CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Comparative and functional genomic analysis of dairy lactococci
Authors	Kelleher, Philip
Publication date	2017
Original Citation	Kelleher, P. 2017. Comparative and functional genomic analysis of dairy lactococci. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2017, Philip Kelleher. - http://creativecommons.org/licenses/by-nc-nd/3.0/
Download date	2024-03-28 22:11:32
Item downloaded from	https://hdl.handle.net/10468/4522

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Comparative and Functional Genomic Analysis of Dairy Lactococci

A thesis presented to the National University of Ireland, Cork

by

Philip Kelleher

BSc Pharmaceutical Biotechnology, MSc Bioinformatics & Systems Biology

For the degree of

PhD in Microbiology

School of Microbiology,

National University of Ireland, Cork

January 2017

Supervisor: Prof. Douwe van Sinderen

Head of School: Prof. Gerald Fitzgerald

General Table of Contents

Abbreviations …….……. 4

Abstract …….…….. 6

Chapter I: General introduction …….. 12

Chapter II: Performance and flavour-based characterisation of lactococcal starter cultures …….……... 93

Chapter III: Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation …….. 130

Chapter IV: Comparative genomic analysis of the *Lactococcus lactis* plasmidome and assessment of its technological properties …….…….. 183

Chapter V: Base modification analysis of *Lactococcus lactis* strains and their corresponding restriction-modification systems ……..… 233

Chapter VI: Assessing functionality and genetic diversity of lactococcal prophages …….……... 273

Chapter VII: General discussion …….……. 321

Appendix A: Large-scale cheese fermentation trial results …….…….. 334

Acknowledgements …….…… 339

Declaration

I hereby declare that the content of this thesis is the result of my own work and has not been submitted for another degree, either at University College Cork or elsewhere

Signed:

_____

Philip Kelleher

Date: 05/01/2017

Abbreviations

AA	=	Amino acid
Abi	=	Abortive infection system
ADI	=	Arginine deanimase pathway
AMC	=	7-amino-4-methyl coumarin
ANI	=	Average nucleotide identity
ARD	=	Amino-proximal recognition domain
BLAST	=	Basic local alignment search tool
BPP	=	Baseplate protein
Cas	=	CRISPR-associated proteins
CDS	=	Coding sequence
CFU	=	Colony forming unit
COG	=	Clusters of Orthologous Groups
CPS	=	Capsular exopolysaccharide
CRD	=	Carboxy-proximal recognition domain
CRISPR	=	Clustered Regularly Interspaced Short Palindromic Repeats
Dit	=	Distal tail protein
dso	=	Double-stranded origin of replication
EPS	=	Exopolysaccharide
FDA	=	Food and Drug Administration
GRAS	=	Generally regarded as safe
HCL	=	Hierarchial clustering
HsdM	=	Methylase subunit
HsdR	=	Restriction endonuclease subunit
HsdS	=	Specificity subunit
IS	=	Insertion sequence elements
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
LAB	=	Lactic acid bacteria
LDH	=	Lactate dehydrogenase
MCL	=	Markov Clustering Algorithm
NCBI	=	National Centre for Biotechnology Information
NGS	=	Next generation sequencing
NICE	=	Nisin-inducible controlled gene expression

NT	=	Nucleotide
ORF	=	Open reading frame
Ori	=	Origin of replication
OriT	=	Origin of transfer
PFGE	=	Pulse field gel electrophoresis
PHAST	=	Phage Search Tool
PTS	=	Phosphotransferase system
qPCR	=	Quantitative PCR
RBP	=	Receptor binding protein
RCR	=	Rolling-circle replication
Rep	=	Replication protein
R-M	=	Restriction modification
RSM	=	Reconstituted skimmed milk
SBS	=	Sequencing-by-synthesis
Sie	=	Superinfection exclusion system
SMRT	=	Single molecule real time
Tal	=	Tail-associated lysin
TETRA	=	Tetranucleotide frequency correlation coefficients
TMP	=	Tail tape measure protein
TRD	=	Target recognition domain
ZMW	=	Zero-mode waveguide

Abstract

Lactococcus lactis has been exploited for thousands of years for the production of fermented dairy products, and from an economic perspective has become one of the most valuable bacteria. *L. lactis* is used predominantly as a starter culture for the production of various hard and soft cheeses. The constant threat of (bacterio)phage infection combined with consumer-driven diversification of product ranges have created an increased need to improve technologies for the rational selection of novel starter culture blends. Whole genome sequencing, spurred on by recent advances in next-generation sequencing (NGS) platforms, is a promising approach to facilitate the rapid identification and selection of such strains based on gene-trait matching. In this thesis the most up-to-date sequencing methodologies were applied to sequence sixteen *L. lactis* isolates to facilitate an in depth comparative and functional genomic analysis of the taxon with particular emphasis placed on dairy traits.

A selection of lactococcal strains were first functionally characterised based on their phenotypic traits and assessed for industrial robustness and flavour formation using a functional approach. The behaviour of the strains under simulated cheese production conditions was monitored, and employed to assess their temperature-induced autolytic properties. This analysis was followed by the determination of activity profiles of enzymes related to key flavour formation pathways, in order to explore proteolytic and lipolytic abilities of each strain. Comparative analysis between our selection of *L. lactis* strains and of four starter cultures currently employed in the Irish dairy industry for the production of half-fat

Cheddar cheese facilitated the identification of potentially novel starter cultures. In total twenty strains were assessed for the activity of twelve separate enzymes related to cheese production. From these strains, eleven were selected for whole genome shotgun sequencing to further investigate their genetic composition, and to explore the possibility of linking genotype to phenotype (also called gene-trait matching).

The genomes of sixteen *L. lactis* subsp. *lactis* and *L. lactis* subsp. *cremoris* dairy strains were sequenced to completion, doubling the number of fully sequenced *L. lactis* genomes currently available from the public National Centre for Biotechnology Information (NCBI) data base. These newly sequenced genomes along with available whole genome sequences were used to perform the largest comparative and functional genomic study to date on the *L. lactis* taxon. Their chromosomal features were assessed with particular emphasis on discerning the *L. lactis* subspecies division, evolution and niche adaptation. This analysis clearly identified a phylogenetic division between subspecies *lactis* and *cremoris* strains, which was further corroborated by hierarchical clustering based on carbohydrate and amino acid metabolic pathways. The pan and core genomes of *L. lactis* were shown to be comprised of 5906 and 1129 genes, respectively. Both were found to be in a closed state, indicating that the representative data sets employed for this analysis are sufficient to fully describe the genetic diversity of the taxon. Niche adaptation appears to play a significant role in governing the genetic content of each *L. lactis* subspecies, while (differential) genome decay and redundancy in the dairy niche was also highlighted. The description of chromosomal adaptations in *L. lactis* has not received the same level of attention compared to plasmid-mediated characteristics due to the perceived biotechnological importance of the latter. Our comparative analysis revealed that the division between plasmid- and chromosome-based traits is

less clear as multiple integration events within the lactococcal chromosome suggests a more fluid genome than previously thought.

The complete genome sequence analysis of sixteen *L. lactis* strains revealed the presence of a total of sixty-seven plasmids, including two megaplasmids representing the first megaplasmids identified in lactococcal strains. Megaplasmids are large autonomous self-replicating extrachromosomal genetic elements greater than 100 Kb. While megaplasmids are not essential for the growth of their host, they may encode additional metabolic capabilities. Comparative genome analysis of these sequences combined with those of publicly available plasmids (eighty one publicly available) allowed the definition of the lactococcal plasmidome based on one hundred and forty eight complete plasmid sequences, and facilitated an investigation into technologically important plasmid-encoded traits. In contrast to the lactococcal chromosomes, the lactococcal pan-plasmidome was found to be in a fluid state implying that continued sequencing efforts will likely expand the diversity of this data set and lead to an increase in the identification of novel plasmid features. In the present study, lactococcal gut adhesion was also investigated identifying potential gut adhesion factors within the lactococcal plasmidome. It is envisioned that this may provide further insights for the application of *L. lactis* as a vector for vaccine and biomolecule delivery. Finally, the frequency of plasmid-encoded phage resistance mechanisms was assessed with particular emphasis on abortive infection (Abi) systems. In total fourteen plasmid-encoded Abi systems were identified, while further analysis also identified frequent occurrences of these systems within the lactococcal chromosomes.

Single molecule real time sequencing (SMRT) was used for the elucidation of finished quality genome sequences in this study, which is the first and only

sequencing technology to allow concomitant detection of base modifications with primary sequence analysis. Here, SMRT technology was applied to determine the methylome of sixteen *L. lactis* strains, which revealed fifty two methylation motifs consisting of N^6-methyladenine base modifications. Five of these motifs were validated as they prevented site-specific cleavage by commercially available restriction enzymes. The sixteen strains were predicted to encode a number of unique Type I, II, III and IV restriction-modification (R-M) systems. These systems are of particular interest in lactococcal strains in terms of the associated bacteriophage-resistance. Sequence analysis and annotation revealed the presence of a conserved type I R-M shufflon system in each of the two identified megaplasmids, consisting of multiple *hsdS* genes arranged around a recombinase gene, thus allowing for the generation of multiple specificity targets. The presumed genetic rearrangement activity of this system was corroborated by the presence of a number of associated type I methylase motifs containing N^6-methyladenine base modifications, while the predicted shuffling patterns were confirmed by qPCR and analysis of the raw sequencing reads. It is envisioned that these systems provide the host with a mechanism of adaptive phage defence in response to infection.

The presence of prophages in lactococcal genomes is widely reported, however only a small number of studies pertaining to the stability of the prophages in the genomes have been performed. In the concluding part of this study, the genomes of thirty lactococcal strains were explored for the presence of potentially intact prophages, so as to assess their genomic diversity and the associated risk (or benefit) of harbouring such prophages. In total we identified fifty nine (potentially) intact prophages, of which most were shown to belong to the so-called P335 phage group, while various (presumed) phage remnants (106) bear similarity to members of

the 936 phage group. The P335 phage group was recently shown to encompass four distinct genetic lineages. In this study a fifth additional lineage was identified, thus expanding the diversity of this industrially significant phage group. Furthermore, the genomic predictions partnered with chemical induction trials revealed that just four strains consistently produced intact phage particles, thus indicating a low risk associated with prophage induction in the fermentation setting. The analysis also revealed the widespread presence of phage-resistance systems encoded by lactococcal prophages including seventeen superinfection exclusion (Sie) systems and twelve phage-encoded Abi systems, highlighting the potential benefits for host fitness. It was found that prophages may represent a relatively low direct risk to cheese production processes but their potential to expedite the evolution of virulent phages and the fitness of the host are key features that should be considered when selecting starter cultures.

The research presented in this thesis has significantly advanced our understanding of *L. lactis* in several ways. Firstly, it has significantly expanded the number of complete lactococcal genomes available for comparative and functional genome analyses, while it has thoroughly scrutinized chromosomal versus plasmid diversity, including the elucidation of both the pan/core genome and the pan-plasmidome. Secondly, the identification of the first lactococcal megaplasmids and undertaking of the first methylome analysis of the *L. lactis* taxon has greatly increased our understanding of host-encoded phage defence systems. The frequency of lactococcal prophages within the chromosomes of *L. lactis* has been thoroughly investigated concomitantly with the risk of prophage excision.

Finally, from the inception of this project, the primary goal was to establish a methodology for the selection of novel dairy starter cultures applicable to low fat

Cheddar cheese fermentations. Phenotypic profile comparisons with four industrial isolates, *L. lactis* JM1-JM4, permitted the selection of similarly performing strains. To test this methodology the closest performing strain *L. lactis* subsp. *cremoris* 158 was selected for large scale cheese trials, with professional cheese grading later applied. The results of these trials indicated a Cheddar cheese with smooth texture reported as "good" overall. The taste was found to be similar to that of a traditional table Cheddar with acidic notes, and most notably it would not be apparent to a consumer that it represented a Cheddar which was low in fat and salt. This one example is encouraging and may pave the way for further explorations using similar genome-based approaches.

Chapter I

Introduction

Excerpts from this chapter have been published in:
Next-generation sequencing as an approach to dairy starter selection. Dairy Science & Technology. 95 (2015) 545-568.
Philip Kelleher, James Murphy, Jennifer Mahony and Douwe van Sinderen

Chapter I contents

1.1 Introduction - *Lactococcus lactis* …….. 14
1.1.2 Taxonomy of *Lactococcus* ……. 15
1.2 Current sequencing strategies ……. 19
1.2.2 Comparison of NGS approaches ……..... 21
1.3 Comparative genomics of *L. lactis* ……..... 25
1.3.1 Dairy industry ……..... 25
1.3.1.1 Lactose metabolism ……...……... 26
1.3.1.2 Citrate metabolism …….. 27
1.3.1.3 Proteolysis and casein metabolism …….. 27
1.3.1.4 Lipolysis …….. 29
1.3.1.5 Matrix formation ……. 29
1.3.2 Biotechnology …… 31
1.3.2.1 Expression systems ……. 31
1.3.2.2 Delivery systems and vectors …….. 33
1.4 Lactococcal plasmids …….. 35
1.4.1 Plasmid replication ……. 38
1.4.2 Plasmid transfer …….. 39
1.4.3 Bacteriocin production ……... 40
1.5 Phages and host resistance systems …… 42
1.5.1 Lactococcal phage taxonomy ……. 43
1.5.2 Industrially relevant phage; 936, c2 and P335 groups ……... 43
1.5.3 Host defence mechanisms–adsorption inhibition …….. 46
1.5.4 Superinfection exclusion systems (Sie) ……. 47
1.5.5 Restriction-modification (R-M) Systems …….. 48
1.5.6 Abortive infection (Abi) systems ……... 49
1.5.7 CRISPR/Cas Systems …… 50
1.6 Phage-host interactions of lactococci ……. 52
1.6.1 *L. lactis* cell wall polysaccharide ……... 52
1.6.2 Lactococcal prophage …… 55
1.7 Future directions …… 57
1.8 Summary of thesis contents …… 58
1.8.1 Aims and objectives ……... 58
1.9 References ……... 60

1.1 Introduction – *Lactococcus lactis*

Lactic acid bacteria (LAB) encompass a diverse group of organisms which are Gram-positive, acid-tolerant, non-sporulating, microaerophilic cocci and rods capable of producing lactic acid from the degradation of hexose sugars [1]. LAB may employ one of two pathways for the metabolism of hexose sugars, making them either heterofermentative if they utilise the pentose phosphate pathway, or homofermentative if they employ the Embden-Meyerhof-Parnas pathway [2]. As such, these organisms include a variety of genera; *Aerococcus*, *Carnobacterium*, *Enterococcus*, *Lactobacillus*, *Lactococcus*, *Leuconostoc*, *Oenococcus*, *Pediococcus*, *Sporolactobacillus*, *Streptococcus*, *Tetragenococcus*, *Vagococcus* and *Weissella* species of the order *Lactobacillales* (Fig. 1.1A) [1]. Many LAB have been granted so-called "GRAS" (generally regarded as safe) status by the American Food and Drug Administration (FDA) due to a long history of safe use in the production of an extensive array of fermented food products [3], although it is also noteworthy that some pathogenic bacteria such as *Streptococcus pneumonia* and *Streptococcus pyogenes* are closely related to this grouping [4]. LAB are commonly associated with the production of fermented dairy products such as cheese and yogurt where members of the lactococci and *S. thermophilus* are typically employed as starter cultures, while lactobacilli are typically used as adjunct cultures [5]. However, lactobacilli are also employed in the production of fermented meats [6], vegetables [7] and wines [8]. The success of LAB in the production of fermented foodstuffs is due to the rheological and organoleptic properties they impart, in conjunction with their preservative qualities through reduced pH and the production of antimicrobial compounds, which inhibit spoilage organisms [9-11]. In recent years the role of LAB

in functional foods and probiotics has also garnered increasing attention [12], while certain LAB including lactococci are now being assessed as delivery vehicles for oral drugs, bioactive molecules and vaccines [13-15]. Their extensive application in the production of fermented foods, their potential for drug delivery, and their extensive laboratory use in cloning and expression studies have rendered lactococci as one of the best studied and most valuable genera of bacteria used today. To explore the potential novelty of *L. lactis* in the genomics era this review will focus on the applications, genomic studies and phage-host interactions of this species while also assessing future prospects for research in this area.

1.1.2 Taxonomy of *Lactococcus*

Lactococcus lactis is a Gram-positive, catalase-negative, non-motile and coccoid bacterium [16] of the phylum *Firmicutes*; low G+C bacteria, class *Bacilli* order *Lactobacillales*, family *Streptococcaceae* of which *Lactococcus* represents the lactic streptococci. Originally classed as *Streptococcus lactis*, it was reclassified as a separate genus in 1985 [17]. *L. lactis* is one of several lactococcal species including; *Lactococcus chungangensis*, *Lactococcus formosensis*, *Lactococcus fujiensis*, *Lactococcus garvieae*, *Lactococcus hircilactis*, *Lactococcus laudensis*, *Lactococcus nasutitermitis*, *Lactococcus piscium*, *Lactococcus plantarum*, *Lactococcus raffinolactis* and *Lactococcus taiwanensis* (Fig. 1.1B). *L. lactis* species are further defined as one of four subspecies; subsp. *cremoris*, subsp. *lactis* [which includes a biovariant; subsp. *lactis* biovar diacetylactis capable of (plasmid encoded) citrate metabolism], subsp. *hordniae* isolated from the leafhopper *Hordnia circellata* [18], and subsp. *tructae* isolated from the brown trout, *Salmo trutta* [19], both identified as *lactis* species on the basis of 16S rRNA gene sequence similarity [18, 19]. *L. lactis*

subsp. *hordniae* and subsp. *tructae* are underrepresented in genomics studies in comparison to their dairy counterparts and will largely be excluded from the remainder of this review, as the core focus will be on the industrially important subsp. *cremoris* and subsp. *lactis* (Fig. 1.1C).

The taxonomic classification of *L. lactis* has been somewhat controversial in recent years for a number of reasons. Firstly, after the reassignment of the dairy streptococci, *Streptococcus cremoris* and *Streptococcus lactis,* to *L. lactis* subsp. *cremoris* and subsp. *lactis*, respectively, the classification of the subspecies was based singularly on industrially relevant phenotypic traits [17]. Namely, members of the subsp. *lactis* can typically tolerate 4 % salt, pH 9.2 and temperatures of up to 40 °C, while growth of subsp. *cremoris* is typically inhibited under these conditions [17]. Additionally, subsp. *lactis* can metabolise arginine and maltose, whereas subsp. *lactis* biovar. diacetylactis can metabolise citrate [20]. However, genome analysis has shed some uncertainty on these divisions wherein specific cases the phenotypes and genotypes of atypical strains do not conform, such as *L. lactis* subsp. *cremoris* MG1363 which displays a *cremoris* genotype but has a characteristic *lactis* phenotype [3], and various other strains [21]. Furthermore, it should be noted that the designation of the biovariant diacetylactis is based on a plasmid-encoded trait which may easily be transferred from strain to strain and indeed between subspecies by horizontal gene transfer [20, 22]. Finally, while previous studies have challenged the taxonomic *lactis* and *cremoris* subspecies division [23], the most recent in-depth genotypic analysis of the taxon suggests that based on the ANI (average nucleotide identity) and TETRA (tetranucleotide frequency correlation coefficients) of the two subspecies, a re-evaluation of the taxonomic group separating *L. lactis* into two distinct species *Lactococcus lactis* and *Lactococcus cremoris* is required [24]. It is

likely that an increasing number of atypical strain variants will be encountered as the rate of lactococcal genome sequencing efforts has accelerated in recent years and as such the taxonomic groupings are likely to evolve further.

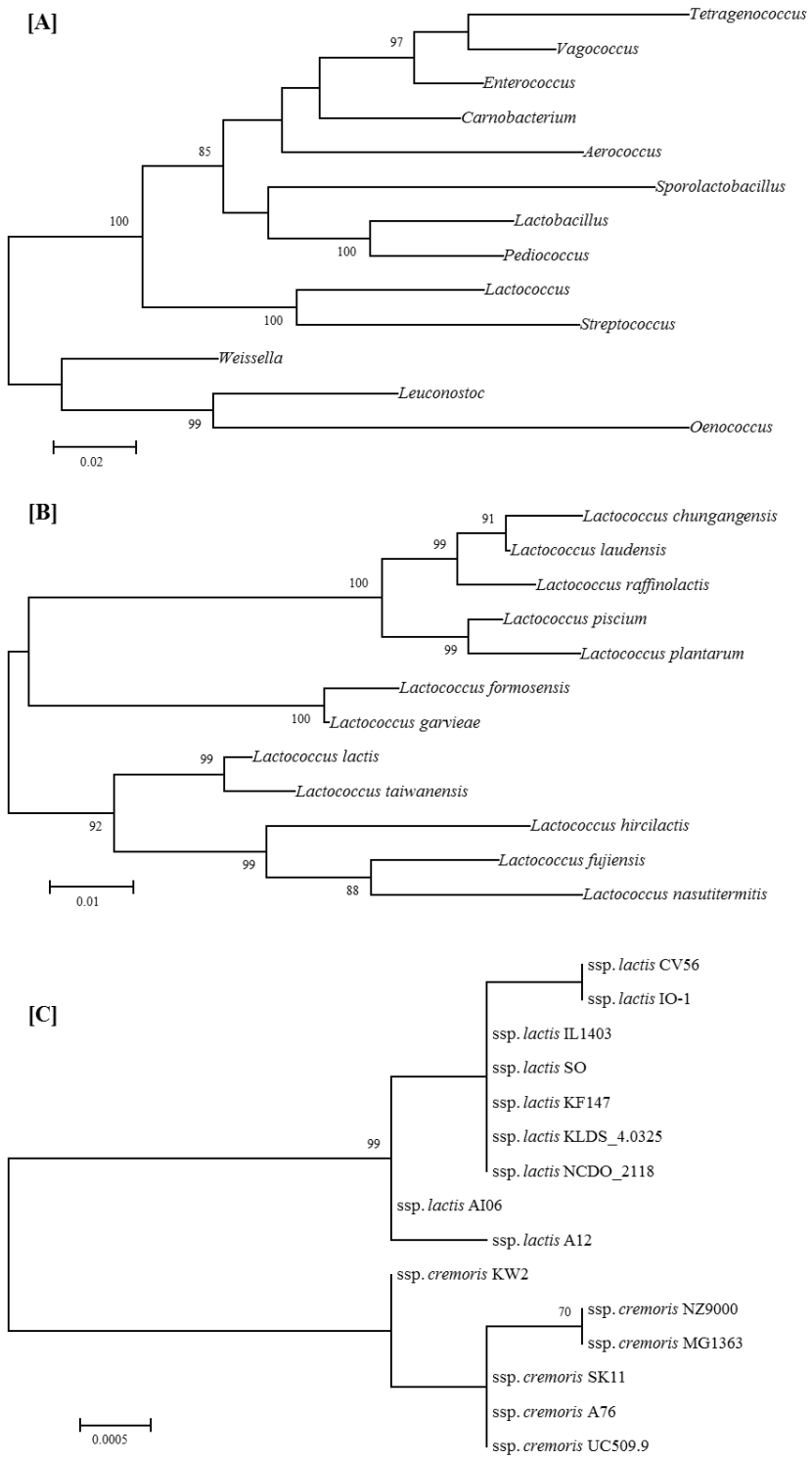


Figure 1: Molecular Phylogenetic analysis by Maximum Likelihood method **[A]** 16S rRNA-based phylogenetic analysis of LAB. **[B]** 16S rRNA-based phylogenetic analysis of *Lactococcus.* **[C]** 16S rRNA-based phylogenetic analysis of the *L. lactis* taxon. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model [25]. The trees are drawn to scale, with branch lengths measured in the number of substitutions per site. All positions with less than 95 % site coverage were eliminated. Evolutionary analyses were conducted in MEGA6 [26].

1.2 Current sequencing strategies

Lactococcal genomes typically possess a GC content of 35 - 36 %, range in size from 2.2 – 2.6 Mbp, and are frequently accompanied by a rich plasmid complement [27, 28]. There are currently eighty five genome assemblies available for the *L. lactis* taxon (correct as of November 2016) from the National Centre for Biotechnology Information (NCBI). These assemblies consist of fifteen finished quality (i.e. gapless) genomes (Table 1) and a further seventy unfinished or draft quality genomes in contigs and scaffolds. The finished quality genomes consist of six subsp. *cremoris* strains and nine subsp. *lactis* strains. The majority (all but two) of the sequenced subsp. *cremoris* strains were isolated from the dairy niche, with the exception of *L. lactis* MG1363 and its derivative NZ9000, which are employed as laboratory strains, although their parent strain NCDO712 was originally isolated from the dairy niche [29]. Conversely, greater diversity is observed within the *lactis* subspecies which are frequently isolated from plant-based niches (Table 1).

Lactococcal isolates of plant origin generally possess a broader carbohydrate utilization profile in comparison to their dairy counterparts, and are frequently capable of metabolising raffinose, sucrose, xylose and arabinose [30, 31]. The extra metabolic abilities of plant-derived strains are reflected in their respective genome sequences. For example, *L. lactis* KF147 is predicted to encode gene complements for the degradation and metabolism of xylan, arabinan, glucans and fructans, which represent plant-associated sugars [31]. The isolation of strains from these non-dairy sources may provide novel cultures for dairy fermentations and deliver desirable capabilities in terms of flavour production and industrial robustness [32, 34]. The application of next-generation sequencing for the screening of such strains offers a valuable avenue of research.

Table 1: Current complete lactococcal sequences (correct as of November 2016)

Strain	Genbank accession	Ecological niche	Sequencing technology	Year	Citation
ssp. *lactis* IL1403	AE005176	Dairy isolate	Sanger	2001	[33]
ssp. *cremoris* SK11	CP000425	Dairy isolate	Sanger	2006	[1]
ssp. *cremoris* MG1363	AM406671	Laboratory derivative of a dairy strain	Sanger	2007	[3]
ssp. *lactis* KF147	CP001834	Plant isolate	Combined 454-pyrosequencing & Illumina	2009	[35]
ssp. *cremoris* NZ9000	CP002094	Laboratory derivative of a dairy strain	Illumina	2010	[29]
ssp. *cremoris* A76	CP003132	Dairy isolate	Sanger	2011	[36]
ssp. *lactis* CV56	CP002365	Human isolate	454-pyrosequencing	2011	[37]
ssp. *lactis* IO-1	AP012281	Drain water	Sanger	2012	[38]
ssp. *cremoris* UC509.9	CP003157	Dairy isolate	Combined 454-pyrosequencing & Illumina	2012	[39]
ssp. *cremoris* KW2	CP004884	Dairy isolate	454-pyrosequencing	2013	[40]
ssp. *lactis* KLDS 4.0325	CP006766	Koumiss	Illumina	2013	[41]
ssp. *lactis* NCDO 2118	CP009054	Frozen peas	SOLiD, Ion PGM & Ion Torrent PGM	2014	[42]
ssp. *lactis* SO	CP010050	Dairy isolate	Ion Torrent PGM	2014	[43]
ssp. *lactis* AI06	CP009472	Açaí palm	454-pyrosequencing	2014	[44]
ssp. *lactis* A12	LT599049	Wheat sourdough	454 GS FLX platform	2016	[45]

1.2.2 Comparison of NGS approaches

There are a number of next-generation techniques available with associated advantages and disadvantages to each technique depending on the desired application (Table 1.2). For the study of lactococcal starter cultures, any of the sequencing techniques mentioned in Table 1.2 may be applied to obtain finished genome sequences due to the small genome size of these species: strains of *L. lactis* typically possess a ~2.5 Mb chromosome.

454-pyrosequencing is a next generation high throughput sequencing method based on the "sequence by synthesis" approach and is useful due to its longer read length compared to read lengths generated by the Illumina or Ion-torrent platforms: 700 bp compared to 300 and 400 bp, respectively. While 454-pyrosequncing has been used extensively over the last 10 years, it was announced in 2013 that Roche will begin phasing out the technology by mid-2016. The discontinuation of 454-pyrosequencing is primarily due to the advent of lower cost, high(er) through-put sequencing technologies, along with increasing read-lengths of the alternative NGS technologies [46]. Errors in homopolymer sequence tracts have also been reported with the 454-pyrosequencing method [47] and Ion torrent technology [48].

The Ion-Torrent PGM "Personal Genome Machine" represents a low-cost and rapid sequencing methodology generating approximately 80 million sequence reads in a single run of ninety minutes. The Illumina system is one of the most widely used sequencing approaches in recent years and can generate a large volume of sequencing data [49], although, the average read length is relatively low, in particular when compared to the newer PacBio SMRT platform. Current Illumina sequencing-by-synthesis (SBS) instruments are capable of generating over 1 terrabase of data in a single run and can sequence bacterial genomes in a matter of

hours. These properties, combined with low sequencing costs, have underpinned the success and current dominance of the Illumina sequencing technology.

