
SK11	*L. cremoris*	C_3	CP000425	(Vos et al., 1989)
MF12	*L. cremoris*	C_4	Unpublished	This study
MA10	*L. cremoris*	C_4	Unpublished	This study
MC12	*L. cremoris*	C_4	Unpublished	This study
MG12	*L. cremoris*	C_4	Unpublished	This study
W34	*L. cremoris*	C_4	CP032430	UCC
1196	*L. cremoris*	C_4	CP032148	UCC
UD4	*L. cremoris*	C_5	Unpublished	This study
UG1	*L. cremoris*	C_5	Unpublished	This study
IO-1	*L. lactis*	C_5	AP012281	(Kato et al., 2012)
UC06	*L. lactis*	C_5	CP015902	UCC
NGA6	*L. cremoris*	C_6	Unpublished	This study
NGF5	*L. cremoris*	C_6	Unpublished	This study
AO45	*L. cremoris*	C_6	Unpublished	This study
A76	*L. cremoris*	C_6	CP003132	(Bolotin, Quinquis, Ehrlich, & Sorokin, 2012)
184	*L. lactis*	D	CP015895	UCC
18-16	*L. lactis*	D	Unpublished	UCC

Supplementary Table S2. Information regarding the whey samples used in this study.

Sample code	Date	Country
T1	06/2019	Spain
BH1	07/2021	
BR	01/2022	
BT	04/2022	
AD1	12/2019	Austria
AF3	02/2020	France
AF7	02/2020	
AF9	02/2020	
BQ*	11/2022	
BA6	10/2020	Poland
BB4	10/2020	
BZ	10/2022	
AO	05/2020	Sweden
V1	06/2019	Denmark
V2	06/2019	
Z1	09/2019	
AE3	02/2020	
AI3	03/2020	
AM2	03/2020	
AN3	04/2020	
AN4	04/2020	
AV6	09/2020	
AZ7	11/2020	
BC9	02/2021	
BD	03/2021	
BE9	03/2021	
BU	06/2022	
BV	07/2022	
CA	11/2022	
CC	01/2023	
CD	01/2023	
BP	10/2021	Czechia

* Different factory in France

Supplementary Table S3. List of 92 lactococcal (*L. lactis* or *L. cremoris*) isolates used in phage screening, first described in Chapter 2, (1).

Isolate	Origin	CWPS-type
MA1	M1	C_1
MB1	M1	
MC1	M1	
MD1	M1	
ME1	M1	
MF1	M1	
MG1	M1	
MH1	M1	
MA2	M1	
MB2	M1	
MC2	M1	
MD2	M1	
ME2	M1	
MF2	M1	
MG2	M1	
MH2	M1	
MA3	M1	
MB3	M1	
MC3	M1	
MD3	M2	
ME3	M2	
MF3	M2	
MG3	M2	
MA4	M2	C_2
MB4	M2	
MC4	M2	
MD4	M2	
ME4	M2	
MF4	M2	
MG4	M2	
MH4	M1	
MA5	M1	
MB5	M1	
MC5	M1	
MD5	M1	
ME5	M1	
MF5	M1	
MG5	M1	
MH5	M1	
MA6	M1	
MB6	M1	
MC6	M1	
MD6	M1	
ME6	M1	
MF6	M1	
MG6	M1	

Isolate	Origin	CWPS-type
MA7	M1	C_3
MB7	M1	
MC7	M1	
MD7	M2	
ME7	M2	
MF7	M2	
MG7	M2	
MH7	M2	
MA8	M2	
MB8	M2	
MC8	M2	
MD8	M2	
ME8	M2	
MF8	M2	
MG8	M2	
MH8	M2	
MA9	M2	
MB9	M2	
MC9	M2	
MD9	M2	
ME9	M2	
MF9	M2	
MG9	M2	
MA10	M1	C_4
MB10	M2	
MC10	M2	
MD10	M2	
ME10	M2	
MF10	M2	
MG10	M2	
MH10	M2	
MA11	M2	
MB11	M2	
MC11	M2	
MD11	M2	
ME11	M2	
MF11	M2	
MG11	M2	
MH11	M2	
MA12	M2	
MB12	M2	
MC12	M2	
MD12	M2	
ME12	M2	
MF12	M1	
MG12	M1	

Supplementary Table S4. *Skunavirus* core genes. The gene IDs of the core genes correspond to *Skunavirus* sk1. The table lists the twenty-nine core genes that were previously determined to make up the *Skunavirus* core genes Murphy *et al*. (3), and whether that were determined to still make up the core genome based on the current MCL analysis of 238 *Skunavirus* genomes.

Gene ID	Predicted protein function	Present or missing from core genome (genome(s) missing from)
sk1p01	terminase small subunit	Present
sk1p02	terminase large subunit	Present
sk1p03	HNH endonuclease	Present
sk1p04	Portal protein	Present
skp05	Prohead protease	Present
sk1p06	Major capsid protein	Present
sk1p07	Hypothetical protein	Present
sk1p08	Connector	Present
sk1p09	**Stopper**	**Missing (CHPC964)**
sk1p10	putative structural protein 4	Present
sk1p11	major tail protein	Present
sk1p12	putative structural protein 5	Present
sk1p13	putative structural protein 6	Present
sk1p14	Tape measure protein	Present
sk1p15	**Distal tail protein (Dit)**	**Missing (CHPC965, MV16)**
sk1p16	Tail associated lysin (Tal)	Present
sk1p17	Putative structural protein 7	Present
sk1p18	receptor binding protein	Present
sk1p19	Holin	Present
sk1p20	Lysin	Present
sk1p22	Hypothetical protein (early expressed)	Present
sk1p24	Hypothetical protein (early expressed)	Present
sk1p34	**Hypothetical protein (early expressed)**	**Missing (CHPC965)**
sk1p35	Hypothetical protein (early expressed)	Present
sk1p36	Single strand DNA binding protein	Present
sk1p37	Homologous recombinase Sak	Present
sk1p40	**Hypothetical protein (early expressed)**	**Missing (05601, CHPC129, CHPC361, CHPC52, CHPC781, CHPC958, CHPC959, CHPC964, CHPC965, P475, P656)**
sk1p54	**Middle expressed protein 2**	**Missing (BD-MC7, BQ-1, BR-MC7)**
sk1p55	Holliday junction resolvase	

Supplementary Table S5. Analysis of methyltransferase encoded within the DNA packaging and replication regions using REBASE.

Genome location	Mtases ORF	Length	Match to	Type	Meth	Specificity	Identity (%)
Packaging	AO-MA1_02	249	M.PspTAD18ORF83P	2		CTSAG	97
	BB4-MG12_02	140	M.Lph93Dam	-1		GATC	97
	T1-MF3_02	249	M.PspTAD18ORF83P	2		CTSAG	69
	BD-MC7_02, BR-MC7_02, BQ-1_02	249	M.PspTAD18ORF83P	2		CTSAG	32
	BD-MC7_03, BR-MC7_03, BQ-1_03	285	M.Lph145I	2		GGWGA	27
Replication	V1-MA6_32c	298	M.Lph93II	2	N6A	CYAG	78
	T1-MH4_32c	293	M.Pma2631ORF2660P	2		UNK	88
	V2-MB6_32c, CA-MA6_32c, BD-MH5_32c, AV6-MC5_32c, AN4-MF5_31c, AD1-ME6_32c	313	M.Lph93II	2	N6A	CYAG	79
	T1-MF3_33c	220	M.Lrh6ORF13185P	1	N6A	UNK	30
	BA6-B_35c	287	M.Pma2631ORF2660P	2		UNK	81
	BB4-SC34_037	306	M.Lla7CORF5045P	2	C5	UNK	57
	BB4-SC34_039	294	M.Sag9828ORF1988P	2		UNK	56
	AM2-MF12_35c, BA6-MF12_36c	302	M.Lph15I	2	N6A	CCAG	81
	AO-4_36	190	M.Lph15I	2	N6A	CCAG	54
	AO-4_35	201	M.Lph93II	2	N6A	CYAG	35
	AO-MA1_34c	298	M.Lph93II	2	N6A	CYAG	75
	BB4-D_31c	220	M.Lrh6ORF13185P	1	N6A	UNK	31
	BD-MC7_31c, BR-MC7_32c, BQ-1_31c,	305	M.Lph93II	2	N6A	CYAG	79
	BZ-MH2_35c	310	M.Lph93II	2	N6A	CYAG	83

Supplementary Table S6. NPS Groups.

Group	Representative Phage	HHpred Hit	Corresponding Protein
A	PhiA.16	5E7T	Tuc2009, BppA
B	Phi155	5E7T	Tuc2009, BppA
C	PhiB1127	5E7T	Tuc2009, BppA
D	Phi19.3	1GQ8, 2PYG	Pectin
E	ASCC365	2PYG, 4MXN	Alginate
F	Phi114	2WAA	Carbohydrate esterase
G	ASCC191	2WAA	Carbohydrate esterase
H*	V2-MF8	5E7T, 4V96_AC, 6IAB_A	Tuc2009, BppA; TP901-1, BppA; P68; tail fibre
I*	AO-MA1	3GRH_A	Hydrolase
J*	BQ-1	3WX7_A	Hydrolase

* Novel NPS Groups described in this study.

Supplementary Table S7. TEM measurements.

Values in nm	Head,	Tail,	Tail,	Base,	Base,
T1-MH4	**diameter**	**length**	**width**	**length**	**width**
	53.5	136	11.8	14	15.4
	54.5	144	13.5	14.7	15.3
	53.8	158	11.5	14.5	13.5
	55.3	146	12.3	10.5	14.7
	55.5	141	12.5	14.8	14.8
	54.1	148	13.4	14.3	16.4
	55.1	156	12.4	14.6	14.8
	53.2	151	12.9	14.5	15.8
	55.5	154	12.8	16.3	17.3
	54.7	155	13.8		
unit	10	10	10	9	9
mean	54.52	148.9	12.69	14.2444444	15.3333333
Stand. dev.	0.84169142	7.17170366	0.74004504	1.54443445	1.09316056

Values in nm	Head	Tail	Tail	Base	Base	Whisker	Whisker ball	Collar	Collar
V2-MF8	**diameter**	**length**	**width**	**length**	**width**	**length**	**width**	**length**	**width**
	54	143	13.1	13.4	13.6	15.8	10.3	15.7	4.43
	53.9	146	11.6	11.4	14	25,2	11,8	12.6	3.23
	54.6	143	11.2	11.8	13.6	24,5	11,5	13.5	4.98
	54	140	12.6	13.1	17.8	27,5	7,23	14.9	3.26
	54.8	139	12.8	14.1	16.1	24,7	7,54	11.3	2.93
	54.4	144	12.5	14.4	17.7	26,1	9,30	12.7	2.26
	55.6	141	12.7	14	19.3	26,5	9,70	18.8	3.08
	54.7	140	12.2	17.9	19.6	22,4	8,54		
	54.7	140	12.5	9.96	16.2	23,6	9,19		
	54.5	139	12.8	14.4	14.4	28,2	9,36		
	53.5	141	12.6	13.3	15.1	31,0	9,23		
	53	135	12.9	16,3	16,6	28,5	6,84		
	54.4	139	12.9	14.2	19.3	27.1	9.3		
	55.1	145	12.8			22.1	7.47		
						25.7	10.1		
						24.5	10.2		
						28,9	10,2		
						24,7			
						22,5			
unit	14	14	14	13	13	19	17	7	7
mean	54.37 14286	141.0 71429	12.51 42857	13.49 66667	16.39 16667	23.04	9.474	14.21 42857	3.452 85714
Stand. dev.	0.659 17031	2.894 67112	0.526 02595	1.956 84037	2.295 63156	4.443 87219	1.1882 2557	2.503 61643	0.931 76739

Supplementary Table S8. TpeX Subgroups.

Subgroup	Representative phage	HHpred hit	Corresponding protein
I	P113G	5E7T	Tuc2009, BppA
II	Phi4.2	5E7T	Tuc2009, BppA
III	Phi91127	5E7T	Tuc2009, BppA
IV	Phi2R06A	5E7T	Tuc2009, BppA
V*	T1-MH4	5E7T	Tuc2009, BppA
VI*	BA6-ME1	5E7T	Tuc2009, BppA

* Novel TpeX subgroup described in this study.

Supplementary Table S9. Detailed EOP information from host range analysis of twenty-six skunaviruses.

(A) Subgroup IA RBP skunaviruses

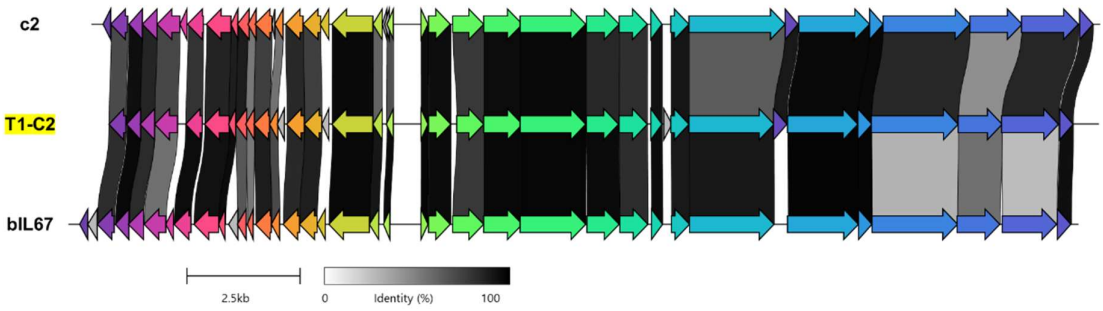
Strains	*cwps* type	V1-MA6	AN3-MA5	BB4-MB5	V2-MB6	AN4-MF5	AV6-MC5	AD1-ME6	BD-MH5	CA-MA6	T1-MH4	AO-MH4
3107	C_2	-	3.79 (± 4.44) × 10^{-1}	1.11 (± 2.29) × 10^{1}	-	-	-	-	1.10 (± 0.916) × 10^{-1}	-	5.28 (± 1.36) × 10^{-4}	1.23 (± 1.40) × 10^{-1}
MH5		8.19 (± 2.657 × 10^{-2}	-	-	-	7.11 (± 4.25) × 10^{-2}	5.31 (± 4.10) × 10^{-2}	4.68 (± 1.22) × 10^{-2}	1	1.14 (± 1.07) × 10^{-1}	-	-
MA6		1	-	-	-	1.27 (± 1.15)	1.60 (± 1.02)	6.24 (± 3.25) × 10^{-2}	3.08 (± 1.66) × 10^{1}	1	-	-
MH4		-	-	-	-	-	-	-	-	-	1	1
MA5		5.95 (± 0.434) × 10^{-2}	1	5.13 (± 5.28)× 10^{-1}	1.32 (± 0.826)	5.00 (± 4.33) × 10^{-3}	2.93 (± 2.53) × 10^{-3}	3.32 (± 0.976) × 10^{-2}	2.27 (± 0.814) × 10^{-3}	1.18 (± 0.792) × 10^{-3}	-	-
MB5		6.26 (± 0.105) × 10^{-2}	8.40 (± 6.59) × 10^{-2}	1	9.92 (± 1.13) × 10^{-1}	3.33 (± 2.31) × 10^{-4}	1.23 (± 0.180) × 10^{-3}	1.11 (± 1.92) × 10^{-3}	5.82 (± 2.07) × 10^{-3}	1.30 (± 0.928) × 10^{-3}	-	-
MB6		-	8.96 (± 4.37) × 10^{-6}	7.18 (± 2.37) × 10^{-3}	1	-	-	-	9.74 (± 0.955) × 10^{-4}	-	-	-
MF5		2.81 (± 0.718) × 10^{-2}	-	-	-	1	4.34 (± 3.72) × 10^{-2}	2.43 (± 0.184) × 10^{-3}	1.18 (± 0.399)	3.60 (± 1.06) × 10^{-2}	-	-
ME5		-	-	-	-	-	-	-	-	-	-	-
MC5		5.12 (± 1.09)× 10^{-1}	-	-	-	5.00 (± 3.48) × 10^{-1}	1	1.67 (± 0.296) × 10^{-3}	9.25 (± 0.718)	4.79 (± 2.70) × 10^{-1}	-	-
MD5		-	-	-	-	-	-	-	-	-	-	-
MC6		-	-	-	-	-	-	-	-	-	6.43 (± 3.11) × 10^{-1}	4.76 (± 1.04) × 10^{-1}
ME6		7.42 (± 1.48)× 10^{-3}	-	-	-	-	2.22 (± 0.776) × 10^{-2}	1	4.28 (± 1.55) × 10^{-1}	2.31 (± 1.14) × 10^{-2}	-	-