The PacBio SMRT sequencing approach has the advantage of the longest read lengths of any sequencing technology currently in use [50], with Pacific Biosciences reporting N50 read lengths of > 14,000 bp and maximum read lengths of > 40,000 bp, which is extremely useful for covering repetitive regions of genomes, particularly so in lactococcal genomes where a large number of insertion sequence (IS) elements cause problems during sequence read assembly [51-53]. The SMRT sequencing approach also moves beyond traditional detection of the four DNA bases as it is the first high-throughput approach to directly detect DNA base modifications [54]. This allows SMRT sequencing to differentiate between unmodified bases and those with m6A, m4C or m5C base modifications [55]. One drawback of the PacBio SMRT platform which should be considered is the higher single read error rate compared to other NGS platforms. Since launching the SMRT platform, Pacific Biosciences have addressed this issue by incorporating circular consensus sequencing (CCS), which has led to greatly reduced error rates [56] and a consensus accuracy, currently reported at 99.999 %. While some early studies reported sequencing inaccuracy of ~13-18 % [57, 58], more recent studies are reporting a large reduction in these rates [59]. As mentioned, PacBio sequencing offers longer read lengths and faster runs than other methods but it also has a lower throughput and higher cost per base. The advantages of PacBio sequencing and other technologies such as Illumina are complementary, and pairing these technologies may be a useful approach for whole genome sequencing.

Table 1.2: Comparison of next-generation sequencing technologies adapted from [49]

Platform	Library Preparation	Chemistry	Consensus Accuracy	Average Read Length (bp)	Reads per Run	Run time	Pros	Cons	Application	References
Roche 454 GS FLX Titanium+	Fragment, Mate Pair/ emPCR	Pyrosequencing	99.997%	~700, maximum 1000	~1,000,000	23 hours	Long read length	High reagent cost, homopolymer repeat errors	De novo assemblies, metagenomics	http://454.com/products/gs-flx-system/index.asp
Illumina-Solexa MiSeq	Fragment, Mate Pair/ Solid phase	Reversible terminator	98%	2 x 300	~25,000,000	~55 hours 300bp reads	Widely used platform	Short read length	Small genomes, 16s amplicon, improving coverage	http://systems.illumina.com/systems/miseq.html
Life Technologies SOLiD 5500 Series	Fragment, Mate Pair/ emPCR	Sequencing by ligation (Cleavable probe)	99.99%	Mate-paired 2 x 60, Paired-end 75x 35 Fragment 75	1.2-1.4 billion	1-2 weeks	Low-cost	Slow, issues with palindromic sequences reported	Whole genome re-sequencing, variant analysis	http://www.lifetechnologies.com/ [60]
Life Technologies Ion Torrent	Fragment, Mate Pair/ emPCR	Sequential ion detection	98%	35-400	80,000,000	90 minutes	Fast and inexpensive	Reported homopolymer errors	Small genomes, gene expression, ChiP-SEQ	http://www.lifetechnologies.com/
Pacific Biosciences SMRT *RS* II	Fragment only/ Single molecule	Real-time	99.999%	N50 14,000, maximum >40,000	~50,000	30mins - 4 hours	Longest read length, detects base modifications, fast	Low single sequence accuracy 87%	De novo assemblies, Base modification detection, Transcriptome sequencing	[50, 61] http://www.pacificbiosciences.com/

As discussed, the PacBio SMRT platform currently possesses a number of unique advantages over other NGS methods. However, a noteworthy new single molecule sequencing method currently in development is Nanopore sequencing which may challenge PacBio's dominance in this area. Nanopore sequencing is predicted to deliver long read lengths and base modification data, while the simple sample preparation and possibility of label-free DNA sequencing is expected to reduce sequencing costs dramatically [62].

Genotypes of lactococcal strains derived from genomics (Table 1.1) can provide significant information about industrially important traits. There is an impressive array of tools available for post-sequencing and comparative genomic analysis (readers requiring more information should refer to [63]). Here we discuss some of the key genetic markers derived from genomic analysis with particular emphasis on their respective industrial applications.

1.3 Comparative genomics of *L. lactis*

1.3.1 Dairy industry

L. lactis is one of the dominant starter cultures employed globally by the dairy industry [64], particularly for the production of soft and hard cheeses, making it one of the most economically important bacteria today [65]. Consequently, genomic research in this area is skewed towards dairy-derived strains [51]. It is widely accepted that the original niche environment of *L. lactis* is plant-based [32, 66, 67] and that the majority of dairy strains in use today are derived from a small number of closely related lineages. Therefore, extensive redundancies are thought to exist in strain collections throughout the world [20]. One of the prominent findings from previous work in lactococcal genomics is the extent of genome decay and reductive evolution in dairy lactococci particularly within the *cremoris* subspecies [1, 28, 68], as is obvious from a substantial number of deletions and pseudogenes. Dairy strains are relatively auxotrophic and have significantly diminished carbohydrate metabolic abilities in comparison to plant-derived isolates which is attributed to continuous passage in rich growth media (milk) [31, 69, 70]. While adaptation to the dairy environment has resulted in significant decay within the chromosomes of dairy lactococci, it has also resulted in the acquisition of a number of novel features (primarily associated with plasmids) within these strains, including the ability to metabolise lactose, citrate and casein.

1.3.1.1 Lactose metabolism

The gene products of the *lac* operon facilitate and govern lactose utilisation in LAB and provide dairy strains with the ability to rapidly ferment lactose required for growth in milk. In *L. lactis* the plasmid-borne *lac* operon consists of the genes *lacABCDEFGX* and is regulated by a repressor, encoded by the adjacent *lacR* gene [71]. Loss of the *lac* operon has been reported due to the instability of the large extra-chromosomal element on which it is located [72], resulting in spontaneous mutants that are incapable of growth in milk. Interestingly, the plasmid-cured laboratory strain *L. lactis* MG1363, which does not harbour the *lac* operon, is capable of growth on lactose-supplemented media following prolonged adaptation due to the activity of a cellobiose-specific phosphotransferase system (PTS), which can act as an alternative lactose utilisation pathway [73]. Another example of an alternative lactose metabolic pathway is found in the slow lactose fermenter *L. lactis* NCDO2054 which metabolises lactose via the Leloir pathway [74]. This occurs as a result of *lacA*, which encodes a galactoside acetyltransferase, and *lacZ*, which encodes a β-galactosidase, being integrated into the *gal* (galactose) operon [75]. Such data suggests that phenotypic growth on lactose may not be an absolutely reliable indicator for the presence of the *lac* operon within lactococcal strains. Further studies have suggested that certain PCR-based techniques may be unreliable in indicating the lactose utilisation phenotype.

A recent study by Ferrario and colleagues [76] reported on the screening for isolates of *L. garvieae* in the dairy environment using primers targeting the *lacG* gene. They found that *lacG* is variably present among *L. garvieae* isolates from meat and is not limited to dairy isolates, demonstrating the need for complete genome sequences for the correct identification of dairy isolates.

1.3.1.2 Citrate metabolism

Citrate metabolism in dairy fermentations conducted by citrate-positive (Cit^+) lactococci and *Leuconostoc* spp. is important as it leads to the production of a number of volatile flavour compounds [77]. Citrate uptake and subsequent diacetyl production is governed by the plasmid-encoded *citQRP* operon in lactococcal species [78]. It has been demonstrated that the *citP* gene is well conserved amongst LAB with approximately 98 % amino acid identity making it a useful screening target for Cit^+ starters [78]. Lactococci capable of metabolising citrate are classified as *L. lactis* subsp. *lactis* biovar diacetylactis [20], a classification that has led to confusion since plasmid-encoded characteristics such as citrate and arginine metabolism can be transferred to subsp. *cremoris* strains leading to incorrect characterisation based on phenotype [20]. It is also noteworthy that recent studies have indicated potential adverse health effects associated with diacetyl production, which may lead to the removal of diacetyl-producing LAB from starter cultures [79]. These adverse effects are predoimantly associated with the exposure of factory workers to vapour-phase diacetyl while diacetyl in dairy products is still considered safe at levels of 6 – 9 ug / g [79].

1.3.1.3 Proteolysis and casein metabolism

Proteolysis and the degradation of casein from milk is one of the most important contributors to flavour development in cheese [80]. Lactococcal strains contribute to proteolysis through the hydrolysis of casein by peptidases and proteases, and the catabolism of peptides and amino acids from casein breakdown [81]. There are a number of genes which contribute to this function, including (i) various (and mostly) chromosomally-specified peptidase-encoding genes (e.g. *pepC*,

pepN, *pepX*, *pepP*, *pepA*, *pepF2*, *pepDA1*, *pepDA2*, *pepQ*, *pepT*, *pepM*, and *pepO1*), (ii) the plasmid-encoded *opp* operon, which specifies an oligopeptide-uptake system, and (iii) the plasmid-borne gene that specifies the *L. lactis* cell wall-associated protease PrtP, required for the proteolytic phenotype [82]. The majority of the genes mentioned above are monocistronic elements (e.g. *pepC*, *pepN* and *prtP*) or co-transcribed, such as *opp* and *pepO1*, while *pepF2*, *pepM* and *pepT* are transcribed with genes that are (apparently) unrelated to proteolysis [83]. There are also a number of uncharacterised proteins which contain peptidase-associated domains, many of which are strain-specific and their roles may become more clear as more genome sequences become available [84].

As discussed, proteolysis contributes greatly to cheese flavour development, however, high levels of proteolysis can also cause bitterness defects in cheese [85]. The *L. lactis* extracellular cell wall proteinase (lactocepin) is directly involved in bitterness flavour defects in Cheddar cheese varieties, specifically starters which produce group a, e, or h lactocepin [85]. Broadbent et al. [85] concluded that the bitterness defect in cheese could be altered through gene exchange or replacement in the starter culture. These findings highlight the benefits of subsp. *cremoris* strains in lactococcal starter cultures in comparison to subsp. *lactis*. A recent study by Liu *et al*. [86] demonstrated that our knowledge of the proteolytic system in LAB can be enhanced by systematic genome-wide studies of the regions encoding proteins involved in proteolysis. They indicated that comparative genomics can be used to distinguish various sub-groups within protein superfamilies involved in proteolysis where the generated information predicts the proteolytic ability of LAB strains. A major finding from this study was the confirmation of proteolytic diversity among ssp. *lactis* and ssp. *cremoris* strains and the provision of a genetic basis for this

diversity, linked to distinct patterns in the presence or absence of genes encoding proteolytic functions [86].

1.3.1.4 Lipolysis

Lipolysis involves the breakdown of milk fats and hydrolysis of triglycerides into lipids and fatty acids, activities that are considered to be crucial for flavour development in cheese production, particularly in the production of Cheddar cheese varieties [77]. Lipolytic enzymes produced by LAB are mainly represented by esterases and lipases that belong to a class of enzymes called the carboxylic ester hydrolases [87]. Apparently, *estA* is the only esterase-encoding gene in *L. lactis*, being capable of hydrolysing short chain fatty acid esters [88]. However, this research area of cheese flavour development remains considerably under-represented in lactococcal studies compared to those related to proteolysis [77]. Therefore, a genomics approach may prove to be beneficial in broadening our scope of knowledge on lipolysis in lactococcal strains.

1.3.1.5 Matrix formation

Exopolysaccharides (EPS) produced by LAB are secreted polysaccharides which may remain attached to the cell envelope as capsular EPS (CPS) or released in the surrounding medium [89]. Producing strains are generally described as “ropy” or “non-ropy”, a term which describes the threads drawn with a needle from the surface of the colonies or fermented liquid [90]. The EPS produced by certain dairy LAB can impact on the protein matrix of fermented dairy products by affecting the casein gel structure and acting as a filler [91]. A common assay for the differentiation of ropy and non-ropy colonies utilises ruthenium red stain in milk agar plates.

Ruthenium red stains the cell walls, thereby producing red colonies in case of non-ropy, non-EPS producing cells, yet is unable to stain cell walls of ropy, EPS producers, which therefore remain white [92].

EPS production by *L. lactis* is a characteristic trait of strains isolated from viscous Scandinavian fermented milk products and is widely reported as a plasmid-encoded trait [93-96]. EPS production by *L. lactis* strains is of particular importance for fermented dairy products, as EPS is considered to be a food-grade additive, and contributes significantly to properties such as mouth-feel and texture [97].

L. lactis strains may produce two types of EPS: homo-polysaccharides and hetero-polysaccharides, being comprised of a repeating monosaccharide, or a repeating oligosaccharide, respectively [98, 99]. EPS biosynthesis of the hetero-polysaccharide type in *L. lactis* takes place by the Wzy-dependent mechanism [99], beginning with the transfer of a sugar-1-phosphate from a Uridine diphosphate-sugar to a lipid undecaprenyl-phosphate acceptor via a priming glycosyltransferase [100]. Subsequent sugars are then incorporated by additional glycosyltransferases, typically encoded downstream of the gene that specifies the priming glycosyltransferase [101]. The oligosaccharide repeating units are then transported across the cytoplasmic membrane by a flippase, and undergo polymerization by the Wzy polymerase [101]. A number of studies have reported on EPS production by *L. lactis* strains, and in each case the EPS produced was (predicted to be) a plasmid-encoded, hetero-polysaccharide whose biosynthesis occurred according to a Wzy-dependent mechanism [98, 99, 102].

1.3.2 Biotechnology

To date a number of bacteria have been exploited for the production and expression of recombinant proteins, both in research and industry. *Escherichia coli* is the most intensively used bacterium for this purpose and this bacterium allows the highest levels of production for certain proteins [103], however *E. coli* has a number of downstream processing issues associated with its use. *E. coli* produces the vast majority of proteins in the cytoplasm or periplasm, which requires cell harvesting, lysis and purification, *E. coli* furthermore produces endotoxins which can cause issues for proteins which are to be administered to people as biopharmaceuticals [104]. *Bacillus* species have also been widely used with the advantages of "GRAS" status and the option of extracellular secretion of the desired protein [105]. However, heterologous proteins produced by *Bacillus* are frequently degraded by its complex extracellular proteolytic system [106]. *L. lactis* is becoming an increasingly employed alternative to the aforementioned species as it also has "GRAS" status, secretes extracellularly, has a relatively simple metabolism [33] and secretes only one extracellular protein, Usp45, at significant levels [107], simplifying downstream processing [104].

1.3.2.1 Expression systems

A number of different expression systems are currently described in *L. lactis* for heterologous protein expression and have been reviewed thoroughly [104, 108, 109]. One of the most successful and extensively employed expression systems is the nisin-inducible controlled gene expression (NICE) system [110] (extensively reviewed in [111]). Briefly, the NICE system is derived from the lactococcal nisin-producing operon (*nisABTCIPRKEFG*) [112] and utilises the regulatory elements of

the operon for controlled expression, P_{nisA} (promoter) and nisRK (regulator-sensor) (Fig. 1.1) [104]. The advantages of such a system are the ease of use, tight control and suitability for large scale production [111]. While extremely useful for research interests, there are however some disadvantages for industrial application of the NICE system, primarily associated with the cost and downstream removal of the nisin for certain applications [104, 111].

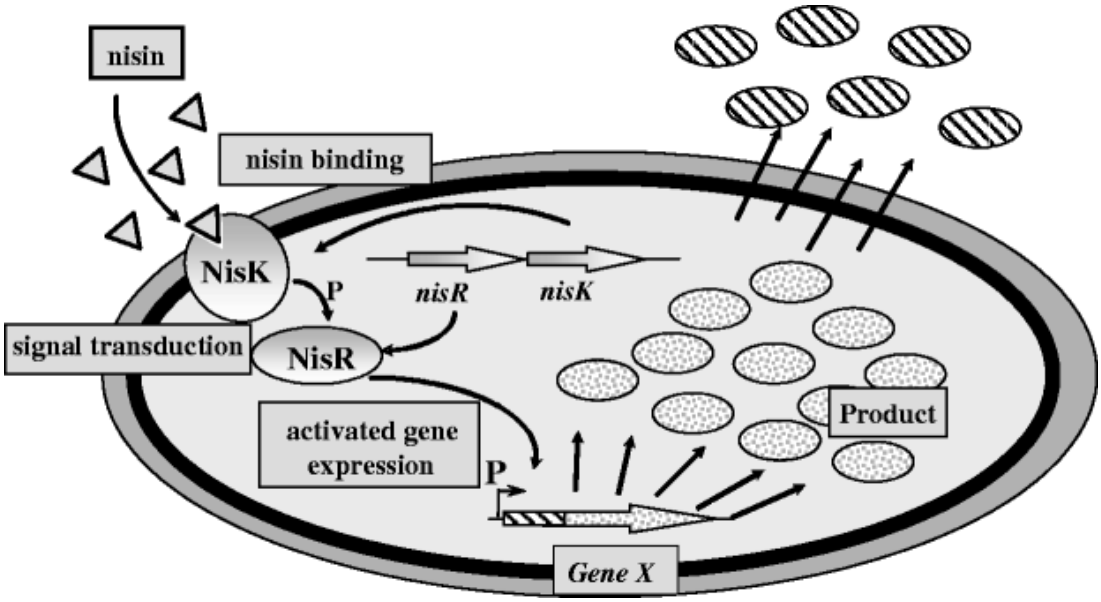


Fig 1.1: NICE system in *L. lactis* (Image taken from [111])
Nisin-controlled gene expression; *nisK* (encodes nisin-responsive histidine–protein kinase), *nisR* (encodes response regulator), Gene X (cloned target gene to be expressed) and *nisA* (promoter, indicated here as P) [111].

1.3.2.2 Delivery systems and vectors

L. lactis used as cell factories for the production of heterologous proteins have a number of industrial and clinical applications with many of those focused on utilising *L. lactis* as a vector for the delivery of specific proteins either directly to a fermentation process or in clinical situations. *L. lactis* strains with increased proteolytic capabilities have previously been used in dairy fermentations by heterologous expression of the peptidases; PepN, PepC, PepX and PepI from *Lactobacillus helveticus* [113] and PepI, PepL, PepW and PepG from *Lactobacillus delbrueckii* [114] for improved cheese ripening. In addition, food fermentations employing *L. lactis* strains that are engineered to (over)produce riboflavin (vitamin B2), acetaldehyde, diacetyl or folate (vitamin B11) have also been developed [115].

Applications in health and medicine are an emerging area of interest for *L. lactis* and a number of studies have demonstrated the potential of *L. lactis* for the production and/or delivery of such pharmaceutical products. Hyaluronic acid, used in medicines, drug delivery systems and vaccine aids, was produced in *L. lactis* by incorporating the hyaluronic acid biosynthesis operon, NICE system and the *lacF* selective marker through chromosomal integration [116]. *L. lactis* has also been engineered to produce Interleukin 10, used as a treatment for inflammation in mouse colitis models [117], and has since been approved for small-scale clinical trials in humans with inflammatory bowel disease [118]. One of the major advantages of using lactococcal strains in human health, particularly for the delivery of vaccines is that *L. lactis* does not belong to the normal human microbiota making oral and gut colonisation unlikely [119]. This is beneficial as mucosal vaccination with gut-colonizing bacteria could lead to increased antigen tolerance [119]. Consequently, a

number of studies are currently investigating lactococcal strains as mucosal delivery vectors for both therapeutic proteins and DNA vaccines [120-124].

1.4 Lactococcal Plasmids

Plasmids are semi-autonomously replicating extrachromosomal genetic elements, which are typically not essential for growth [125], but which may encode traits that confer important niche-specific phenotypes to their respective host [66]. Large plasmid complements are frequently carried by lactococcal strains, particularly those isolated from the dairy niches, these appeared to have acquired such an elaborate plasmid complement in order to adapt to the nutrient-rich environment, milk [20, 27, 34]. Many of the conferred phenotypic traits include industrially important adaptations such as stress tolerance, (bacterio)phage-resistance mechanisms, and enhanced proteolytic and carbohydrate metabolic capabilities [65, 126, 127].

Current sequencing efforts have resulted in the availability of 86 lactococcal plasmids from the NCBI (correct as of December 2016) detailed in Table 1.3. Lactococcal plasmids range in size from 1 – 72 Kbp (the latter being pCIS8 of UC509.9), with various sequenced complements containing up to eight individual plasmids [27], while it has been estimated that some strains may contain up to fourteen [20]. The majority of the current data set has been obtained from the dairy niche (64/86); dairy strains typically contain four or five plasmids, while their plant equivalents generally harbour just one or two, or none at all [40, 128]. The persistence of larger plasmid complements in dairy strains is due to a number of factors; the *lac* and *cit* operon as discussed previously [71, 78] are predominantly plasmid-encoded, as is the *opp* operon and a number of casein specific peptidases [82, 83], as well as a host of phage defence mechanisms [65, 129]. Many of these traits are mobilisable and transmissible by horizontal gene transfer via conjugation or transduction [126].

Table 1.3: Lactococcal plasmids sequenced to date*

Name	Accession	Size (Kbp)	GC (%)	Genes	Niche	Replication mode	Citation
KLDS 4.0325 p1	CP006767	4.094	30.02	4	Fermented food	RCR	[41]
KLDS 4.0325 p2	CP007042	0.870	32.64	2	Fermented food	Undetermined	[41]
KLDS 4.0325 p3	CP007043	1.278	32.63	4	Fermented food	Undetermined	[41]
pAF04	JQ821353	3.801	32.02	4	Dairy	Theta	[225]
pAF07	JQ821354.1	7.435	36.44	6	Dairy	Theta	[225]
pAF12	JQ821355.1	12.067	33.30	11	Dairy	Theta	[225]
pAF14	JQ821356.1	14.419	34.07	11	Dairy	Theta	[225]
pAF22	JQ821357.1	22.388	34.95	23	Dairy	Theta	[225]
pAG6	AB198069	8.663	33.70	8	Unknown	Theta	-
pAH33	AF207855	6.159	35.85	7	Dairy	Theta	[183]
pAH82	AF243383	20.331	34.44	17	Dairy	Theta	[226]
pAR141	DQ288662	1.594	36.14	2	Dairy	RCR	[227]
pAW153	HQ646604.1	7.122	31.35	8	Unknown	Theta	[228]
pAW601	AJ132009.2	4.752	31.42	1	Unknown	Theta	-
pBL1	AF242367	10.899	32.62	8	Dairy	Theta	[229]
pBM02	AY026767	3.854	35.73	6	Dairy	RCR	[230]
pCD4	AF306799	6.094	33.43	5	Dairy	Theta	[134]
pCI305	AF179848	8.694	32.41	8	Dairy	Theta	[231]
pCIS1	CP003165	4.263	31.97	2	Dairy	Theta	[39]
pCIS2	CP003164	5.461	30.07	4	Dairy	Theta	[39]
pCIS3	CP003163	6.159	35.85	5	Dairy	Theta	[39]
pCIS4	CP003162	7.045	38.42	10	Dairy	Theta	[39]
pCIS5	CP003161	11.676	34.06	10	Dairy	Theta	[39]
pCIS6	CP003160	38.673	37.12	30	Dairy	Theta	[39]
pCIS7	CP003159	53.051	32.40	48	Dairy	Theta	[39]
pCIS8	CP003158	80.592	33.97	72	Dairy	Theta	[39]
pCL2.1	U26594	2.047	33.95	2	Unknown	RCR	[232]
pCRL1127	AF409136	8.278	34.82	7	Unknown	Theta	-
pCRL291.1	AF380336	4.640	33.51	3	Unknown	Theta	-
pCV56A	CP002366	44.098	32.08	41	Human	Theta	[37]
pCV56B	CP002367	35.934	34.54	31	Human	Theta	[37]
pCV56C	CP002368	31.442	32.49	27	Human	Theta	[37]
pCV56D	CP002369	5.543	32.24	6	Human	Theta	[37]
pCV56E	CP002370	2.262	33.82	4	Human	Theta	[37]
pDBORO	DQ089807	16.404	35.16	15	Unknown	Theta	-
pDR1-1	AB079381	7.412	33.70	6	Dairy	Theta	-
pDR1-1B	AB079380	7.344	33.74	6	Dairy	Theta	[233]
pFI430	DQ011112.1	59.474	34.63	57	Dairy	Theta	[234]
pGdh442	AY849557	68.319	35.11	63	Plant	Theta	[235]

pHP003	AF247159	13.433	40.05	6	Dairy	Theta	[236]
pIL1	HM021326	6.382	32.28	7	Dairy	Theta	[237]
pIL105	AF116286	8.506	29.79	7	Dairy	Theta	[238]
pIL2	HM021327	8.277	34.82	10	Dairy	Theta	[237]
pIL3	HM021328	19.244	35.11	20	Dairy	Theta	[237]
pIL4	HM021329	48.978	35.11	47	Dairy	Theta	[237]
pIL5	HM021330	23.395	34.49	22	Dairy	Theta	[237]
pIL6	HM021331	28.434	33.64	25	Dairy	Theta	[237]
pIL7	HM197723	28.546	34.10	26	Dairy	Theta	[237]
pK214	X92946	29.871	32.45	29	Unknown	Theta	[239]
pKF147A	CP001835	37.510	32.38	32	Plant	Theta	[35]
pKL001	EU289287	6.068	32.86	4	Unknown	Theta	-
pKP1	FR872378	16.181	35.94	7	Dairy	Theta	[240]
pL2	DQ917780	5.299	32.46	5	Dairy	Theta	[241]
pLP712	FJ649478.1	55.395	37.39	44	Dairy	Theta	[135]
pMN5	AF056207	5.670	30.26	4	Dairy	RCR	[242]
pMRC01	AE001272	60.232	30.11	63	Dairy	Theta	[243]
pNCDO2118	CP009055	37.571	32.33	32	Plant	Theta	[42]
pND324	U44843	3.602	33.37	3	Unknown	Theta	-
pNP40	DQ534432	64.980	32.33	62	Dairy	Theta	[137]
pNZ4000	AF036485	42.810	33.31	45	Dairy	Theta	[93]
pQA504	CP003136	3.978	37.83	3	Dairy	Undetermined	[36]
pQA518	CP003135	17.661	37.40	13	Dairy	Theta	[36]
pQA549	CP003134	49.219	35.14	44	Dairy	Theta	[36]
pQA554	CP003133	53.630	34.86	54	Dairy	Theta	[36]
pS7a	AJ550509	7.302	33.43	5	Dairy	Theta	[244]
pS7b	AJ550510	7.264	33.65	5	Dairy	Theta	[244]
pSRQ700	U16027	7.784	34.19	9	Dairy	Theta	[245]
pSRQ800	U35629	7.858	31.33	7	Dairy	Theta	[245]
pSRQ900	AF001314	10.836	31.13	11	Dairy	Theta	[245]
pVF18	JN172910	18.977	33.90	21	Dairy	Theta	[246]
pVF21	JN172911	21.728	33.59	14	Dairy	Theta	[246]
pVF22	JN172912	22.166	35.14	19	Dairy	Theta	[246]
pVF50	JN225497	53.876	34.50	41	Dairy	Theta	[246]
pWC1	L75827	2.846	29.48	1	Dairy	RCR	-
pWV01	X56954	2.178	33.43	4	Dairy	RCR	[130]
pWVO2	NC_002193.1	3.826	31.34	1	Unknown	Theta	[131]
SK11 p1	CP000426	14.041	34.37	13	Dairy	Theta	[1]
SK11 p2	CP000427	9.554	30.44	10	Dairy	Theta	[1]
SK11 p3	CP000428	74.750	35.41	69	Dairy	Theta	[1]
SK11 p4	CP000429	47.208	34.84	42	Dairy	Theta	[1]
SK11 p5	CP000430	14.206	33.55	10	Dairy	Theta	[1]
pIBB477a	CM007354	66.364	33.18	66	Dairy	Theta	[247]
pIBB477b	CM007355	64.760	35.99	56	Dairy	Theta	[247]

pIBB477c	CM007356	48.496	32.97	42	Dairy	Theta	[247]
pIBB477d	CM007357	16.577	31.78	17	Dairy	Theta	[247]
pIBB477e	CM007358	11.987	39.60	15	Dairy	Theta	[247]

***(Correct as of November 2016)**

1.4.1 Plasmid replication

Most of the lactococcal plasmids that have been isolated to date replicate by the theta mechanism, while in a small number of cases the rolling-circle replication (RCR) mechanism of replication is used (Table 1.3) [130, 131]. RCR replication relies on a replication protein and a double-stranded origin of replication (*dso*), which contains a *nic* site composed of one or more inverted repeats, and a Rep-binding site consisting of two to three direct repeats or an inverted repeat [127, 132]. Replication initiates when a single-stranded break is introduced to the *nic* site of the *dso* by the replication protein, which results in a free 3′ single strand DNA used in leading strand synthesis [133]. The parental strand is then displaced by the replicating strand until the new *dso* is reached. Lagging-strand replication occurs on the displaced parental strand from a non-coding region, which generates a stem loop structure, termed the single-stranded origin (*sso*) [133]. Just seven plasmids in the current data set are predicted to utilise RCR (Table 1.3). The finding that only a relatively small number of plasmids utilise RCR may be due to a number of factors, such as the limited replicon size (<10 Kb), incompatibility with other RCR type plasmids [130], and/or intrinsic structural and segregational instability [65].

Replication via the theta method requires a replication (initiation) protein (encoded by *rep*), an origin of replication (*ori*) comprised of an AT-rich region with (typically) three and a half iterons of 22 bp in length and two inverted repeats overlapping the -35 site of the *rep* promoter. Replication in theta plasmids may be uni- or bi-directional from multiple origins [127, 132]. Theta type replicons have a

limited host range but are significantly more stable in comparison to RCR plasmids [134] and perhaps for this reason represent the majority of sequenced lactococcal plasmids (76 out of 86 known plasmids) (Table 1.3).

1.4.2 Plasmid transfer

Conjugation and transduction are believed to be the dominant mechanisms of plasmid transfer in *L. lactis* [65]. Transduction is a phage-mediated method of DNA transfer, where the plasmid DNA is packaged (instead of phage DNA) into the phage head and transferred to a new host upon infection, the size of the transduced plasmid is limited by the internal capacity of the capsid of the transducing phage [135, 136]. Previous studies have observed high frequency transduction in *L. lactis* NCDO712, where small (< 5 kb) plasmids were transduced at a frequency of 2.1×10^{-3} to 2×10^{-4} transductants per plaque-forming unit (PFU) [135].

Particular emphasis has been placed on conjugation as it is considered a naturally occurring DNA transfer process without the issues of host specificity associated with phages and for this reason may be used in food-grade applications to confer beneficial traits to industrial strains [127]. During conjugation, plasmid DNA is passed from a donor cell to a recipient through the formation of a conjugative channel or pilus [126]. Generally, during conjugation the AT-rich, so-called ‘origin of transfer’ or *oriT* of the conjugative plasmid is nicked by a nickase, and the resulting ssDNA strand is passed on to a recipient cell [126], though the precise mechanistic details of the conjugation process in *L. lactis* remain unclear. The *tra* (transfer) locus is believed to be responsible for the donor-to-recipient DNA transfer process of conjugation [137]. Previous studies have identified *traF* as encoding a membrane-spanning protein involved in channel formation and membrane fusion

[22, 137]. In addition, *traE* and *traG* have been proposed to encode proteins involved in the formation of the conjugal pilus similar to type IV secretion systems [22, 137]. Typically, the three *tra* genes (i.e. *traE*, *traF* and *traG*) are part of a larger gene cluster (consisting of up to fifteen genes), including *traA*, which encodes a relaxase. However, precise functions for the remainder of the genes in the *tra* gene cluster have yet to be elucidated.

While the *tra* operon is believed to be involved in the formation of the conjugative apparatus and physical transfer of plasmid DNA, another set of mobilisable genes termed *mob* are thought to be responsible for the mobilisation of other (non-conjugatable) plasmids in *L. lactis* [127, 137, 138]. Variants of four main *mob* genes are distributed throughout the lactococcal plasmidome; *mobA* and *mobD* encode nickases, and *mobB* and *mobC*, whose protein products are thought to form a relaxosome with an associated nickase (*mobA* or *mobD*) at the origin of transfer (*oriT*); usually in the genetic conformation *mobABC* or *mobDC* [139]. Although many lactococcal plasmids appear to be non-conjugatable, *mob* genes appear in high frequency throughout the plasmidome and may be a reflection of plasmid acquisition/transfer events by mobilisation in the past [129].

1.4.3 Bacteriocin Production

Bacteriocins are a diverse group of ribosomally synthesized peptides, produced by some bacteria and archaea, which have a bactericidal or bacteriostatic effect on other bacteria when secreted [140]. Bacteriocin production is a double-edged and important consideration in selecting starter cultures, as producing strains may inhibit other desirable strains in mixed starter cultures or adjunct cultures added later in the fermentation process; however, they also offer the benefit of inhibiting the growth of spoilage bacteria in food products. Traditionally, a range of culture-

based methods have been used in screening for bacteriocin producers, most commonly based on the principles of diffusion in agar plates and cell-free supernatants [141-143]. In recent years, sequence-based analysis tools have become a valuable aid in the identification of novel bacteriocins with the availability of databases and search-tools such as the BAGEL3 web-based bacteriocin mining tool [144].

1.5 Phages and host resistance systems

While the technological attributes of dairy starter cultures are essential to achieve the desired flavours and characteristics in the final product, the phage robustness of these strains is also an important consideration. Since the discovery of lactococcal phages (i.e. viruses that infect bacterial cells) by Whitehead and Cox in 1935, phage infection has been recognized as the main cause of commercial fermentation problems with concomitant economic impact [145]. The selection of a suitable starter culture would normally include the assessment of susceptibility to phage infection (and of course acidification and flavour/texture formation), but with the advent of modern sequencing technologies, starter culture suppliers can now also screen strains for the presence of prophages as well as the arsenal of plasmid- and/or chromosomally-encoded phage-resistance mechanisms [27, 146, 147].

Phages may follow one of two possible life cycles, i.e. the lytic or temperate/lysogenic life cycle, depending on the phage and the environmental circumstances. Phages entering the lytic cycle subvert the host DNA/protein synthesising machinery in order to multiply themselves intracellularly, which is then followed by host cell lysis and consequent release of progeny phages. However, conditions may not favour the lytic life cycle and as a result some phages engage in a lysogenic life cycle by incorporating their genomes within the chromosome of their host, thereby allowing phage genome replication *in situ* with that of the host's chromosome. This process allows the phage to replicate 'silently' each time the bacterium undergoes cell division by binary fission. Under certain (stress) conditions the lysogenic phage will excise from the host chromosome and enter the lytic cycle.

1.5.1 Lactococcal phage taxonomy

Lactococcal phages belong to the order *Caudovirales* which encompasses over 95 % of all known phages into an extremely large, genetically and morphologically diverse group [148]. The order *Caudovirales* is composed of three major phage families, namely the *Myoviridae*, *Siphoviridae* and *Podoviridae*. Lactococcal phages belong to one of two of these phage families, i.e. the *Siphoviridae* and *Podoviridae* [149]. Most lactococcal phages belong to the *Siphoviridae* family, and in fact eight of the ten known lactococcal phage groups/species (i.e. 936, P335, c2, 1358, Q54, P087, 1706, and 949), while the two remaining groups/species, P034 and KSY1, represent members of the *Podoviridae* family and are rarely encountered in dairy facilities [149]. The *Siphoviridae* are recognised by their characteristic long non-contractile tail, while the *Podoviridae* have short tails (Fig. 1.3).

1.5.2 Industrially relevant phage; 936, c2 and P335 groups

The 936 group of lactococcal phages represents the most prevalent of the ten groups of lactococcal phages found in commercial dairy environments [150-152]. Phages belonging to the 936 group possess a double-stranded (ds) DNA genome with a size of ~26 - 32 Kb, and a modular genetic structure similar to other *Siphoviridae* phages composed of late-, early- and middle-expressed gene modules [151, 153]. To date, a total of ninety 936 group phages have been fully sequenced and both the core and pan genomes of this group have been resolved [150]. This latter work suggested that a link exists between gene complement/phylogeny and geographical origin of the isolated 936 phages, and that the distribution of the non-core genome can also be linked to these groups [150]. Infection by 936 phages

constitutes the single most significant risk for dairy fermentations and these phages have consequently received significant research attention [150, 151, 154-156], resulting in the identification of a number of adaptive genetic features including (orphan) methyltransferases [157].

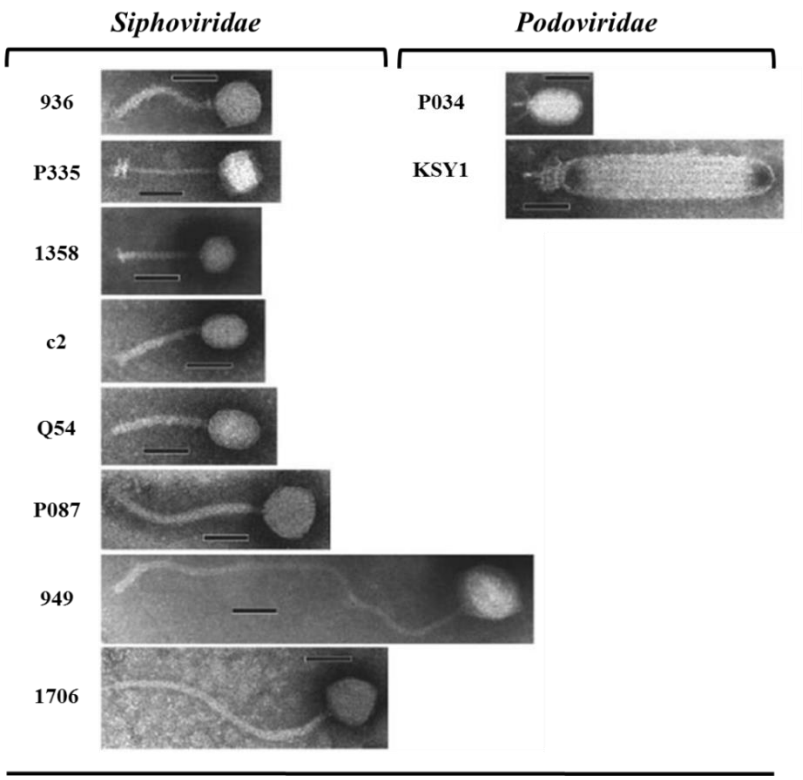


Fig 1.3: Overview of lactococcal phage morphology and biodiversity. Electron micrograph images of representative phages from each of the known species/groups of lactococcal phages. Phage family is also indicated; *Siphoviridae* or *Podoviridae*. Adapted from [149].

The c2 group of lactococcal phages is represented by two subgroups based on their host receptor preference, i.e. the c2- and bIL67-like subgroups, with a total of ten isolates sequenced to date: two isolates that belong to the c2-like subgroup and eight isolates that are members of the bIL67-like subgroup [158]. They are characterised by highly conserved genome sequences of approximately 22 Kb, which share 80 % nucleotide identity across their genome length. The c2 phages have a highly diverse host range believed to be determined by the structural region of the phage which is one of the few regions with relatively low sequence conservation [159]. The open reading frames (ORFs) involved were identified by Millen and colleagues who demonstrated that swapping ORF14–15–16 (found in the c2-like subgroup) and ORF34–35–36 (found in the bIL67-like subgroup) resulted in phage recombinants with an altered host range [158].

The P335 group of lactococcal phages is an extremely diverse group of both temperate [160] and lytic phages [161], and is characterised by its extreme genome plasticity [162]. There is no single gene conserved within the entire P335 population and current subgroupings are based in part on the level of amino acid identity in the structural region [162]. The P335 baseplate is a large heteropolymeric organelle located at the tip of the tail used for host recognition [163]. There are currently ten sequenced members of this group (4268, BK5-T, LC3, P335, r1t, TP901-1, Tuc2009, ul36, Q33 and BM13) alongside a plethora of integrated P335-like phage present within host genome sequences [162]. Lysogenic P335 phages are of particular concern to fermentations as they may pose the risk of becoming activated during the fermentation process leading to partial or complete culture lysis [53]. When selecting appropriate starter cultures for the production of various dairy products the presence of prophages can be determined by phage induction assays whereby the bacterium is

exposed to chemical, thermal or environmental treatments or conditions (chemical treatment or exposure to UV-light) to stimulate the excision and transcriptional activation of the integrated phage, which may ultimately cause lysis of the host cell [53]. However, such approaches are time-consuming and require the assessment of large collections of strains. In addition, 'true' prophage induction can only be determined using additional methods such as confirmation of the presence of prophages by performing phage sensitivity assays (upon identification of a sensitive host strain), PCR or flow cytometry [164]. Whole genome sequencing can readily identify the presence of temperate phages within the host genome, although it cannot with absolute certainty determine if a phage is genuinely inducible and thus a threat during fermentation. Furthermore, the availability of programmes such as Phage_Finder as well as gene annotation tools, aid in the determination of the presence of intact or cryptic prophage elements [165], which is important as the presence of prophages is common in *L. lactis* with some strains harbouring up to six prophages [3, 53].

1.5.3 Host defence mechanisms–adsorption inhibition

Strains of *L. lactis* may encode multiple phage resistance mechanisms which target specific steps in the phage life cycle. The initial step of phage infection that can be targeted is phage adsorption, which may be blocked by as yet uncharacterised host cell surface modifications. After attachment DNA entry occurs via an injection process which is targeted by so-called Superinfection exclusion (Sie) systems. The injected DNA may be targeted for cutting after entry by restriction modification (R-M) systems or Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) systems. The final step of phage infection that can be targeted by host

defence is represented by the phage DNA replication and transcription, and phage protein production and assembly processes, which can be disrupted by so-called abortive infection (Abi) systems that cause programmed cell death [166].

Bacteria can take a multifaceted approach to adsorption inhibition either by blocking/modifying phage receptors, or by producing extracellular matrices or compounds which may act as competitive inhibitors [166]. In *L. lactis* the receptors for the three main groups of infecting phages have received substantial attention. The c2 group of lactococcal phage follow conventional reversible saccharide binding prior to irreversible binding to the membrane protein termed Pip (phage infection protein) or its homologue YjaE [158, 167-170]. Members of both the 936 and P335 phage groups possess complex multi-protein organelles, termed baseplates, at the distal end of their tails which bind to carbohydrates that are present in the surface-exposed lactococcal cell wall-associated polysaccharide (CWPS) [154, 171]. To block adsorption of these phages, lactococcal strains employ a number of different native inhibition systems. The plasmids pSK112 and pCI528 have been shown to produce a galactosyl-containing lipoteichoic acid and a galactose/rhamnose-rich polymer, respectively, in both cases capable of inhibiting the attachment of phages [172-174], while plasmid pCI658 encodes the biosynthetic machinery for an EPS that is thought to mask phage receptor(s) [102].

1.5.4 Superinfection exclusion systems (Sie)

The presence of prophages in commercial strains has generally been considered an undesirable trait due to the risk of phage excision, however, some prophage elements encode superinfection exclusion (Sie) systems [175-177]. Sie systems block the entry of phage DNA to the host cell, thus preventing infection

[166]. The best characterised Sie system in *L. lactis* is Sie_{2009} encoded by the temperate phage Tuc2009 which confers resistance against the 936 group of lactococcal phages [160, 166, 177]. These Sie proteins were found to be associated with the lactococcal cell membrane and to confer resistance by inhibiting DNA injection into the host cell [160, 177].

1.5.5 Restriction-Modification (R-M) Systems

Genes encoding R-M systems are present on approximately 90 % of currently available bacterial and archaeal genome sequences [178]. These systems can be plasmid- or chromosomally-encoded, and their general role is to recognize and target invading foreign DNA with restriction enzymes, while simultaneously protecting the host DNA by methyltransferase (MTase) activity. Four types of R-M systems (I, II, III & IV) are currently recognized and have been extensively reviewed [178-181]. Briefly, Type I R-M systems are multi-subunit proteins that function as a single protein complex, usually composed of one or two REase subunits (HsdR), one or two MTase subunits (HsdM) and (typically) one specificity (S) subunit (HsdS) [178, 182]. However, instances of the intergenic shuffling of multiple HsdS-encoding genes belonging to a single Type I R-M system have been reported [183-185]. Type I R-M systems recognize long, (mostly) non-palindromic motifs [186]. Type II R-M systems are composed of separate REase and MTase activities. Type II REases act as homodimers to target specific DNA sequences and act independently of their cognate MTase [178]. Type II R-M systems are among the most thoroughly studied due to their importance in molecular biology [187, 188]. Type III R-M systems are composed of two subunits that function either in DNA recognition and modification (Mod) or restriction (Res) [189]. Type III systems require ATP hydrolysis to

function [178] and are frequently found in prokaryotic genomes [187, 190]. Type IV R-M systems are those which, unlike Types I-III, only target methylated DNA. Type IV systems are composed of two genes and their target motifs are not well defined [178].

The presence of various R-Ms in industrial starter cultures is an important technological property to help in phage defence, as invading phage DNA, if unmethylated (except in the case of type IV R-M systems), will be subject to endonuclease activity. The advent of accessible sequencing technologies allowing for the characterisation of chromosomal- and plasmid-encoded R-Ms, e.g. the *L. lactis* systems LlaJI, LldI and LlaI [191-193], has been helped by the availability of online resources such as the REBASE search platform [187, 188]. In recent years, the emergence of SMRT sequencing technology (as discussed above) has revolutionised the identification of whole genome modification and the functionalities of R-Ms. Combining whole genome sequencing and MTase motif analysis, the functions of one or more bacterially-encoded R-M can be predicted which can then be confirmed using heterologous gene expression coupled with restriction endonuclease assays. This approach has been applied to both bacteria and bacteriophages alike [155, 194], though it has not yet been applied to *Lactococcus*.

1.5.6 Abortive infection (Abi) systems

Abortive infection (Abi) systems are host-encoded resistance mechanisms that disrupt critical stages in the lytic phage life cycle such as transcription, translation, DNA replication or phage DNA packaging, and have been extensively studied in *L. lactis* [129, 195]. Abi-mediated resistance typically culminates in the death of the infected host cell in order to limit the release of progeny particles, thus

protecting the neighbouring bacterial population. Currently, twenty-three Abi systems (AbiA-AbiZ) are known for *L. lactis*, which, with the exception of AbiN and AbiV, are all plasmid-encoded [129, 195-197]. The presence of Abi systems was first identified due to the protective effect that certain lactococcal plasmids (such as pTR2030 and pIL105) have against phage infection, by causing a decreased burst size and an altered phage plaque morphology [198, 199]. Subsequently, plasmids that conferred such resistance to infecting phages were digested with restriction endonucleases and the fragments cloned into suitable shuttle vectors. The various recombinant derivatives were then screened to determine if a particular fragment provided phage resistance as observed for AbiE and AbiF encoded on the lactococcal plasmid pNP40 [200].

1.5.7 CRISPR/Cas Systems

CRISPR and CRISPR associated genes (Cas) form an acquired adaptive immunity system against foreign genetic elements in prokaryotes [201-203]. CRISPR systems are composed of a series of conserved repeats which are separated by protospacers, variable sequences involved in target recognition, an A-T rich leader region located at the 5' end of the CRISPR locus and *cas* genes [204]. CRISPR systems play an important role in phage-resistance in dairy starter strains [205] and furthermore, CRISPR systems can be used as a tool for the typing and comparative analyses of strains of *S. thermophilus* [201]. CRISPR typing of *S. thermophilus* performed by Horvath *et al*. [201], based on a combination of primers targeting conserved regions and Sanger sequencing resulted in the identification of CRISPR3 and demonstrated the diversity of CRISPR systems across 124 *S. thermophilus* strains. To date, there have been four different types of CRISPR loci

identified in *S. thermophilus*, CRISPR(1-4) [206]. In *L. lactis* only one CRISPR/Cas locus has been identified, being present on plasmid pKLM (though it is unable to incorporate new spacers) [201]. However, PCR-based screening of 400 lactococcal strains in this latter study also identified a further four strains with putative CRISPR systems indicating that continued genome sequencing is likely to result in future identification of CRISPRs in *L. lactis* [138].

1.6 Phage-host interactions of lactococci

The infection of lactococcal strains by phages, whether temperate or lytic results in the co-evolution of both the phage and the host populations. One of the major drivers of this co-evolution is the presence of bacterial host defence mechanisms (discussed above), which forces the phage population to circumvent these systems, and which in turn promotes further innovations within the bacterial lineage [208 - 209]. The initial interaction between an infecting phage and the host strain involves the attachment of the phage to its cognate receptor on the host cell surface. As discussed previously for the c2 group, this is represented by either of two membrane proteins termed Pip and YjaE [167 - 170]. In contrast, the host receptors for the P335, 936, P087, 949, 1358 and P087 groups have all been identified as saccharides that are covalently bound to the lactococcal cell wall [154, 171, 210-213] and has led to significant research interest in the area of lactococcal CWPS biochemistry and genetics [154, 171].

1.6.1 *L. lactis* cell wall polysaccharide

The LAB cell wall represents a complex structure comprised of a thick peptidoglycan layer, teichoic acids, cell wall polysaccharides (CWPS) and various surface carbohydrates [214]. *L. lactis* displays a smooth cell surface with CWPS homogeneously distributed across it, whereas mutants lacking CWPS show periodic bands of peptidoglycan running parallel to the short axis of the cell [215, 216]. The *L. lactis* genetic locus that encodes the CWPS biosynthetic machinery (called *cwps*) varies between strains and this diversity has allowed the classification of *cwps* and the associated strains into three types: A, B, and C [154] and a fourth class U composed of unknown CWPS types. Type C strains have been further classified into

five subtypes, C_{1-5}, based on variability in the amino acid identity of encoded glycosyltransferases in the variable region of the *cwps* locus (Fig. 1.3) [171]. The three major CWPS types are differentiated based on genotype differences in the *cwps* locus [154]. The *cwps* locus in *L. lactis* is typically ~25-30 Kbp in length, and comprises a conserved region and a variable region, the latter governing the CWPS type. The variable region typically contains a number of genes encoding predicted glycosyltransferases, variations in which are believed to govern both the type of sugar to be incorporated and the glycosidic connection to preceding sugars, resulting in the glycan diversity [101, 217].

As mentioned above lactococcal strains can be divided into three types based on their variability of their respective *cwps* locus (type A, B and C) [154]. Using a multiplex PCR, a collection of lactococcal strains can be classified to one of the three CWPS groups with primers based on the type-specific genetic elements including a glycosyltransferase-encoding gene (type A), NAD dependent epimerase-encoding gene (type B) and a surface membrane protein-encoding gene (type C). This rapid approach is useful in classifying the CWPS biosynthesis cluster, in particular from a phage sensitivity prediction standpoint. It may also be used for the purpose of selecting a blend of strains of varying CWPS types so as to avoid phages infecting multiple strains of that blend. The biochemical CWPS structures of some strains have been characterised and show consistency with the genetic differentiation of the strains based on the predicted variable glycosyltransferase-encoding gene composition of the corresponding *cwps* locus [154].

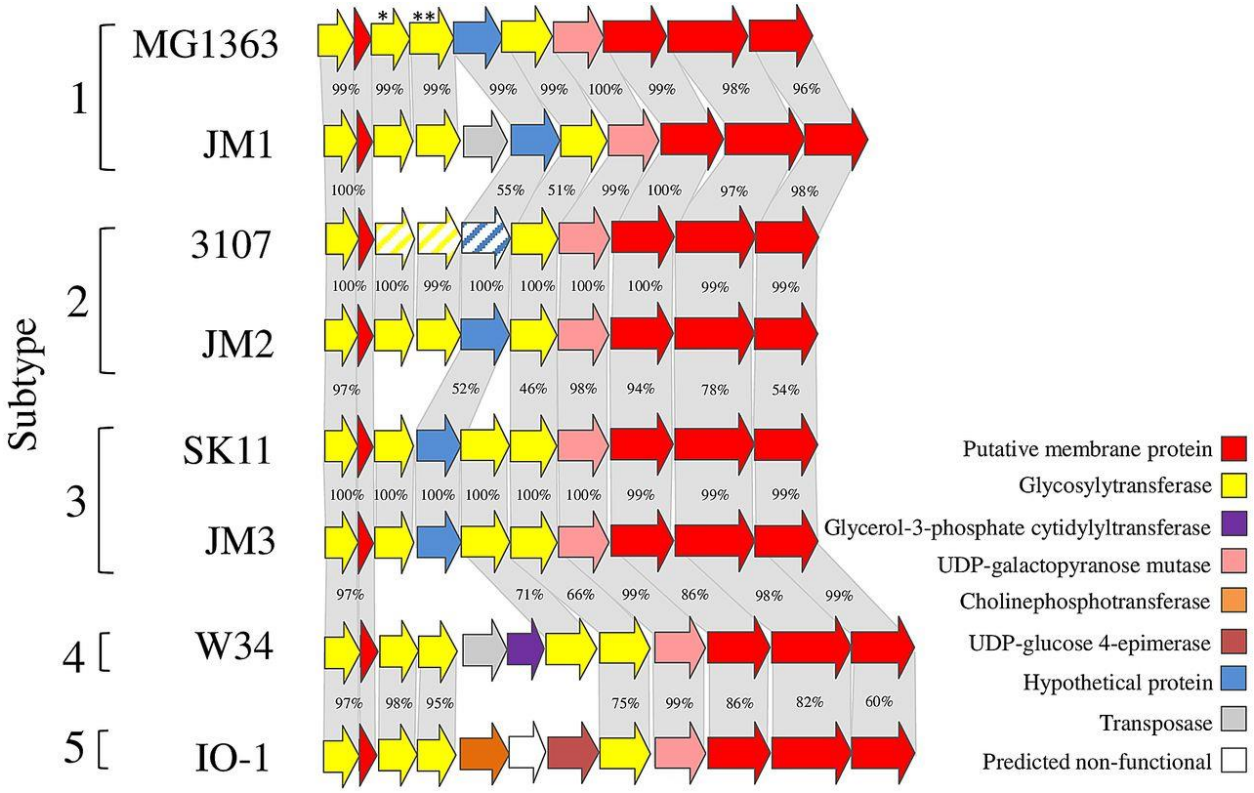


Figure 1.3: Comparison of the variable regions of the type C CWPS biosynthesis in *L. lactis*. (Taken from [171]).
Variable regions from the type C CWPS biosynthesis cluster of lactococcal strains MG1363, JM1, 3107, JM2, SK11, JM3, W34, and IO-1. Shaded boxes indicate homology based on nucleotide identity. The five subtypes (C1 to C5) of the C genotype are also indicated.

1.6.2 Lactococcal prophage

As discussed in (Section 1.5.2), prophages are of particular concern to fermentations as these may pose the risk of becoming active during the fermentation process leading to partial or complete culture lysis [53]. These phages integrate and silently replicate within the host's chromosome, posing the risk of excision. Prophage induction can culminate in both positive and negative effects within fermentations: it may cause unwanted/premature lysis leading to poor quality or loss of product, or, conversely, it may provoke phage-mediated cell lysis at specific points in the ripening process that can be favourable due to the release of intracellular enzymes involved in flavour development [218].

Prophages are widespread within the currently sequenced lactococcal genomes. For example, strains MG1363 and IL1403 each possess six prophage-encoding regions [3, 33, 219]. Previous work has indicated variable prophage induction profiles for *L. lactis* MG1363 [33, 219, 220] and positive induction of two prophage elements from *L. lactis* IL1403 [221]. Induction of the lactococcal strains ASCC890310 and ASCC890049 resulted in the release of phages with similarity to P335 sub-group I (BK5-T-like) and sub-group II (TP901-1-like) phages, respectively, among others [220].

While prophages are considered a threat to the dairy production process, their presence may also confer some competitive advantages on the host. The provision of prophage-encoded phage resistance systems is one of the best examples, with systems such as Sie_{2009} identified in the temperate phage Tuc2009 conferring resistance against certain members of the 936 group of lactococcal phages [160, 166, 177]. However, the opposite may also be true with previous studies demonstrating that the presence of prophages may lead to a competitive advantage for infecting

lytic phage. The P335 phage ul36 was previously observed to circumvent two Abi systems resident on the genome of *L. lactis* SMQ86, AbiK and AbiT, by recombining with a resident prophage to produce progeny with altered receptor binding proteins and baseplate components [222]. Genome sequencing of starter cultures will allow us to readily identify prophages and assess potential risks, so as to provide a rational basis for starter culture selection.

1.7 Future Directions

Metagenomics is a useful tool to assess the diversity of complex microbial communities and functional properties of their dominant populations [223]. In dairy applications such as the production of cheeses, these populations are often complex and not well characterized [224]. While metagenomics has previously been applied to all manner of niches (human microbiota, soil, water) it has played a limited role in dairy fermentations with a small number of studies published to date [223]. Future work with these technologies should expand our knowledge of the complex communities of bacterial hosts, phages and prophages within dairy fermentations.

While it is likely that "omics"-based technologies will never completely replace traditional culture-based methods, there is a vast array of knowledge to be gained from integrating these disciplines. Small-scale trial fermentations will continue to be the only genuine test to determine the performance of starter cultures within an industrial setting, though it is an impractical technique for screening large culture banks. Recent advances in NGS technologies have ensured that sequencing is a suitable approach in order to limit the number of potential candidates for such trials, and to reduce screening times and labour intensive cultivation techniques.