(B) Subgroup IB RBP skunaviruses

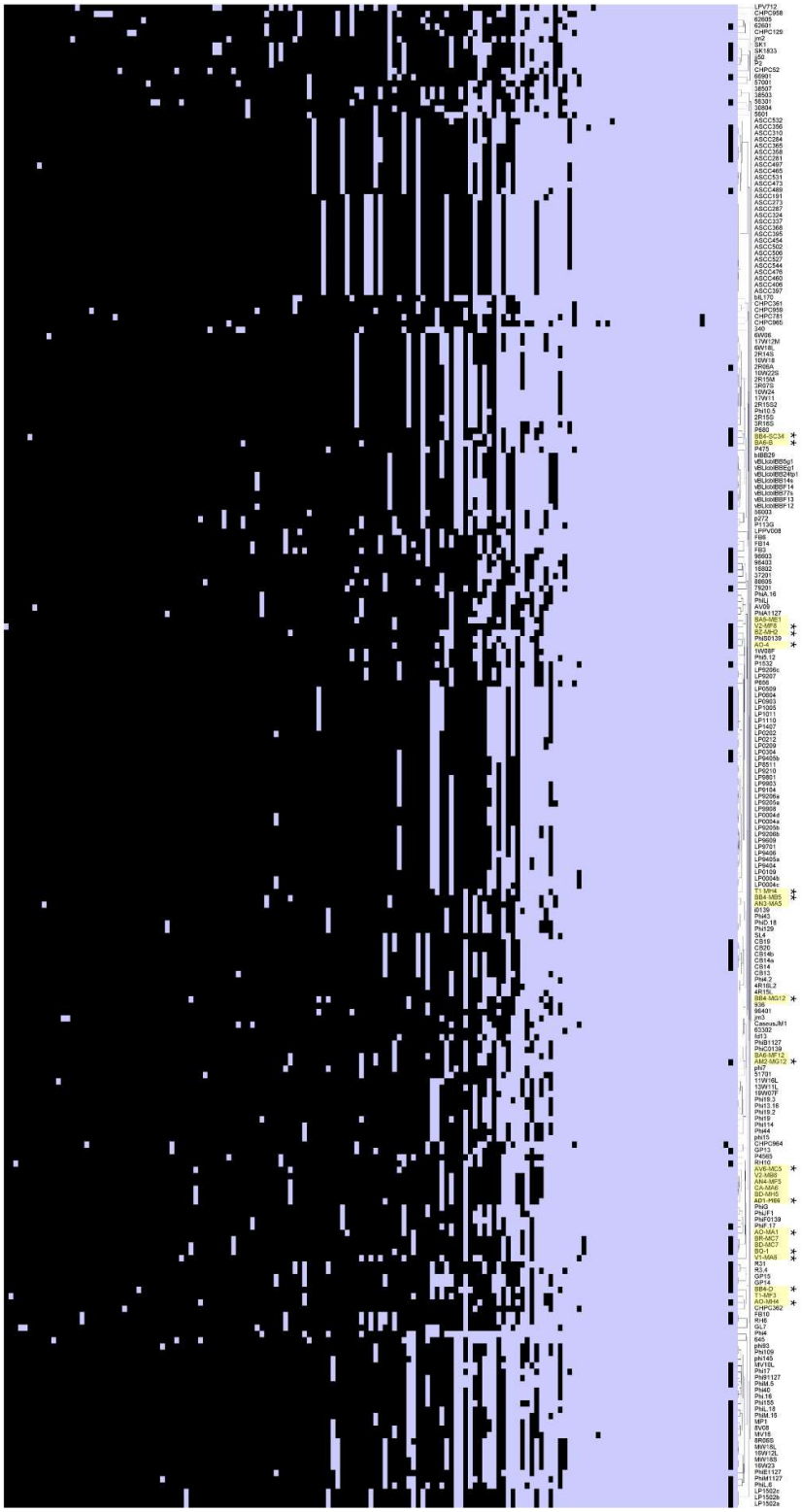
Strains	*cwps* type	BA6-ME1	T1-MF3	BZ-MH2	AO-MA1
ME1	C_1	1	-	-	-
SC74		1.66 (± 0.683) × 10^{-4}	-	-	-
MF3		-	1	-	-
MA1		-	-	1.39 (± 0.202)	1
MH2		-	-	1	3.74 (± 2.03) × 10^{-1}
3107	C_2	-	-	-	-
MH5		-	-	-	2.53 (± 0.759) × 10^{-2}
MA6		-	-	-	2.48 (± 0.518) × 10^{-2}
MC5		-	-	-	5.42 (± 0.993) × 10^{-1}

(C) Subgroup IC RBP, group III, group VI, group VIII, and group X RBP skunaviruses

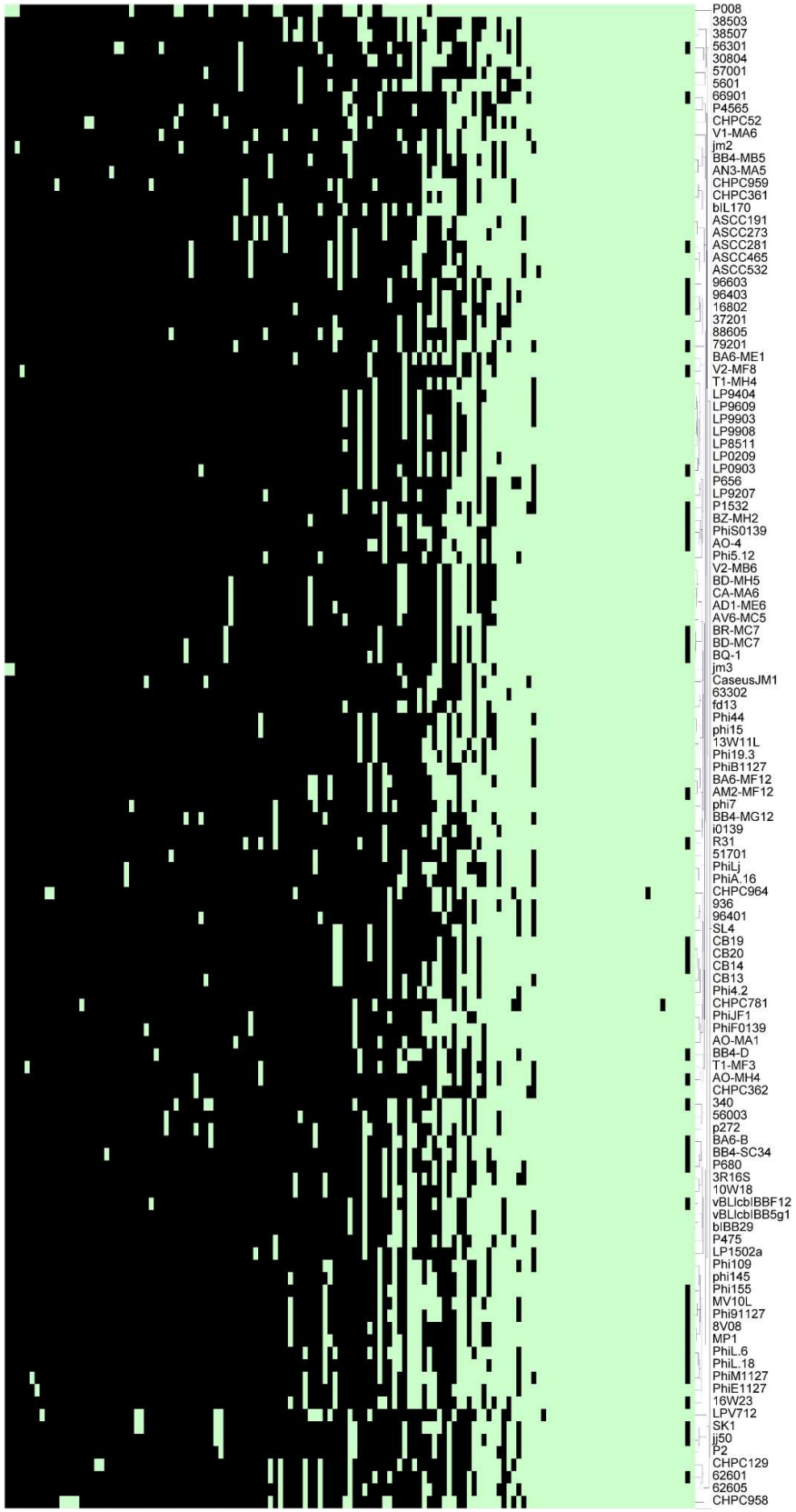
Strains	*cwps* type	Subgroup IC			Group III		Group VI	Group VIII	Group X			
		BB4-MG12	AM2-MF12	BA6-MF12	BA6-B	BB4-SC34	BB4-D	AO-4	V2-MF8	BQ-1	BD-MC7	BR-MC7
IL1403	B	-	-	-	1	1.14 (± 0.236)	-	-	-	-	-	-
SC34		-	-	-	7.49 (± 0.907) × 10^{-1}	1	-	-	-	-	-	-
MC7	C_3	-	-	-	-	-	-	-	-	2.11 (± 0.848)	1	1
BA		-	-	-	-	-	-	-	-	1	1.29 (± 0.541)	9.72 (± 5.42) × 10^{-2}
W34	C_4	6.60 (± 0.318) × 10^{-2}	2.26 (± 1.27) × 10^{-1}	6.92 (± 5.14) × 10^{-1}	-	-	-	-	-	-	-	-
MF12		7.51 (± 2.75) × 10^{-1}	1	1	-	-	-	-	-	-	-	-
MA10		4.04 (± 0.315) × 10^{-1}	1.69 (± 0.667)	2.50 (± 3.25)	-	-	-	-	-	-	-	-
MC12		-	3.75 (± 0.899) × 10^{-5}	-	-	-	-	-	-	-	-	-
MG12		1	2.06 (± 0.517)	3.24 (± 1.52)	-	-	-	-	-	-	-	-
NGF5	C_6	-	-	-	-	-	-	2.85 (± 4.03) × 10^{-6}	-	-	-	-
AO45		-	-	-	-	-	-	1	-	-	-	-
18-16	D	-	-	-	-	-	1	-	-	-	-	-



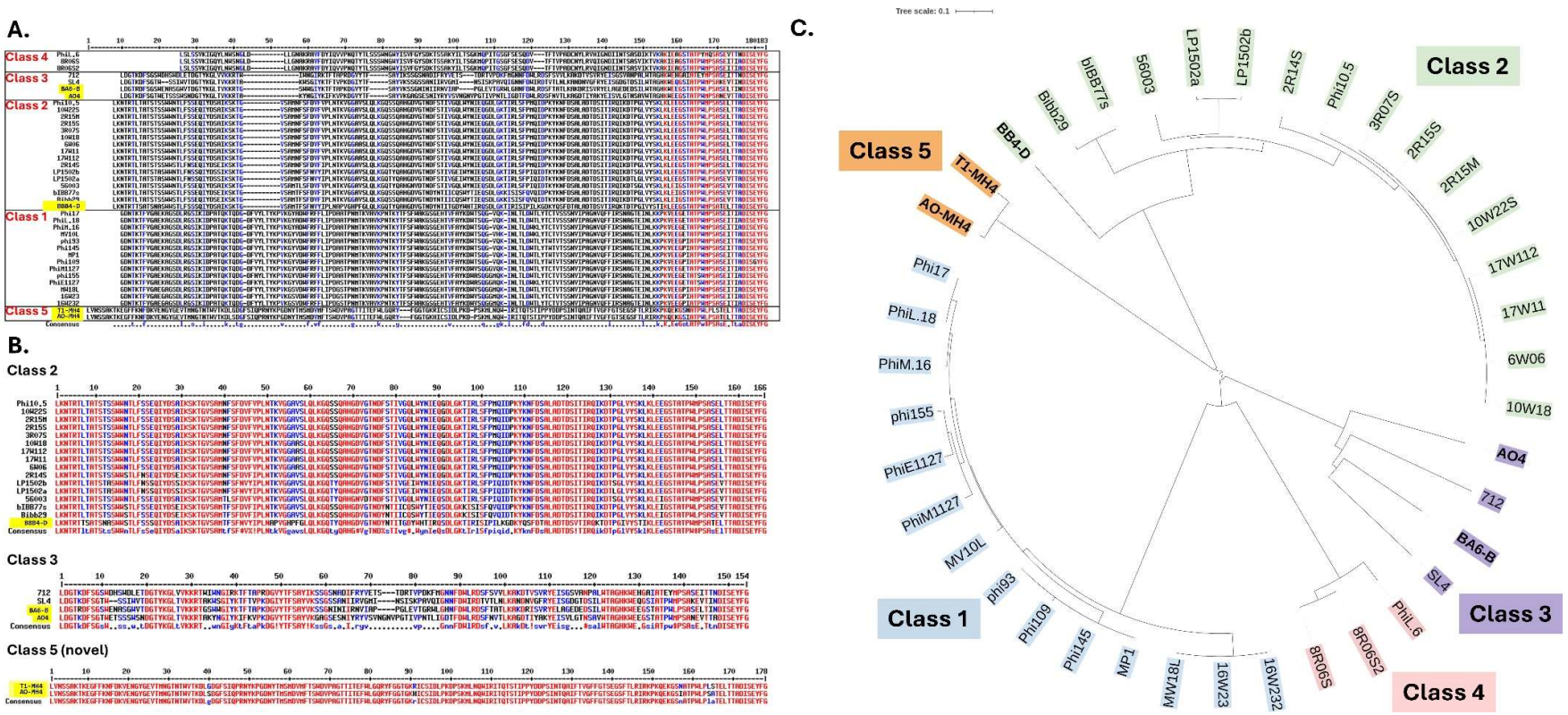
Supplementary Figure S1. Comparative genome analysis of *Ceduovirus* T1-C2 type phages c2 and bIL67.



Supplementary Figure S2. Combined MCL and HCL analysis of all 239 available Skunavirus genomes. Phages sequenced within this study are highlighted. The 16 *Skunavirus* genomes used in the schematic overview of the whey-isolated *Skunavirus* genomes (Figure 3) have a star beside them.



Supplementary Figure S3. Combined MCL and HCL analysis of 121 *Skunavirus* genome of high quality and complete assemblies.



Supplementary Figure S4. Sequence alignment of the classes of insertions in evolved Dit proteins. High consensus (>90%) amino-acids are in red, partially (50%) conserved in blue, and the rest (neutral consensus <50%) in black. (A) Alignment of all five Dit insertion class. (B) Alignment of individual Dit classes in which at least one of the skunaviruses isolated (which are highlighted in yellow) and sequenced in this study falls into. (C) Phylogenetic tree of evolved Dit inserts.

SUPPLEMENTARY MATERIALS REFERENCES

1. White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* **90**:e02152-23.
2. Millen AM, Romero DA. 2016. Genetic determinants of lactococcal *C2viruses* for host infection and their role in phage evolution. *J Gen Virol* **97**:1998-2007.
3. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, Van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci Rep* **6**:21345.

Chapter 5. Genomic analysis and technological assessment of lactococcal strains isolated from undefined mesophilic starter cultures

Experiments, analyses, and first draft preparations were performed by the candidate.

ABSTRACT

Undefined mesophilic starter cultures that incorporate lactic acid bacterial species such as *Lactococcus lactis* and *Lactococcus cremoris* are used in the production of a variety of fermented dairy products. Strain diversity within these undefined starter cultures is typically uncharacterised, however, it is important that the composition of these starter cultures remains stable during milk fermentation to ensure consistency of the final product. In the current study, the genomes of 30 lactococcal strains originating from two mesophilic undefined starter cultures (M1 and M2) were sequenced to investigate the strain-level diversity present in these complex cultures. Genome analysis of the strains revealed significant diversity associated with technologically relevant genomic elements such as cell wall polysaccharide biosynthesis gene clusters, plasmid sequences, and phage-defense systems. In addition, the genomes of strains belonging to the emerging dairy-associated lactococcal species, *Lactococcus laudensis*, were analysed.

INTRODUCTION

Lactic acid bacteria (LAB) include the genera *Lactococcus, Streptococcus, Pediococcus, Leuconostoc, Lactobacillus, Enterococcus,* and *Lactiplantibacillus*, among others, and have been extensively linked to food fermentations due to their ability to produce lactic acid through the metabolism of hexose carbohydrates (1). LAB can be classified as starter LAB, which are primarily responsible for acidification and the primary organolepsis of a fermented product, or non-starter LAB which play a role in maturation and ripening processes (2). Mesophilic starter cultures are largely comprised of two lactococcal species, *Lactococcus lactis* and *Lactococcus cremoris,* in conjunction with *Leuconostoc* spp. in some cases (3). Strains of these mesophilic species grow optimally between 30 to 37 °C (but are capable of fermenting lactose at temperatures between 10 to 40 °C) (4). Starter cultures are furthermore characterised as defined or undefined. The composition of undefined starter cultures is largely unknown as they are composed of a variable number of strains in an unknown composition and abundance (5). Mesophilic undefined starter cultures, for example, can be used to produce both soft and semi-hard cheeses such as Gouda, Edam, and Brie (4). In contrast, defined starter cultures typically comprise a small number of highly characterised strains (typically between two and six) in a known combination. Often these strains are derived from undefined mixed starter cultures (4). Defined starter cultures are typically applied in the production of Cheddar-style cheeses (6, 7).

Considerable research efforts have endeavoured to define the technological properties and genetic composition of the lactococcal species

dominating these starter cultures. Lactococcal chromosomes are typically between ~2.2 to 2.6 Mb (8) and they are reported to have undergone significant decay as a result of adapting from plant-based environments to a nutrient-rich milk-based environment (9–12). The acquisition of certain plasmids has facilitated their adaptation to the dairy environment as they typically harbour genes associated with lactose utilisation, casein metabolism, citrate metabolism, flavour formation, bacteriocin production, and bacteriophage resistance mechanisms (13–16). Lactococcal dairy strains have been reported to contain up to twelve plasmids, while plant-derived isolates typically harbour fewer plasmids (17). One of the major and persistent threats to dairy fermentation processes are bacteriophage (or phage) attacks (18). Phage infection can have significant impacts on starter culture acidification activity, which has negative economic and environmental (i.e. food waste) impacts (19, 20). Plasmid-encoded phage defense mechanisms such as restriction-modification (RM) and abortive infection (Abi)-like systems account, at least in part, for differences in phage susceptibility observed between lactococcal strains (21, 22).

Phage-host interactions between lactococcal phages and their hosts have been extensively investigated (23–27). The most problematic lactococcal phage genus in dairy fermentations is *Skunavirus* (formerly termed the 936 group phages), whose members are obligately virulent (28, 29). Phage infection commences when a *Skunavirus* receptor binding protein recognises and binds to a surface-exposed cell wall polysaccharide (CWPS) receptor structure (27, 30–32). A lactococcal CWPS is composed of a peptidoglycan-embedded rhamnan and surface-exposed polysaccharide pellicle (PSP) (33,

34). The latter structures (PSP) of lactococcal strains are quite diverse and represent the RBP binding target. The chemical diversity of lactococcal PSPs is consistent with the genetic diversity observed in the 3' region of the *cwps* gene cluster. Currently, there are four *cwps*-genotypes: *cwps* type A-D and 11 C-subtypes (C_1 to C_{11}) (30, 35). However, there is likely more genetic and associated chemical diversity than that which is currently defined (36). The structural diversity of the CWPS is believed to be a key contributor to the specific and narrow host range of skunaviruses. Consequently, the application of strains possessing diverse *cwps*-types in a starter culture may aid in culture robustness.