Genome decay and redundancy, as highlighted in dairy lactococcal isolates [1, 28, 68], coupled to *cremoris* type strains which are believed to be descended from a few closely related lineages [20], are factors likely to limit the selection of novel starter strains in the future. This is exacerbated by the likelihood of large redundancies in culture collections and the differentiation of many of these strains. Additionally, the possibility of incorrect phenotype/genotype association, such as the plasmid-encoded citrate metabolism trait, can only be resolved by complete genome sequencing.

1.8 Summary of thesis contents

In this thesis the so-called SMRT sequencing methodology was applied to sequence sixteen *L. lactis* isolates in order to facilitate an in depth comparative and functional genomic analysis of this LAB taxon with particular emphasis placed on dairy traits.

Chapter II describes the phenotypic characterisation of twenty dairy *L. lactis* strains in terms of their contribution to flavour development in cheese fermentations. Chapter III describes whole genome sequencing of sixteen *L. lactis* strains on which all subsequent chapters are based and details a comparative genomics analysis of these newly sequenced chromosomes combined with fourteen publicly available *L. lactis* genomes. Chapter IV describes the lactococcal plasmidome including sixty seven newly sequenced plasmids and investigates the technologically relevant traits encoded by this plasmidome. Chapter V describes the base modifications and restriction-modification systems of the sixteen newly sequenced *L. lactis* genomes. Chapter VI describes the analysis of predicted prophages of thirty lactococcal strains and investigates the potential risk of phage excision.

1.8.1 Aims and objectives

- Determine the genome sequences of representative lactococcal strains and their plasmid complements
- Conduct a comparative genomic study of various dairy lactococcal starter culture strains, some of which are used for the production of reduced fat Cheddar cheese

- Functional analysis of these strains with particular reference to flavour development, enzymatic activity and growth during the cheese production cycle
- Identify links to the phenotypic characteristics of low fat Cheddar starter cultures with the genomic composition of such lactococcal strains
- Investigate *L. lactis* genomes for the presence of prophages and restriction-modification systems

1.9 References

1. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N *et al*: Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci. 2006, 103(42):15611-15616.
2. Kandler O: Carbohydrate metabolism in lactic acid bacteria. Antonie van Leeuwenhoek 1983, 49(3):209-224.
3. Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP: Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. J Bacteriol. 2007, 189(8):3256-3270.
4. Daniel C, Roussel Y, Kleerebezem M, Pot B: Recombinant lactic acid bacteria as mucosal biotherapeutic agents. Trends Biotechnol. 2011, 29(10):499-508.
5. Leroy F, De Vuyst L: Lactic acid bacteria as functional starter cultures for the food fermentation industry. Trends Food Sci Tech. 2004, 15(2):67-78.
6. Hammes WP, Bantleon A, Min S: Lactic acid bacteria in meat fermentation. FEMS Microbiol Rev. 1990, 7(1-2):165-173.
7. Tamang JP, Tamang B, Schillinger U, Franz CM, Gores M, Holzapfel WH: Identification of predominant lactic acid bacteria isolated from traditionally fermented vegetable products of the Eastern Himalayas. Int J Food Microbiol. 2005, 105(3):347-356.
8. Wibowo D, Eschenbruch R, Davis C, Fleet G, Lee T: Occurrence and growth of lactic acid bacteria in wine: a review. Am J Enol Vitic. 1985, 36(4):302-313.

9. Cotter PD, Hill C, Ross RP: Bacteriocins: developing innate immunity for food. Nat Rev Microbiol. 2005, 3(10):777-788.

10. Nes IF, Johnsborg O: Exploration of antimicrobial potential in LAB by genomics. Curr Opin Biotechnol. 2004, 15(2):100-104.

11. de Vos WM: Systems solutions by lactic acid bacteria: from paradigms to practice. Microb Cell Fact. 2011, 10(1):1.

12. Stanton C, Ross RP, Fitzgerald GF, Van Sinderen D: Fermented functional foods based on probiotics and their biogenic metabolites. Curr Opin Biotechnol. 2005, 16(2):198-203.

13. Spacova I, Seys S, Dockx R, Petrova M, Devos F, Kasran A, Vanoirbeek J, Ceuppens J, Vanderleyden J, Lebeer S: Oral delivery of allergen-specific vaccination using recombinant *Lactobacillus rhamnosus* GG in a murine model of birch pollen allergic asthma. 2015.

14. Guo S, Yan W, McDonough SP, Lin N, Wu KJ, He H, Xiang H, Yang M, Moreira MAS, Chang Y-F: The recombinant *Lactococcus lactis* oral vaccine induces protection against *C. difficile* spore challenge in a mouse model. Vaccine. 2015, 33(13):1586-1595.

15. Rosales-Mendoza S, Angulo C, Meza B: Food-grade organisms as vaccine biofactories and oral delivery vehicles. Trend Biotechnol. 2016, 34(2):124-136.

16. Schleifer K, Kilpper-Bälz R: Molecular and chemotaxonomic approaches to the classification of streptococci, enterococci and lactococci: a review. Sys Appl Microbiol. 1987, 10(1):1-19.

17. Schleifer K, Kraus J, Dvorak C, Kilpper-Bälz R, Collins M, Fischer W: Transfer of *Streptococcus lactis* and related *Streptococci* to the genus *Lactococcus* gen. nov. Sys Appl Microbiol. 1985, 6(2):183-195.

18. Latorre-Guzman BA, Kado CI, Kunkee RE: *Lactobacillus hordniae*, a New Species from the Leafhopper (*Hordnia circellata*). Int J Sys Evol Microbiol. 1977, 27(4):362-370.

19. Perez T, Balcazar JL, Peix A, Valverde A, Velazquez E, de Blas I, Ruiz-Zarzuela I: *Lactococcus lactis* subsp. *tructae* subsp. nov. isolated from the intestinal mucus of brown trout (*Salmo trutta*) and rainbow trout (*Oncorhynchus mykiss*). Int J Sys Evol Microbiol. 2011, 61(Pt 8):1894-1898.

20. Kelly WJ, Ward LJH, Leahy SC: Chromosomal Diversity in *Lactococcus lactis* and the Origin of Dairy Starter Cultures. Gen Biol Evol. 2010, 2:729-744.

21. Rademaker JL, Herbet H, Starrenburg MJ, Naser SM, Gevers D, Kelly WJ, Hugenholtz J, Swings J, van Hylckama Vlieg JE: Diversity analysis of dairy and nondairy *Lactococcus lactis* isolates, using a novel multilocus sequence analysis scheme and (GTG) 5-PCR fingerprinting. Appl Environ Microbiol. 2007, 73(22):7128-7137.

22. Górecki RK, Koryszewska-Bagińska A, Gołębiewski M, Żylińska J, Grynberg M, Bardowski JK: Adaptative Potential of the *Lactococcus Lactis* IL594 Strain Encoded in Its 7 Plasmids. PLoS ONE. 2011, 6(7):e22238.

23. Passerini D, Beltramo C, Coddeville M, Quentin Y, Ritzenthaler P, Daveran-Mingot M-L, Le Bourgeois P: Genes but not genomes reveal bacterial domestication of *Lactococcus lactis*. PLoS One. 2010, 5(12):e15306.

24. Cavanagh D, Casey A, Altermann E, Cotter PD, Fitzgerald GF, McAuliffe O: Evaluation of *Lactococcus lactis* Isolates from Nondairy Sources with Potential Dairy Applications Reveals Extensive Phenotype-Genotype Disparity and Implications for a Revised Species. Appl Environ Microbiol. 2015, 81(12):3961-3972.

25. Tamura K, Nei M: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993, 10(3):512-526.

26. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013, 30(12):2725-2729.

27. Ainsworth S, Mahony J, van Sinderen D: The Plasmid Complement of *Lactococcus lactis* UC509.9 Encodes Multiple Bacteriophage Resistance Systems. Appl Environ Microbiol. 2014, 80(14):4341-4349.

28. Ainsworth S, Zomer A, de Jager V, Bottacini F, van Hijum SAFT, Mahony J, van Sinderen D: Complete Genome of *Lactococcus lactis* subsp. *cremoris* UC509.9, Host for a Model Lactococcal P335 Bacteriophage. Gen Announc.2013, 1(1):e00119-00112.

29. Linares DM, Kok J, Poolman B: Genome sequences of *Lactococcus lactis* MG1363 (revised) and NZ9000 and comparative physiological studies. J Bacteriol.2010, 192(21):5806-5812.

30. Kelly WJ, Davey GP, Ward LJ: Characterization of lactococci isolated from minimally processed fresh fruit and vegetables. Int J Food Microbiol. 1998, 45(2):85-92.

31. Siezen RJ, Starrenburg MJ, Boekhorst J, Renckens B, Molenaar D, van Hylckama Vlieg JE: Genome-scale genotype-phenotype matching of two *Lactococcus lactis* isolates from plants identifies mechanisms of adaptation to the plant niche. Appl Environ Microbiol. 2008, 74(2):424-436.

32. Cavanagh D, Fitzgerald GF, McAuliffe O: From field to fermentation: the origins of *Lactococcus lactis* and its domestication to the dairy environment. Food Microbiol. 2015, 47:45-61.

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

34. Ayad EH, Verheul A, de Jong C, Wouters JT, Smit G: Flavour forming abilities and amino acid requirements of *Lactococcus lactis* strains isolated from artisanal and non-dairy origin. Int.Dairy J. 1999, (9.10): 725-735.

35. Siezen RJ, Bayjanov J, Renckens B, Wels M, van Hijum SA, Molenaar D, van Hylckama Vlieg JE: Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium. J Bacteriol.2010, 192(10):2649-2650.

36. Bolotin A, Quinquis B, Ehrlich SD, Sorokin A: Complete genome sequence of *Lactococcus lactis* subsp. *cremoris* A76. J Bacteriol.2012, 194(5):1241-1242.

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

38. Kato H, Shiwa Y, Oshima K, Machii M, Araya-Kojima T, Zendo T, Shimizu-Kadota M, Hattori M, Sonomoto K, Yoshikawa H: 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.2012, 194(8):2102-2103.

39. Ainsworth S, Zomer A, de Jager V, Bottacini F, van Hijum SA, Mahony J, van Sinderen D: Complete genome of *Lactococcus lactis* subsp. *cremoris* UC509. 9, host for a model lactococcal P335 bacteriophage. Gen Announc.2013, 1(1):e00119-00112.

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

41. Yang X, Wang Y, Huo G: Complete Genome Sequence of *Lactococcus lactis* subsp. *lactis* KLDS4. 0325. Gen Announc.2013, 1(6):e00962-00913.

42. Oliveira LC, Saraiva TD, Soares SC, Ramos RT, Sá PH, Carneiro AR, Miranda F, Freire M, Renan W, Júnior AF: Genome Sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-Producing Strain. Gen Announc.2014, 2(5):e00980-00914.

43. Zhao F, Ma H, Lu Y, Teng K, Kang X, Wang F, Yang X, Zhong J: Complete genome sequence of *Lactococcus lactis* S0, an efficient producer of nisin. J Biotechnol. 2015, 198:15-16.

44. McCulloch JA, de Oliveira VM, de Almeida Pina AV, Perez-Chaparro PJ, de Almeida LM, de Vasconcelos JM, de Oliveira LF, da Silva DE, Rogez HL, Cretenet M et al: Complete Genome Sequence of *Lactococcus lactis* Strain

AI06, an Endophyte of the Amazonian Acai Palm. Genome Announc 2014, 2(6).

45. Guellerin M, Passerini D, Fontagné-Faucher C, Robert H, Gabriel V, Loux V, Klopp C, Le Loir Y, Coddeville M, Daveran-Mingot M-L: Complete genome sequence of *Lactococcus lactis* subsp. *lactis* A12, a strain isolated from wheat sourdough. Gen Announc.2016, 4(5):e00692-00616.

46. Chaisson MJ, Brinza D, Pevzner PA: *De novo* fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 2009, 19(2):336-346.

47. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin J-F: Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011, 12(1):245.

48. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotech 2012, 30(5):434-439.

49. Metzker ML: Sequencing technologies - the next generation. Nat Rev Genet 2010, 11(1):31-46.

50. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al*: Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth. 2013, 10(6):563-569.

51. Kok J, Buist G, Zomer AL, Van Hijum SAFT, Kuipers OP: Comparative and functional genomics of lactococci. FEMS Microbiol Rev. 2005, 29(3):411-433.

52. Daveran-Mingot M-L, Campo N, Ritzenthaler P, Le Bourgeois P: A Natural Large Chromosomal Inversion in *Lactococcus lactis* Is Mediated by Homologous Recombination between Two Insertion Sequences. J Bacteriol.1998, 180(18):4834-4842.

53. Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M-C: Analysis of six prophages in *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage populations. Nuc Acids Res. 2001, 29(3):644-651.

54. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature methods. 2010, 7(6):461-465.

55. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J: Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nuc Acids Res. 2012, 40(4):e29.

56. Hodkinson BP, Grice EA: Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Advances in wound care. 2015.

57. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012, 13(1):341.

58. Mosher JJ, Bernberg EL, Shevchenko O, Kan J, Kaplan LA: Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S

rRNA genes from environmental samples. J Microbiol Methods. 2013, 95(2):175–181.

59. Schloss PD, Westcott SL, Jenior ML, Highlander SK: Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ PrePrints. 2015.

60. Huang Y-F, Chen S-C, Chiang Y-S, Chen T-H, Chiu K-P: Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol. 2012, 6(Suppl 2):S10.

61. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ: The methylomes of six bacteria. Nuc Acids Res. 2012, 40(22):11450-11462.

62. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H: Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nano. 2009, 4(4):265-270.

63. Edwards DJ, Holt KE: Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp. 2013, 3(1):2.

64. Beresford TP, Fitzsimons NA, Brennan NL, Cogan TM: Recent advances in cheese microbiology. Int Dairy J. 2001, 11(4–7):259-274.

65. Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D: The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. FEMS Microbiol Rev. 2014, 38(5):1066-1088.

66. Siezen RJ, Bayjanov JR, Felis GE, van der Sijde MR, Starrenburg M, Molenaar D, Wels M, van Hijum SA, van Hylckama Vlieg JE: Genome-scale diversity and niche adaptation analysis of *Lactococcus lactis* by comparative

genome hybridization using multi-strain arrays. Microbial biotechnology 2011, 4(3):383-402.

67. Price CE, Zeyniyev A, Kuipers OP, Kok J: From meadows to milk to mucosa – adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. FEMS Microbiol Rev. 2012, 36(5):949-971.

68. Goh YJ, Goin C, O'Flaherty S, Altermann E, Hutkins R: Specialized adaptation of a lactic acid bacterium to the milk environment: the comparative genomics of *Streptococcus thermophilus* LMD-9. Microb Cell Fact. 2011, 10 Suppl 1:S22.

69. Godon J-J, Delorme C, Bardowski J, Chopin M, Ehrlich SD, Renault P: Gene inactivation in *Lactococcus lactis*: branched-chain amino acid biosynthesis. J Bacteriol.1993, 175(14):4383-4390.

70. Delorme C, Godon J-J, Ehrlich SD, Renault P: Gene inactivation in *Lactococcus lactis*: histidine biosynthesis. J Bacteriol.1993, 175(14):4391-4399.

71. Cords BR, McKay LL, Guerry P: Extrachromosomal elements in group N *streptococci*. J Bacteriol 1974, 117(3):1149-1152.

72. McKay LL, Baldwin KA, Zottola EA: Loss of Lactose Metabolism in Lactic *Streptococci*. Appl Microbiol. 1972, 23(6):1090-1096.

73. Solopova A, Bachmann H, Teusink B, Kok J, Neves AR, Kuipers OP: A Specific Mutation in the Promoter Region of the Silent cel Cluster Accounts for the Appearance of Lactose-Utilizing *Lactococcus lactis* MG1363. Appl Environ Microbiol. 2012, 78(16):5612-5621.

74. Bissett DL, Anderson RL: Lactose and d-Galactose Metabolism in Group N *Streptococci*: Presence of Enzymes for Both the d-Galactose 1-Phosphate and d-Tagatose 6-Phosphate Pathways1. J Bacteriol.1974, 117(1):318.

75. Vaughan EE, Pridmore RD, Mollet B: Transcriptional Regulation and Evolution of Lactose Genes in the Galactose-Lactose Operon of *Lactococcus lactis* NCDO2054. J Bacteriol.1998, 180(18):4893-4902.

76. Ferrario C, Ricci G, Borgo F, Rollando A, Fortina MG: Genetic investigation within *Lactococcus garvieae* revealed two genomic lineages. FEMS Microbiol Lett. 2012, 332(2):153-161.

77. McSweeney PLH, Sousa MJ: Biochemical pathways for the production of flavour compounds in cheeses during ripening: A review. Lait 2000, 80(3):293-324.

78. Drider D, Bekal S, Prévost H: Genetic organization and expression of citrate permease in lactic acid bacteria. Genet Mol Res 2004, 3(2):271-281.

79. Shibamoto T: Diacetyl: Occurrence, Analysis, and Toxicity. J Agric Food Chem. 2014, 62(18):4048-4053.

80. McSweeney PLH: Biochemistry of cheese ripening. Int J Dairy Technol. 2004, 57(2-3):127-144.

81. Steele J, Broadbent J, Kok J: Perspectives on the contribution of lactic acid bacteria to cheese flavor development. Curr Opin Biotechnol. 2013, 24(2):135-141.

82. Yu W, Gillies K, Kondo JK, Broadbent JR, McKay LL: Loss of Plasmid-Mediated Oligopeptide Transport System in Lactococci: Another Reason for Slow Milk Coagulation. Plasmid. 1996, 35(3):145-155.

83. Guédon E, Renault P, Ehrlich SD, Delorme C: Transcriptional Pattern of Genes Coding for the Proteolytic System of *Lactococcus lactis* and Evidence for Coordinated Regulation of Key Enzymes by Peptide Supply. J Bacteriol. 2001, 183(12):3614-3622.

84. Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM: Complete sequences of four plasmids of *Lactococcus lactis* subsp. *cremoris* SK11 reveal extensive adaptation to the dairy environment. Appl Environ Microbiol. 2005, 71(12):8371-8382.

85. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: PHAST: a fast phage search tool. Nuc Acids Res. 2011:gkr485.

86. Liu M, Bayjanov J, Renckens B, Nauta A, Siezen R: The proteolytic system of lactic acid bacteria revisited: a genomic comparison. BMC Genomics. 2010, 11(1):36.

87. Verger R: 'Interfacial activation'of lipases: facts and artifacts. Trend Biotechnol. 1997, 15(1):32-38.

88. Nardi M, Fiez-Vandal C, Tailliez P, Monnet V: The EstA esterase is responsible for the main capacity of *Lactococcus lactis* to synthesize short chain fatty acid esters in vitro. J Appl Microbiol. 2002, 93(6):994-1002.

89. Hassan AN, Ipsen R, Janzen T, Qvist KB: Microstructure and Rheology of Yogurt Made with Cultures Differing Only in Their Ability to Produce Exopolysaccharides. J Dairy Sci. 2003, 86(5):1632-1638.

90. Hassan AN: ADSA Foundation Scholar Award: Possibilities and Challenges of Exopolysaccharide-Producing Lactic Cultures in Dairy Foods. J Dairy Sci. 2008, 91(4):1282-1298.

91. Hassan AN, Frank JF, Farmer MA, Schmidt KA, Shalabi SI: Formation of Yogurt Microstructure and Three-Dimensional Visualization as Determined by Confocal Scanning Laser Microscopy. J Dairy Sci. 1995, 78(12):2629-2636.

92. Borucki MK, Peppin JD, White D, Loge F, Call DR: Variation in biofilm formation among strains of *Listeria monocytogenes*. Appl Environ Microbiol. 2003, 69(12):7336-7342.

93. Kranenburg Rv, Marugg JD, Van Swam II, Willem NJ, De Vos WM: Molecular characterization of the plasmid-encoded eps gene cluster essential for exopolysaccharide biosynthesis in *Lactococcus lactis*. Mol Microbiol. 1997, 24(2):387-397.

94. Vedamuthu ER, Neville JM: Involvement of a plasmid in production of ropiness (mucoidness) in milk cultures by *Streptococcus cremoris* MS. Appl Environ Microbiol. 1986, 51(4):677-682.

95. von Wright A, Tynkkynen S: Construction of *Streptococcus lactis* subsp. *lactis* strains with a single plasmid associated with mucoid phenotype. Appl Environ Microbiol. 1987, 53(6):1385-1386.

96. Neve H, Geis A, Teuber M: Plasmid-encoded functions of ropy lactic acid streptococcal strains from Scandinavian fermented milk. Biochimie. 1988, 70(3):437-442.

97. Kleerebezem M, van Kranenburg R, Tuinier R, Boels IC, Zoon P, Looijesteijn E, Hugenholtz J, de Vos WM: Exopolysaccharides produced by *Lactococcus lactis*: from genetic engineering to improved rheological properties? In: Lactic Acid Bacteria: Genetics, Metabolism and Applications. Springer; 1999: 357-365.

98. Nakajima H, Hirota T, Toba T, Itoh T, Adachi S: Structure of the extracellular polysaccharide from slime-forming *Lactococcus lactis* subsp. *cremoris* SBT 0495. Carbohydr Res. 1992, 224:245-253.

99. Van Kranenburg R, Van Swam II, Marugg JD, Kleerebezem M, de Vos WM: Exopolysaccharide biosynthesis in *Lactococcus lactis* NIZO B40: functional analysis of the glycosyltransferase genes involved in synthesis of the polysaccharide backbone. J Bacteriol.1999, 181(1):338-340.

100. Cartee RT, Forsee WT, Bender MH, Ambrose KD, Yother J: CpsE from type 2 *Streptococcus pneumoniae* catalyzes the reversible addition of glucose-1-phosphate to a polyprenyl phosphate acceptor, initiating type 2 capsule repeat unit formation. J Bacteriol.2005, 187(21):7425-7433.

101. Yother J: Capsules of *Streptococcus pneumoniae* and other bacteria: paradigms for polysaccharide biosynthesis and regulation. Annu Rev Microbiol. 2011, 65:563-581.

102. Forde A, Fitzgerald GF: Analysis of exopolysaccharide (EPS) production mediated by the bacteriophage adsorption blocking plasmid, pCI658, isolated from *Lactococcus lactis* ssp. *cremoris* HO2. Int Dairy J. 1999, 9(7):465-472.

103. Makrides SC: Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiol Rev. 1996, 60(3):512-538.

104. Morello E, Bermudez-Humaran L, Llull D, Sole V, Miraglio N, Langella P, Poquet I: *Lactococcus lactis*, an efficient cell factory for recombinant protein production and secretion. J Mol Microbiol. and biotechnology 2007, 14(1-3):48-58.

105. Li W, Zhou X, Lu P: Bottlenecks in the expression and secretion of heterologous proteins in *Bacillus subtilis*. Res Microbiol. 2004, 155(8):605-610.

106. Westers L, Westers H, Quax WJ: *Bacillus subtilis* as cell factory for pharmaceutical proteins: a biotechnological approach to optimize the host organism. BBA Mol Cell Res.2004, 1694(1):299-310.

107. van Asseldonk M, de Vos WM, Simons G: Functional analysis of the *Lactococcus lactis* usp45 secretion signal in the secretion of a homologous proteinase and a heterologous α-amylase. Mol Gen Genet. 1993, 240(3):428-434.

108. de Vos WM: Gene expression systems for lactic acid bacteria. Curr Opin Microbiol. 1999, 2(3):289-295.

109. de Vos WM, Kleerebezem M, Kuipers OP: Expression systems for industrial Gram-positive bacteria with low guanine and cytosine content. Curr Opin Biotechnol. 1997, 8(5):547-553.

110. De Ruyter P, Kuipers OP, Beerthuyzen MM, van Alen-Boerrigter I, De Vos W: Functional analysis of promoters in the nisin gene cluster of *Lactococcus lactis*. J Bacteriol.1996, 178(12):3434-3439.

111. Mierau I, Kleerebezem M: 10 years of the nisin-controlled gene expression system (NICE) in *Lactococcus lactis*. Appl Microbiol Biotechnol. 2005, 68(6):705-717.

112. Kuipers OP, de Ruyter PG, Kleerebezem M, de Vos WM: Quorum sensing-controlled gene expression in lactic acid bacteria. J Biotechnol. 1998, 64(1):15-21.

113. Joutsjoki V, Luoma S, Tamminen M, Kilpi M, Johansen E, Palva A: Recombinant *Lactococcus* starters as a potential source of additional peptidolytic activity in cheese ripening. J Appl Microbiol. 2002, 92(6):1159-1166.

114. Wegmann U, Klein J, Drumm I, Kuipers O, Henrich B: Introduction of peptidase genes from *Lactobacillus delbrueckii* subsp. *lactis* into *Lactococcus lactis* and controlled expression. Appl Environ Microbiol. 1999, 65(11):4729-4733.

115. Sybesma W, Starrenburg M, Kleerebezem M, Mierau I, de Vos WM, Hugenholtz J: Increased Production of Folate by Metabolic Engineering of *Lactococcus lactis*. Appl Environ Microbiol. 2003, 69(6):3069-3076.

116. Sheng J, Ling P, Wang F: Constructing a recombinant hyaluronic acid biosynthesis operon and producing food-grade hyaluronic acid in *Lactococcus lactis*. J Ind Microbiol Biotechnol. 2015, 42(2):197-206.

117. Steidler L, Neirynck S, Huyghebaert N, Snoeck V, Vermeire A, Goddeeris B, Cox E, Remon JP, Remaut E: Biological containment of genetically modified *Lactococcus lactis* for intestinal delivery of human interleukin 10. Nat Biotechnol. 2003, 21(7):785-789.

118. Landete JM: A review of food-grade vectors in lactic acid bacteria: from the laboratory to their application. Crit Rev Biotechnol. 2016:1-13.

119. Nouaille S, Ribeiro LA, Miyoshi A, Pontes D, Le Loir Y, Oliveira SC, Langella P, Azevedo V: Heterologous protein production and delivery systems for *Lactococcus lactis*. Genet Mol Res. 2003, 2(1):102-111.

120. Xin K-Q, Hoshino Y, Toda Y, Igimi S, Kojima Y, Jounai N, Ohba K, Kushiro A, Kiwaki M, Hamajima K: Immunogenicity and protective efficacy

of orally administered recombinant *Lactococcus lactis* expressing surface-bound HIV Env. Blood. 2003, 102(1):223-228.

121. Bermúdez-Humarán LG: *Lactococcus lactis* as a live vector for mucosal delivery of therapeutic proteins. Hum Vaccin. 2009, 5(4):264-267.

122. Bermúdez-Humarán LG, Aubry C, Motta J-P, Deraison C, Steidler L, Vergnolle N, Chatel J-M, Langella P: Engineering lactococci and *lactobacilli* for human health. Curr Opin Microbiol. 2013, 16(3):278-283.

123. Bermúdez-Humarán LG, Kharrat P, Chatel J-M, Langella P: Lactococci and *lactobacilli* as mucosal delivery vectors for therapeutic proteins and DNA vaccines. Microb Cell Fact. 2011, 10(1):1.

124. Bermúdez-Humarán LG, Cortes-Perez NG, Le Loir Y, Alcocer-González JM, Tamez-Guerra RS, de Oca-Luna RM, Langella P: An inducible surface presentation system improves cellular immunity against human papillomavirus type 16 E7 antigen in mice after nasal administration with recombinant lactococci. J Med Microbiol. 2004, 53(5):427-433.

125. Pinto UM, Pappas KM, Winans SC: The ABCs of plasmid replication and segregation. Nat Rev Microbiol. 2012, 10(11):755-765.

126. Grohmann E, Muth G, Espinosa M: Conjugative Plasmid Transfer in Gram-Positive Bacteria. Microbiol Mol Biol Rev. 2003, 67(2):277-301.

127. Mills S, McAuliffe OE, Coffey A, Fitzgerald GF, Ross RP: Plasmids of lactococci – genetic accessories or genetic necessities? FEMS Microbiol Rev. 2006, 30(2):243-273.

128. Nomura M, Kobayashi M, Narita T, Kimoto-Nira H, Okamoto T: Phenotypic and molecular characterization of *Lactococcus lactis* from milk and plants. J Appl Microbiol. 2006 Aug 1;101(2):396-405.

129. Hill CO, Miller LA, Klaenhammer TR: In vivo genetic exchange of a functional domain from a type II A methylase between lactococcal plasmid pTR2030 and a virulent bacteriophage. J Bacteriol. 1991, Jul 1;173(14):4363-70.

130. Leenhouts KJ, Tolner B, Bron S, Kok J, Venema G, Seegers JF: Nucleotide sequence and characterization of the broad-host-range lactococcal plasmid pWVO1. Plasmid. 1991, 26(1):55-66.

131. Kiewiet R, Bron S, Jonge K, Venema G, Seegers JF: Theta replication of the lactococcal plasmid pWVO2. Mol Microbiol. 1993, 10(2):319-327.

132. del Solar G, Moscoso M, Espinosa M: In vivo definition of the functional origin of replication ori^{+} of the promiscuous plasmid pLS1. Mol Gen Genet. 1993, 237(1-2):65-72.

133. Del Solar G, Giraldo R, Ruiz-Echevarría MJ, Espinosa M, Díaz-Orejas R: Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev. 1998, 62(2):434-464.

134. Émond É, Lavallée R, Drolet G, Moineau S, LaPointe G: Molecular characterization of a theta replication plasmid and its use for development of a two-component food-grade cloning system for *Lactococcus lactis*. Appl Environ Microbiol. 2001, 67(4):1700-1709.

135. Wegmann U, Overweg K, Jeanson S, Gasson M, Shearman C: Molecular characterization and structural instability of the industrially important composite metabolic plasmid pLP712. Microbiol. 2012, 158(12):2936-2945.

136. Ammann A, Neve H, Geis A, Heller KJ: Plasmid transfer via transduction from *Streptococcus thermophilus* to *Lactococcus lactis*. J Bacteriol. 2008, 190(8):3083-3087.

137. O'Driscoll J, Glynn F, Fitzgerald GF, Sinderen Dv: Sequence Analysis of the Lactococcal Plasmid pNP40: a Mobile Replicon for Coping with Environmental Hazards. J Bacteriol. 2006, 188(18):6629-6639.

138. Millen AM, Horvath P, Boyaval P, Romero DA: Mobile CRISPR/Cas-Mediated Bacteriophage Resistance in *Lactococcus lactis*. PLoS ONE. 2012, 7(12):e51663.

139. Hofreuter D, Haas R: Characterization of two cryptic *Helicobacter pylori* plasmids: a putative source for horizontal gene transfer and gene shuffling. J Bacteriol. 2002, 184(10):2755-2766.

140. Dobson A, Cotter PD, Ross RP, Hill C: Bacteriocin Production: a Probiotic Trait? Appl Environ Microbiol. 2012, 78(1):1-6.

141. Kékessy DA, Piguet JD: New Method for Detecting Bacteriocin Production. Appl Microbiol. 1970, 20(2):282-283.

142. Barefoot SF, Klaenhammer TR: Detection and activity of lactacin B, a bacteriocin produced by *Lactobacillus acidophilus*. Appl Environ Microbiol. 1983, 45(6):1808-1815.

143. Yang R, Johnson MC, Ray B: Novel method to extract large amounts of bacteriocins from lactic acid bacteria. Appl Environ Microbiol. 1992, 58(10):3355-3359.

144. van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP: BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nuc Acids Res. 2013, 41(W1):W448-W453.

145. Whitehead HR, Cox G: The Occurrence of Bacteriophage in Cultures of Lactic *Streptococci*: A Preliminary Note. New Zealand J Dairy Sci Technol.1935:313-320.

146. Allison GE, Klaenhammer TR: Phage resistance mechanisms in lactic acid bacteria. Int Dairy J. 1998, 8(3):207-226.

147. Hyman P, Abedon ST: Bacteriophage host range and bacterial resistance. Adv Appl Microbiol. 2010 Dec 31;70:217-48.

148. Maniloff J, Ackermann H-W: Taxonomy of bacterial viruses: establishment of tailed virus genera and the other *Caudovirales*. Arch Virol. 1998, 143(10):2051-2063.

149. Deveau H, Labrie SJ, Chopin M-C, Moineau S: Biodiversity and classification of lactococcal phages. Appl Environ Microbiol. 2006, 72(6):4338-4346.

150. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D: Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. Sci Rep.2016, 6:21345.

151. Mahony J, Murphy J, van Sinderen D: Lactococcal 936-type phages and dairy fermentation problems: from detection to evolution and prevention. Front Microbiol. 2012, 3:335.

152. Castro-Nallar E, Chen H, Gladman S, Moore SC, Seemann T, Powell IB, Hillier A, Crandall KA, Chandry PS: Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage. Gen Biol Evol. 2012, 4(3):382-393.

153. Brüssow H, Desiere F: Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. Mol Microbiol. 2001, 39(2):213-223.

154. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C: Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. Appl Environ Microbiol. 2013, 79(14):4385-4392.

155. Murphy J, Klumpp J, Mahony J, Mary O, Nauta A, van Sinderen D: Methyltransferases acquired by lactococcal 936-type phage provide protection against restriction endonuclease activity. BMC Genomics. 2014, 15(1):831.

156. Murphy J, Royer B, Mahony J, Hoyles L, Heller K, Neve H, Bonestroo M, Nauta A, van Sinderen D: Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type cheese-producing plants. J Dairy Sci. 2013, 96(8):4945-4957.

157. Murphy J, Mahony J, Ainsworth S, Nauta A, van Sinderen D: Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. Appl Environ Microbiol. 2013, 79(24):7547-7555.

158. Millen AM, Romero DA: Genetic determinants of lactococcal C2 viruses for host infection and their role in phage evolution. J Gen Virol. 2016, 97(8):1998-2007.

159. Lubbers MW, Waterfield NR, Beresford T, Le Page R, Jarvis AW: Sequencing and analysis of the prolate-headed lactococcal bacteriophage c2

genome and identification of the structural genes. Appl Environ Microbiol. 1995, 61(12):4348-4356.

160. McGrath S, Fitzgerald GF, Sinderen Dv: Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Mol Microbiol. 2002, 43(2):509-520.

161. Moineau S, Fortier J, Ackermann H-W, Pandian S: Characterization of lactococcal bacteriophages from Quebec cheese plants. Can J Microbiol. 1992, 38(9):875-882.

162. Mahony J, Martel B, Tremblay DM, Neve H, Heller KJ, Moineau S, van Sinderen D: Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. Appl Environ Microbiol. 2013, 79(14):4401-4409.

163. Collins B, Bebeacua C, Mahony J, Blangy S, Douillard FP, Veesler D, Cambillau C, van Sinderen D: Structure and Functional Analysis of the Host Recognition Device of Lactococcal Phage Tuc2009. J Virol. 2013, 87(15):8429-8440.

164. Sozhamannan S, Chute MD, McAfee FD, Fouts DE, Akmal A, Galloway DR, Mateczun A, Baillie LW, Read TD: The *Bacillus anthracis* chromosome contains four conserved, excision-proficient, putative prophages. BMC Microbiol. 2006, 6(1):34.

165. Fouts DE: Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nuc Acids Res. 2006, 34(20):5839-5851.

166. Labrie SJ, Samson JE, Moineau S: Bacteriophage resistance mechanisms. Nat Rev Micro. 2010, 8(5):317-327.

167. Geller BL, Ivey RG, Trempy JE, Hettinger-Smith B: Cloning of a chromosomal gene required for phage infection of *Lactococcus lactis* subsp. *lactis* C2. J Bacteriol. 1993, 175(17):5510-5519.

168. Monteville MR, Ardestani B, Geller BL: Lactococcal bacteriophages require a host cell wall carbohydrate and a plasma membrane protein for adsorption and ejection of DNA. Appl Environ Microbiol. 1994, 60(9):3204-3211.

169. Stuer-Lauridsen B, Janzen T, Schnabl J, Johansen E: Identification of the host determinant of two prolate-headed phages infecting *Lactococcus lactis*. Virol. 2003, 309(1):10-17.

170. Derkx PM, Janzen T, Sørensen KI, Christensen JE, Stuer-Lauridsen B, Johansen E: The art of strain improvement of industrial lactic acid bacteria without the use of recombinant DNA technology. Microb Cell Fact. 2014, 13(1):1.

171. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, Van Sinderen D: Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. MBio. 2014, 5(3):e00880-00814.

172. Lucey M, Daly C, Fitzgerald GF: Cell surface characteristics of *Lactococcus lactis* harbouring pCI528, a 46 kb plasmid encoding inhibition of bacteriophage adsorption. Microbiol. 1992, 138(10):2137-2143.

173. Sijtsma L, Sterkenburg A, Wouters JT: Properties of the cell walls of *Lactococcus lactis* subsp. *cremoris* SK110 and SK112 and their relation to bacteriophage resistance. Appl Environ Microbiol. 1988, 54(11):2808-2811.

174. Sijtsma L, Jansen N, Hazeleger WC, Wouters JT, Hellingwerf KJ: Cell surface characteristics of bacteriophage-resistant *Lactococcus lactis* subsp.

cremoris SK110 and its bacteriophage-sensitive variant SK112. Appl Environ Microbiol. 1990, 56(10):3230-3233.

175. Gasson MJ, Davies FL: Prophage-Cured Derivatives of *Streptococcus lactis* and *Streptococcus cremoris*. Appl Environ Microbiol. 1980, 40(5):964-966.

176. McGrath S, Fitzgerald GF, van Sinderen D: Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Mol Microbiol. 2002, 43(2):509-520.

177. Mahony J, McGrath S, Fitzgerald GF, van Sinderen D: Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. Appl Environ Microbiol. 2008, 74(20):6206-6215.

178. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SK, Dryden DT, Dybvig K: A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nuc Acids Res. 2003, 31(7):1805-1812.

179. Pingoud A, Wilson GG, Wende W: Type II restriction endonucleases—a historical perspective and more. Nuc Acids Res. 2014:gku447.

180. Loenen WA, Dryden DT, Raleigh EA, Wilson GG: Type I restriction enzymes and their relatives. Nuc Acids Res. 2013:gkt847.

181. Rao DN, Dryden DT, Bheemanaik S: Type III restriction-modification enzymes: a historical perspective. Nuc Acids Res. 2014, 42(1):45-55.

182. Wilson GG, Murray NE. Restriction and modification systems. Annu Rev Genet. 1991 Dec, 25(1):585-627.

183. O'sullivan D, Twomey DP, Coffey A, Hill C, Fitzgerald GF, Ross RP: Novel type I restriction specificities through domain shuffling of HsdS subunits in *Lactococcus lactis*. Mol Microbiol. 2000, 36(4):866-875.

184. Cerdeño-Tárraga AM, Patrick S, Crossman LC, Blakely G, Abratt V, Lennard N, Poxton I, Duerden B, Harris B, Quail MA *et al*: Extensive DNA Inversions in the *B. fragilis* Genome Control Variable Gene Expression. Science. 2005, 307(5714):1463-1465.

185. Claesson MJ, Li Y, Leahy S, Canchaya C, van Pijkeren JP, Cerdeño-Tárraga AM, Parkhill J, Flynn S, O'Sullivan GC, Collins JK *et al*: Multireplicon genome architecture of *Lactobacillus salivarius*. Proc Natl Acad Sci. 2006, 103(17):6718-6723.

186. Galli D, Lottspeich F, Wirth R: Sequence analysis of *Enterococcus faecalis* aggregation substance encoded by the sex pheromone plasmid pAD1. Mol Microbiol. 1990, 4.

187. Roberts RJ, Vincze T, Posfai J, Macelis D: REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nuc Acids Res. 2015, 43(Database issue):D298-D299.

188. Roberts RJ, Vincze T, Posfai J, Macelis D: REBASE—enzymes and genes for DNA restriction and modification. Nuc Acids Res. 2007, 35(suppl 1):D269-D270.

189. Källström H, Blackmer Gill D, Albiger B, Liszewski MK, Atkinson JP, Jonsson AB: Attachment of *Neisseria gonorrhoeae* to the cellular pilus receptor CD46: identification of domains important for bacterial adherence. Cell Microbiol. 2001, 430.

190. Bae T, Schnewind O: The YSIRK-G/S motif of staphylococcal protein A and its role in efficiency of signal peptide processing. J Bacteriol. 2003, 185.

191. O'driscoll J, Glynn F, Cahalane O, O'Connell-Motherway M, Fitzgerald GF, Van Sinderen D: Lactococcal plasmid pNP40 encodes a novel, temperature-

sensitive restriction-modification system. Appl Environ Microbiol. 2004, 70(9):5546-5556.

192. Deng Y-M, Liu C-Q, Dunn NW: Lld I, a Plasmid-Encoded Type I Restriction and Modification System in *Lactococcus lactis*. Mitochondrial DNA. 2000, 11(3-4):239-245.

193. Hill C, Pierce K, Klaenhammer T: The conjugative plasmid pTR2030 encodes two bacteriophage defense mechanisms in lactococci, restriction modification R/M and abortive infection Hsp. Appl Environ Microbiol. 1989, 55(9):2416-2419.

194. O'Connell-Motherway M, Watson D, Bottacini F, Clark TA, Roberts RJ, Korlach J, Garault P, Chervaux C, van Hylckama Vlieg JET: Identification of Restriction-Modification Systems of *Bifidobacterium animalis* subsp. *lactis* CNCM I-2494 by SMRT Sequencing and Associated Methylome Analysis. PloS One. 2014, 9(4):e94875.

195. Chopin M-C, Chopin A, Bidnenko E: Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol. 2005, 8(4):473-479.

196. Prevots F, Ritzenthaler P: Complete Sequence of the New Lactococcal Abortive Phage Resistance Gene *abiO*. J Dairy Sci. 1998, 81(6):1483-1485.

197. Twomey DP, De Urraza PJ, McKay LL, O'Sullivan DJ: Characterization of AbiR, a Novel Multicomponent Abortive Infection Mechanism Encoded by Plasmid pKR223 of *Lactococcus lactis* subsp. *lactis* KR2. Appl Environ Microbiol. 2000 Jun 1, 66(6):2647-51.

198. Sing WD, Klaenhammer TR: Conjugal transfer of bacteriophage resistance determinants on pTR2030 into *Streptococcus cremoris* strains. Appl Environ Microbiol. 1986, 51(6):1264-1271.

199. Gautier M, Chopin M-C: Plasmid-determined systems for restriction and modification activity and abortive infection in *Streptococcus cremoris*. Appl Environ Microbiol. 1987, 53(5):923-927.

200. Garvey P, Fitzgerald G, Hill C: Cloning and DNA sequence analysis of two abortive infection phage resistance determinants from the lactococcal plasmid pNP40. Appl Environ Microbiol. 1995, 61(12):4321-4328.

201. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R: Diversity, Activity, and Evolution of CRISPR Loci in *Streptococcus thermophilus*. J Bacteriol. 2008, 190(4):1401-1412.

202. Wittenberger CL, Angelo N: Purification and Properties of a Fructose-1,6-Diphosphate-Activated Lactate Dehydrogenase from *Streptococcus faecalis*. J Bacteriol. 1970, 101(3):717-724.

203. Dherbecourt J, Falentin H, Canaan S, Thierry A: A genomic search approach to identify esterases in *Propionibacterium freudenreichii* involved in the formation of flavour in Emmental cheese. Microb Cell Fact. 2008, 7(1):16.

204. Kato T, Nagatsu T, Kimura T, Sakakibara S: Fluorescence assay of *X*-prolyl dipeptidyl-aminopeptidase activity with a new fluorogenic substrate. Biochem Med. 1978, 19(3):351-359.

205. Lloyd RJ, Pritchard GG: Characterization of *X*-prolyl dipeptidyl aminopeptidase from *Lactococcus lactis* subsp. *lactis*. J Gen Microbiol. 1991, 137(1):49-55.

206. Siezen RJ, Bachmann H: Genomics of dairy fermentations. Microb Biotechnol. 2008, 1(6):435-442.

207. Casey MG, Meyer J: Presence of *X*-Prolyl-Dipeptidyl-Peptidase in Lactic Acid Bacteria. J Dairy Sci. 1985, 68(12):3212-3215.

208. Spus M, Li M, Alexeeva S, Wolkers-Rooijackers JC, Zwietering MH, Abee T, Smid EJ: Strain diversity and phage resistance in complex dairy starter cultures. J Dairy Sci. 2015 Aug 31, 98(8):5173-82.

209. Mahony J, Ainsworth S, Stockdale S, van Sinderen D: Phages of lactic acid bacteria: The role of genetics in understanding phage-host interactions and their co-evolutionary processes. Virol. 2012, 434(2):143-150.

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

211. Farenc C, Spinelli S, Vinogradov E, Tremblay D, Blangy S, Sadovskaya I, Moineau S, Cambillau C: Molecular Insights on the Recognition of a *Lactococcus lactis* Cell Wall Pellicle by the Phage 1358 Receptor Binding Protein. J Virol. 2014, 88(12):7005-7015.

212. Villion M, Chopin M-C, Deveau H, Ehrlich SD, Moineau S, Chopin A: P087, a lactococcal phage with a morphogenesis module similar to an *Enterococcus faecalis* prophage. Virol. 2009, 388(1):49-56.

213. Tremblay DM, Tegoni M, Spinelli S, Campanacci V, Blangy S, Huyghe C, Desmyter A, Labrie S, Moineau S, Cambillau C: Receptor-Binding Protein of *Lactococcus lactis* Phages: Identification and Characterization of the Saccharide Receptor-Binding Site. J Bacteriol. 2006, 188(7):2400-2410.

214. Chapot-Chartier M-P: Interactions of the cell-wall glycopolymers of lactic acid bacteria with their bacteriophages. Front Microbiol. 2014, 5.

215. Chapot-Chartier M-P, Vinogradov E, Sadovskaya I, Andre G, Mistou M-Y, Trieu-Cuot P, Furlan S, Bidnenko E, Courtin P, Péchoux C: Cell surface of *Lactococcus lactis* is covered by a protective polysaccharide pellicle. J Biol Chem. 2010, 285(14):10464-10471.

216. Andre G, Kulakauskas S, Chapot-Chartier M-P, Navet B, Deghorain M, Bernard E, Hols P, Dufrêne YF: Imaging the nanoscale organization of peptidoglycan in living *Lactococcus lactis* cells. Nat Comm. 2010, 1:27.

217. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA: Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. 2006, 2(3):e31.

218. O'Sullivan D, Ross RP, Fitzgerald GF, Coffey A: Investigation of the relationship between lysogeny and lysis of *Lactococcus lactis* in cheese using prophage-targeted PCR. Appl Environ Microbiol. 2000, 66(5):2192-2198.

219. Ventura M, Zomer A, Canchaya C, O'Connell-Motherway M, Kuipers O, Turroni F, Ribbera A, Foroni E, Buist G, Wegmann U et al: Comparative analyses of prophage-like elements present in two *Lactococcus lactis* strains. Appl Environ Microbiol. 2007, 73(23):7771-7780.

220. Ho CH, Stanton-Cook M, Beatson SA, Bansal N, Turner MS: Stability of active prophages in industrial *Lactococcus lactis* strains in the presence of heat, acid, osmotic, oxidative and antibiotic stressors. Int J Food Microbiol. 2016, 220:26-32.

221. Chopin MC, Chopin A, Rouault A, Galleron N: Insertion and amplification of foreign genes in the *Lactococcus lactis* subsp. *lactis* chromosome. Appl Environ Microbiol. 1989, 55(7):1769-1774.

222. Labrie SJ, Moineau S: Abortive infection mechanisms and prophage sequences significantly influence the genetic makeup of emerging lytic lactococcal phages. J Bacteriol. 2007, 189(4):1482-1487.

223. Almeida M, Hébert A, Abraham A-L, Rasmussen S, Monnet C, Pons N, Delbès C, Loux V, Batto J-M, Leonard P et al: Construction of a dairy microbial genome catalog opens new perspectives for the metagenomic analysis of dairy fermented products. BMC Genomics. 2014, 15(1):1101.

224. O'Sullivan DJ, Cotter PD, O'Sullivan O, Giblin L, McSweeney PLH, Sheehan JJ: Temporal and Spatial Differences in Microbial Composition during the Manufacture of a Continental-Type Cheese. Appl Environ Microbiol. 2015, 81(7):2525-2533.

225. Fallico V, Ross R, Fitzgerald G, McAuliffe O: Novel conjugative plasmids from the natural isolate *Lactococcus lactis* subspecies *cremoris* DPC3758: a repository of genes for the potential improvement of dairy starters. J Dairy Sci. 2012, 95(7):3593-3608.

226. O'Sullivan D, Ross RP, Twomey DP, Fitzgerald GF, Hill C, & Coffey A: Naturally occurring lactococcal plasmid pAH90 links bacteriophage resistance and mobility functions to a food-grade selectable marker. Appl Environ Microbiol. 2001, 67(2):929-937.

227. Raha AR, Hooi WY, Mariana NS, Radu S, Varma NR, Yusoff K. DNA sequence analysis of a small cryptic plasmid from *Lactococcus lactis* subsp. *lactis* M14. Plasmid. 2006, 56(1):53-61.

228. Madsen A, Josephsen J. Characterization of LlaCI, a new restriction-modification system from *Lactococcus lactis* subsp. *cremoris* W15. Biol Chem. 1998, 379(4-5):443-50.

229. Sánchez C, de Rojas AH, Martínez B, Argüelles ME, Suárez JE, Rodríguez A, Mayo B. Nucleotide sequence and analysis of pBL1, a bacteriocin-producing plasmid from *Lactococcus lactis* IPLA 972. Plasmid. 2000, 44(3):239-49.

230. Sánchez C, Mayo B. Sequence and analysis of pBM02, a novel RCR cryptic plasmid from *Lactococcus lactis* subsp. *cremoris* P8-2-47. Plasmid. 2003, 49(2):118-29.

231. Hayes F, Vos P, Fitzgerald GF, de Vos WM, Daly C. Molecular organization of the minimal replicon of novel, narrow-host-range, lactococcal plasmid pCI305. Plasmid. 1991, 25(1):16-26.

232. Chang HC, Do Choi Y, Lee HJ. Nucleotide Sequence of a Plasmid pCL2. 1 from *Lactococcus lactis* ssp. *lactis* ML 8. Plasmid. 1995, 34(3):234-5.

233. Kobayashi M, Nomura M, Kimoto H. Manipulation for plasmid elimination by transforming synthetic competitors diversifies *Lactococcus lactis* starters applicable to food products. Biosci Biotechnol Biochem. 2007, 71(11):2647-54.

234. Gasson MJ, Godon JJ, Pillidge CJ, Eaton TJ, Jury K, Shearman CA. Characterization and exploitation of conjugation in *Lactococcus lactis*. Int Dairy J. 1995, 5(8):757-62.

235. Tanous C, Chambellon E, Yvon M. Sequence analysis of the mobilizable lactococcal plasmid pGdh442 encoding glutamate dehydrogenase activity. Microbiol. 2007, 153(5):1664-75.

236. Christensson C, Pillidge CJ, Ward LJ, O' Toole P. Nucleotide sequence and characterization of the cell envelope proteinase plasmid in Lactococcus lactis subsp. cremoris HP. J Appl Microbiol. 2001, 2;91(2):334-43.

237. Górecki RK, Koryszewska-Bagińska A, Gołębiewski M, Żylińska J, Grynberg M, Bardowski JK: Adaptative Potential of the *Lactococcus lactis* IL594 Strain Encoded in Its 7 Plasmids. PLOS ONE. 2011, 6(7):e22238.

238. Anba J, Bidnenko E, Hillier A, Ehrlich D, Chopin MC. Characterization of the lactococcal abiD1 gene coding for phage abortive infection. J Bacteriol. 1995, 177(13):3818-23.

239. Perreten V, Schwarz F, Cresta L, Boeglin M, Dasen G, Teuber M. Antibiotic resistance spread in food. Nature. 1997, 389(6653):801-2.

240. Kojic M, Jovcic B, Strahinic I, Begovic J, Lozo J, Veljovic K, Topisirovic L: Cloning and expression of a novel lactococcal aggregation factor from *Lactococcus lactis* subsp. *lactis* BGKP1. BMC Microbiol. 2011, 11(1):265.

241. Chang SM, Yan TR. DNA sequence analysis of a cryptic plasmid pL2 from *Lactococcus lactis* subsp. *lactis*. Biotechnol Letters. 2007, 29(10):1519-27.

242. Gajic O, Buist G, Kojic M, Topisirovic L, Kuipers OP, Kok J. Novel mechanism of bacteriocin secretion and immunity carried out by lactococcal multidrug resistance proteins. J Biol Chem. 2003, 278(36):34291-8.

243. Dougherty BA, Hill C, Weidman JF, Richardson DR, Venter JC, Ross RP. Sequence and analysis of the 60 kb conjugative, bacteriocin-producing plasmid pMRC01 from *Lactococcus lactis* DPC3147. Mol Microbiol. 1998, 29(4):1029-38.

244. Strahinic I, Kojic M, Tolinacki M, Fira D, Topisirovic L. Molecular characterization of plasmids pS7a and pS7b from *Lactococcus lactis* subsp. *lactis* bv. *diacetylactis* S50 as a base for the construction of mobilizable cloning vectors. J Appl Microbiology. 2009, 106(1):78-88.

245. Boucher I, Emond E, Parrot M, Moineau S. DNA sequence analysis of three *Lactococcus lactis* plasmids encoding phage resistance mechanisms. J Dairy Sci. 2001, 84(7):1610-20.

246. Fallico V, McAuliffe O, Fitzgerald GF, Ross RP. Plasmids of raw milk cheese isolate *L. lactis* subsp. *lactis* biovar. *diacetylactis* DPC3901 suggest a plant-based origin for the strain. Appl Environ Microbiol. 2011, AEM-00661.

247. Radziwill-Bienkowska JM, Le DT, Szczesny P, Duviau MP, Aleksandrzak-Piekarczyk T, Loubière P, Mercier-Bonin M, Bardowski JK, Kowalczyk M. Adhesion of the genome-sequenced *Lactococcus lactis* subsp. *cremoris* IBB477 strain is mediated by specific molecular determinants. Appl Microbiol Biotechnol. 2016, 100(22):9605-17.

Chapter II

Performance and flavour-based characterisation of lactococcal starter cultures

Chapter II contents

2 Abstract ……... 95
2.1 Introduction ……96
2.2 Materials & Methods …….. 98
2.2.1 Strain growth conditions and media …… 98
2.2.2 Nessler's arginine broth assay ……. 98
2.2.3 Growth profile analysis by (a modification of) the Pearce activity test 99
2.2.4 Enzymatic assays ……. 100
2.2.5 Determination of lactate dehydrogenase activity (LDH) …… 100
2.2.6 Determination of amino acid transferase activity (Phenylalanine and Methionine) ……. 101
2.2.7 Determination of specific peptidase activities by fluorescence ….. 101
2.2.8 Determination of short chain esterase activity …… 102
2.3 Results ……103
2.3.1 Strain differentiation …… 103
2.3.2 Performance testing ……. 105
2.3.3 Performance testing - Lactate dehydrogenase (LDH) ……. 111
2.3.4 Flavour Capabilities – Aminotransferase activity ……... 113
2.3.5 Flavour capabilities – Peptidase activity ……. 115
2.3.6 Flavour Capabilities – Esterase activity (lipolysis) ……. 120
2.4 Discussion ……... 122
2.5 Conclusion …….. 125
2.6 References ……... 126

2 Abstract

In the current study we describe the characterisation of a selection of lactococcal strains in terms of industrial robustness and flavour formation using a functional genomics approach. Comparison with four starter cultures currently employed in the Irish dairy industry for the production of half fat Cheddar cheese facilitated the identification of potentially applicable novel starter cultures within the assessed collection. In principle, this methodology represents a useful tool to expand the biodiversity of starter cultures in a rational manner.

2.1 Introduction

Lactic acid bacteria (LAB) represent a diverse group of Gram-positive organisms which produce lactic acid from the degradation of hexose sugars, and which for this reason are widely used in food fermentations. A typical LAB member is *Lactococcus lactis*, a Gram-positive, catalase-negative, non-motile and coccoid bacterium [1]. The *L. lactis* species can be further divided into subspecies (subsp.) *cremoris*, subsp. *lactis* or subsp. *lactis* biovar diacetylactis, the latter having the distinctive characteristic of being capable of metabolising citrate. *L. lactis* is extensively employed as a starter culture for the manufacture of various fermented dairy products, such as sour cream and many cheese varieties [2].

Lactococcal starter cultures, used in commercial food fermentations, are frequently composed of defined strains, selected for their desirable traits in relation to industrial robustness and flavour development [3]. Industrial robustness is generally focused on the stresses encountered by strains during manufacture, such as oxidative, temperature-mediated (e.g. due to spray- and freeze-drying), osmotic and/or solvent stress [4]. In addition, starter performance qualities such as growth rate [5, 6], acidification rate [7] and phage insensitivity [8] are equally important technological traits.

Cheese flavour development occurs predominantly during ripening, and in many cheese types is mainly due to the addition of adjunct cultures [9]. Starter cultures significantly impact on cheese flavour development through the proteolytic breakdown of caseins [10]. Casein proteolysis is the most complex and possibly the most important process in terms of primary flavour development in cheese. Proteolysis is responsible for the liberation of peptides and subsequently their component amino acids, thereby supplying substrates for various secondary

pathways of amino acid catabolism [9]. A balance between proteolysis and peptidolysis is desirable as it helps to prevent the formation of bitterness and off-flavours in cheese [10]. *L. lactis* strains produce aroma compounds through amino acid catabolism, which further contributes to cheese flavour development [11]. Amino acid transamination is catalysed by aminotransferases which transfer the amino group of an amino acid to an α-ketoacid, with α-ketoglutarate representing the α-ketoacid acceptor in LAB [11]. Parallel quantification of activity levels of peptidases and aminotransferases may help to generate a detailed biochemical profile of the flavour-forming abilities of a particular strain.