In the current study, the technological properties and genetic diversity of lactococcal strains previously isolated from two undefined starter cultures (M1 and M2) employed in mesophilic dairy fermentations were investigated. The complete genomes of ten *L. cremoris* strains isolated from the two mesophilic starter cultures were sequenced with a combined Illumina and long read (Nanopore or PacBio) sequencing approach. The genetic diversity of the *cwps* gene clusters of the strains was evaluated while the genomes were also assessed for the presence of phage defense systems. An additional, 18 *L. lactis/cremoris* genomes were sequenced using an Illumina platform to gain further insights into the extent of strain diversity present in these mesophilic starter cultures. Furthermore, the genomes of two strains belonging to the emerging lactococcal species *Lactococcus laudensis* were analysed.

MATERIALS AND METHODS

Lactococcal strains growth conditions

Bacterial strains used in this study are listed in Table 1 and Supplementary Table S1. Bacterial strains were grown overnight at 30 °C in M17 broth (Oxoid, Hampshire, United Kingdom) supplemented with 0.5 % lactose (Sigma Aldrich, Gillingham, UK) (LM17).

Table 1. Functional characterisation of strains isolated from mesophilic undefined starter cultures (either M1 or M2).

Strains	Origin	CWPS type or species	Acidification profile	Proteolytic activity	Phage sensitivity
SC34	M1	B	5.8 ± 0.1	-	2/26
ME1	M1	C_1	5.0 ± 0.2	$++^O$	1/26
SC74	M1		4.9 ± 0.1	$++^O$	1/26
MB2	M1		4.2 ± 0.1	$++^O$	0/26
MG3	M2		4.4 ± 0.1	$++^O$	0/26
MF1	M1		4.8 ± 0.1	$++^O$	0/26
MF3	M2		5.3 ± 0.2	$++^O$	1/26
MA1	M1		4.3 ± 0.1	$++^O$	2/26
MH2	M1		4.3 ± 0.2	$++^O$	2/26
MH5	M1	C_2	4.5 ± 0.1	$++^O$	7/26
MA6	M1		4.5 ± 0.1	$++^O$	7/26
MH4	M1		4.8 ± 0.2	$++^O$	2/26
MA5	M1		5.4 ± 0.2	$++^O$	9/26
MB5	M1		4.8 ± 0.2	$++^O$	9/26
MB6	M1		4.9 ± 0.1	$++^O$	4/26
MF5	M1		4.6 ± 0.1	$++^O$	6/26
ME5	M1		4.5 ± 0.1	$++^O$	0/26
MC5	M1		4.6 ± 0.1	$++^O$	7/26
MD5	M1		5.2 ± 0.1	$++^O$	0/26
MC6	M1		4.7 ± 0.2	$++^O$	2/26
ME6	M1		4.4 ± 0.1	$++^O$	5/26
MC7	M1	C_3	4.1 ± 0.1	$++^O$	3/26
MF8	M2		4.9 ± 0.1	$++^O$	1/26
MA9	M2		6.1 ± 0.1	-	1/26
BA	M1		4.6 ± 0.1	$++^O$	3/26
MF12	M1	C_4	5.2 ± 0.2	-	3/26
MA10	M1		5.0 ± 0.1	-	3/26
MC12	M2		4.4 ± 0.2	-	1/26
MG12	M1		5 ± 0.2	-	3/26
UD4	M2	C_5	5.5 ± 0.2	$+^C$	0/26
UG1	M2		5.7 ± 0.1	$+^C$	0/26
NGA6	M1	C_6	4.7 ± 0.1	$+^C$	0/26
NGF5	M1		5.5 ± 0.2	$+^C$	1/26
AO45	M1		4.7 ± 0.1	$+^C$	1/26
UD2	M2	*L. laudensis*	5.8 ± 0.1	$+^C$	0/26
UC8	M2		5.6 ± 0.1	$+^C$	0/26

Good acidifiers are highlighted green, moderate highlighted yellow, and poor acidifiers are highlighted in red. In proteolysis assays, strains that produced small clear zones are labelled $+^C$, strains that only produced opaque zones are labelled $+^O$, opaque zones surrounded by a small clear zone are marked $++^O$, and strains that did not exhibit any proteolytic activity are labelled -.

Functional characterisation of strains

To test the milk acidification capability of each strain, sterilized (115 °C for 20 minutes) 10 % (w/v) reconstituted skimmed milk (RSM) was inoculated with overnight cultures (at 1 %) and incubated for 16 hr at 32 °C. Following incubation, the pH of the culture was measured by a HANNA HI-98103 pH meter (Hanna Instruments, Woonsocket, RI, USA). Proteolytic activity was qualitatively examined by spotting 5 µL of each strain (from a fresh overnight culture) on agar plates containing 1 % reconstituted skimmed milk (RSM) and 1 % agar (w/v). Plates were incubated at 30 °C for 48 hr and strains producing clear or strong opaque zones around colony growth was determined to be positive for proteolytic activity.

DNA extraction, genome sequencing, assembly and annotations for short-reads sequence assembly

For Illumina-based sequencing (sequencing methods attributed to each genome listed in Supplementary Table S2), bacterial DNA was extracted from 2-10 mL of fresh overnight culture of selected lactococcal strains (Table 2) using the Invitrogen PureLink™ Genomic DNA Mini Kit (Thermo Fisher, Waltham, M.A., U.S.A.) according to the manufacturer's instructions with some modifications. Cell pellets were resuspended and incubated in TE buffer [10 mM Tris-HCl (pH 7.5) and 1 mM Ethylenediaminetetraacetic acid (EDTA)] supplemented with 25 % sucrose (Thermo Fisher) and 30 mg/mL lysozyme (Sigma Aldrich). Bacterial genome libraries were prepared using an Illumina Nextera XT DNA Library Preparation Kit (Illumina). Libraries were quantified

using a fluorometric Qubit quantification system (Life Technologies, USA), and loaded on a 2200 TapeStation instrument (Agilent Technologies, USA). Sequencing was performed by GenProbio srl (Parma, Italy) using the Illumina MiSeq platform with a 600-cycle flow cell version 3 (Illumina).

Fastq files of Illumina paired-end reads (250 bp were used as input for genome assembly through the MEGAnnotator2 pipeline (assembly methods attributed to each genome listed in Supplementary Table S2) (37). The SPAdes program v3.14 was used for *de novo* assembly of bacterial sequences with the pipeline option "--careful" and a list of k-mer sizes of 21, 33, 55, 77, 99, 127 (38). Contigs were employed by MEGAnnotator2 for the prediction of protein-encoding open reading frames (ORFs) using Prodigal (39). Predicted ORFs were functionally annotated by means of DIAMOND against the NCBI RefSeq database and INTERPRO (40) against the PFAM database. Furthermore, tRNA genes were identified using tRNAscan-SE version 2.0 (41) and rRNA loci were identified with barrnap (https://github.com/tseemann/barrnap).

Table 2. General features of *L. cremoris* genomes (of strains originating from M1 or M2 starter cultures) sequenced through combined short and long read sequencing approach.

CWPS-type	Origin	Strain	No. of total bases	Chromosome size (bp)	Average GC % (Chromosome, Total)	Contigs	No. of predicted ORFs (Chromosome, Total)	Putative intact prophages	No. of plasmids	Genome completeness* (%)
C_1	M1	**ME1**	2 592 845	2 479 946	35.93, 35.85	6	2594, 2702	1	5	100
	M2	**MG3**	2 722 407	2 555 506	35.88, 35.81	6	2666, 2831	3	5	100
C_2	M1	**MF5**	2 710 142	2 433 009	36.89, 35.72	11	2559, 2836	1	10	100
	M1	**ME5**	2 625 943	2 431 405	35.77, 35.71	11	2573, 2761	2	10	100
C_3	M2	**MF8**	2 589 375	2 372 813	35.89, 35.73	10	2458, 2687	2	9	100
	M2	**MA9**	2 573 593	2 389 246	35.89, 35.74	9	2485, 2682	2	8	100
C_4	M1	**MC12**	2 630 179	2 435 792	35.98, 35.83	10	2548, 2760	1	9	100
	M2	**MG12**	2 420 463	2 579 108	36.05, 35.93	8	2535, 2704	1	7	100
C_5	M2	**UG1**	2 460 012	2 349 458	35.89, 35.81	4	2466, 2584	-	3	100
C_6	M1	**NGA6**	2 549 422	2 395 627	35.97, 35.82	9	2493, 2643	1	8	100

*Genome completeness as calculated by MiGA webserver considering 106 essential genes.

DNA extraction, genome sequencing, assembly and annotations for hybrid assembly

Genomic DNA from strains selected for short (Illumina) and long (Nanopore or PacBio) reads hybrid sequencing assembly (sequencing methods attributed to each genome listed in Supplementary Table S2) was isolated from 40 mL of fresh overnight cultures using Nucleobond® AXG columns and the Nucleobond ® buffer set III (Macherey-Nagel Gmbh, Düren, Germany) according to manufacturer's instructions with some modifications. For cell disruption, 40 µL of lysozyme (100 mg/mL), 50 µL mutanolysin (50 kU/mL) and 25 µL proteinase K (supplied with kit) was added to the resuspended cell pellet. In the final precipitation step, the obtained DNA pellet was dissolved in 10 mM Tris-HCl buffer (pH 7.5) and incubated overnight at room temperature and/or at 55 °C for 1 hr prior to sequencing. Long reads were filtered based on quality using the Filtlong tool (https://github.com/rrwick/Filtlong), while short reads were filtered through the fastq-mcf script (https://github.com/ExpressionAnalysis/ea-utils). The filtered fastq files of MinION long reads obtained from genome sequencing efforts were used as the input for genome assembly through either CANU software (42), Unicycler hybrid assembly pipeline (43), or TryCyler (44) (assembly methods attributed to each genome listed in Supplementary Table S2).

The resulting assemblies were polished through Polypolish (45) using Illumina paired-end reads (250 bp). Polished contigs were processed by MEGAnnotator (46) for the prediction of protein-encoding open reading frames (ORFs) using Prodigal (47). Predicted ORFs were functionally annotated by means of BLASTp using the software DIAMOND (48) performed against the

NCBI nr database coupled with hidden Markov model profile (HMM) searches using InterProScan (49) performed against the manually curated Pfam-A database. Furthermore, tRNA genes were identified using tRNAscan-SE version 2.0 (50), while rRNA genes were detected using Barrnap (https://github.com/tseemann/barrnap).

Genome analysis of strains

ANI distances as well as the completeness of the genome assemblies was determined with the microbial atlas (MiGA) webserver (https://disc-genomics.uibk.ac.at/miga/) (51). Species assignment was based on phylogenetic analysis of the 16S-ITS-23S rRNA against other typed lactococcal strains (Supplementary Table S1) (52). In addition, phylogenetic analysis of the 16S rRNA, *rpoB*, *recA*, and *pheS* genes that had been previously applied in the multi-locus sequence typing and differentiation of *L. lactis* and *L. cremoris* species were used to confirm speciation results (3, 53). The presence of prophage(s) (intact, incomplete, and questionable) was determined using PHASTEST (PHAge Search Tool with Enhanced Sequence Translation) (54). The presence of intact prophage sequence(s) was manually confirmed based on the presence of genes required to produce functional phage particles (i.e., the presence of genetic modules encoding genes associated with lysogeny, replication, packaging, head and tail morphogenesis, and lysis).

To predict the presence of phage defense systems, genome sequences were submitted to the web platform PADLOC (Prokaryotic Antiphage Defence

LOCator) v2.0.0 (55), DefenseFinder V0.1.0 (56–58), and NCBI Translated BLAST (tblastn) (for recently described Abi-like systems that are not currently present on either database (59)).

Comparative genome analysis of cell wall-associated gene clusters

To determine the *cwps* genotype of each strain, the gene cluster responsible for CWPS biosynthesis was identified through inspection of the assembled genomes and the nucleotide sequence of the gene cluster was compared against known *cwps* gene clusters of reference lactococcal strains (Supplementary Table S1) using BLASTn (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The *cwps* gene clusters of the sequenced *L. cremoris* and *L. lactis* strains were compared to other publicly available lactococcal *cwps* gene clusters through the presence/absence of identified protein families with at least 50 % identity over 50 % of the protein using the Markov Clustering (MCL) algorithm via the mclblastline pipeline v12-0678 as described previously (60) followed by hierarchical clustering (HCL) analysis performed and viewed in the multi experiment viewer (MeV) (61). Comparative analysis of *cwps* and *eps* gene clusters compared to those gene clusters of other lactococcal reference strains was generated using CAGECAT version 1.0 web platform (https://cagecat.bioinformatics.nl/tools/clinker) (62). The predicted function of *eps* and *cwps* genes was determined through BLASTp and HHpred (Homology detection and structure production by HMM-HMM comparison) using default settings.

RESULTS

Functional characterisation of lactococcal strains

The mesophilic starter cultures M1 and M2 had previously been shown to harbour a diversity of lactococcal strains (see Chapter 2 of this thesis). A panel of strains isolated from these cultures was used to evaluate the presence of phages in whey samples (see Chapter 4 of this thesis). Based on their distinct phage sensitivity profiles, a total of 36 lactococcal strains isolated from mesophilic undefined starter cultures M1 and M2 were selected for further technological characterisation in the context of this study.

The *cwps* genotypes of the isolated strains was evaluated in the context of previous studies [see Chapters 2 & 3 of this thesis; (63), (64)] using an established multiplex PCR approach. This approach facilitated the genotyping of 32 (out of the 36 assessed) strains with eight lactococcal strains possessing a C_1-subtype *cwps* genotype, 12 C_2-type strains, four C_3-type, four C_4, three C_6, and one B-type strain. Among the four remaining strains, two were identified as C_5 type strains based on sequence analysis of the complete *cwps* gene cluster, following genome sequencing (which will be discussed further below). While the two remaining strains were identified as strains of *L. laudensis* through 16S rRNA gene sequence analysis [see Chapter 2 of this thesis; (63)].

Subsequently, the phenotypic properties of the strains were evaluated including their acidification capabilities. The lactococcal strains were classified as good (where the pH was reduced by 2.0 or more), moderate (where the pH drop was between 1.5-1.9 units), or poor acidifiers (where the pH reduction was less than 1.5 pH units) based on their ability to reduce the pH of milk (from

6.5) after 18 hr of incubation at 32 °C (Table 1). Using these parameters, 13, 12, and 11 strains were identified as poor, moderate and good acidifiers, respectively. The proteolytic activity of all lactococcal isolates was assessed qualitatively through milk (1 %) agar hydrolysis assay (Table 1) (68). Proteolytic activity was evaluated based on the development of opaque or clear zones surrounding the colony, indicating protease activity of a strain (Table 1). Six (of the 36 strains) did not exhibit any proteolytic activity. The majority (23/36) produced opaque zones, surrounded by a small clear zone. The remaining seven strains produced only clear zones. Phage sensitivity profiling performed in the context of Chapter 4 revealed that ten (of the 36) lactococcal isolates were phage insensitive (Table 1) against the evaluated panel of 26 phages. Of note, two of the four C_3-type strains that were found be good acidifiers were also found to be more phage sensitive than their more poorly acidifying C_3-type strain counterparts.

Genomic analysis of lactococcal strains

To assess the genomic diversity of lactococcal strains isolated from the mesophilic undefined M1 and M2 starter cultures, the complete genomes of ten lactococcal strains were sequenced (shown in Table 2) as they possess distinguishing biotechnological features including *cwps* profile, phage sensitivity, acidification profile, and proteolytic activity. A combined sequencing approach of long (Nanopore or PacBio) and short (Illumina) reads sequencing followed by a hybrid assembly of the resulting sequence data was utilised to yield a complete, circularised chromosome and corresponding

plasmid complement for each of the strains (Table 2). The ten selected *L. cremoris* strains included two representatives of C_1 (ME1 and MG3), C_2 (MF5 and ME5), C_3 (MF8 and MA9), and C_4 (MC12 and MG12) *cwps*-type strains. These *cwps* genotypes were found to be the most dominant in the M1 and M2 starter cultures [see Chapter 2 of this thesis; (63)]. Both a phage sensitive and phage robust strain was chosen as a representative for the C_1, C_2, and C_4-type strains, while both C_3 representative were sensitive to just a single phage (Table 1). The remaining two strains were chosen as representatives of the C_5 (strain UG1) and C_6-type (strain NGA6) *cwps* strains (both of which were found to be phage insensitive).

Due to the high level of sequence homology between genomes of different lactococcal species, ANI analysis and phylogenetic analyses (against other well characterised lactococcal strains as species-specific reference strains) of the 16S-ITS-23S rRNA, 16S rRNA, *rpoB*, *RecA*, and *pheS* genes were performed. The latter four genes have been previously applied in the multi-locus sequence typing and differentiation of *L. lactis* and *L. cremoris* (3, 53). All applied methods of speciation resulted in all ten strains being identified as members of the *L. cremoris* species. Chromosome size of the assessed *L. cremoris* strains ranged from 2.35 to 2.58 Mbp (average of 2.44 Mbp), with a GC content between 35.77 to 36.89 % (average of 36.01 %) and encoding between 2,458 and 2,666 (2537.7 on average) open reading frames. This is consistent with other previously sequenced genomes of the *L. cremoris* species (65). Prophages are commonly present in lactococcal genomes, with most strains harbouring at least one such (mobile) genetic element (66). PHASTEST was used to identify potential prophage regions within the

genomes of the analysed lactococcal strains (Table 2). Between one and three (1.4 on average) putative intact prophage sequence(s) were identified in nine of the ten *L. cremoris* strains (Table 2) and all strains also appeared to encode (between one to four) incomplete prophage regions (Supplementary Table S3).

Lactococcal skunaviruses utilise the host surface-exposed component of CWPS as their cognate receptor. To establish the genetic diversity of the receptor encoding region i.e. the CWPS biosynthetic gene cluster, the *cwps* loci of all ten sequenced *L. cremoris* strains was compared to lactococcal *cwps* clusters extracted from publicly available (Supplementary Table S1). The C_1, C_2, C_3, C_4, and C_5-type strains were all nearly identical to a previously established reference strain of their same *cwps*-type (Figure 1). And the most notable difference between the *cwps* loci of NGA6 and the C_6 reference strains A76 is the presence of a transposase-encoding gene between the two rhamnan side-chain transferase genes in NGA6 *cwps* gene cluster.

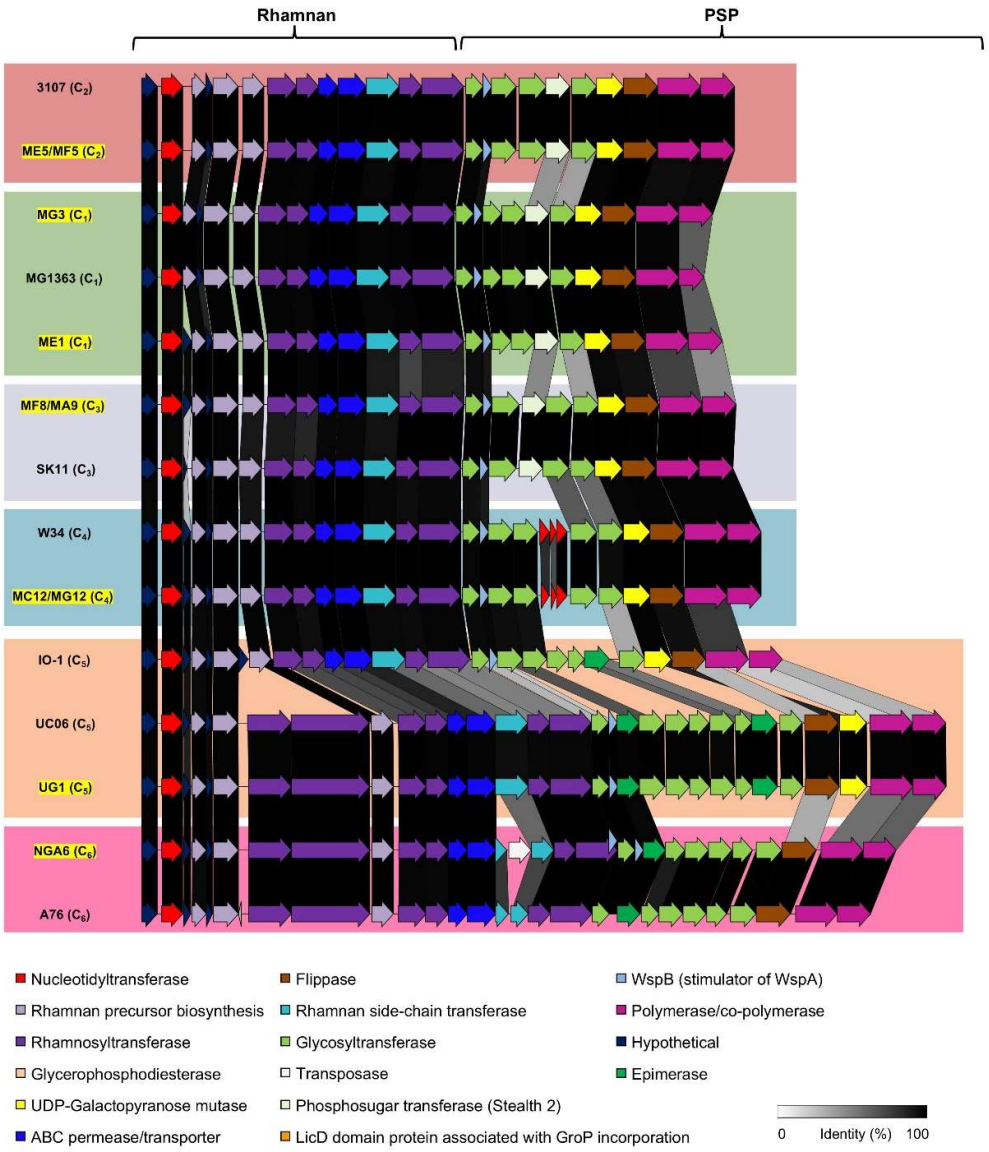


Figure 1. Schematic representation of the *cwps* gene clusters of the *L. cremoris* strains isolated in this study compared to reference strains. Arrows represent open reading frames which are coloured according to their proposed function (indicated in the figure). The strains isolated and sequenced within the study are highlighted yellow.

Evaluation of plasmid-associated functionalities of lactococcal strains

Dairy-associated lactococcal strains typically contain multiple plasmids, which is believed to be the result of their adaptation to dairy environments. Not surprisingly, these plasmids are commonly associated with industrially

relevant phenotypic traits such as lactose utilization, casein metabolism, and phage resistance. The ten *L. cremoris* strains were found to harbour between three and ten plasmids (seven on average), with both C_2-type *L. cremoris* strains encoding the highest number of plasmids and *L. cremoris* UG1 (C_5) encoding the lowest. Pertinent features of the 74 identified plasmids were investigated (Table 3), with a particular focus on technologically relevant, plasmid-encoded characteristics, such as phage resistance, conjugation, exopolysaccharide production (EPS), lactose utilisation, and oligopeptide uptake (Opp). The examined plasmids range in size from 1.98 to 76.03 Kbp (23.5 Kbp average) encoding between two and 83 open reading frames (24.2 ORFs average) with a GC content between 30.85 and 37.83 % (34.1 % average).

Table 3. Characteristics and features of *L. cremoris* plasmids identified through hybrid assembly.

Strain (origin)	Plasmid name	Size (bp)	GC (%)	Genes	Relevant functionality
***L. cremoris* ME1 (M1)**	pME1A	53,517	34.54	50	*Lac* operon, *Opp* operon, mobilization genes
	pME1B	22,294	34.67	23	Mobilization genes, bacteriocin, AbiB
	pME1C	15,714	33.77	16	Type I RM orphan specificity subunit
	pME1D	12,273	33.77	10	Type IIG RM system
	pME1E	9,101	32.04	9	Hypothetical genes (unknown function)
***L. cremoris* MG12 (M1)**	pMG12A	47,204	35.87	42	*Lac* operon, *Opp* operon, oligopeptidase PepB, competence-induced protein (CoiA)
	pMG12B	39,169	36.13	43	EPS production, mobilization genes
	pMG12C	20,851	32.73	25	Metallo-peptidase & -phosphoesterase, *mobilization*, Type I RM orphan specificity subunit, Peptidase E
	pMG12D	19,900	33.81	24	Mobilization genes, pyroglutamyl-peptidase I, intramembrane glutamic endopeptidase

	pMG12E	15,680	31.93	19	Mobilization gene, AbiB, bacteriocin immunity protein, metallophosphoesterase
	pMG12F	9,164	32.01	9	AbiB
	pMG12G	6,676	33.24	7	Hypothetical genes (unknown function)
***L. cremoris* MC12 (M2)**	pMC12A	63,960	34.52	81	Type I RM orphan restriction subunit, bacteriocin, lactococcin, Competence genes (ComC), PII-type proteinase, Peptidase E, mobilization genes
	pMC12B	47,215	34.77	43	*Lac* operon, *Opp* operon, protease, mobilization genes
	pMC12C	20,850	32.73	24	metallophosphoesterase/peptidase, mobilization genes, Peptidase E
	pMC12D	19,048	32.84	20	Type I RM system (subunits R, M & S), AbiB, mobilization genes
	pMC12E	16,884	33.14	16	Type I RM subunits (subunits R, M & S), AbiB
	pMC12F	16,464	34.99	16	Type I RM orphan restriction subunit
	pMC12G	4,373	33.07	3	Type I RM orphan specificity subunit
	pMC12H	3,474	33.68	6	Hypothetical genes (unknown function)
	pMC12I	2,119	33.41	3	Hypothetical genes (unknown function)
***L. cremoris* NGA6**	pNGA6A	44,522	33.65	41	Mobilization genes, conjugal transfer protein (Type IV secretion system), Type III RM system (subunits R and M)
	pNGA6B	32,714	36.30	27	Type I RM system (subunits S, R & M), metallopeptidase, pyroglutamyl-peptidase I, mobilization gene
	pNGA6C	25,487	31.51	28	Mobilization genes (*MobC*), peptidase E, EPS production
	pNGA6D	18,745	33.13	18	Type I RM system (subunits S, R & M), AbiB, mobilization genes
	pNGA6E	13,297	33.23	14	Type IIG RM system, methylase
	pNGA6F	9,547	32.89	9	Type I RM orphan specificity subunit
	pNGA6G	7,500	32.17	10	Mobilization genes, Hypothetical genes (unknown function)
	pNGA6H	1,983	31.58	3	Hypothetical genes (unknown function)
***L. cremoris* MA9 (M2)**	pMA9A	67,759	33.36	83	Conjugation (Type IV secretion system), mobilization genes, Type I RM orphan restriction subunit, competence genes, Hesat (Abi system), Type III toxin-

					antitoxin system (ToxN/AbiQ family toxin), peptidase
	pMA9B	40,068	36.74	39	Type III RM system (R & M subunits), mobilization genes, glutamin endopeptidase, lactococcin, bacteriocin immunity gene
	pMA9C	26,599	33.75	26	Peptidase E, mobilization genes
	pMA9D	16,798	36.94	16	Conjugation (Type IV secretion system), Aminohydrolase peptidase, metallopeptidase
	pMA9E	12,336	34.25	10	Type IIG RM system
	pMA9F	10,530	32.00	12	Mobilization genes, metallophosphoesterase
	pMA9G	6,210	35.93	4	Type II RM system
	pMA9H	4,047	33.56	7	Type IV conjugative transfer protein, hypothetical genes (unknown function)
***L. cremoris* MF8 (M2)**	pMF8A	67,759	32.78	83	Type I RM orphan restriction subunit, competence genes, Hesat (Abi), Type III toxin-antitoxin system (ToxN/AbiQ family toxin), peptidase
	pMF8B	44,552	34.72	42	*Lac* operon, *Opp* operon, peptidase
	pMF8C	40,068	35.31	39	Type III RM system (R & M subunits), intramembrane glutamic endopeptidase, mobilization genes, metallopeptidase, bacteriocin immunity protein, lactococcin
	pMF8D	21,682	33.47	22	Peptidase E, mobilization genes
	pMF8E	16,798	35.56	16	Conjgation (Type IV conjugative transfer protein), metallopeptidase
	pMF8F	10,530	31.96	12	Mobilization genes, metallophosphoesterase
	pMF8G	6,210	35.93	4	Type IV conjugative transfer protein
	pMF8H	4,917	35.93	4	Mobilization gene, hypothetical genes (unknown function)
	pMF8I	4,046	33.54	7	Type IV conjugative transfer protein, hypothetical genes (unknown function)
***L. cremoris* ME5 (M1)**	pME5A	64,109	33.60	72	*Lac* operon, *Opp* operon, Type I RM (orphan R subunit), AbiL (AbiLi & AbiLii), AbiD, bacteriocin, peptidase, lactococcin, metallopeptidase, mobilization genes
	pME5B	53,322	35.39	52	Peptidases, competence genes (*tra*), mobilization genes, *Lac* operon, *Opp* operon

	pME5C	23,600	37.19	17	AbiB, mobilization genes, PII-type proteinase
	pME5D	18,830	34.38	18	Type IIG RM system, AbiJ, Type IV secretory system conjugative protein
	pME5E	17,217	35.20	14	Type IV secretory system conjugative protein, metallophosphoesterase
	pME5F	4,468	32.95	3	Type I RM orphan specificity subunit
	pME5G	4,157	36.64	2	Mobilization genes and hypothetical genes (unknown function)
	pME5H	4,103	37.83	4	Mobilization and hypothetical genes (unknown function)
	pME5I	2,749	32.92	4	Hypothetical genes (unknown function)
	pME5J	1,983	31.01	2	Hypothetical genes (unknown function)
***L. cremoris* MG3 (M2)**	pMG3A	64,884	36.39	55	*Lac* operon, *Opp* operon, mobilization genes, PII-type proteinase
	pMG3B	51,579	33.48	61	Mobilization genes
	pMG3C	32,901	35.42	34	AbiB, Type I RM system (R, M, & S subunits), competence genes
	pMG3D	13,296	33.75	12	Type IIG RM system
	pMG3E	4,241	35.68	3	Hypothetical function (unknown)
***L. cremoris* MF5 (M1)**	pMF5A	76,036	36.21	68	*Lac* operon, *Opp* operon, Type I RM system (R, M, & S subunits), PII-type proteinase, NisI/SpaI family lantibiotic immunity, peptidase, competence genes, peptidase B
	pMF5B	56,418	34.69	46	Type III RM system (M & R subunits), Stk2 (Abi), mobilization genes, conjugation genes
	pMF5C	49,283	33.77	60	Mobilization genes, metallopeptidase, lactoococcin , peptidase
	pMF5D	22,442	31.92	24	Type I RM orphan specificity subunit, mobilization genes, metallophosphoesterase
	pMF5E	21,772	30.85	23	Type II RM system (R and M subunits), conjugal transfer proteins, mobilization genes
	pMF5F	18,743	33.09	21	Type I RM system (R, M, and S subunits), peptidase E
	pMF5G	14,095	32.96	15	Type I RM orphan specificity subunit
	pMF5H	10,727	34.47	12	Mobilization genes, pyroglutamyl-peptidase I
	pMF5I	4,023	37.76	4	Mobilization gene, hypothetical (unknown function)
	pMF5J	3,594	32.69	4	Mobilization gene,hypothetical (unknown function)

***L. cremoris* UG1 (M2)**	pUG1A	57,076	33.67	67	PrrC Abi-like system (in the presence of full Type I RM system), lactococcin, pyroglutamyl-peptidase I, metallophosphoesterase, mobilization genes
	pUG1B	27,052	35.87	26	Bacteriocin, Type I RM orphan specificity subunit, metallopeptidase, mobilization genes
	pUG1C	26,426	33.72	25	Type I RM system (R, M, and S subunits), mobilization genes

Potential technologically relevant functional genes identified in the plasmid complement of the *L. cremoris* strains such as those associated with bacteriocin production, peptidases, conjugation, EPS production, lactose utilisation and oligopeptide uptake, and mobilization are listed in Table 3. Of note, gene clusters responsible for the biosynthesis of EPS were identified in two of the ten *L. cremoris* strains: MG12 and NGA6 (Figure 2). Both *eps* gene clusters possess a conserved transcriptional regulator (*epsR*) at the 5' end of the gene cluster, with plasmid pNGA6C encoding the smaller of the two gene clusters at 17.4 Kbp, while pMG12B encodes the larger *eps* cluster at 30.0 Kbp. BLASTN analysis of the two gene clusters reveals that both appear to be distinct from each other and those in previously sequenced strains' genomes/plasmidomes. A schematic depicting the organisation of the *eps* clusters encoded on pMG12B and pNGA6C and sequence relatedness of each cluster to each other and the *eps* clusters of the most similar *L. lactis/cremoris eps* cluster (pIBB109_AgSa_5 and pJM4A, respectively) is shown in Figure 2.

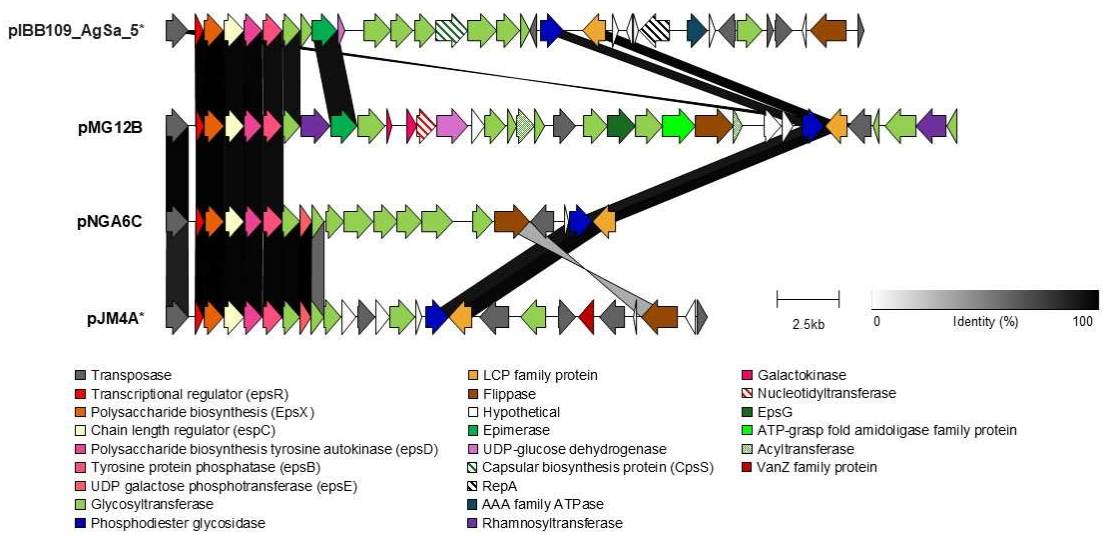


Figure 2. Schematic comparison of the *eps* gene clusters of MG12 and NGA6. Comparative analysis of identified *eps* gene clusters to those of representative strains.

To evade phage infection, lactococcal strains are known to encode various phage defense systems within their genomes (many of which are plasmid-encoded) (10). Genes encoding (components of) RM systems were the most frequently observed phage-resistance genes identified in the 74 studied plasmids, a number of which are orphan genes. Type I RM systems (with a complete complement of three distinct genes encoding subunits for restriction, modification, and specificity activities) were identified in five of the *L. cremoris* strains genomes (three of which carried more than one system). Type IIG RM systems (single protein with an apparent fusion of restriction and modification functional domains) were identified in the plasmid complement of five (of the ten) strains. Type III and Type II RM systems were encoded in the plasmids of four and one strain, respectively. Abi-like systems were identified widely within the plasmid complement of all ten lactococcal strains, with AbiB being the most common (identified in six of the ten strains). The two C_3-type strains (MF8 and MA9) were found to encode both an AbiQ and the recently

described lactococcal Abi-like system Hesat (59). Lastly, several Abi-like systems (AbiL, AbiD, AbiB, and AbiJ) were encoded within in the plasmid complement of the phage-resistant strain ME5.