Degradation of milk fats and hydrolysis of triglycerides into lipids and fatty acids by lipolysis is also an important contributor to cheese flavour development [12]. In lactococci the dominant lipolytic enzymes involved in lipolysis belong to a class of enzymes called the carboxylic ester hydrolases, mainly represented by esterase and lipase activities [13]. Lipolysis is of particular concern in low-fat cheeses made from skimmed milk, as the reduced fat content can lead to incorrect flavour development.

In the current work we assessed the performance and flavour production capabilities of a selection of lactococcal strains, and compared these quantitative parameters to a number of commercially employed starter cultures. This indirect flavour profile analysis was performed so as to establish if a correlation exists between genome content and measured flavour production abilities.

2.2 Materials & Methods

2.2.1 Strain growth conditions and media

Bacterial strains used in this study are detailed in Table 2.2. *L. lactis* strains were routinely cultured at 30 °C in 10 % RSM (reconstituted skimmed milk) at 30 °C without agitation. Cells were prepared via a 1.5 % inoculum into 10 % RSM and grown overnight (16 hours) at 30 °C. Cells were then plated on M17 (Oxoid) agarose supplemented with 0.5 % lactose to determine a viable plate count in cfu/ml.

2.2.2 Nessler's arginine broth assay

Phenotypic sub-speciation of lactococcal cultures was performed by the arginine broth assay [14]. 1 ml of overnight culture was added to 5 ml of arginine broth; 1 % tryptone, 0.5 % yeast extract, 0.1 % glucose, 0.4 % potassium hydrogen phosphate, 0.6 % L-arginine. 1 ml of Nessler's reagent [15] (14.3 % sodium hydroxide, 5 % mercuric iodide and 4 % potassium iodide) was added after incubation overnight at 30 °C. A colour change from yellow to red indicates the production of ammonia due to arginine hydrolysis. Strain identification was based on a colour change following incubation, with red colour development being characteristic of subspecies *lactis*, while yellow being characteristic of subsp. *cremoris*.

2.2.3 Growth profile analysis by (a modification of) the Pearce activity test

A modification of the Pearce activity test was performed to evaluate growth profiles of lactococcal strains under simulated cheese production conditions [16]. This was carried out in order to assess the effect of the Cheddar cheese cooking

temperature regime on growth and intracellular enzyme release. Cultures were prepared from stocks stored at -80 °C in 10 % RSM, which were thawed and incubated at 21 °C until coagulation (~16 hours). The coagulated culture was used to make a 1.5 % inoculum into 100 ml of 10 % RSM (~16 hours). This culture was used to prepare 500 ml of 10 % RSM with a 1.5 % inoculum for the Pearce activity test. The test was performed according to the temperature cycle displayed in Figure 2.1. Samples were taken at 60 min intervals throughout the incubation and plated on M17 (Oxoid) agarose supplemented with 0.5 % lactose to determine a viable plate count (expressed as cfu/ml). The point of temperature induced autolysis was assessed from a decrease in viable plate counts.

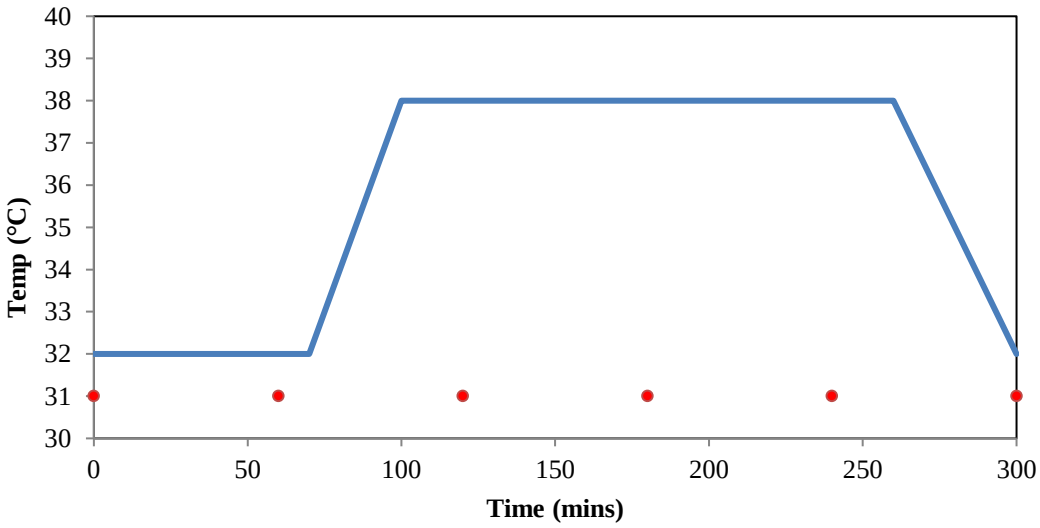


Figure 2.1: Temperature cycle for Pearce activity test
Blue line indicates temperature profile, red markers indicate sampling points.

2.2.4 Enzymatic assays

Assays to measure lactate dehydrogenase, amino acid transferase, peptidase and esterase activities were performed in triplicate, dH_2O was used as a blank unless otherwise indicated. Reagents were from Sigma-Aldrich (MO, USA), except in the case of fluorescent coupled peptidase substrates (Bachem AG, Switzerland).

2.2.5 Determination of lactate dehydrogenase activity (LDH)

A 0.2 M Tris-Maleate buffer was prepared and adjusted to pH 7.0 with the addition of 0.2 M NaOH. A solution, containing 45 mM NADH, 30 mM fructose 1,6-bis-phosphate (FBP) and 300 mM Na-pyruvate, was prepared in 10 ml aliquots and stored at 4 °C, wrapped in aluminium foil for maximum of 1 week. All reagents were sourced from Sigma-Aldrich, MO, USA.

A mixture of 2.7 ml of Tris-maleate (pH 7.5), 0.1 ml of NADH (4.5 mM), 0.1 ml of FBP (30 mM) and 0.1 ml of sample was added to a cuvette and the absorbance recorded at 340 nm for 30 sec on a DU Series 730 spectrophotometer (Beckman Coulter). 0.1 ml of 300 mM pyruvate was then added and the decrease in A340 was monitored for 90 sec. LDH activity was calculated using the following equation:

$$\text{Units per 0.1 ml sample} = \frac{\Delta\ \text{ABS@340 Nm x 3 x Dilution factor x 1000}}{\text{6270 x 0.1 x 1}}$$

The change in absorbance per minute **(Δ340/min**) was calculated from the spectrophotometer readings, **3** ml final volume in the cuvette, **1000** converts NADH from nM to µM, **6270** is the extinction co-efficient for NADH, **0.1** ml of sample in solution and **1** cm is the path length. Activity was then expressed as unit per ml of

extract, where one 1 unit was defined as the amount of enzyme that is required to catalyse the oxidation of 1 µM of NADH/min.

2.2.6 Determination of amino acid transferase activity (Phenylalanine and Methionine)

The reaction mixture for L-phenylalanine (L-Phe) contained the following solutes: 50 µM pyridoxal phosphate, 5 mM α-keto glutaric acid and 5 mM L-Phe, 0.5 mM sodium arsenate and 500 mM EDTA prepared in 50 mM sodium tetra-borate buffer, pH 8.5. The reaction mixture for methionine (Met) contained the following chemical ingredients: 50 µM pyridoxal phosphate, 5 mM α-keto glutaric acid, 5 mM Met, 0.5 mM sodium arsenate and 500 mM EDTA prepared in 50 mM sodium tetra-borate buffer, pH 8.5. α-ketoacid standards; phenylpyruvate and α-ketomethylthiobutyrate were prepared in dH_2O to final concentrations; 100, 200, 300, 400, 500 & 1000 mM. dH_2O was used as a blank.

100 µl of sample and 1 ml of reaction mixture were incubated at 30 °C. After 30 min the reaction was stopped by the addition of 1 ml of 10 % Trichloroacetic acid (TCA) solution. The reaction mixture was centrifuged at 10,000 rpm for 2 min to remove precipitated proteins and the absorbance was read at a wave length of 300 nm. The specific activity of amino acid transferase was expressed as µM/min/mg of protein.

2.2.7 Determination of specific peptidase activities by fluorescence

Specific peptidase activities were assessed by detection of fluorescence release using 7-amino-4-methyl coumarin (AMC)-coupled substrates (Table 2.1). Peptidase substrates were sourced from Bachem AG, Switzerland. 0.111 mM (X)-AMC substrates were prepared by dissolving the particular substrate in 100 µl of

DMF (dimethylformamide), and then adding 50 mM Tris-HCl (pH 7) to a final volume of 50 ml. 0.5 mM AMC standards were prepared using the same procedure. An AMC standard curve was obtained across the concentration range 0-1 µM, while 50 mM Tris-HCl (pH 7) was used as a blank. The protocol was as previously described [17], except when using reduced volumes for high throughput screening in 96-well plates. Released fluorescence was measured on a SpectraMax M3 Multi-Mode Microplate Reader (Molecular Devices). Enzyme activity was calculated in RFU PPDA (1 RFU = the amount of µM of AMC released min^{-1} by 1 mg of protein).

Table 2.1: AMC-coupled peptidase substrates

Substrate	Target peptidase
H-Lys-AMC.acetate (Lys-AMC)	PepN and PepC
H-Asp (AMC)-OH (Asp-AMC)	PepA
H-Pro-AMC.HBr (Pro-AMC)	Proline imino peptidase
H-Gly-Pro-AMC. HBr (Gly-Pro-AMC)	PepX
CBZ-Gly-Pro-AMC (Z-Gly-Pro-AMC)	Carboxypeptidase
N-Suc-Gly-Pro-Leu-Gly-Pro-AMC (Gly-Pro-Leu-Gly-Pro-AMC)	Endopeptidase

2.2.8 Determination of short chain esterase activity

Short chain esterase activity was detected via a previously described spectrophotometric assay [18], utilising p-nitrophenyl butyrate as a substrate. Absorbance was measured on a DU Series 730 spectrophotometer (Beckman Coulter).

2.3 Results

2.3.1 Strain differentiation

The phenotypic characteristics and enzymatic abilities of 20 lactococcal strains from the UCC culture collection were assessed, in order to ascertain their potential performance as starter cultures for the production of Cheddar type cheeses. Firstly, we wanted to assign each of these 20 strains to either subsp. *lactis* or subsp. *cremoris* by means of the arginine broth assay, which allows subspecies identification based on the strain's (in)ability to release ammonia from arginine via the arginine deanimase pathway (ADI) [19]. *L. lactis* subsp. *lactis* strains utilise this pathway for arginine metabolism resulting in ammonia release, while *L. lactis* subsp. *cremoris* strains are unable to use this pathway [19]. This allows for strain differentiation based on an indicator colour change due to the presence/absence of ammonia (see Materials and Methods). The assay resulted in the identification of twelve subsp. *cremoris* and eight subsp. *lactis* strains out of the 20 strains tested (Table 2.2), and where possible this was confirmed by genotypic analysis of the 16S rRNA-encoding gene if a corresponding genome sequence was available from the National Centre for Biotechnology Information (NCBI).

Table 2.2: Lactococcal strains used in this study

Strain[$]	Species (Arginine Broth)	Genome sequenced (Accession Number)
JM1 *	*cremoris*	CP015899
JM2 *	*cremoris*	CP015900
JM3 *	*cremoris*	CP015901
JM4 *	*cremoris*	CP015909
158 *	*cremoris*	CP015894
Bu260	*lactis*	-
303	*lactis*	-
SK11 *	*cremoris*	CP000425
3107 *	*cremoris*	(Unpublished)
UC23	*cremoris*	-
HP *	*cremoris*	Draft assembly (JAUH00000000)
F7/2	*lactis*	-
UC109 *	*cremoris*	CP015907
UC77 *	*lactis*	CP015906
275 *	*lactis*	CP015897
AM2	*cremoris*	-
R1	*cremoris*	-
UC063 *	*lactis*	CP015905
184 *	*lactis*	CP015895
229 *	*lactis*	CP015896

***Indicates strains whose subspeciation was confirmed via genotypic analysis**
[$]All strains are dairy isolates from the UCC culture collection

2.3.2 Performance testing

An important aspect of strain selection for starter culture use is performance testing, during which growth rate, acid production and responses to temperature and salt are assessed as an indication as to how strains behave in a fermentation process [20]. The Pearce activity test is commonly employed in the dairy industry as an indicator for growth and temperature-induced autolysis of starter strains [16]. This test mimicks the temperature cycles used in the relevant (i.e. Cheddar-type) cheese production process and allows the point of temperature-induced autolysis to be determined. The lactococcal strains used in this study were assessed by (a modification of; see Materials and Methods) the Pearce activity test to determine the point of autolysis for each strain and to assess release of intracellular peptidase and aminotransferase activities (Fig. 2.2[A]-[S]).

The main observation made during the Pearce activity test was the higher cell viability obtained for strains belonging to subsp. *lactis* compared to their subsp. *cremoris* counterparts. The *L. lactis* subsp. *lactis* controls grown at 30 °C for the duration of the test (Fig. 2.2[E], [G], [J-L], [P] & [R-S]) regularly reached ~9 log cfu/ml, while their *cremoris* couterparts displayed a comparably lower viable count at ~7-8 log cfu/ml. The point of induced autolysis was also more prominent in the subsp. *cremoris* strains. This is unsurprising as these strains are generally more temperature sensitive than their *lactis* counterparts which can grow at temperatures up to 40 °C [21]. The largest observed reductions in cell viability were for *L. lactis* subsp. *cremoris* UC23 and 158, strains that exhibited more than a log reduction in cell counts when the temperature was increased from 32 to 38 °C. A substantial reduction in cell viability and consequent cell lysis during the cheese cooking process can lead to a significant increase in available intracellular peptidases for the

degradation of milk proteins, although enhanced proteolytic activity may also lead to a bitter taste in certain cheese types [22].

To test the significance of the observed levels of lysis between the subspecies *cremoris* and subspecies *lactis* strains used in the analysis, a paired t-test was applied. The hypothesis used for the test was; is there sufficient evidence to sugest that greater cell lysis occurs during the Pearce activity test than during controlled growth at 30 ℃? This was represented mathematically as $H_O : \mu_d = 0$ while the alternative hypothesis was $H_A : \mu_d > 0$, with a significance p-value cut-off of $\alpha = 0.05$. The analysis indicated a p-value of 0.987 for the subspecies *cremoris* strains indicating that the hypothesis should be accepted and a significant level of lysis has occurred during the Pearce activity test. In the case of the subspecies *lactis* strains the p-value was calculated to be 0.00018 indicating that the hypothesis should be rejected and no significant lysis has occurred. This demonstrates that subspecies *cremoris* strains are better suited to these fermentation conditions where autolysis is desirable, then the equivalent subspecies *lactis* strains.

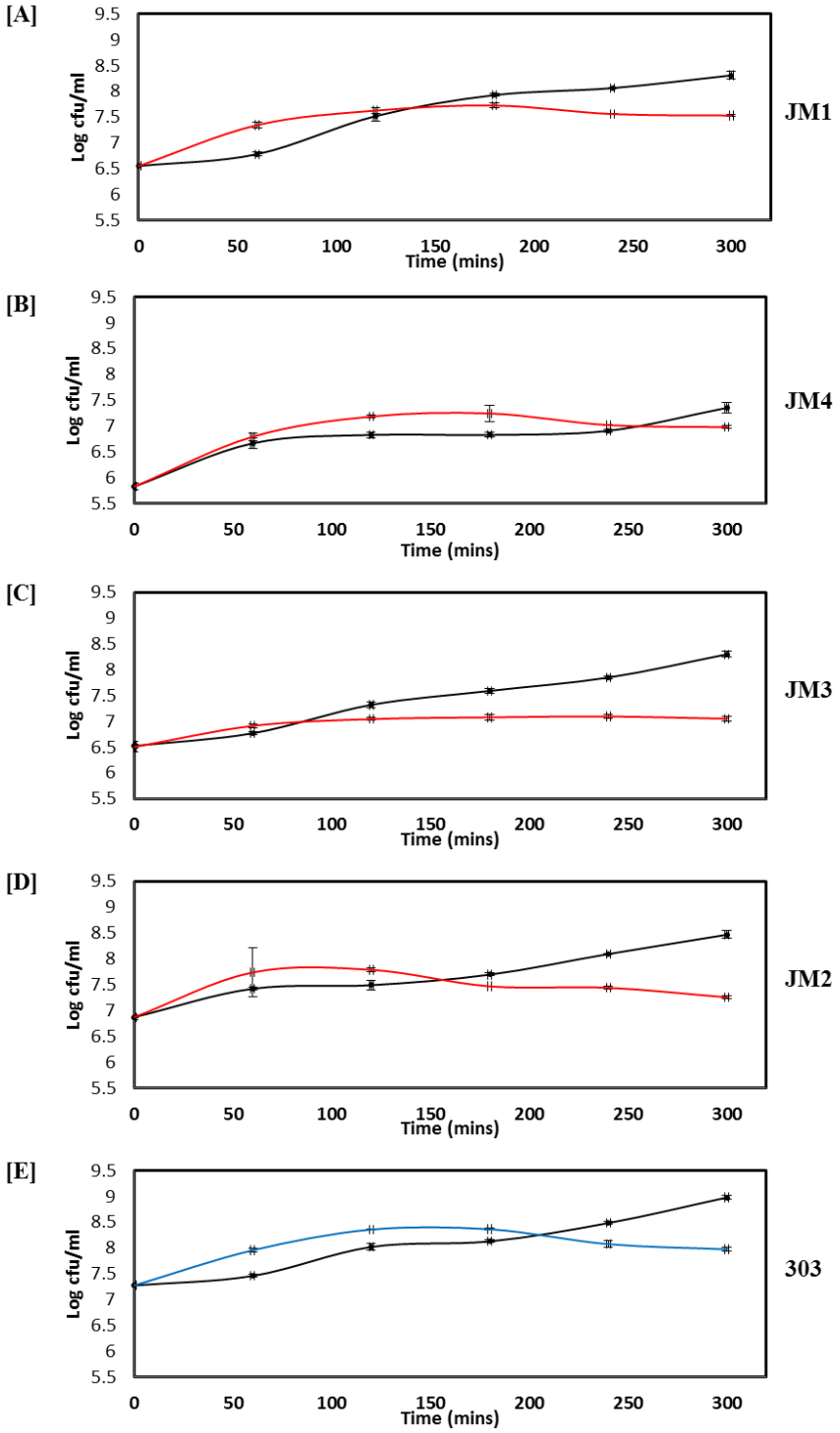
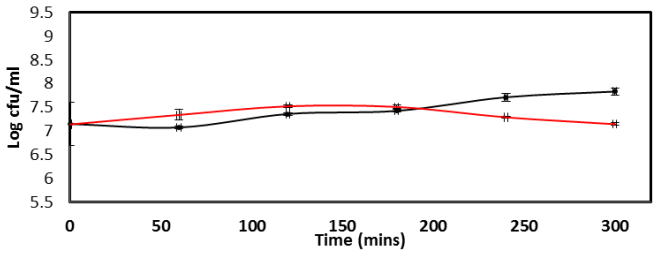


Figure 2.2: Modified Pearce activity test growth curves

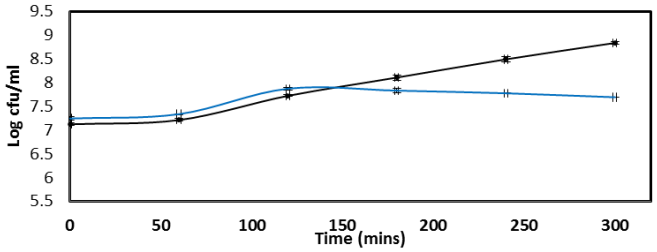
Control growth profiles of cultures grown at 30 °C are indicated as a black line. Strains grown under Pearce assay conditions are indicated as a red (subsp. *cremoris*) or blue (subsp. *lactis*) line. Strains were plated in triplicate.

[F]



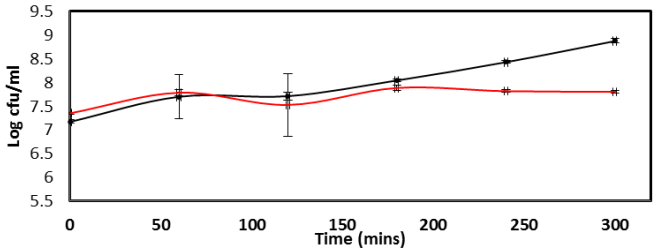
HP

[G]



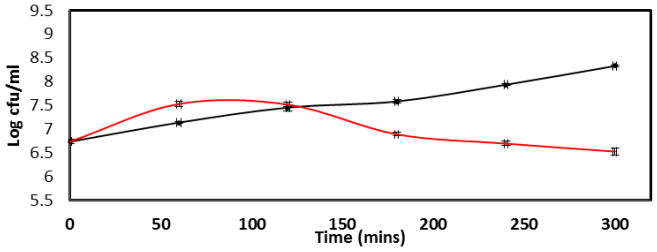
229

[H]



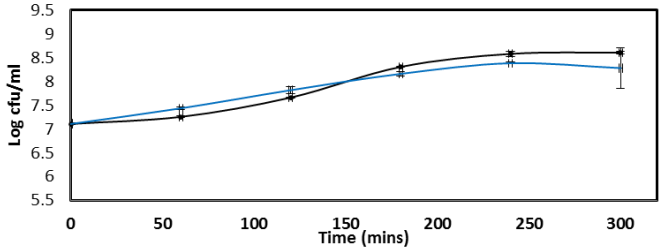
3107

[I]



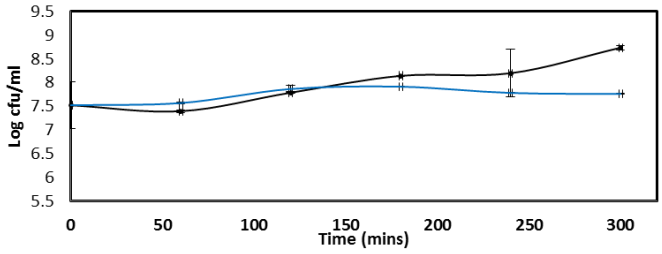
UC23

[J]



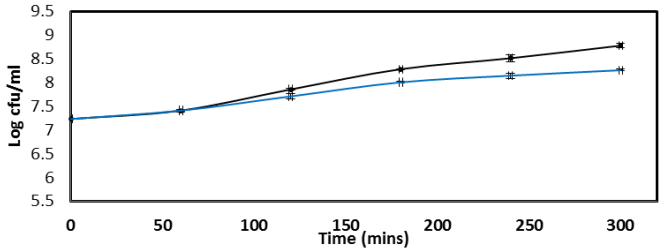
F7/2

[K]



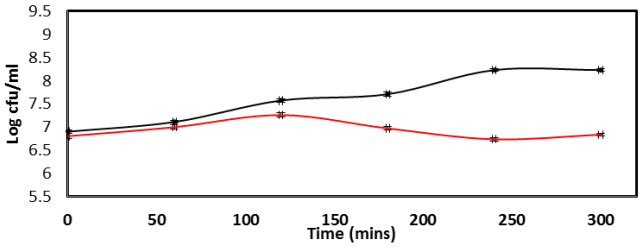
UC063

[L]



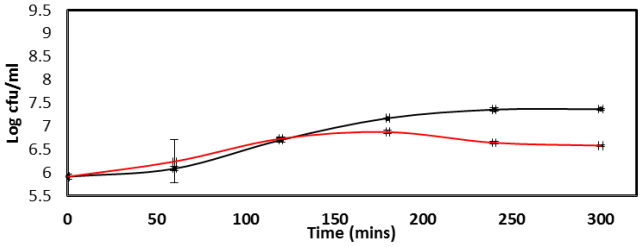
Bu260

[M]



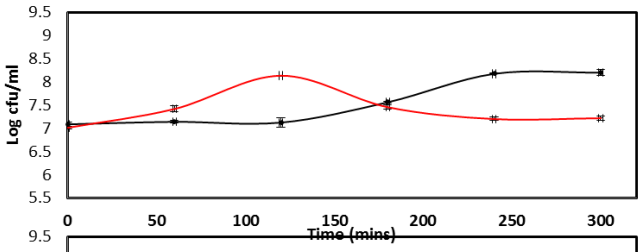
SK11

[N]



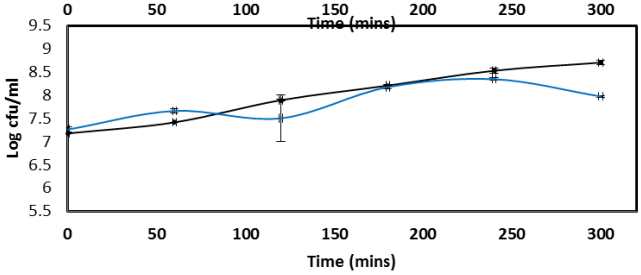
UC109

[O]



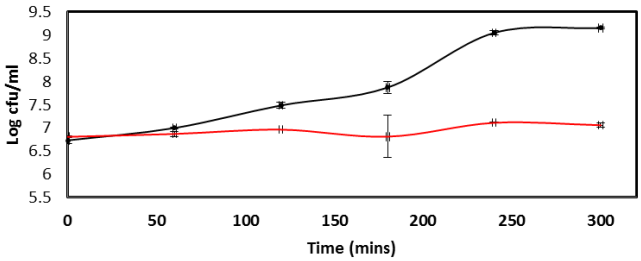
158

[P]



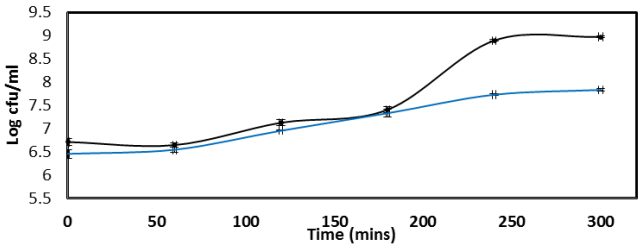
UC77

[Q]



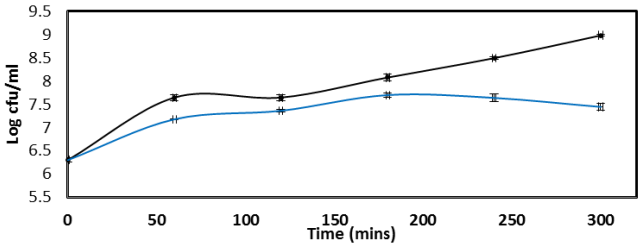
R1

[R]



275

[S]



184

Table 2.3: Paired t-test analysis of [A] subspecies *cremoris* and [B] subspecies *lactis*

[A]

	Final cell count (cfu/ml) after 300 minute incubation		
	Control (μ1)	**Pearce assay (μ2)**	**Difference (μ_d)**
JM1	2.07E+08	3.39E+07	1.73E+08
JM4	7.00E+07	9.50E+06	6.05E+07
JM3	2.00E+08	1.13E+07	1.89E+08
JM2	2.93E+08	1.80E+07	2.75E+08
HP	6.90E+07	1.40E+07	5.50E+07
UC23	2.13E+08	3.40E+06	2.10E+08
3107	7.53E+08	6.45E+07	6.89E+08
SK11	1.70E+08	6.93E+06	1.63E+08
UC109	2.33E+07	3.83E+06	1.95E+07
158	1.60E+08	1.66E+07	1.43E+08
R1	1.43E+09	1.15E+07	1.42E+09

Sample mean (xbar)	3.09E+08
Sample standard deviation (s)	3.91E+08
Sample size (n)	11
Degrees of freedom (df)	10
T-test statistic (t)	2.62132
P-value (p)	0.98723

*Accept hypothesis (> 0.05)

[B]

	Final cell count (cfu/ml) after 300 minute incubation		
	Control (μ1)	**Pearce assay (μ2)**	**Difference (μ_d)**
303	9.50E+08	9.43E+07	8.56E+08
229	6.87E+08	4.91E+07	6.38E+08
F7/2	4.10E+08	2.76E+08	1.34E+08
UC063	5.47E+08	5.75E+07	4.89E+08
Bu260	6.03E+08	1.82E+08	4.21E+08
UC17	5.07E+08	9.40E+07	4.13E+08
275	9.37E+08	6.77E+07	8.69E+08
184	9.63E+08	2.80E+07	9.35E+08

Sample mean (xbar)	5.94E+08
Sample standard deviation (s)	2.62E+08
Sample size (n)	8
Degrees of freedom (df)	7
T-test statistic (t)	6.42648
P-value (p)	0.00018

*Reject hypothesis (< 0.05)

*Hypothesis; is there sufficient evidence to suggest that greater cell lysis occurs during the Pearce activity test than during controlled growth at 30 ℃ for the respective subspecies?