Predicting the phage defense-ome of lactococcal strains

Although lactococcal phage defense systems are often found to be plasmid-encoded they may also be chromosomally- or prophage-encoded. Therefore, to gain insights into the overall phage "defense-ome" of these ten *L. cremoris* strains, potential antiviral defense systems were identified in their corresponding genome sequences using PADLOC (55), DefenseFinder (56–58), and tblastn (for recently described Abi-like systems that are not yet incorporated in either database (59)). Overall, from the ten *L. cremoris* genomes (both chromosome and plasmidome) a total of 165 individual phage defense-associated genes (16.5 per strain on average) were identified (Figure 3). 90 of these 165 identified phage defense-associated genes were found to be chromosomally-encoded (16 of which were associated with prophages), while the remaining 75 are encoded on plasmid sequences. These defense-associated genes were grouped into three broad categories: RM and RM-like systems (e.g. BREX), Abi-like, or genes belonging to other systems (which includes candidate systems and systems with an undefined mechanism). Among the defense genes, 48 were determined to be Abi-like, 89 RM or RM-like systems, and the remaining 28 predicted phage defense genes were identified as phage defense candidate (PDC) systems or represent those with an undefined mechanism of activity. PDC systems identified by PADLOC are proteins whose predicted defense function was inferred due to the proximity of their encoding genes to genes of previously defined phage-resistance

systems. The breakdown of the distribution of the identified phage defense genes across the chromosome, plasmid complement, and prophage regions is schematically displayed in Figure 3. The vast majority of candidate or systems with unknown modes of action were found on the chromosome (23, ten of which were within prophage regions), with the remaining five defense genes encoded on plasmid sequences. RM (and RM-like) genes were more often encoded within the plasmidome (49/89 genes) of the strains than the chromosome (40/89 genes). Conversely, Abi-like genes were more commonly found on the chromosome (27 of the 48 genes, six of which were identified within predicted prophage regions) and the remaining 21 were located within the plasmid complement of the strains. It is important to note that both PADLOC and DefenseFinder do not consider orphan RM-associated genes as defense systems, therefore the number of RM genes presented here is likely an underestimation of the actual number.

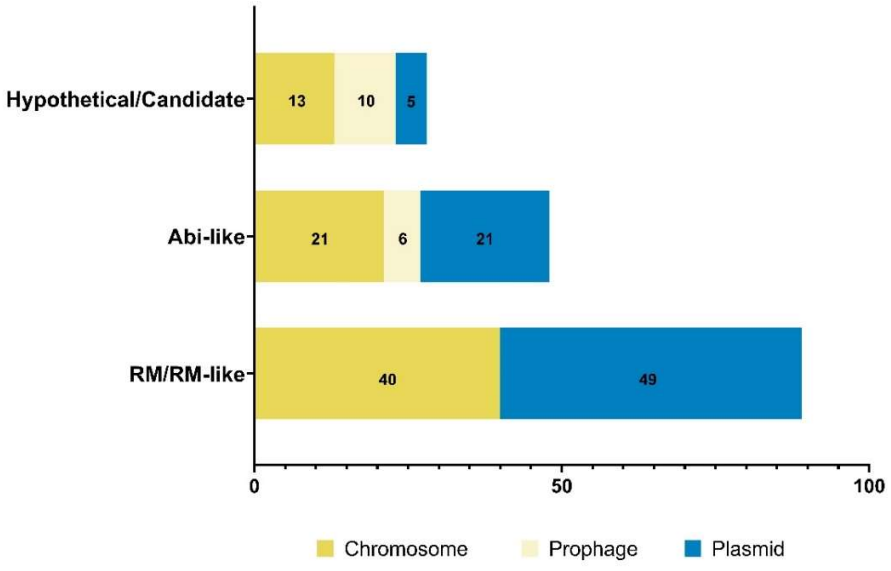


Figure 3. Distribution of Abi-like, RM and RM-like, and those with unknown modes of action (hypothetical) or candidate defense genes across the genome of the ten lactococcal strains.

The RM systems identified through PADLOC and DefenseFinder are classified as either Type I, Type II, Type IIG, or Type III systems. In lactococcal strains, RM systems are frequently encoded on plasmid sequences, with Type I systems being the most common (10). A summary of the presence or absence of both complete and incomplete RM and RM-like systems identified in the examined lactococcal strains (and whether they were found on the chromosome or encoded within a plasmid sequence) is outlined in Figure 4. Type I RM systems were the most frequently identified RM system, with 12 apparently complete systems (consisting of all three subunits) and six incomplete systems (consisting of at least two of the three components) identified amongst the ten *L. cremoris* genomes (Figure 4). All Type II (three systems), Type IIG (five systems), and Type III (four systems) RM systems were plasmid-encoded and complete. In addition, a RM-like, Type I BREX (bacteriophage exclusion) anti-phage system, consisting of eight genes, was identified in the chromosome of ME1. Other DNA modification systems encoding *drmD* (DISARM system helicase), *sspA*, and/or HEases (homing endonucleases) were also identified in five *L. cremoris* genomes (Figure 4).

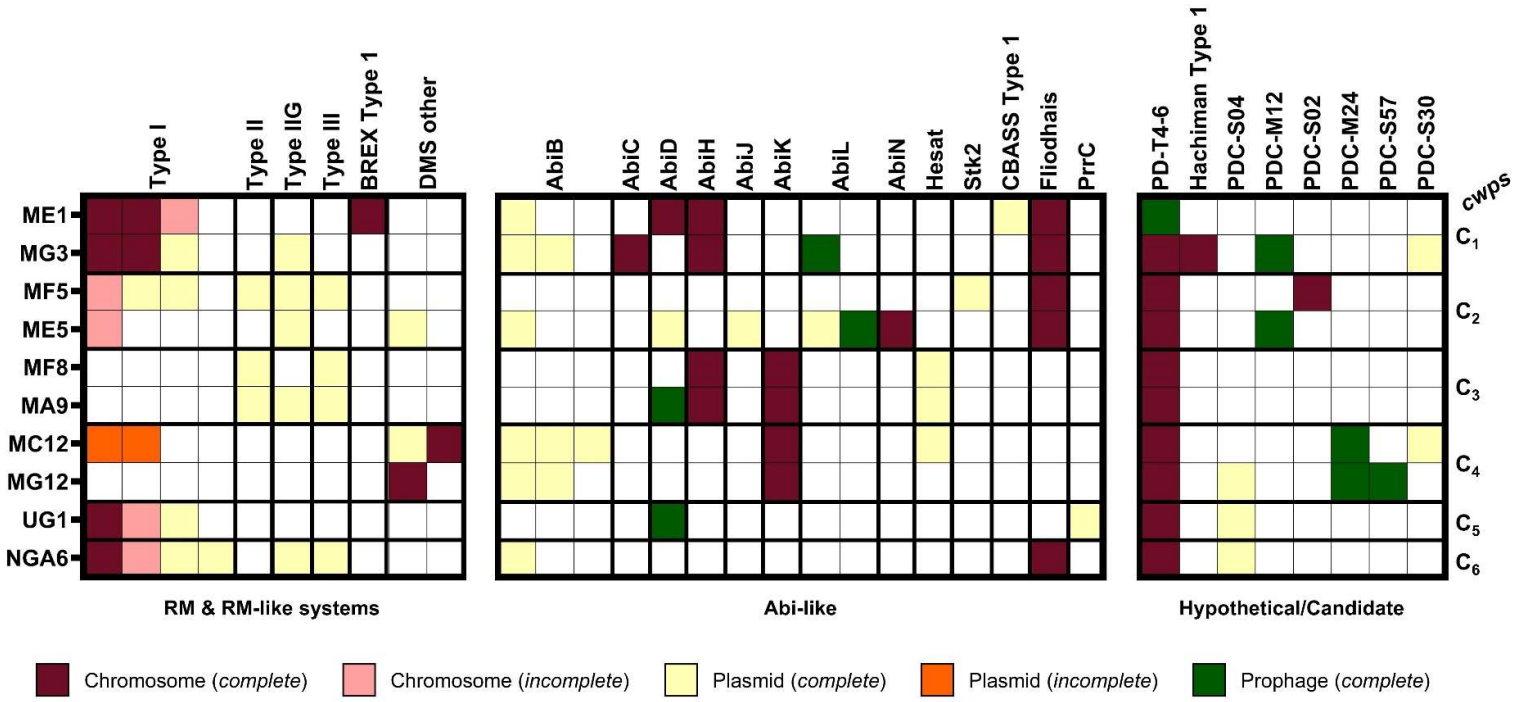


Figure 4. Predicted defense-ome of the ten *L. cremoris* strains isolated from mesophilic undefined starter cultures and sequenced within this study. Presence or absence of defense systems identified by PADLOC and DefenseFinder is indicated by coloured or white boxes, respectively. Defense systems are coloured based on their completeness (as indicated by PADLOC and/or DefenseFinder, based on the presence of all essential genes) and whether they were identified on a plasmid sequence, the chromosome, or within a prophage as indicated by the figure legend.

FURTHER INSIGHTS IN LACTOCOCCAL DIVERSITY OF THE ANALYSED STARTER CULTURES

To complement the sequencing and subsequent hybrid assembly of ten *L. cremoris* genomes, an additional 18 *L. lactis*/*L. cremoris* strains were selected for Illumina-based sequencing (Table 4). Speciation (conducted as described above) revealed that 17 of these strains belong to *L. cremoris* species while the remaining strain was found to be *L. lactis.* The general features of the 18 genomes are outlined in Table 4.

First, to establish the genetic diversity of the gene cluster responsible for receptor biosynthesis, i.e. the CWPS gene cluster, a presence/absence matrix of protein families (at least 50 % identity over 50 % of the protein) encoded within the *cwps* gene cluster of all sequenced *L. cremoris*/*L. lactis* strains in this study (including the ten fully sequenced *L. cremoris* genomes described above) and 65 publicly available *L. cremoris* and *L. lactis* strains was assessed using the MCL algorithm followed by HCL analysis (Figure 5). A subset of *cwps* gene clusters used in a previous MCL/HCL analysis by Mahony *et al.* (30) were included as references, as well the gene clusters representing subtypes C_7 and C_9-C_{11} recently published by Parlindungan *et al*. (35) (a list of strains and corresponding GenBank accession numbers can be found in Supplementary Table S1). The *cwps* genotype of these 18 additional *L. cremoris/L. lactis* strains sequenced was consistent with that previously established through multiplex PCR with the *L. lactis* strain found to possess a B-type *cwps*, while from the 17 *L. cremoris* strains, three were found to be C_1-type, eight C_2-type, one C_3-type, two C_4-type, one C_5-type, and two C_6-type (Figure 5).

Table 4. General genome features of additional *L. cremoris/L. lactis* strains sequenced (with Illumina-based sequencing) in this study.

Species	Strain	Origin	CWPS-type	No. of bases	Contigs	No. of predicted ORFs	Average GC %	Putative intact prophages	Genome completeness* (%)
L. lactis	SC34	M1	B	2 298 151	49	2303	36.10	2	100
L. cremoris	SC74	M1	C_1	2 428 822	181	2460	35.51	1	100
	MB2	M1		2 365 138	212	2383	36.71	1	100
	MF1	M1		2 439 630	178	2490	35.53	1	100
	MH4	M1	C_2	2 462 059	129	2526	35.51	1	100
	MA5	M1		2 384 344	117	2452	35.58	1	100
	MB5	M1		2 384 069	121	2458	35.57	1	100
	MB6	M1		2 506 707	146	2585	35.53	-	100
	MC5	M1		2 563 557	149	2640	35.54	1	100
	MD5	M1		2 540 106	147	2623	35.56	1	100
	MC6	M1		2 459 396	133	2540	36.61	1	100
	ME6	M1		2 557 285	139	2627	35.53	-	100
	BA	M1	C_3	2 460 507	168	2561	35.53	-	100
	MF12	M1	C_4	2 352 304	179	2418	35.61	1	100
	MA10	M1		2 346 101	186	2410	35.61	-	100
	UD4	M2	C_5	2 301 533	160	2352	36.69	-	100
	NGF5	M1	C_6	2 435 497	200	2471	36.63	-	100
	AO45	M1		2 400 837	195	2431	35.54	-	100

*Genome completeness as calculated by MiGA webserver considering 106 essential genes.

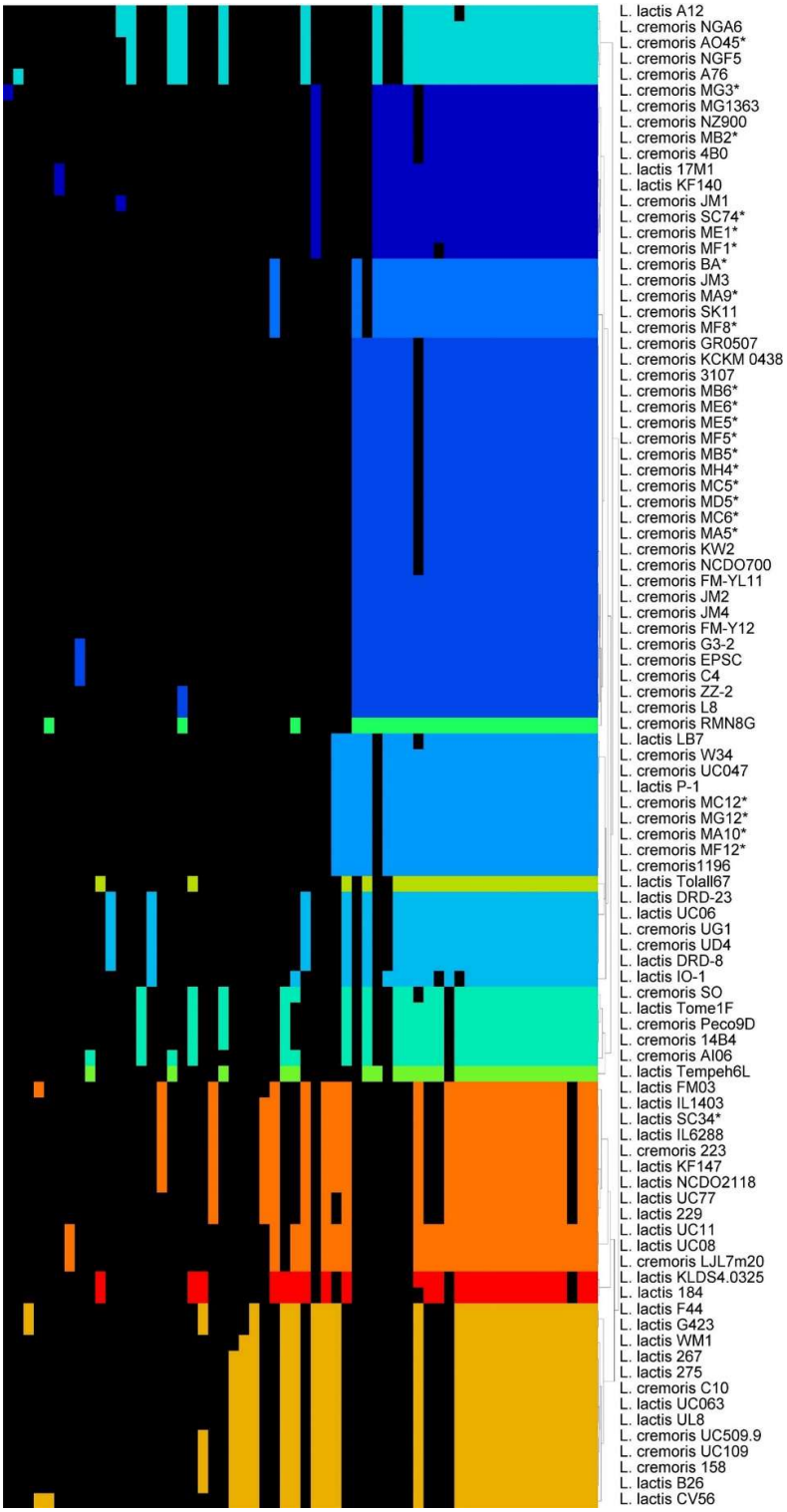


Figure 5. Combined MCL and HCL analysis of the *cwps* gene cluster of 95 *L. cremoris* and *L. lactis* strains. The presence (colour) or absence (black) of individual protein families is indicated in the heatmap. The assigned *cwps* genotypes are indicated and colour-coded within the heatmap. Strains sequenced within this study are indicated with an asterisk beside their name.

Predicted phage-defense systems in these 18 additional *L. lactis/L. cremoris* genomes were assessed to determine if additional defense genes/systems may be present in these strains (beyond those described above). The latter may indicate that M1 and M2 starter cultures possess a higher strain-level diversity compared to that uncovered in the ten lactococcal genomes as described above. The diversity and abundance of the defense systems identified in these genomes is summarised in Figure 6.

Overall, these 18 *L. cremoris/L. lactis* Illumina-assembled genomes add an additional 263 individual defense genes to those total defense-ome of these starter cultures. However, it is important to note, that as most of the genomes are assembled into over 100 contigs this may be an underrepresentation of the complete phage defense-ome especially in the case of multi-component systems, such as RM systems. RM and RM-like system genes were yet again the most commonly identified defense-associated genes with 156 genes identified across the 18 strains' genomes, followed by Abi-like genes (60 genes), and those with unknown modes of action or candidate defense genes (47 genes). These 18 lactococcal strains harbour between four and thirteen defense systems, 8.3 on average. The more recently described two component Fliodhais Abi-like system was the most commonly identified Abi-like system identified in 13 of the 18 strains, although the antiphage effect of this system was found to only be limited to plaque size reduction (59). The Abi-like systems AbiG, Paris, and SEFIR were identified in one of these 18 lactococcal strains (*L. cremoris* strain BA) but not in any the ten *L. cremoris* strains described above. Other defense systems identified in at least one of these 18 strains (but not the previous ten) include: Viperin solo (Virus Inhibitory

Protein, Endoplasmic Reticulum-associated, Interferon-inducible) and PDC-M33. This indicates that although a considerable number of (97) defense systems were identified in the ten-hybrid assembled *L. cremoris* strain, there is likely to be additional systems and more strain-level diversity present in these starter cultures than what is discussed in this study.

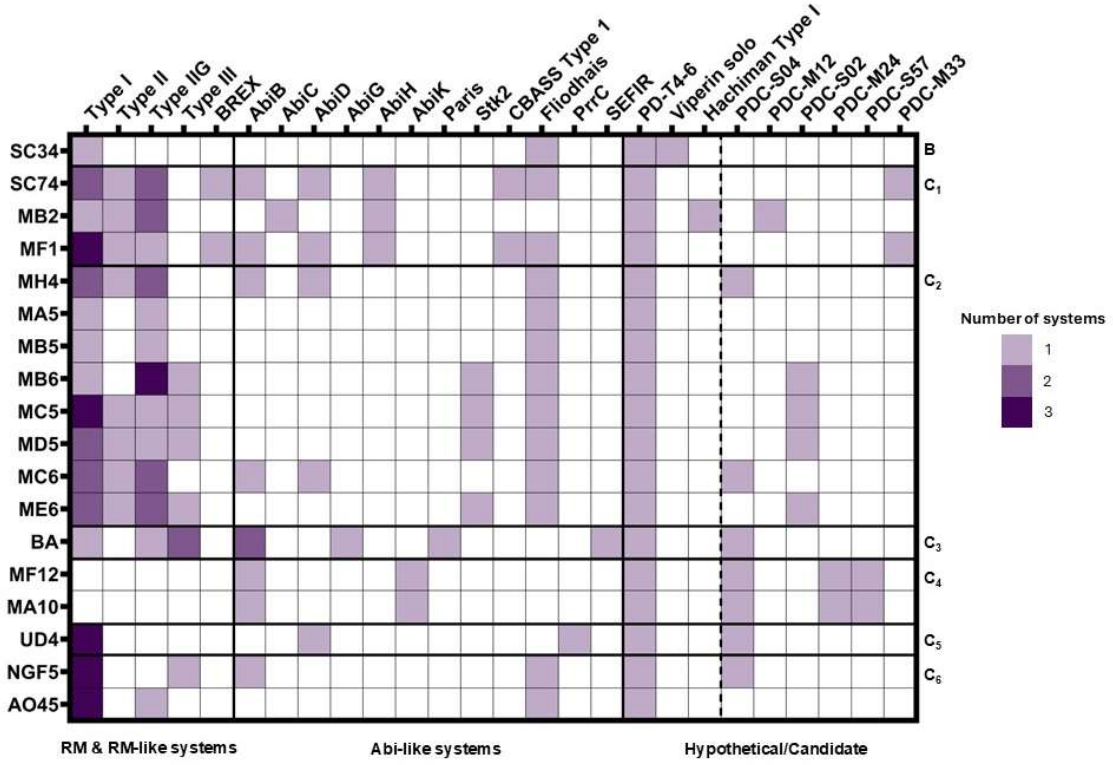


Figure 6. Overview of the diversity of defense systems identified in 18 *L. lactis* and *L. cremoris* genomes. Presence or absence of defense systems identified by PADLOC and DefenseFinder is indicated by coloured or white boxes, respectively. Coloured boxes vary in shades based on the number of defense systems identified in a particular genome, as indicated by the figure legend.

Characterisation of *Lactococcus laudensis* strains within mesophilic starter cultures

In addition to the dominant *L. cremoris* component of undefined starter culture M2, *L. laudensis* strains were also identified in the starter culture in low

abundance [see Chapter 2 of this thesis; (63)]. There are currently just eight *L. laudensis* genome sequences available in public repositories, with an average genome size of ~2.3 Mb. The genome sizes of the two sequenced *L. laudensis* strains UD2 and UC8 were consistent with previous reports at 2.4 and 2.3 Mb, respectively (Table 5). Presumed intact prophages were identified in both *L. laudensis* strains (two in UD2 and one in UC8), however since *L. laudensis* phages have not been reported or studied to date, it is difficult to confidently predict if the identified prophage sequences are truly 'intact' and functional.

Table 5. General genomic features of the two *L. laudensis* strains. Bolded genome (UD2) was assembled through hybrid assembly using short- (Illumina) and long- (Nanopore) reads.

Strain	No. of bases	Contigs	No. of predicted ORFs	Average GC %	Putative intact prophages	Genome completeness (%)
UD2	2 398 433	3	2,505	38.67	2	100
UC8	2 245 654	138	2,324	38.72	1	100

The two *L. laudensis* strains harboured an identical *cwps* gene cluster and a schematic representation of the *cwps* gene clusters of *L. laudensis*, *L. raffinolactis, L. lactis,* and *L. cremoris* strains is presented in Figure 7. This analysis highlights that there appears to be a degree of sequence conservation at the 5' end of the *cwps* loci (responsible for biosynthesis of rhamnan component of the CWPS) between lactococcal strains of different species. However, there appears to be a considerable amount of diversity in the 3' end of the *cwps* loci (which encodes the PSP/oligosaccharide side chain) of the *L. laudensis* strains not only when compared to other species, but also between strains of the same species. Of note, it appears that the *cwps* loci of UD2 and

UC8 more closely resemble the *cwps* of *L. raffinolactis* Lr_19_17 than those of any of the other *L. laudensis* strains analysed.

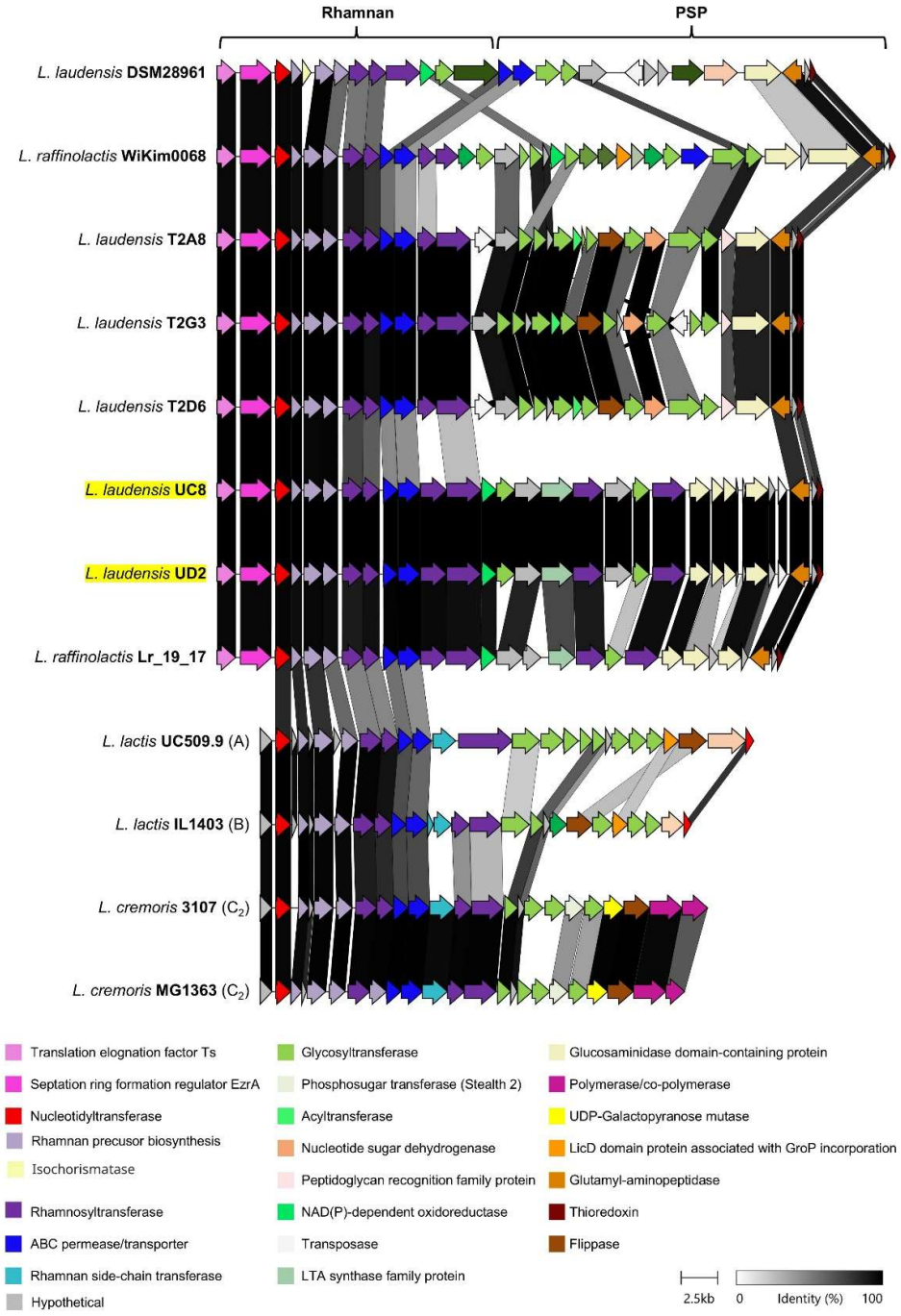


Figure 7. Schematic overview of the organization and sequence relatedness of the putative CWPS-encoding gene clusters of lactococcal (*L. lactis, L. cremoris, L. laudensis,* and *L. raffinolactis*) strains. The *cwps* gene cluster corresponding to the two *L. laudensis* strains (UC8 and UD2) sequenced within the current study are

highlighted yellow. Arrows represent open reading frames which are coloured according to their proposed function, determined through BLASTP and HHpred analysis, (indicated in the figure).

PADLOC and DefenseFinder analysis of the two *L. laudensis* genomes revealed that both strains' genomes possess a nearly identical repertoire phage defense system (and PDCs). The defense repertoire of these strains consists of an intact Type I RM system, SoFic, Hma-embedded candidate HEC-06, Hachiman type 1, and three PDC systems (PDC-S07, PDC-S04, and PDC-S14). In addition, *L. laudensis* UD2 was found to encode a DNA modification system encoding a Type II methyltransferase and a *drmD* gene. No known lactococcal Abi-like systems were identified in either genome.

A combined sequencing approach (using Nanopore and Illumina-based methods) was used to determine both the complete chromosomal and plasmid complement of *L. laudensis* UD2. This strain was found to harbour two plasmids (Figure 8; Supplementary Table S4). The larger plasmid (pUD2A) is 45.4 Kbp in size (45 ORFs) and harbours a *lac* operon (Figure 8A) while the smaller plasmid (pUD2B) is 24.1 Kbp and encodes *eps* genes. This *eps* gene cluster encompasses 10.6 Kbp beginning at a gene encoding a hypothetical protein that shares little amino acid identity with an XRE family transcriptional regulator found in *L. lactis* (41 % identity across 47.30 % of the protein) and ending at a transposase-encoding gene. However, this is unlikely to be a functional *eps* cluster as it only consists of multiple repeats of two glycosyltransferase genes (*epsF* and *epsJ*) genes interspaced by several transposases (Figure 8B).

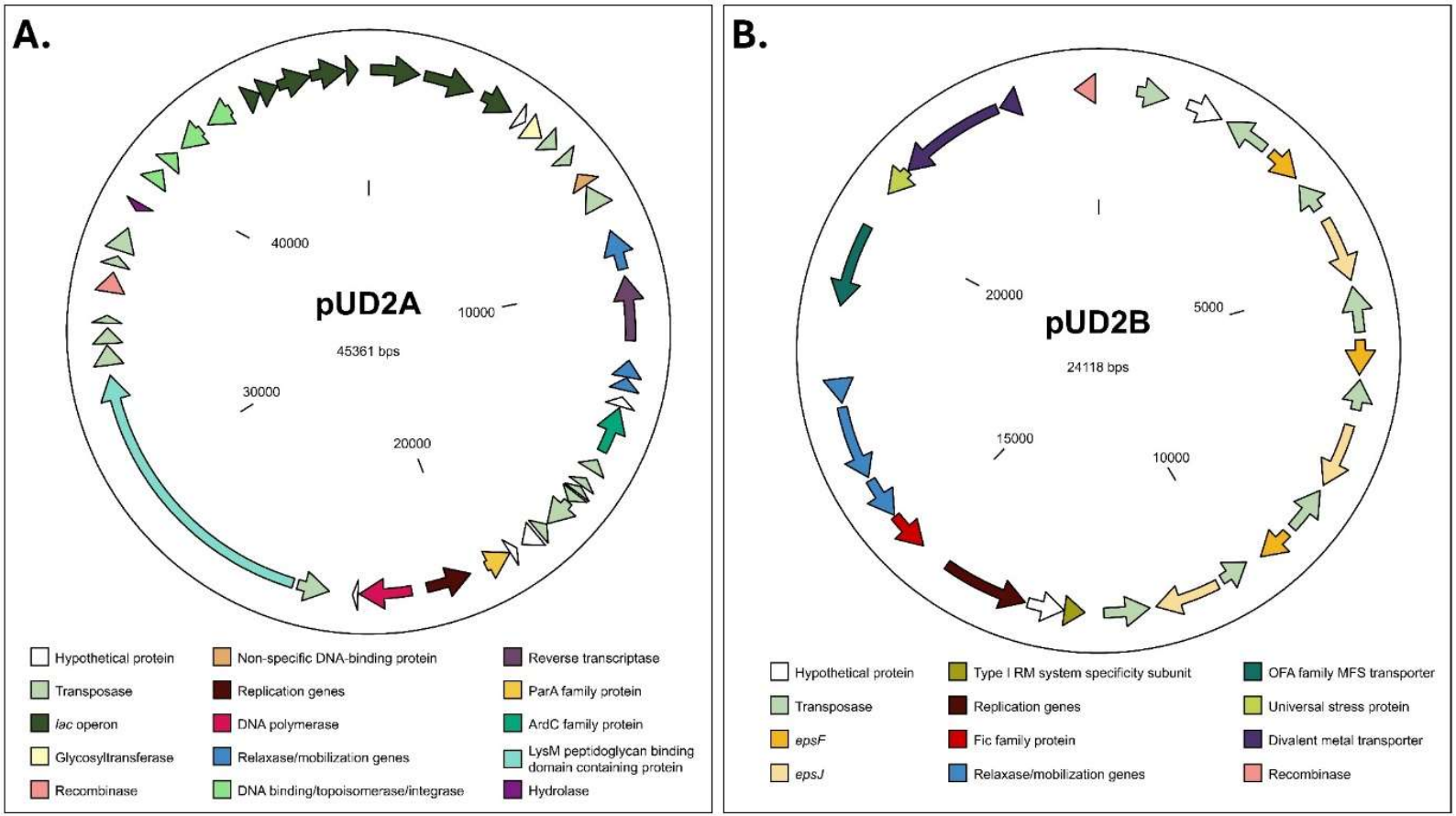


Figure 8. Schematic representation depicting the general features of *L. laudensis* UD2 plasmids. Circular maps of (A) pUD2A and (B) pUD2B. Genes of known predicted function encoded on each plasmid are coloured, function/predicted domains of each gene/operon are coloured according to the figure legend.

DISCUSSION

In the current study lactococcal strain diversity present in two mesophilic undefined starter cultures (M1 and M2) used as starter cultures in dairy fermentation factories across Europe was assessed. Through this analysis, a total of 30 lactococcal genomes belonging to three species (27 *L. cremoris,* one *L. lactis*, and two *L. laudensis* strains) were sequenced and analysed. Genomic and technological assessment (acidification, proteolytic activity, and phage sensitivity) revealed a functionally and genomically diverse pool of lactococcal strains (11, 65, 67). In addition, a combined sequencing approach using Nanopore/Pacbio and Illumina-based sequencing methods was used to generate ten *L. cremoris* and one *L. laudensis* genomes to investigate both the complete, circular chromosome and the plasmid composition of these strains. By assembling the data outputs from both sequencing platforms, it was possible to benefit from the low error rates of Illumina sequencers and production of reads long enough to span plasmid sequences with Nanopore sequencing (68).

The first indicator of whether or not a *Skunavirus* is compatible with a particular lactococcal strain is the host's CWPS, as an infecting phage must first recognise and bind to this specific host receptor. A significant amount of diversity was found within the *cwps* loci of the *L. lactis* and *L. cremoris* strains sequenced in this study, representing seven *cwps* genotypes (B and C_1-C_6). In addition, a notable genetic diversity was observed between the *cwps* gene clusters of C_5-type *cwps* (Figure 1), with both C_5-type *L. cremoris* strains sequenced within this study (UG1 and UD4) sharing high nucleotide identity to previously typed C_5 strain *L. lactis* UC06 (98.98 % identity, across 100 % of

the cluster) but all three sharing only limited similarity to another C_5-type strain IO-1 (98 % identity across 31 % of the cluster) (30). It now appears likely that the CWPS produced by IO-1 and the UC06-like members of this genotype is distinct and that the UC06-like genotype should be considered a novel C-subtype. Further studies of the CWPS composition and structure would be required to validate this hypothesis.

Following binding of a phage to a cognate host, lactococcal strains typically encode an arsenal of phage defense systems such as RM and Abi-like systems. The most common defense system in the genomes of the analysed strains PD-T4-6 (found in all genomes) however, this is a single protein phage defense system identified in *E. coli* (and was identified as an Abi-like system) and has not been functionally evaluated in lactococci yet (69). Often in previously studied lactococcal genomes, phage defense systems have been found to be plasmid-encoded (10). Specifically, the majority of currently identified lactococcal Abi-like systems have been found to be plasmid-encoded (59, 70). However, the research emphasis placed on identifying defense systems on plasmid sequences is likely due to their mobility. This holds particular industrial significance as derivative strains generated through the natural transfer methods (e.g. conjugation) of beneficial plasmids are not considered to be genetically modified organisms (GMOs) and can therefore be readily used in dairy fermentation culture systems (71). As an unintended ramification of this plasmid-centric emphasis, chromosomes may be an understudied domain with respect to lactococcal phage defense systems. In the current study, over half (54.5 %) of the defense genes identified in the ten hybrid-assembled *L. cremoris* genomes were found to be

chromosomally-encoded. More specifically, more than half of the 48 identified Abi-like genes were found on the chromosome (27 genes, six of which were within prophage regions) versus the plasmid sequenced of the ten strains. The diversity and abundance of phage defense systems encoded within the additional 18 Illumina-based *L. cremoris/lactis* genome sequences was also determined (Figure 5) however, it is not possible to definitively predict whether these defense systems are chromosomal, prophage, or plasmid in origin as the genomes were assembled into multiple (typically > 100) contigs. Overall, a diverse phage defense-ome was uncovered in the 28 total *L. cremoris* and *L. lactis* genomes and an increase in defense systems was typically correlated to reduced phage sensitivity. For example, of the C_2-type *L. cremoris* strains the two most phage sensitive strains (MA5 and MB5) encoded the lowest number of anti-phage systems (four each) whereas one of the phage resistant C_2-type *L. cremoris* strains (ME5) encoded significantly more (12 systems). Overall constant interactions with the ubiquitous phages present in dairy fermentation facilities likely supports and drives the maintenance of significant strain-level diversity in the starter cultures. In addition to Abi-like systems, RM and RM-like genes and corresponding systems were identified in abundance in the lactococcal genomes. However, there is likely additional RM genes than those identified here by PADLOC or DefenseFinder, as these platforms do not identify orphan RM genes as complete defense systems. In addition to the known defense systems a significant amount of a PDC systems were identified in a comparatively (to the diversity likely present in the undefined starter cultures) small subset of strains.