Hypothesis statement uses μ1 - μ2 and a significance level = 0.05:

H_O: $\mu_d = 0 \Rightarrow \mu_1 - \mu_2 = 0$

H_A: $\mu_d > 0 \Rightarrow \mu_1 - \mu_2 \neq 0$ $\alpha = 0.05$

2.3.3 Performance testing - Lactate dehydrogenase (LDH)

LDH is an intracellular enzyme found in LAB which converts lactate to pyruvate in the presence of NAD^+, via a reversible reaction. LDH cannot be measured directly therefore the activity of LDH is measured by the decrease in NADH (NADH is a stoichiometric equivalent to LDH). The enzyme in lactococcal strains requires fructose 1,6 bis-phosphate for activation and can be used as an indicator of autolysis in dairy starter strains [23].

$$\textbf{Lactate} + \textbf{NAD}^+ \underset{}{\overset{\textbf{LDH}}{\rightleftharpoons}} \textbf{Pyruvate} + \textbf{NADH}$$

Analysis of the LDH released from strains grown under Pearce activity test conditions compared to strains grown under standard culture conditions, did not show a significant increase in the level of LDH released (Table 2.4). A paired t-test was applied to the data (as in section 2.3.2), using the hypothesis; is there sufficient evidence to suggest an increase in released LDH during the Pearce activity test compared to during controlled growth at 30 ℃? In this case there was found to be no significant increase in the levels of LDH activity detected in strains of either subspecies which underwent the Pearce activity test, suggesting that this method may not be suitable for monitoring autolysis in cheese starter cultures.

Table 2.4: Lactate dehydrogenase activity expressed by lactococcal strains

	Activity (enzyme units)*		
	Pearce assay (μ1)	**Control (μ2)**	**Difference (μ_d)**
JM1	0.542213333	0.541776167	0.00043717
JM4	0.5421198	0.5419124	0.0002074
JM3	0.542532567	0.541971367	0.0005612
JM2	0.543024633	0.542994133	3.05E-05
158	0.5419795	0.541641967	0.00033753
Bu260	0.542428867	0.541959167	0.0004697
303	0.542079133	0.542428867	-0.0003497
229	0.5419978	0.5425712	-0.0005734
SK11	0.5429494	0.541888	0.0010614
3107	0.542201133	0.542209267	-8.133E-06
UC23	0.5417294	0.541792433	-6.303E-05
HP	0.5433398	0.5419551	0.0013847
F7/2	0.541959167	0.5420893	-0.0001301
UC109	0.541989667	0.541831067	0.0001586
UC77	0.541719233	0.5416562	6.3033E-05
184	0.541707033	0.541690767	1.6267E-05
275	0.5416501	0.541774133	-0.000124
AM2	0.543219833	0.542182833	0.001037
R1	0.542361767	0.5417599	0.00060187
UC063	0.5419246	0.5419612	-3.66E-05

Sample mean (xbar)	0.000254065
Sample standard deviation (s)	0.000479236
Sample size (n)	20
Degrees of freedom (df)	19
T-test statistic (t)	2.3108
P-value (p)	0.0161
	**Reject hypothesis (< 0.05)

* 1 enzyme unit = amount of enzyme required to catalyse the oxidation of 1 μmole of NADH/min

Hypothesis;** is there sufficient evidence to suggest an increase in released LDH during the Pearce activity test compared to during controlled growth at 30 °C?

Hypothesis statement uses μ1 - μ2 and a significance level = 0.05:

H_O: $\mu_d = 0 \Rightarrow \mu1 - \mu2 = 0$

H_A: $\mu_d > 0 \Rightarrow \mu1 - \mu2 \neq 0$ $\alpha = 0.05$

2.3.4 Flavour Capabilities – Aminotransferase activity

Proteolysis is a very complex and possibly the most important process in terms of primary flavour development in cheese, being responsible for the liberation of peptides and amino acids, which in turn represent substrates for secondary pathways of amino acid catabolism [9]. Proteolysis also indirectly contributes to cheese flavour and aroma formation via transamination, dehydrogenation, decarboxylation and reduction of amino acids giving rise to a wide range of aromatic compounds [9]. A number of different amino acid transferase activities have been identified in different LAB that work on various aromatic, branched and sulphur-containing amino acids. In the current analysis we quantified the activity towards L-phenylalanine (L-Phe; an aromatic amino acid) and methionine (Met; a sulphur-containing amino acid), both of which are common in milk and important in terms of cheese production [24].

All strains demonstrated aminotransferase activity using methionine as a substrate (Fig. 2.4[A]), while a considerably lower level of activity was obtained when phenylalanine was used as a substrate (Fig. 2.4[B]). The subsp. *lactis* strains Bu260, 303 and 229 expressed a high level of aminotransferase activity suggesting that these strains are promising candidates for milk fermentations, in particular Bu260, which also performed well in terms of LDH production. Combined, the analysis indicates that Bu260 has starter culture potential both in terms of industrial robustness (LDH), and flavour and aroma development.

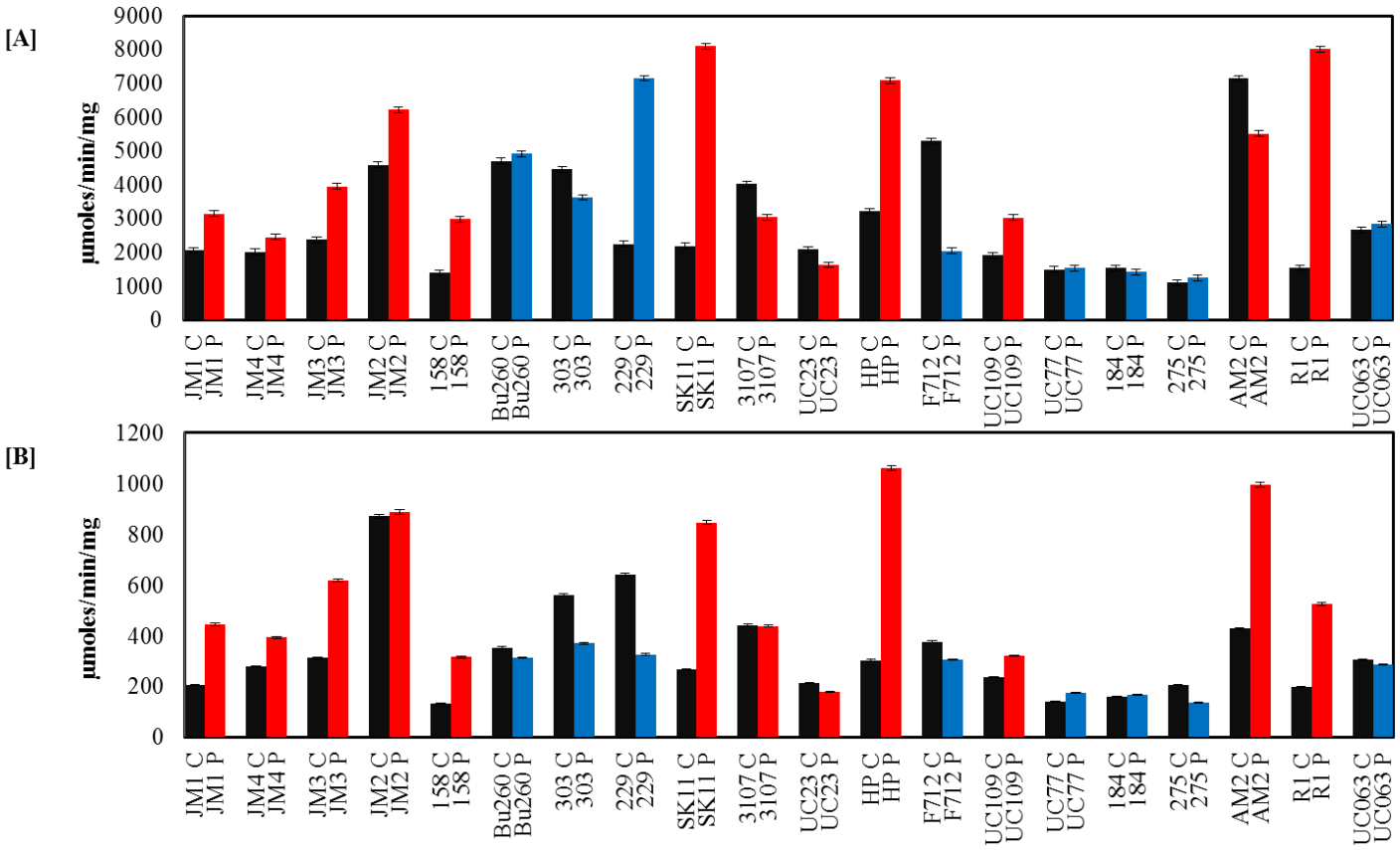


Figure 2.4: Aminotransferase activity expressed by lactococcal strains

Aminotransferase activity expressed by lactococcal strains against [A] methionine and [B] phenylalanine after growth at 30 °C for 5 hours, (black bars, names marked with 'C') or following the Pearce Activity test (subsp. *cremoris* and subsp. *lactis* indicated in red and blue, respectively, names marked with 'P').

2.3.5 Flavour capabilities – Peptidase activity

The main peptidase categories that contribute to proteolysis in *L. lactis* are aminopeptidases, endopeptidases, di/tri-peptidases, proline peptidases, endopeptidases and carboxypeptidases [9]. To assess the level of peptidase activity within *L. lactis*, a number of enzymatic analyses were undertaken. Quantitative assays utilising fluorescently labelled substrates (see Materials and Methods section) were used to determine the activity levels of PepN/C, PepA, PepX, PepI, carboxypeptidase and endopeptidase produced by each strain (Fig. 2.5[A-H]).

The dominant peptidase activities expressed by the analysed strains were the proline peptidase PepX (Fig. 2.5[A]) and the aminopeptidase PepA (Fig. 2.5[B]). PepX is of particular importance in milk fermentations due to the high proline content of *β*-casein [25] and has been reported to influence proteolysis in cheese ripening [26]. PepX activity was also observed to be higher for subsp. *cremoris* strains isolated from the dairy environment as compared to PepX levels observed for subsp. *lactis* strains. The other assessed peptidase activity levels appear to be similar across all strains in the analysis, except in the cases of PepN and PepC (Fig. 2.5[F-H]). The expressed levels of PepN and PepC activity appears to be lower in strains which have high levels of PepX activity, indicating that strains may have a preference towards either the X-prolyldipeptidyl aminopeptidase (PepX), or general aminopeptidases such as PepN or PepC.

Figure 2.5: Peptidase activity expressed by lactococcal strains

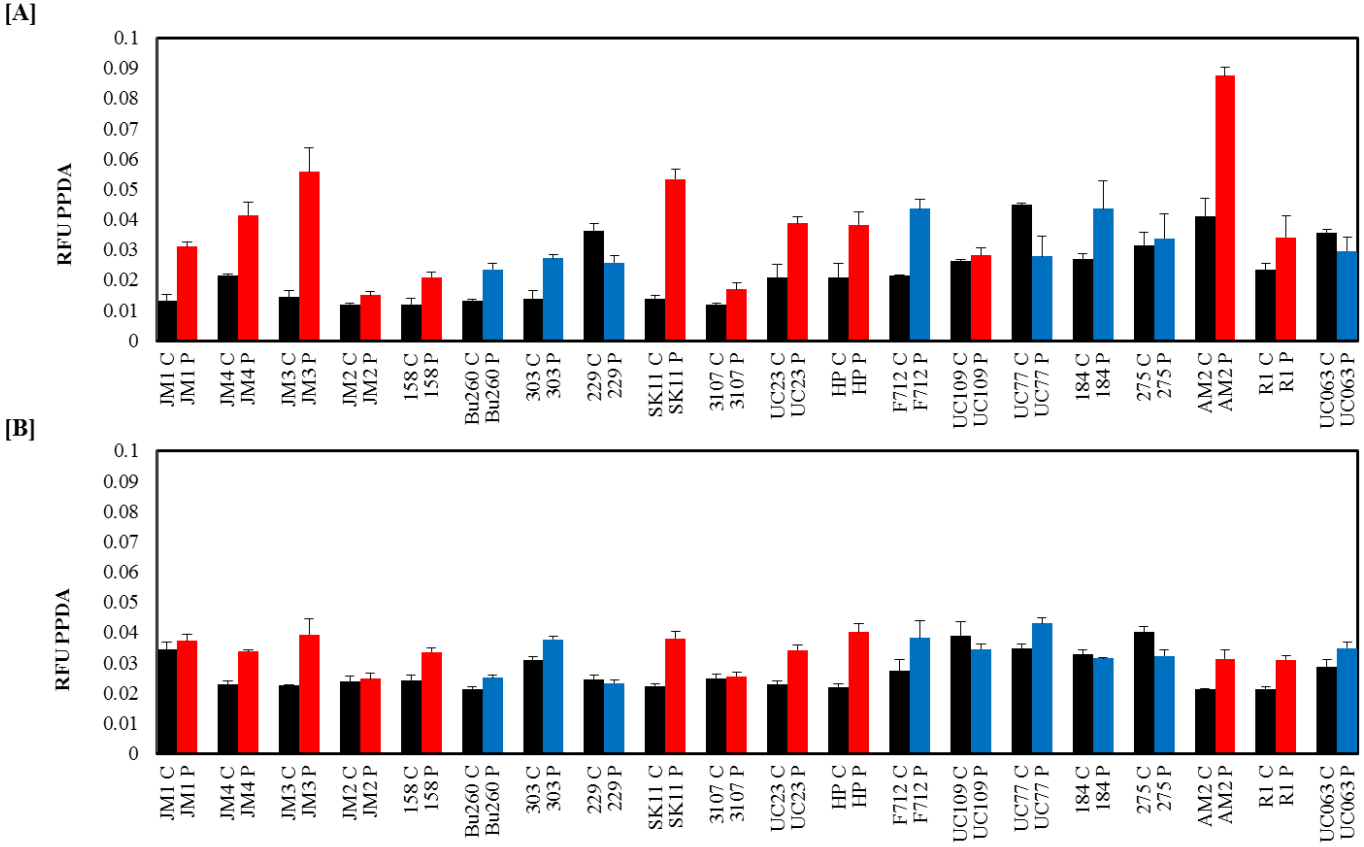
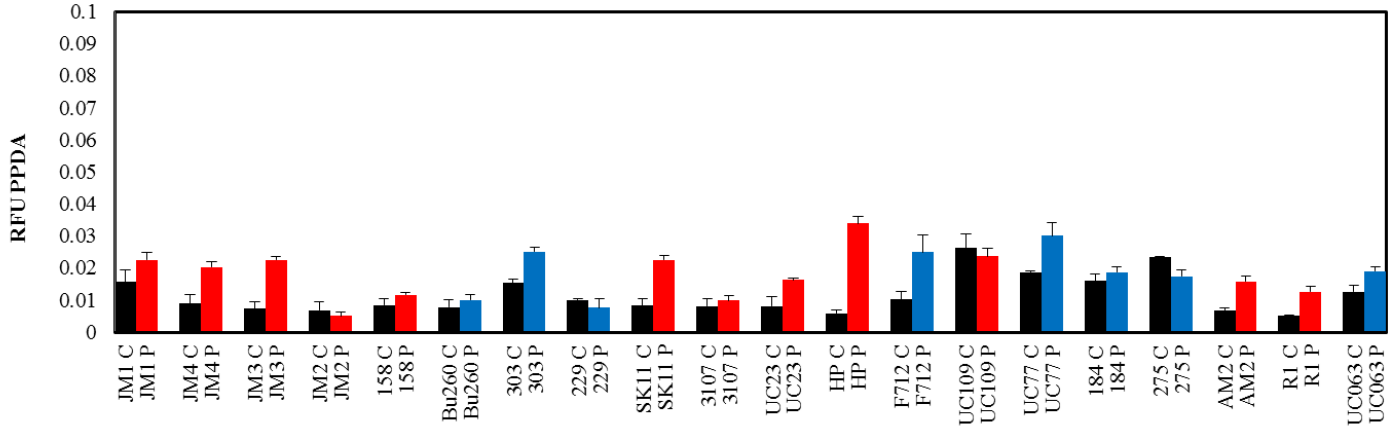


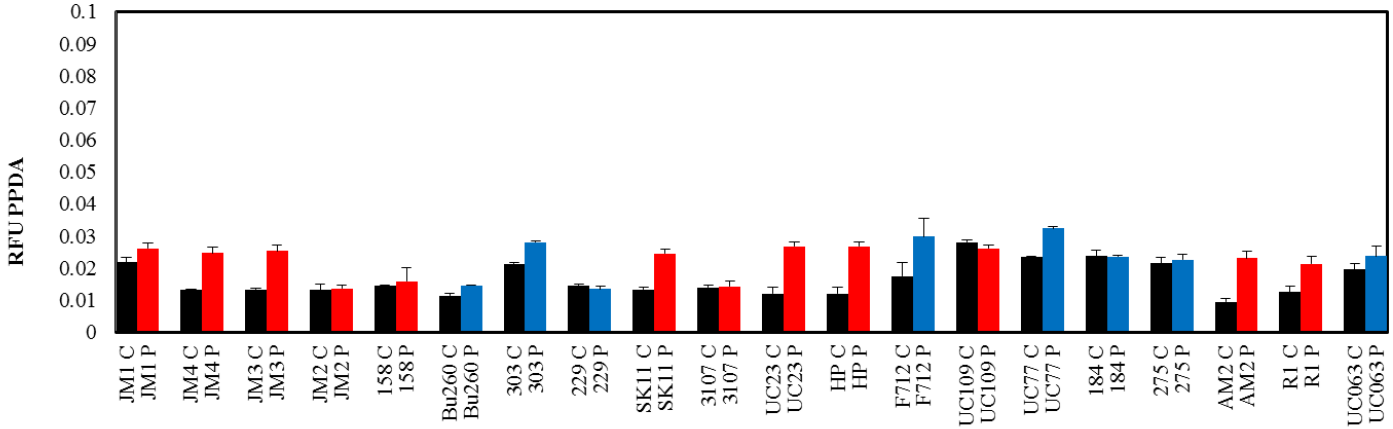
Figure 2.5: Peptidase activity expressed by lactococcal strains

Peptidase activity expressed by lactococcal strains against; [A] PepX, [B] PepA, [C] proline imino peptidase, [D] endopeptidase, [E] carboxy-peptidase, [F] PepN/C, [G] PepC and [H] PepN after growth at 30 °C for 5 hours (black, names marked with 'C'), and following the Pearce Activity test (red (*cremoris*) and blue (*lactis*), names marked with 'P')

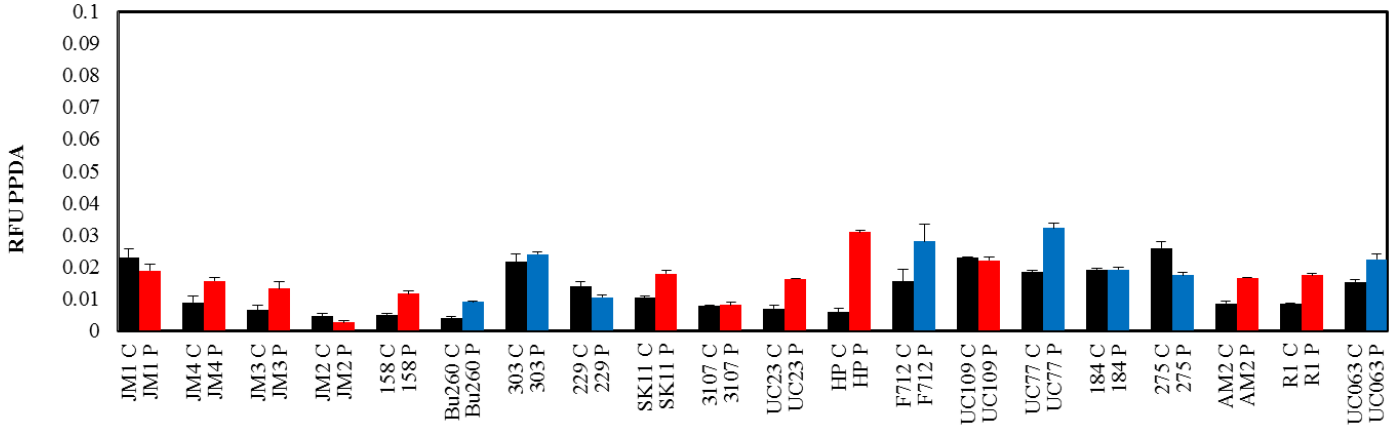
[C]


RFU PPDA
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
JM1 C JM1 P
JM4 C JM4 P
JM3 C JM3 P
JM2 C JM2 P
158 C 158 P
Bu260 C Bu260 P
303 C 303 P
229 C 229 P
SK11 C SK11 P
3107 C 3107 P
UC23 C UC23 P
HP C HP P
F712 C F712 P
UC109 C UC109 P
UC77 C UC77 P
184 C 184 P
275 C 275 P
AM2 C AM2 P
R1 C R1 P
UC063 C UC063 P

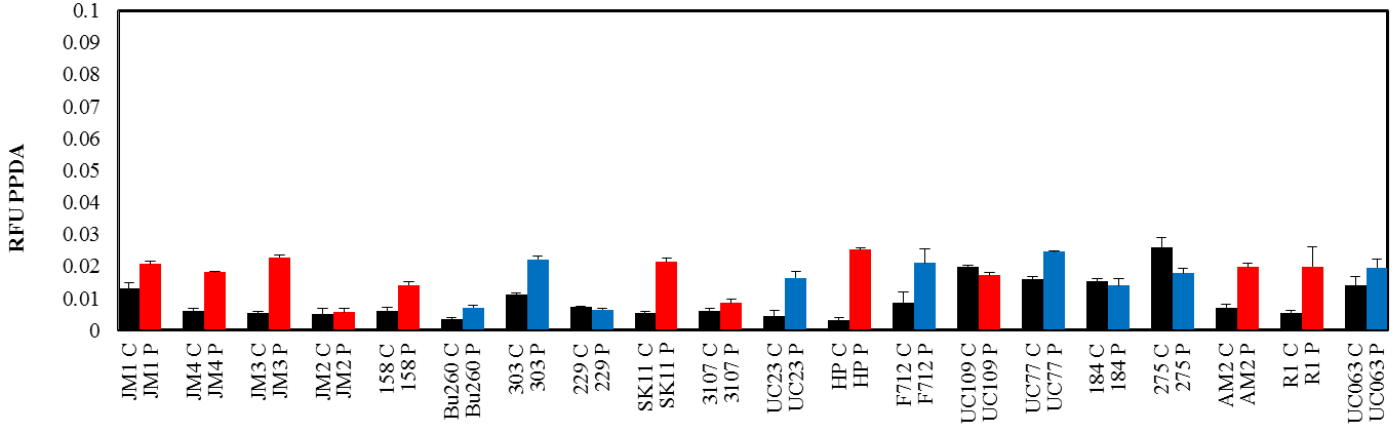
[D]


RFU PPDA
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
JM1 C JM1 P
JM4 C JM4 P
JM3 C JM3 P
JM2 C JM2 P
158 C 158 P
Bu260 C Bu260 P
303 C 303 P
229 C 229 P
SK11 C SK11 P
3107 C 3107 P
UC23 C UC23 P
HP C HP P
F712 C F712 P
UC109 C UC109 P
UC77 C UC77 P
184 C 184 P
275 C 275 P
AM2 C AM2 P
R1 C R1 P
UC063 C UC063 P

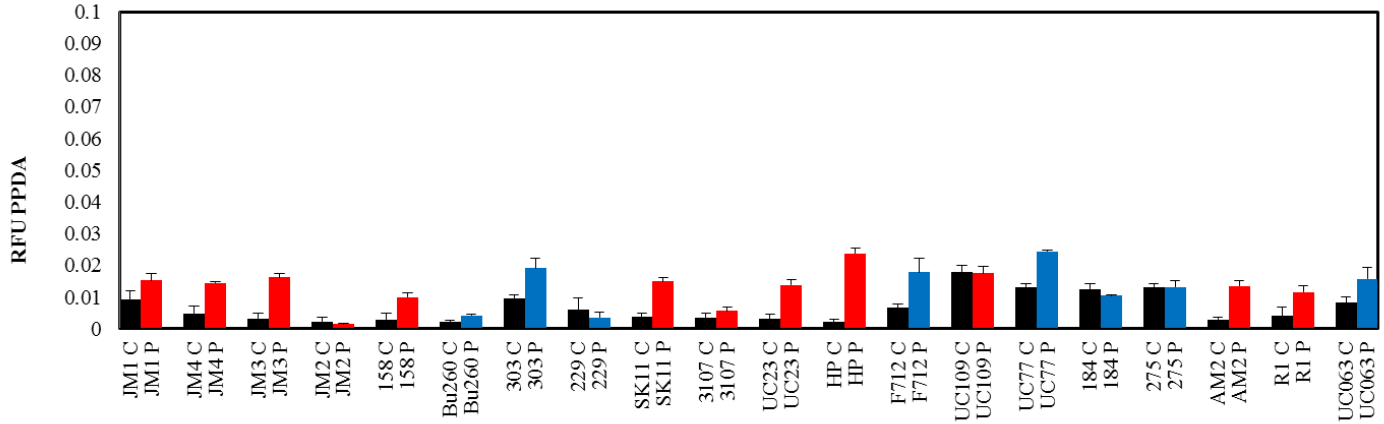
[E]


RFU PPDA
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
JM1 C JM1 P
JM4 C JM4 P
JM3 C JM3 P
JM2 C JM2 P
158 C 158 P
Bu260 C Bu260 P
303 C 303 P
229 C 229 P
SK11 C SK11 P
3107 C 3107 P
UC23 C UC23 P
HP C HP P
F712 C F712 P
UC109 C UC109 P
UC77 C UC77 P
184 C 184 P
275 C 275 P
AM2 C AM2 P
R1 C R1 P
UC063 C UC063 P

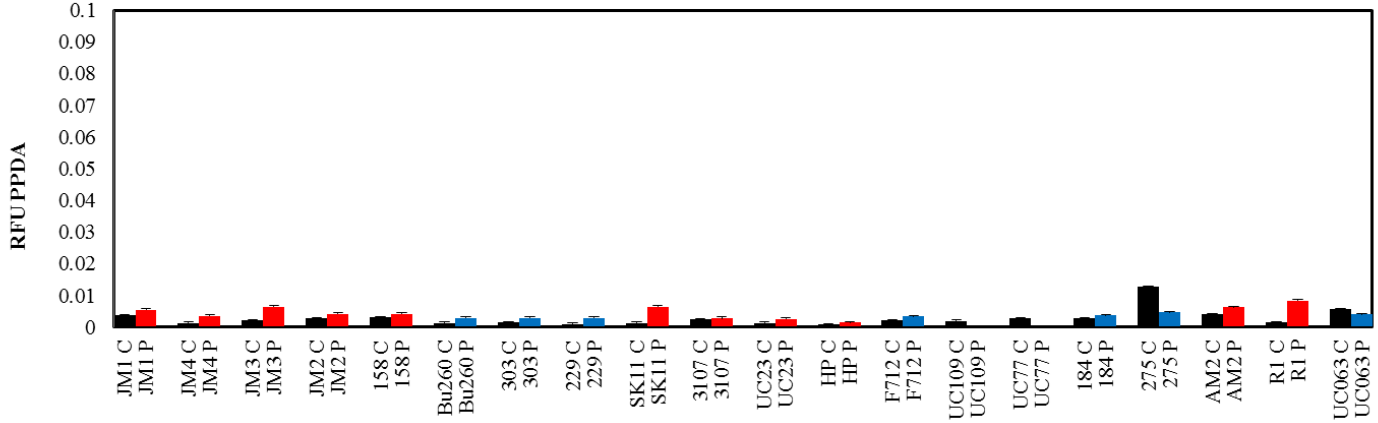
[F]


RFU PPDA
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
JM1 C JM1 P
JM4 C JM4 P
JM3 C JM3 P
JM2 C JM2 P
158 C 158 P
Bu260 C Bu260 P
303 C 303 P
229 C 229 P
SK11 C SK11 P
3107 C 3107 P
UC23 C UC23 P
HP C HP P
F712 C F712 P
UC109 C UC109 P
UC77 C UC77 P
184 C 184 P
275 C 275 P
AM2 C AM2 P
R1 C R1 P
UC063 C UC063 P

[G]


RFU PPDA
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
JM1 C JM1 P
JM4 C JM4 P
JM3 C JM3 P
JM2 C JM2 P
158 C 158 P
Bu260 C Bu260 P
303 C 303 P
229 C 229 P
SK11 C SK11 P
3107 C 3107 P
UC23 C UC23 P
HP C HP P
F712 C F712 P
UC109 C UC109 P
UC77 C UC77 P
184 C 184 P
275 C 275 P
AM2 C AM2 P
R1 C R1 P
UC063 C UC063 P

[H]


RFU PPDA
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
JM1 C JM1 P
JM4 C JM4 P
JM3 C JM3 P
JM2 C JM2 P
158 C 158 P
Bu260 C Bu260 P
303 C 303 P
229 C 229 P
SK11 C SK11 P
3107 C 3107 P
UC23 C UC23 P
HP C HP P
F712 C F712 P
UC109 C UC109 P
UC77 C UC77 P
184 C 184 P
275 C 275 P
AM2 C AM2 P
R1 C R1 P
UC063 C UC063 P

2.3.6 Flavour Capabilities – Esterase activity (lipolysis)

Lipolysis involves the breakdown of milk fats and hydrolysis of triglycerides into lipids and fatty acids, activities that are considered to be crucial for flavour development in cheese production [12]. Lipolytic assays utilizing p-nitrophenyl-butyrate for the detection of short chain esterase activity revealed a trend showing higher levels of esterase activity in strains used routinely in cheese fermentations, compared to their counterparts (Fig. 2.6). A marked increase in the level of esterase activity expressed by each of the strains was also observed between the cultures grown at 30 °C and those which underwent the Pearce assay, indicating that esterase activity is predominantly expressed intracellularly, therefore autolysis of the starter culture would appear to be a necessary prerequisite for these strains to contribute to lipolysis in cheese. Strains JM3, JM1 and AM2 were shown to exhibit the highest levels of activity, while the remaining strains appear to express this activity at a far lower level.