These starter cultures were previously found to mainly be comprised of *L. cremoris* strains of varying *cwps* genotypes [see Chapter 2 of this thesis; (63)], however *L. laudensis* strains were also isolated from the mesophilic starter culture M2. Two of these *L. laudensis* strains (which appear to contain nearly identical genomes) were sequenced adding to the eight *L. laudensis* genomes currently available in the GenBank public database. These strains are likely to be present in very low abundance in the undefined starter culture [as they were not identified in through metagenomic analysis, see Chapter 2 of this thesis; (63)] and were not found to contribute to acidification (Table 1). However, future studies into genetic and metabolic diversity of these strains will no doubt provide further insights into the potential role played by these strains as part of the starter culture and the functionality of this dairy-associated species in dairy fermentations. In particular, future studies into the structural diversity that likely correspond to the genetically diverse *cwps* genotypes observed in *L. laudensis* genomes may render insight into these potential phage receptors (once phages against these strains are isolated). The applicability and role of these species along with other emerging dairy-associated lactococcal species (such as *Lactococcus hircilactis* and *Lactococcus raffinolactis*) in starter culture systems remains largely unclear (65, 67, 72).

CONCLUSION

Through the characterisation of a total of 30 lactococcal (*L. lactis*, *L. cremoris*, and *L. laudensis*) genomes significant insights into the strain-level diversity present in two mesophilic undefined starter cultures (M1 and M2) were garnered. The diversity of the technologically relevant (in regard to phage-host interactions) gene clusters (*cwps* and *eps*) encoding surface-exposed, cell wall or exo-polysaccharide structures was examined. In addition, the plasmidome of several lactococcal strains was thoroughly assessed which may be used to identify potential candidate strains that can be exploited to achieve more phage robust strains and starter cultures due to the presence of industrially relevant traits encoded alongside natural DNA transfer systems. Furthermore, the sequencing and assembly of a complete *L. laudensis* genome (both chromosome and plasmid complement) adds to the currently limited body of knowledge surrounding the genetic diversity of this emerging dairy lactococcal species. Overall, a better understanding of the genetic diversity of strains forming, in part, undefined mesophilic starter cultures allows for a glimpse into the fascinating and diverse microbial community present in these fermentation systems.

REFERENCES

1. Tamang JP, Watanabe K, Holzapfel WH. 2016. Diversity of microorganisms in global fermented foods and beverages. *Front Microbiol* **7**:377.
2. Gobbetti M, De Angelis M, Di Cagno R, Mancini L, Fox PF. 2015. Pros and cons for using non-starter lactic acid bacteria (NSLAB) as secondary/adjunct starters for cheese ripening. *Trends Food Sci Technol* **45**:167–178.
3. Li TT, Tian WL, Gu CT. 2021. Elevation of *Lactococcus lactis* subsp. *cremoris* to the species level as *Lactococcus cremoris* sp. nov. and transfer of *Lactococcus lactis* subsp. *tructae* to *Lactococcus cremoris* as *Lactococcus cremoris* subsp. *tructae* comb. nov. *Int J Syst Evol Microbiol* **71**:004727.
4. Johnson ME. 2013. Mesophilic and Thermophilic Cultures Used in Traditional Cheesemaking. *Microbiol Spectr* **1**:10.1128/microbiolspec.cm-0004-2012.
5. McGrath S, Fitzgerald GF, Sinderen D van. 2004. Starter Cultures: Bacteriophages, p. 163-190. *In* Fox PF, McSweeney PLH, Cogan T, Guinee T (ed), Cheese: Chemistry, Physics and Microbiology, Third Edition. Academic Press, San Diego.
6. Timmons P, Hurley M, Drinan F, Daly C, Cogan TM. 1988. Development and use of a defined strain starter system for Cheddar cheese. *Int J Dairy Technol* **41**:49–53.
7. Farkye NY. 2014. CHEESE | Microbiology of Cheesemaking and Maturation, p. 395-401. *In* Batt CA, Tortorello M Lou (ed),

Encyclopedia of Food Microbiology, Second Edition. Academic Press, Oxford.

8. Kelleher P, Bottacini F, Mahony J, Kilcawley KN, van Sinderen D. 2017. Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation. *BMC Genomics* **18**:267.
9. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee J-H, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W, Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B, Mills D. 2006. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci USA* **103**:15611-15616.
10. Kelleher P, Mahony J, Bottacini F, Lugli GA, Ventura M, Van Sinderen D. 2019. The *Lactococcus lactis* pan-Plasmidome. *Front Microbiol* **10**:1–20.
11. Nomura M, Kobayashi M, Narita T, Kimoto-Nira H, Okamoto T. 2006. Phenotypic and molecular characterization of *Lactococcus lactis* from milk and plants. *J Appl Microbiol* **101**:396–405.
12. Torres Manno M, Zuljan F, Alarcón S, Esteban L, Blancato V, Espariz M, Magni C. 2018. Genetic and phenotypic features defining

industrial relevant *Lactococcus lactis*, *L. cremoris* and *L. lactis* biovar. *diacetylactis* strains. *J Biotechnol* **282**:25–31.

13. Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D. 2014. The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. *FEMS Microbiol Rev* **38**:1066–1088.
14. Price CE, Zeyniyev A, Kuipers OP, Kok J. 2012. From meadows to milk to mucosa - adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. *FEMS Microbiol Rev* **36**:949-971.
15. Wels M, Siezen R, van Hijum S, Kelly WJ, Bachmann H. 2019. Comparative genome analysis of *Lactococcus lactis* indicates niche adaptation and resolves genotype/phenotype disparity. *Front Microbiol* **10**:4.
16. Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D. 2014. The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. *FEMS Microbiol Rev* **38**:1066–1088.
17. van Mastrigt O, Di Stefano E, Hartono S, Abee T, Smid EJ. 2018. Large plasmidome of dairy *Lactococcus lactis* subsp. *lactis* biovar *diacetylactis* FM03P encodes technological functions and appears highly unstable. *BMC Genomics* **19**:620.
18. McGrath S, Fitzgerald GF, van Sinderen D. 2007. Bacteriophages in dairy products: Pros and cons. *Biotechnol J* **2**:450–455.
19. Mahony J, van Sinderen D. 2022. Virome studies of food production systems: time for ‘farm to fork’ analyses. *Curr Opin Biotechnol* **73**:22–27.

20. Samson JE, Moineau S. 2013. Bacteriophages in Food Fermentations: New Frontiers in a Continuous Arms Race. *Annu Rev Food Sci Technol* **4**:347–368.
21. Labrie SJ, Samson JE, Moineau S. 2010. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**:317–327.
22. Kelleher P, Mahony J, Bottacini F, Lugli GA, Ventura M, van Sinderen D. 2019. The *Lactococcus lactis* pan-Plasmidome. *Front Microbiol* **10**:1-20.
23. Mahony J, Murphy J, van Sinderen D. 2012. Lactococcal 936-type phages and dairy fermentation problems: From detection to evolution and prevention. *Front Microbiol* **3**:1–9.
24. Chmielewska-Jeznach M, Bardowski JK, Szczepankowska AK. 2020. *Lactococcus Ceduovirus* phages isolated from industrial dairy plants—from physiological to genomic analyses. *Viruses* **12**:280.
25. Szczepańska AK, Hejnowicz MS, Kołakowski P, Bardowski J. 2007. Biodiversity of *Lactococcus lactis* bacteriophages in Polish dairy environment. *Acta Biochim Pol* **54**:151–158.
26. Murphy J, Royer B, Mahony J, Hoyles L, Heller K, Neve H, Bonestroo M, Nauta A, van Sinderen D. 2013. Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type cheese-producing plants. *J Dairy Sci* **96**:4945–4957.
27. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C. 2013. Investigation of the relationship between lactococcal host cell wall polysaccharide

genotype and 936 phage receptor binding protein phylogeny. *Appl Environ Microbiol* **79**:4385–4392.

28. Deveau H, Labrie SJ, Chopin MC, Moineau S. 2006. Biodiversity and classification of lactococcal phages. *Appl Environ Microbiol* **72**:4338–4346.
29. Oliveira J, Mahony J, Hanemaaijer L, Kouwen TRHM, van Sinderen D. 2018. Biodiversity of bacteriophages infecting *Lactococcus lactis* starter cultures. *J Dairy Sci* **101**:96–105.
30. Mahony J, Frantzen C, Vinogradov E, Sadovskaya I, Theodorou I, Kelleher P, Chapot-Chartier M-P, Cambillau C, Holo H, van Sinderen D. 2020. The CWPS Rubik's cube: linking diversity of cell wall polysaccharide structures with the encoded biosynthetic machinery of selected *Lactococcus lactis* strains. *Mol Microbiol* **114**:582–596.
31. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, van Sinderen D. 2014. Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. *mBio* **5**:e00880-14.
32. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, Van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. *Sci Rep* **6**:21345.
33. Sadovskaya I, Vinogradov E, Courtin P, Armalyte J, Meyrand M, Giaouris E, Palussière S, Furlan S, Péchoux C, Ainsworth S, Mahony J, van Sinderen D, Kulakauskas S, Guérardel Y, Chapot-Chartier M-

P. 2017. Another brick in the wall: a rhamnan polysaccharide trapped inside peptidoglycan of *Lactococcus lactis*. *mBio* **8**: e01303-17.

34. Chapot-Chartier MP, Vinogradov E, Sadovskaya I, Andre G, Mistou MY, Trieu-Cuot P, Furlan S, Bidnenko E, Courtin P, Péchoux C, Hols P, Dufre YF, Kulakauskas S. 2010. Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. *J Biol Chem* **285**:10464-10471.
35. Parlindungan E, Sadovskaya I, Vinogradov E, Lugli GA, Ventura M, van Sinderen D, Mahony J. 2024. Novel cell wall polysaccharide genotypes and structures of lactococcal strains isolated from milk and fermented foods. *Int J Food Microbiol* **424**:110840.
36. Romero DA, Magill D, Millen A, Horvath P, Fremaux C. 2020. Dairy lactococcal and streptococcal phage–host interactions: an industrial perspective in an evolving phage landscape. *FEMS Microbiol Rev* **44**:909–932.
37. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura M. 2021. METAnnotatorX2: a comprehensive tool for deep and shallow metagenomic data set analyses. *mSystems* **6**:e00583-21.
38. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**:455-477.

39. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.

40. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240.

41. Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**:9077–9096.

42. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**:722–736.

43. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**:e1005595.

44. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B, Wyres KL, Holt KE. 2021. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* **22**:266.

45. Wick RR, Holt KE. 2022. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* **18**:e1009802.

46. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. 2016. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. *FEMS Microbiol Lett* **363**:fnw049.

47. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.

48. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**:366–368.

49. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240.

50. Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**:9077–9096.

51. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. 2018. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res* **46**:W282–W288.

52. Walsh CJ, Srinivas M, Stinear TP, van Sinderen D, Cotter PD, Kenny JG. 2024. GROND: a quality-checked and publicly available database of full-length 16S-ITS-23S rRNA operon sequences. *Microb Genom* **10**:001255.

53. Meier-Kolthoff JP, Göker M. 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat Commun* **10**:2182.

54. Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, Stothard P, Gautam V. 2023. PHASTEST: faster than PHASTER, better than PHAST. *Nucleic Acids Res* **51**:W443–W450.

55. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, Jackson SA. 2022. PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res* **50**:W541–W550.

56. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, Bernheim A. 2022. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**:2561.

57. Néron B, Denise R, Coluzzi C, Touchon M, Rocha EPC, Abby SS. 2023. MacSyFinder v2: improved modelling and search engine to identify molecular systems in genomes. *Peer Community J* **3**:e28.

58. Tesson F, Planel R, Egorov A, Georjon H, Vaysset H, Brancotte B, Néron B, Mordret E, Atkinson G, Bernheim A, Cury J. 2024. A comprehensive resource for exploring antiphage defense: DefenseFinder webservice, wiki and databases. *Peer Community J* **4**:e91.

59. Grafakou A, Mosterd C, Beck MH, Kelleher P, McDonnell B, de Waal PP, van Rijswijck IMH, van Peij NNME, Cambillau C, Mahony J, van Sinderen D. 2024. Discovery of antiphage systems in the lactococcal plasmidome. *Nucleic Acids Res* **52**:9760-9776.

60. Enright AJ, Dongen S Van, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**:1575–1584.

61. Saeed Ai, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M. 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**:374–378.

62. Gilchrist CLM, Chooi Y-H. 2021. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**:2473–2475.

63. White K, Eraclio G, McDonnell B, Bottacini F, Lugli GA, Ventura M, Volontè F, Dal Bello F, Mahony J, van Sinderen D. 2024. A multifaceted investigation of lactococcal strain diversity in undefined mesophilic starter cultures. *Appl Environ Microbiol* **90**:e02152-23.

64. Casey E, McDonnell B, White K, Stamou P, Crowley T, O'Neill I, Lavelle K, Hayes S, Lugli GA, Arboleya S, James K, Ventura M, Martinez I, Gueimonde M, Dal Bello F, Nally K, Mahony J, van Sinderen D. 2022. Needle in a Whey-Stack: PhRACS as a Discovery Tool for Unknown Phage-Host Combinations. *mBio* **13**:e03334-21.

65. Mahony J, Bottacini F, van Sinderen D. 2023. Towards the diversification of lactococcal starter and non-starter species in mesophilic dairy culture systems. *Microb Biotechnol*. **16**:1745-1754.

66. Kelleher P, Mahony J, Schweinlin K, Neve H, Franz CM, van Sinderen D. 2018. Assessing the functionality and genetic diversity of lactococcal prophages. *Int J Food Microbiol* **272**:29-40.

67. Meucci A, Zago M, Rossetti L, Fornasari ME, Bonvini B, Tidona F, Povolo M, Contarini G, Carminati D, Giraffa G. 2015. *Lactococcus*

hircilactis sp. nov. and *Lactococcus laudensis* sp. nov., isolated from milk. *Int J Syst Evol Microbiol* **65**:2091–2096.

68. Xie H, Yang C, Sun Y, Igarashi Y, Jin T, Luo F. 2020. PacBio long reads improve metagenomic assemblies, gene catalogs, and genome binning. *Front Genet* **11**:1-8.
69. Vassallo CN, Doering CR, Littlehale ML, Teodoro GIC, Laub MT. 2022. A functional selection reveals previously undetected anti-phage defence systems in the *E. coli* pangenome. *Nat Microbiol* **7**:1568–1579.
70. Chopin MC, Chopin A, Bidnenko E. 2005. Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol* **8**:473–479.
71. Sybesma W, Hugenholtz J, De Vos WM, Smid EJ. 2006. Safe use of genetically modified lactic acid bacteria in food. Bridging the gap between consumers, green groups, and industry. *Electron J Biotechnol* **9**:242-448.
72. Tidona F, Meucci A, Povolo M, Pelizzola V, Zago M, Contarini G, Carminati D, Giraffa G. 2018. Applicability of *Lactococcus hircilactis* and *Lactococcus laudensis* as dairy cultures. *Int J Food Microbiol* **271**:1–7.

SUPPLEMENTARY MATERIALS

Supplementary Table S1. Bacteria genomes/strains used in this study.

Lactococcal strains*	CWPS type	Genbank Accession
158	A	CP015894
267	A	CP032058
275	A	CP015897
B26	A	CP138309
C10	A	CP015898
CV56	A	CP002365
F44	A	CP024954
G423	A	CP024958
UC063	A	CP015905
UC109	A	CP015907
UC509.9	A	CP003157
UL8	A	CP015908
WM1	A	CP032500
SC34	B	Unpublished
223	B	CP031926
229	B	CP015896
FM03	B	CP020604
IL1403	B	AE005176
IL6288	B	CP033606
CP001834	B	CP001834
NCDO2118	B	CP009054
UC08	B	CP015903
UC11	B	CP015904
UC77	B	CP015906
KF147	B	CP001834
17M1	C_1	CP092748
4B0	C_1	CP032526
JM1	C_1	CP015899
KF140	C_1	CP070382
MG1363	C_1	AM406671
NZ9000	C_1	CP002094
ME1	C_1	Unpublished
SC74	C_1	Unpublished
MB2	C_1	Unpublished
MG3	C_1	Unpublished
MF1	C_1	Unpublished
MF3	C_1	Unpublished
MA1	C_1	Unpublished
MH2	C_1	Not sequenced
3107	C_2	CP031538
C4	C_2	AP018499
EPSC	C_2	AP024222
FM-YL11	C_2	CP071729
FM-YL12	C_2	CP071728
G3-2	C_2	AP024210
GR0507	C_2	CP070856
JM2	C_2	CP015900
JM4	C_2	CP015909
KCKM 0438	C_2	CP133787
KW2	C_2	CP004884
LJL7m20	C_2	CP104388
NCDO700	C_2	CP069179
ZZ-2	C_2	CP129526
L8	C_2	AP025700

MH5	C_2	Not sequenced
MA6	C_2	Not sequenced
MH4	C_2	Unpublished
MA5	C_2	Unpublished
MB5	C_2	Unpublished
MB6	C_2	Unpublished
MF5	C_2	Unpublished
ME5	C_2	Unpublished
MC5	C_2	Unpublished
MD5	C_2	Unpublished
MC6	C_2	Unpublished
ME6	C_2	Unpublished
JM3	C_3	CP015901
SK11	C_3	CP000425
MC7	C_3	Not sequenced
MF8	C_3	Unpublished
MA9	C_3	Unpublished
BA	C_3	Unpublished
1196	C_4	CP032148
LB7	C_4	CP117409
P-1	C_4	CP118738
UC047	C_4	CP068658
W34	C_4	CP032430
MF12	C_4	Unpublished
MA10	C_4	Unpublished
MC12	C_4	Unpublished
MG12	C_4	Unpublished
DRD-23	C_5	NZ_CP133257
DRD-8	C_5	NZ_CP133258
IO-1	C_5	AP012281
UC06	C_5	CP015902
UD4	C_5	Unpublished
UG1	C_5	Unpublished
A12	C_6	LT599049
A76	C_6	CP003132
NGA6	C_6	Unpublished
NGF5	C_6	Unpublished
AO45	C_6	Unpublished
14B4	C_7	CP028160
AI06	C_7	CP009472
SO	C_7	CP010050
Peco9D	C_7	JAHIBV000000000
Tome1F	C_7	JAHIBU000000000
RMN8G	C_9	JAHIBS000000000
Tempeh6L	C_{10}	JAHIBT000000000
Tolla67	C_{11}	JAHIBR000000000
184	D	CP015895
KLDS 4.0325	D	CP006766
UD2	*na*	Unpublished
UC8	*na*	Unpublished
DSM 28961	*na*	NZ_JACBNY000000000
T2D6	*na*	NZ_JAZHKZ000000000
T2G3	*na*	NZ_JAZHKX000000000
T2A8	*na*	NZ_JAZHKW000000000
WiKim0068†	*na*	CP023392
Lr_19_17†	*na*	CP047614

*Species *cremoris* strains are indicated in normal face text, species *lactis* strains in bold face text, and species *laudensis* in italicised text.
†Strains belong to the species *Lactococcus raffinolactis*.

Supplementary Table S2. Sequencing and assembly methods attributed to each sequenced genome.

Strain	Sequencing method	Assembly method
L. cremoris ME1	Illumina/Nanopore	TryCycler
L. cremoris MG3	Illumina/Nanopore	TryCycler
L. cremoris MF5	Illumina/Nanopore	TryCycler
L. cremoris ME5	Illumina/Nanopore	TryCycler
L. cremoris MF8	Illumina/Nanopore	Unicyler
L. cremoris MA9	Illumina/Nanopore	TryCycler
L. cremoris MC12	Illumina/Nanopore	TryCycler
L. cremoris MG12	Illumina/PacBio	CANU
L. cremoris UG1	Illumina/Nanopore	TryCycler
L. cremoris NGA6	Illumina/PacBio	CANU
L. lactis SC34	Illumina	SPAdes
L. cremoris SC74	Illumina	SPAdes
L. cremoris MB2	Illumina	SPAdes
L. cremoris MF1	Illumina	SPAdes
L. cremoris MH4	Illumina	SPAdes
L. cremoris MA5	Illumina	SPAdes
L. cremoris MB5	Illumina	SPAdes
L. cremoris MB6	Illumina	SPAdes
L. cremoris MC5	Illumina	SPAdes
L. cremoris MD5	Illumina	SPAdes
L. cremoris MC6	Illumina	SPAdes
L. cremoris ME6	Illumina	SPAdes
L. cremoris BA	Illumina	SPAdes
L. cremoris MF12	Illumina	SPAdes
L. cremoris MA10	Illumina	SPAdes
L. cremoris UD4	Illumina	SPAdes
L. cremoris NGF5	Illumina	SPAdes
L. cremoris AO45	Illumina	SPAdes

Supplementary Table S3. Breakdown of PHASTEST results.

Strain	No. prophage regions detected by PHASTEST			No. prophage regions confirmed manually	
	Intact	Incomplete	Questionable	Intact	Incomplete
L. lactis SC34	6	-	-	2	4
L. cremoris ME1	4	1	-	1	3
L. cremoris SC74	2	1	1	1	2
L. cremoris MB2	3	2	-	1	2
L. cremoris MG3	4	-	1	3	2
L. cremoris MF1	1	2	-	1	2
L. cremoris MH4	3	1	-	1	3
L. cremoris MA5	1	3	2	1	5
L. cremoris MB5	2	2	1	1	4
L. cremoris MB6	4	-	-	-	4
L. cremoris MF5	3	1	-	1	3
L. cremoris ME5	3	1	1	2	3
L. cremoris MC5	2	4	-	1	3
L. cremoris MD5	1	3	1	1	4
L. cremoris MC6	1	1	-	1	1
L. cremoris ME6	2	3	2	-	5
L. cremoris MF8	2	1	-	2	1
L. cremoris MA9	2	2	-	2	2
L. cremoris BA	-	2	2	-	3
L. cremoris MF12	1	3	-	1	3
L. cremoris MA10	1	4	1	-	5
L. cremoris MC12	2	1	1	1	3
L. cremoris MG12	2	2	1	1	4
L. cremoris UD4	-	1	1	-	2
L. cremoris UG1	1	2	1	-	4
L. cremoris NGA6	2	1	-	1	2
L. cremoris NGF5	1	1	-	-	2
L. cremoris AO45	2	1	1	-	2
L. laudensis UD2	4	1	-	2	3
L. laudensis UC8	3	3	-	1	5

Supplementary Table S4. Characteristics and features of *L. laudensis* UD2 plasmids.

Plasmid name	Size (bp)	GC (%)	Genes	Notable technological relevance
pUD2A	45361	35.25	45	*Lac* operon, mobilization genes, LysM (Lysin motif) peptidoglycan-binding domain-containing protein
pUD2B	24118	34.05	27	EPS genes (*epsF* and *epsJ*) interspaced by transposases, Type I RM orphan specificity subunit, mobilization genes

Chapter 6. General discussion and future perspectives

Despite extensive efforts and decades of research, phage infection of lactic acid bacteria used in starter cultures persists as a significant challenge to the dairy fermentation industry. It is generally accepted that complete eradication of phages in fermentation facilities is unlikely, if not impossible. However, through rigorous monitoring of both bacterial and phage populations in dairy fermentations it may be possible to mitigate the risk of fermentation failures/inconsistencies. This thesis describes efforts to improve current understanding of interactions that occur between prevalent lactococcal phages and their cognate hosts in dairy fermentation factories utilizing undefined mesophilic starter cultures. This was achieved through a combination of culture-dependent and -independent approaches which allowed the assessment of both the diversity of lactococcal strains encompassing undefined mesophilic starter cultures as well as the prevalent phage populations present in various dairy fermentation facilities.

Therefore, a multifaceted investigation (employing both culture-dependent and -independent approaches) of the lactococcal strain-level diversity in two undefined mesophilic starter cultures, named M1 and M2, was undertaken (Chapter 2). Consistent with previous studies, *L. cremoris* strains were identified as the dominant mesophilic LAB species present in both starter cultures. Compositional analysis of the M1 and M2 starter cultures, based on the diversity and abundance of *cwps* types, yielded different outcomes based on culture-based and culture-independent (both metagenomic reads mapping and qPCR) compositional profiles. This is possibly due to the cultivation conditions or the presence of viable but non-culturable (VBNC) cells. Ultimately, these results demonstrate the importance of employing both

culture-dependent and -independent methods when defining complex microbial community systems to strain level in order to establish a holistic view of the microbiota. To ensure consistent product quality, it is essential that the starter cultures are robust to withstand a phage challenge and consequently perform reliably (i.e. consistently produce the desired high-quality fermentation products), which can be achieved through diligent microbial monitoring efforts. While the traditional approach to achieve this is by means of milk acidification assessment in fermentation factories, a rapid approach to tracking culture compositional changes is described in Chapter 2, in which a qPCR assay capable of discerning nine industrially relevant lactococcal *cwps* genotypes was established. As each of the corresponding nine primer pairs is specific to a single strain or genetic lineage, they have the potential to be more widely applicable to other *L. lactis* and *L. cremoris*-dominant cultures or environments compared to other currently described lactococcal qPCR assays. Therefore, this qPCR assay may be useful as a classification and risk evaluation tool from a perspective of lactococcal strain-level diversity (as based on *cwps* profiles), which is essential when implementing effective starter culture rotation strategies. The lactococcal *cwps* gene cluster is constantly evolving, likely as a result of constant phage pressure. Recently the diversity of defined *cwps* genotypes has been expanded to include several additional C-subtypes, therefore the herein described qPCR assay is expected to require expansion in the future.

Despite the benefits of culture-independent approaches including the above-described qPCR system, the importance of culture-based analysis should not be diminished as it is necessary to isolate and characterise

individual strains to evaluate the role and functionality of both dominant and minor components of starter cultures. As described in Chapter 5, strains isolated from M1 and M2 were technologically (acidification, proteolysis, and phage sensitivity) and genomically analysed. This analysis provided further insight into the genetic diversity of several strain's CWPS biosynthesis gene clusters, plasmid content, and phage defense systems. One example of this is the observed genetic diversity among current C_5-subtype members, suggesting the need to reclassify this subtype (likely by defining a novel C-subtype). However, this would require validation through CWPS compositional and structural analyses of these strains. Knowledge of the genetic diversity of strains isolated from these starter cultures will assist in identifying strains that may act as adjunct cultures or be applicable in the generation of non-GMO, phage-resistant derivatives of certain strains either through the generation of bacteriophage-insensitive mutants (BIMs) or generated through natural conjugation. In addition, the genomic diversity of *L. laudensis* strains that were isolated through culture-based analysis but not identified through metagenomic analysis, was assessed. These *L. laudensis* isolates represent strains of an emerging dairy-associated lactococcal species, for which currently just eight genome sequences are available in public repositories (the majority of which have only been released this year). Analysis of the *cwps* gene clusters of the sequenced *L. laudensis* strains revealed significant genetic diversity. Future studies pertaining to this genetic diversity and likely structural diversity of the corresponding CWPS of these strains facilitate insights into their role as a potential phage receptor.

In addition to assessing the diversity of bacterial communities that make up undefined starter cultures employed in mesophilic fermentation facilities, it is also essential to understand the extent of the phage problem within factories. Phage diversity in dairy fermentation factories employing mesophilic undefined starter cultures was first assessed (in Chapter 3) through phageome analysis of 27 whey samples. Members of the strictly lytic *Skunavirus* genus were the principal phage group present in almost all assessed whey samples. The diversity of identified phage-encoded anti-receptors or RBPs, which are responsible for the initial binding and recognition of phages to a cognate host, was thoroughly scrutinised during phageome analysis. This ultimately led to the refinement and expansion of the *Skunavirus* RBP grouping system, including the identification of several subgroups (particularly for the group I RBPs). Furthermore, through the generation of representative ^{His}GFP-RBPs it was possible to determine the binding specificity of many of the RBP groups/subgroups through binding assays against lactococcal strains representing 15 distinct CWPSs (A, B, C_1-C_7, C_9-C_{11}, and D). However, the binding affinity of certain RBP groups (groups IV, VII, and IX) remains unclear as representative ^{His}GFP-RBPs of these groups did not appear to bind to any of the CWPS-type strains tested with particularly high affinity. This may indicate that these RBPs recognise CWPS(s) that were not included in this study. Consequently, it is anticipated that as more lactococcal phages and strains are isolated the RBP grouping system (and links between groups and host CWPS structures) will need to be further refined and/or expanded. Despite the knowledge gaps that remain, it is now possible to better predict potential hosts of skunaviruses through an analysis of RBP sequences within

a population. Additionally, the recently described PhRACS approach was successfully implemented to exploit this specific interaction between *Skunavirus* RBPs and host cell surface receptors to facilitate the isolation of phage-host combinations from whey samples and a mesophilic undefined starter culture. This demonstrates the ability of PhRACS to bridge the gap between virome and culture-based analysis. Overall, it is proposed that by combining *cwps* analysis of starter strains/cultures and RBP analysis of phageomes occurring within complex microbial communities, it is now possible to improve starter culture blends and rotation strategies. Beyond the practical aspects in the case lactococci, this may have implications for understanding phage-host interactions of tailed phages that infect Gram-positive ovococcoid bacteria.

In addition to virome sequencing, 26 skunaviruses were isolated from various whey samples and their host range and genome characteristics evaluated (Chapter 4). Through this analysis it was possible to validate RBP phylogeny-based host predictions and to define the associated CWPS specificity of the updated RBP grouping system initially described in Chapter 3. In addition, with access to complete phage genomes it was possible to investigate predicted CBMs contained within apparently auxiliary binding proteins of assessed skunaviruses. This analysis led to the identification of novel CBMs in the phage's NPSs, (evolved) Dits, and major tail extension proteins. Of note, TEM analysis and AlphaFold modelling revealing a unique NPS with an associated collar structure and whiskers ending with small, bobble-like structures on one *Skunavirus* isolate. Furthermore, identification of novel CBMs among this relatively small pool of phages indicates a need to

revisit the current classification systems of the ubiquitous CBMs encoded within the auxiliary binding proteins of skunaviruses. Moreover, a thorough investigation into whether these proteins always possess the same CWPS specificity as their corresponding RBP (which has been previously suggested) or if they possibly may aid in expanding a phage's host range, demonstrating that there may be further evolutionary benefits to encoding these CBMs.

Overall, the work described in this thesis significantly expands current knowledge of phage-host interactions between skunaviruses and their lactococcal hosts. In addition, it incorporates both culture-based and -independent analyses of bacteria and phage populations in dairy fermentation facilities, thereby generating a more comprehensive picture of the microbial composition of these complex communities. However, there remain areas that require further investigation particularly regarding CBM diversity in phages and CWPS diversity of lactococcal strains (including the emerging dairy-associated species *L. laudensis, L. raffinolactis,* etc). Despite the pervasive phage problem in food fermentations, it may now be possible to outmanoeuvre infecting phage populations through more guided, rational and preventative starter culture blends and rotation practices.

Acknowledgements

First and foremost, I would like to wholeheartedly thank my supervisors Professors Douwe van Sinderen and Jennifer Mahony. Thank you for all you have done for me over the past four (and a bit) years. I genuinely couldn't have asked for better supervisors, you both are the most encouraging, patient, and supportive mentors. Your knowledge, expertise, and guidance has been invaluable. I particularly want to thank Jenn for always being there for me, answering all my questions, and calming my spells of panic (apologies for always showing up at your office but thank you for always letting me in). I will forever be grateful to both of you for this opportunity; it has been an absolute pleasure.

Secondly, I want to thank everyone at Sacco Srl: my enterprise partner mentor Fabio Dal Bello, as well as Giovanni Eraclio and Federica Vólante. For the constant support and communication and plethora of samples. I greatly appreciate everything you have done for me.

I am ever grateful for and acknowledge the funding awarded to me by the Irish Research Council and Sacco Srl under the Enterprise Partner Scheme that made it possible for me to pursue this PhD, through the School of Microbiology and the APC Microbiome research institute.

I would also like to recognise and thank all the collaborators that have assisted me throughout my PhD. Thank you to Christian Cambillau for the time and effort you put into generating structural predictions and openness to questions and discussion. And to Tadhg Crowley, who allowed me to run (and rerun) samples through the cell sorter whenever I needed to. To Francesca

Bottacini for her help with the bioinformatics aspects of this PhD, particularly when I was starting out. A special thanks to everyone that has contributed to the server command ‘bible’ over the years (Phillip Kelleher, Keith Coughlan, and Brian McDonnell in particular). To the GenProbio team for all the sequencing data they have provided me. I would also like to thank all the staff of the School of Microbiology, in particular John O’Callaghan especially in regard to getting me access to working qPCR machines, fluorescent plate reader, etc. when I desperately needed it. And additionally, thanks (for everything) to Catherine O’Leary.

To all the friends and colleagues, I’ve had between labs 335 to 340 over the course of the last four years. Thank you for making the lab such as wonderful place to work. And thank you (and sorry) for listening to me chat and rant constantly. Especially, to the people who were unlucky enough to be stuck around me in lab 339/340: Elvina (for keeping me company through my first year), Sofia and Katherine (for the coffee and discussions), Zoe (for entertaining my bouts of chatter and the numerous tips and media I “borrowed” from you, I still owe you), and to Jun and Andriana (who started and are completing their own PhDs alongside me). Thanks also to Ortensia and Lisa for their willingness to entertain my love for the odd quiet pint. And to Keith for his patience (that I definitely tested) when answering any and all questions regarding bioinformatics/the server. A special thank you to Brian, who introduced me to the lab at the start of the PhD and has been a constant and invaluable help in all aspects of my time here. Overall, I am overwhelmingly grateful to everyone that has been in or passed through the lab over these last four years, it has been an absolute joy meeting and working with you all.

To my friends outside of the lab, thank you for your friendship, understanding, and willingness to travel thousands of miles to come visit me. Particularly Veronica, Sara & Katie, and Carolynn, for listening to me rant about science and excusing my bouts of distance when things get stressful (and for checking in when I need it), especially these last few months. Your friendship is so important to me and greatly appreciated.

Finally, to my family. To my Uncle Paudie, Aunt Nuala, and cousins Darragh and Ella, for opening your door to me when I first arrived in Ireland and allowing me to show up unannounced whenever I'm in Clare (An Clár Abú). Your love and care have been a large part of what makes Ireland feel like home. To my dad, Tomás (Mossy), you have supported me throughout the entirety of my education, I will never be able to repay you for that but hope you know your constant love and support means the world to me. To my brother Thomas, you may not know what my PhD is on, but I appreciate your willingness to let me explain it to you every time we speak; thank you for always visiting dad and me, I can't wait to see you again soon for your wedding. I couldn't have done this without you both. Lastly, and most of all, to my mom (Liz) and brother Brian whose memory and boundless love has shaped me and carried me to where I am today. I hope you both would have been proud; this thesis and everything I might accomplish is always because of and dedicated to you both. Inár gcroí go deo.