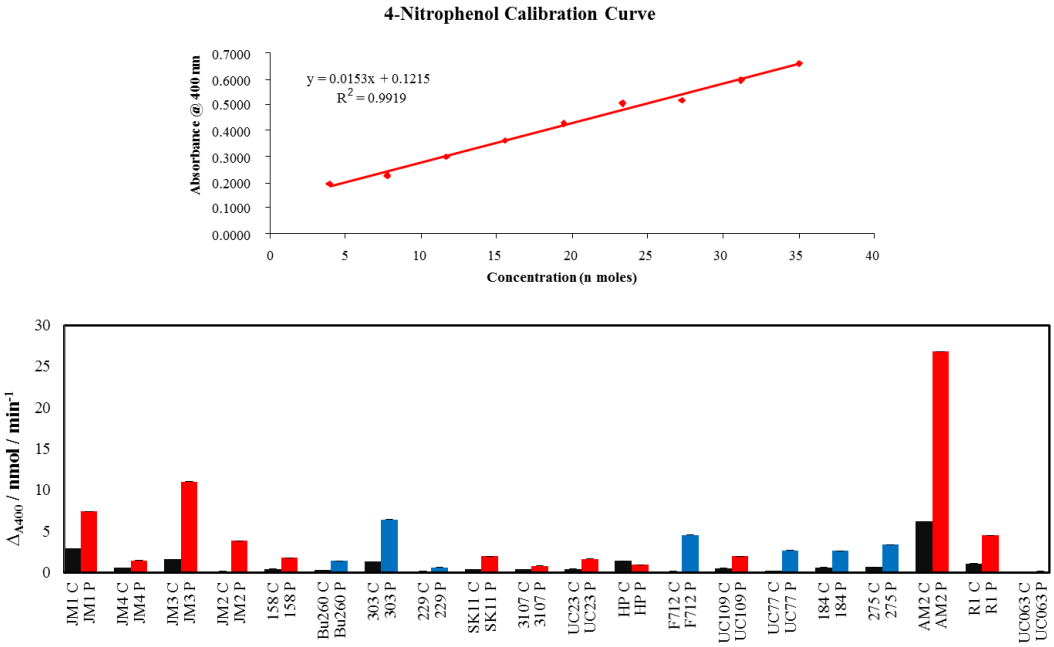


Figure 2.6: Short chain esterase activity

Standard curve for 4-nitrophenol and the deduced mathematical function used to calculate esterase activity for each of the starter cultures. Histogram of esterase activity; expressed by each strain grown under standard culture conditions at 30 °C (black) and after the Pearce activity test red (*cremoris*) and blue (*lactis*).

2.4 Discussion

The contribution of lactococcal starter strains to cheese flavour development is predominantly through the major flavour pathways of lactose, lactate and citrate utilisation, lipolysis, proteolysis and the catabolism of free amino acids [9]. The current report assessed the levels of key enzymes, including LDH, amino acid transferase, peptidases (PepX, PepN, PepC, PepA, endopeptidase, carboxypeptidase and PepI) and esterase for twenty strains which were quantified for extracellular activity and enzyme released through temperature-induced autolysis. This was used to produce a biochemical profile of these strains to be used in conjunction with the Pearce activity test data for the selection of strains to be sequenced for functional genomic analysis. Included in this study were four lactococcal starter cultures used in the Irish dairy industry for the production of low-fat Cheddar cheese, namely *L. lactis* subsp. *cremoris* JM1, JM2, JM3 and JM4. These strains are useful for this type of fermentation due to their relatively slow-growth rate, prominent autolysis and strong flavour performance. Comparative analysis of these starter cultures with a selection of strains was envisioned to be a practical method for the selection of novel starter cultures to be used for low-fat Cheddar cheese production.

The Pearce activity test was conducted to assess how these strains would behave under fermentation conditions; this test can be employed to assess the level of autolysis, as previously described for two dairy starter strains, *L. lactis* subsp. *cremoris* HP and *L. lactis* subsp. *cremoris* AM2 [6]. The level of induced autolysis was found to be greater in subsp. *cremoris* strains indicating that strains from this subspecies are potentailly more useful for Chedder-type fermentations. Higher cell viability was observed for subsp. *lactis* strains in comparison to their *cremoris* counterparts due to the higher temperature tolerance of *lactis* strains [14]. The Pearce

assay is used to simulate cheese making conditions which vary substantially with cheese type. The conditions used here were designed to replicate those of a Cheddar type cheese cooking cycle which is generally associated with subspecies *cremoris* strains. Further modifications of this assay would therefore appear to be necessary to improve its suitability for subspecies *lactis* strains. Furthermore, the intracellular enzyme LDH has been used as an indicator of autolysis in dairy starter strains [5, 27]. However, for the strains which underwent the Pearce activity test no significant increase in LDH was observed as compared to corresponding control cultures, which did not undergo temperature-induced autolysis. As such monitoring of viable cell counts via plating appears to be a more reliable method of estimating autolysis.

To assess the contribution of lactococcal strains to cheese flavour development, these strains were assayed for aminotransferase and peptidase activities. All strains were shown to exhibit aminotransferase activity against phenylalanine and methionine (in the latter case reaching a considerably higher specific activity level). Subspecies *cremoris* strains proved to demonstrate the highest activity levels, while three subsp. *lactis* strains Bu2-60, 303 and 229 also expressed a high level of aminotransferase activity indicating that these represent good candidates as starter cultures for commercial milk fermentations. Strains were assessed for peptidase activity utilising AMC substrates as previously demonstrated for lactococci [28-30]. The dominant peptidase activities expressed by each strain were those represented by the aminopeptidases PepA and the proline peptidase PepX which is of particular importance due to the high proline content of β-casein [25]. PepX activity was also observed to be higher for subsp. *cremoris* strains. Expressed levels of PepN/C activity were found to be lower in strains which had high levels of PepX activity, indicating that strains may exhibit a preference towards either the X-

prolyl dipeptidyl aminopeptidase PepX, or general aminopeptidases PepN/C. Significant increases in the level of esterase activity expressed by each of the strains were observed between the cultures grown at 30 °C and those which underwent the Pearce assay. This finding indicates that esterase activity is predominantly expressed intracellularly and that autolysis of the starter culture is therefore necessary for these strains to contribute to lipolysis in cheese.

While a number of useful parameters are explored in this chapter particularly related to cheese flavour development, these should be considered in parallel with several other properties related to dairy fermentation. For instance, *L. lactis* subsp. *cremoris* AM2, which was one of the strongest performing strains in this analysis in terms of flavour development, has been excluded from dairy fermentations in recent years due to its bacteriophage sensitivity. Therefore, further sequence-based analysis in the subsequent chapters will attempt to assess these factors and determine an overall genetic/genomic blueprint for starter culture selection.

2.5 Conclusion

In conclusion, the phenotypic analysis of four lactococcal starter cultures used in the Irish dairy industry for the production of low-fat Cheddar cheese allowed for the selection of potential novel starter cultures from the UCC starter culture collection which may be useful for the same type of fermentation. The *L. lactis* subsp. *cremoris* strains were found to perform in a similar manner to the industrial isolates JM1-JM4. Therefore, strains 158 and UC109 were selected for whole genome sequencing in addition to the industrial strains JM1-JM4, to further investigate their genetic composition. Furthermore, these data may permit genotype–phenotype links to be derived. In addition, five *L. lactis* subsp. *lactis* strains, namely UC77, 275, 229, 184 and UC063 were also selected for sequencing to increase the genetic diversity of the analysis.

2.6 References

[1] Schleifer K, Kilpper-Bälz R: Molecular and chemotaxonomic approaches to the classification of streptococci, enterococci and lactococci: a review. Syst Appl Microbiol. 1987, 10:1-19.

[2] Kelly W, Ward L: Genotypic vs. phenotypic biodiversity in *Lactococcus lactis*. Microbiol. 2002, 148:3332-3333.

[3] Limsowtin G, Powell I, Parente E: Types of starters. Dairy Starter Cultures. 1996, 101-129.

[4] Bron PA, Kleerebezem M: Engineering lactic acid bacteria for increased industrial functionality. Bioeng Bugs. 2011, 2:80-87.

[5] Sheehan A, O'Loughlin C, O'Cuinn G, Fitzgerald RJ, Wilkinson MG: Cheddar cheese cooking temperature induces differential lactococcal cell permeabilization and autolytic responses as detected by flow cytometry: implications for intracellular enzyme accessibility. J Appl Microbiol. 2005, 99:1007-1018.

[6] Wilkinson MG, Guinee TP, O'Callaghan DM, Fox PF: Autolysis and proteolysis in different strains of starter bacteria during Cheddar cheese ripening. J Dairy Res. 1994, 64:249-262.

[7] Høier E, Janzen T, Rattray F, Sørensen K, Børsting M, Brockmann E, Johansen E: The production, application and action of lactic cheese starter cultures. Technology of Cheesemaking. 1999, 99-131.

[8] Kleppen HP, Bang T, Nes IF, Holo H: Bacteriophages in milk fermentations: diversity fluctuations of normal and failed fermentations. Int Dairy J. 2011, 21:592-600.

[9] McSweeney PLH: Biochemistry of cheese ripening. Int J Dairy Technol. 2004, 57:127-144.

[10] Smit G, Verheul A, van Kranenburg R, Ayad E, Siezen R, Engels W: Cheese flavour development by enzymatic conversions of peptides and amino acids. Food Res Intl. 2000, 33:153-160.

[11] Tanous C, Kieronczyk A, Helinck S, Chambellon E, Yvon M: Glutamate dehydrogenase activity: a major criterion for the selection of flavour-producing lactic acid bacteria strains, Lactic acid bacteria: Genetics, Metabolism and Applications, Springer, 2002.

[12] McSweeney PLH, Sousa MJ: Biochemical pathways for the production of flavour compounds in cheeses during ripening: A review. Lait. 2000, 80:293-324.

[13] Verger R: 'Interfacial activation'of lipases: facts and artifacts. Trends Biotechnol. 1997, 15:32-38.

[14] Harrigan WF: Laboratory methods in food microbiology (3rd ed). Academic Press, San Diego, CA. 1998.

[15] Koch F, McMeekin T: A new direct nesslerization micro-Kjeldahl method and a modification of the Nessler-Folin reagent for ammonia. J Am Chem Soc. 1924, 46:2066-2069.

[16] Feirtag JM, McKay LL: Thermoinducible Lysis of Temperature-Sensitive *Streptococcus cremoris* Strains. J Dairy Sci. 1987, 70:1779-1784.

[17] Kato T, Nagatsu T, Kimura T, Sakakibara S: Fluorescence assay of *X*-prolyl dipeptidyl-aminopeptidase activity with a new fluorogenic substrate. Biochem Med. 1978, 19:351-359.

[18] Vorderwülbecke T, Kieslich K, Erdmann H: Comparison of lipases by different assays. Enz Microb Technol. 1992, 14:631-639.

[19] Martinussen J, Hammer K: The *carB* gene encoding the large subunit of carbamoylphosphate synthetase from *Lactococcus lactis* is transcribed monocistronically. J Bacteriol. 1998, 180:4380-4386.

[20] Kelleher P, Murphy J, Mahony J, Van Sinderen D: Next-generation sequencing as an approach to dairy starter selection. Dairy Sci. & Technol. 2015, 95:545-568.

[21] Schleifer K, Kraus J, Dvorak C, Kilpper-Bälz R, Collins M, Fischer W, Transfer of *Streptococcus lactis* and Related Streptococci to the Genus *Lactococcus* gen. nov. Sys Appl Microbiol. 1985, 6:183-195.

[22] Smit G, Smit BA, Engels WJM: Flavour formation by lactic acid bacteria and biochemical flavour profiling of cheese products. FEMS Microbiol Rev. 2005, 29:591.

[23] Wittenberger CL, Angelo N: Purification and Properties of a Fructose-1,6-Diphosphate-Activated Lactate Dehydrogenase from *Streptococcus faecalis*. J Bacteriol. 1970, 101:717-724.

[24] Smit BA, Engels WJM, Wouters JTM, Smit G: Diversity of l-leucine catabolism in various microorganisms involved in dairy fermentations, and identification of the rate-controlling step in the formation of the potent flavour component 3-methylbutanal. Appl Microbiol Biotechnol. 2004, 64:396-402.

[25] Law BA: Microbiology and biochemistry of cheese and fermented milk. Springer Science & Business Media, 2012.

[26] Meyer J, Spahni A: Influence of *x*-prolyl-dipeptidylaminopeptidase of *Lactobacillus delbrueckii* subsp. *lactis* on proteolysis and taste of swiss gruyere cheese. Milchwissenschaft. 1998, 53:449-453.

[27] Hannon JA, Wilkinson MG, Delahunty CM, Wallace JM, Morrissey PA, Beresford TP: Use of autolytic starter systems to accelerate the ripening of Cheddar cheese. Int Dairy J. 2003, 13:313-323.

[28] Casey MG, Meyer J: Presence of *X*-Prolyl-Dipeptidyl-Peptidase in Lactic Acid Bacteria. J Dairy Sci. 1985, 68:3212-3215.

[29] Lloyd RJ, Pritchard GG: Characterization of *X*-prolyl dipeptidyl aminopeptidase from *Lactococcus lactis* subsp. *lactis*. J Gen Microbiol. 1991, 137:49-55.

[30] Booth M, Fhaoláin IN, Jennings PV, O'Cuinn G: Purification and characterization of a post-proline dipeptidyl aminopeptidase from *Streptococcus cremoris* AM2. J Dairy Res. 1990, 57:89-99.

Chapter III

Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation

This chapter is published in BMC Genomics:
Comparative and functional genomics of the *Lactococcus lactis* taxon; insights into evolution and niche adaptation. DOI: 10.1186/s12864-017-3650-5
Philip Kelleher, Francesca Bottacini, Jennifer Mahony, Kieran N. Kilcawley and Douwe van Sinderen

Chapter III contents

3 Abstract 132

3.1 Introduction 133

3.2 Materials & Methods 135

3.2.1 Genome sequencing 135

3.2.2 General feature predictions 137

3.2.3 Comparative genomics 137

3.2.4 Phylogenetic analysis 138

3.2.5 Pan- and core-genome analysis 138

3.2.6 Growth conditions and media 139

3.2.7 Intracellular enzyme assays 139

3.2.8 Nucleotide sequence accession numbers 141

3.3 Results 142

3.3.1 General genome features 142

3.3.2 Phylogenetic analysis and genome synteny 146

3.3.3 Pan/core-genome analysis 151

3.3.4 Comparative analysis of orthologous genes 153

3.3.5 Metabolism and niche adaptation 157

3.3.6 Amino acid transport and metabolism 161

3.3.7 Lipid transport and metabolism 165

3.3.8 Plasmid Integration 167

3.4 Discussion 169

3.5 Conclusions 172

3.6 References 173

3.7 Supplementary Figure S3.1 182

3 Abstract

Lactococcus lactis is among the most widely studied lactic acid bacterial species due to its long history of safe use and economic importance to the dairy industry, where it is exploited as a starter culture in cheese production. In the current study, we report on the complete sequencing of sixteen *L. lactis* subsp. *lactis* and *L. lactis* subsp. *cremoris* genomes. The chromosomal features of these sixteen *L. lactis* strains were assessed with particular emphasis on discerning the *L. lactis* subspecies division, evolution and niche adaptation. The deduced pan-genome of *L. lactis* was found to be closed, indicating that the representative data sets employed for this analysis are sufficient to fully describe the chromosomal diversity of the taxon. Niche adaptation appears to play a significant role in governing the genetic content of each *L. lactis* subspecies, while (differential) genome decay and redundancy in the dairy niche is also highlighted.

3.1 Introduction

Lactococcus lactis is a Gram positive, catalase-negative, non-motile and coccoid bacterium [1]. *L. lactis* has a long history of safe use in the fermented food industry and as such enjoys a so-called "GRAS" (Generally Regarded as Safe) status. Lactococcal strains are particularly important to the dairy industry, where they are employed as starter cultures for cheese production. *L. lactis* has four component subspecies, two of which are routinely employed in the dairy fermentation sector, i.e. subspecies (subsp.) *cremoris* and subsp. *lactis* (and a biovariant; subsp. *lactis* biovar diacetylactis, which distinguishes itself based on citrate metabolism, see also below). The two remaining *L. lactis* subspecies, i.e. *L. lactis* subsp. *hordniae* isolated from the leafhopper *Hordnia circellata* [2], and *L. lactis* subsp. *tructae* isolated from brown trout, *Salmo trutta* [3], are considerably under-represented in both biological and genomic studies compared to their dairy-associated counterparts.

Genetically, a typical *L. lactis* chromosome ranges in size from ~2.2 to 2.6 Mb, often accompanied by a rich plasmid complement [4] and multiple integrated (remnant) prophages [5]. Reductive evolution and genome decay have previously been reported in 'domesticated', dairy *L. lactis* strains, particularly those belonging to subsp. *cremoris* [6, 7]. Niche adaptation by lactococcal strains has been investigated most thoroughly in relation to the dairy environment. In this particular niche, host adaptations appear to be mainly plasmid-encoded and two examples of this are lactose and citrate utilisation. Lactose utilisation in *L. lactis* is performed via the *lac* operon, which consists of the *lacABCDEFGX* genes and which is regulated by the repressor *lacR* [8, 9]. Citrate metabolism by citrate-positive (Cit^+) lactococci is mediated by the *citQRP* operon [10]. The classification of Cit^+ lactococci as *L.*

lactis subsp. *lactis* biovar diacetylactis has led to confusion as plasmid-encoded characteristics can be transferred from one strain to another and may lead to incorrect classification based on phenotype [11], highlighting the importance of genome sequencing for the correct characterisation of members of this taxon.

The advent of modern sequencing technologies has made whole genome analysis more accessible, and as a result there are now eighty-four lactococcal assemblies publicly available in the NCBI (National Centre for Biotechnology Information) database, fourteen of which represent complete genome sequences including the two prototypical stains *L. lactis* subsp. *lactis* IL1403 [12] and *L. lactis* subsp. *cremoris* MG1363 [13]. To date a number of comparative genomic studies have been conducted and have provided novel insights into the lipolysis [14], prophage [5, 6], proteolysis [15], taxonomy [16] and niche adaptation functions of these strains [17].

In the current study we applied one of the latest sequencing technologies, Single-Molecule-Real-Time (SMRT) sequencing developed by Pacific Biosciences [18, 19] to contribute a further sixteen complete lactococcal genomes to the public database. The increased dataset of complete lactococcal genomic sequences allows for the investigation of the corresponding pan-genome, which when closed defines the total number of genes encoded in the *L. lactis* taxon [20-22]. In the current study, the phylogeny, core and non-core genes, metabolism and niche-specific adaptations in terms of the total genetic content of the taxon were examined.

3.2 Materials & Methods

3.2.1 Genome sequencing

All genomes sequenced in this study are dairy isolates of *L. lactis* subsp. *lactis* and subsp. *cremoris*, with the exception of *L. lactis* subsp. *lactis* UC08 and UC11, which were isolated from fermented meat products (Table 3.1). Chromosomal DNA from *L. lactis* subsp. *cremoris* JM1, JM2, JM3 and JM4 was isolated as previously described [42]. Chromosomal DNA extraction from *L. lactis* subsp. *cremoris* 158, UC109, *L. lactis* subsp. *lactis* UC11, C10, UL8 UC08, 275, UC063, UC06 184, 229 and UC77 was performed by commercial sequencing service providers GATC Biotech Ltd. (Germany).

SMRT sequencing was performed on a Pacific Biosciences RS II sequencing platform (executed by GATC Biotech Ltd., Germany). *De novo* genome assemblies were performed using the Pacific Biosciences SMRTPortal analysis platform (version 2.3.1), utilizing the RS_HGAP_Assembly.2 protocol. Remaining low quality regions or sequencing conflicts were resolved by primer walking and Sanger sequencing of PCR products (through sequence service provider Eurofins MWG Operon, (Germany)).

Table 3.1: Lactococcal representative strains used in this study

Strain name	Genbank accession	Ecological niche	Sequencing technology	Year	Citation
subsp. *lactis*					
Il1403	AE005176	Dairy isolate	Sanger	2001	[12]
KF147	CP001834	Plant isolate	454-pyrosequencing & Illumina	2009	[57]
CV56	CP002365	Human isolate	454-pyrosequencing	2011	[58]
IO-1	AP012281	Drain water	Sanger	2012	[59]
KLDS 4.0325	CP006766	Koumiss	Illumina	2013	[60]
NCDO 2118	CP009054	Frozen peas	SOLiD, Ion PGM & Ion Torrent PGM	2014	[61]
SO	CP010050	Dairy isolate	Ion Torrent PGM	2014	[62]
AI06	CP009472	Açaí palm	454-pyrosequencing	2014	[26]
184	CP015895	Dairy isolate	PacBio SMRT	2016	**
229	CP015896	Dairy isolate	PacBio SMRT	2016	**
275	CP015897	Dairy isolate	PacBio SMRT	2016	**
UC06	CP015902	Dairy isolate	PacBio SMRT	2016	**
UC08	CP015903	Fermented meat	PacBio SMRT	2016	**
UC11	CP015904	Fermented meat	PacBio SMRT	2016	**
UC063	CP015905	Dairy isolate	PacBio SMRT	2016	**
UC77	CP015906	Dairy isolate	PacBio SMRT	2016	**
UL8	CP015908	Dairy isolate	PacBio SMRT	2016	**
C10	CP015898	Dairy isolate	PacBio SMRT	2016	**
subsp. *cremoris*					
SK11	CP000425	Dairy isolate	Sanger	2006	[7]
MG1363	AM406671	Dairy isolate	Sanger	2007	[13]
NZ9000	CP002094	Laboratory strain	Illumina	2010	[63]
A76	CP003132	Dairy isolate	Sanger	2011	[64]
UC509.9	CP003157	Dairy isolate	454-pyrosequencing & Illumina	2012	[6]
KW2	CP004884	Dairy isolate	454-pyrosequencing	2013	[65]
158	CP015894	Dairy isolate	PacBio SMRT	2016	**
UC109	CP015907	Dairy isolate	PacBio SMRT	2016	**
JM1	CP015899	Dairy isolate	PacBio SMRT	2016	**
JM2	CP015900	Dairy isolate	PacBio SMRT	2016	**
JM3	CP015901	Dairy isolate	PacBio SMRT	2016	**
JM4	CP015909	Dairy isolate	PacBio SMRT	2016	**

** Sequenced in the framework of this study.

3.2.2 General feature predictions

Following final genome assembly, Open Reading Frame (ORF) prediction was performed employing Prodigal v2.5 prediction software (http://prodigal.ornl.gov) and confirmed using BLASTX v2.2.26 alignments [28]. ORFs were automatically annotated using BLASTP v2.2.26 [28] analysis against the non-redundant protein databases curated by the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nig.gov/). Following automatic annotation, ORFs were manually curated using Artemis v16 genome browser and annotation tool (http://www.sanger.ac.uk/science/tools/artemis). This latter software tool was used for the combination and inspection of ORF-identification results, for adjustment of start codons (where necessary), and for the identification of pseudogenes. Finally ORF annotations were refined further where required using alternative functional searches using Pfam [43], HHpred [44], PHAST [45] and Uniprot/EMBL (http://www.uniprot.org/).

Transfer RNA (tRNA) and ribosomal RNA (rRNA) genes were predicted using tRNA-scan-SE v1.4 (http://lowelab.ucsc.edu/tRNAscan-SE/) and RNAmmer v1.2 (http://www.cbs.dtu.dk/services/RNAmmer/), respectively. Predicted RNA-specifying loci were manually added to each genome using Artemis v16.

3.2.3 Comparative genomics

The Mauve alignment tool was employed in order to perform whole genome alignments at the nucleotide level, and to explore synteny within the genomes and identify potential integration sites [46]. Genome synteny was explored and dotplots generated using Geopard v1.40 [47]. All sequence comparisons at the protein level were performed via all-against-all, bi-directional BLAST alignments [28]. An

alignment cut-off value of E-value 0.0001, > 30 % amino acid identity across 80 % of the sequence length was used. For analysis and clustering of these results, the Markov Clustering Algorithm (MCL) was implemented in the mclblastline pipeline v12-0678 [29]. To further analyse genomic functions, the deduced protein complement was categorised based on COG (clusters of orthologous groups) assignments [48]. Metabolic pathways encoded by *L. lactis* strains were predicted and mapped using KEGG (Kyoto Encyclopaedia of Genes and Genomes)[49, 50]. Logo motifs were produced using WebLogo 3 [51].

3.2.4 Phylogenetic analysis

The lactococcal supertree computation was performed by the BLAST-based comparative approach outlined above to identify a subset of 596 orthologous proteins. The subset was concatenated for each strain and an ungapped alignment was performed using MUSCLE v3.8.31 [52] with *Streptococcus thermophilus* LMG 18311 (Accession: CP000023) used as an outgroup. The phylogenetic tree was computed by the maximum-likelihood method in PhyML v3.0 and bootstrapped employing 1000 replicates [53]. The final tree file was visualised using ITOL (Interactive Tree of Life) (http://itol.embl.de/index.shtml). 16S rRNA trees were prepared in MEGA6. Alignments were performed using MUSCLE. The evolutionary history was inferred by the Neighbour-joining method [54].

3.2.5 Pan- and core-genome analysis

For the 30 available lactococcal genomes in this study, PGAP v1.0 [27] was used to perform the pan-genome analysis according to Heaps law pan-genome model [20]. The ORF content of each genome is organised in functional gene clusters via

the Gene Family method. ORFs which produce an alignment with a minimum of 50 % sequence identity across 50 % of the gene/protein length are clustered and a pan/core genome profile was subsequently generated.

3.2.6 Growth conditions and media

Bacterial strains used in this study are detailed in Table 3.1. *L. lactis* strains were routinely cultured at 30 °C in M17 broth (Oxoid) supplemented with 0.5 % glucose/lactose without agitation. Alternatively, and where indicated, strains were grown in 10 % RSM (reconstituted skimmed milk) at 30 °C without agitation.

3.2.7 Intracellular enzyme assays

Cells were prepared via a 1.5 % inoculum into 10 % RSM and grown overnight (16 hours) at 30 °C. Cells were then plated on M17 agarose supplemented with lactose to determine a viable plate count in cfu/ml. 50 ml of an overnight culture was added to 450 ml of borate buffer (0.05 M EDTA and 0.5 M borate pH8 with NaOH) and cells were collected by centrifugation (7000 rpm for 9 min). Cells were then washed in imidazole buffer (50 mmol/l imidazole and 10 mmol/l calcium chloride pH6.5) and pelleted by centrifugation (7000 rpm for 9 min). Cell pellet was re-suspended in 5 ml of lysis buffer (10 mM Tris-HCL, 50 mM $CaCl_2$, 300 mM NaCl, 10 mM imidazole, 25 mg/ml of lysozyme, pH 7.5). Cells were then sonicated five times (30 seconds each) with 30 seconds on ice in between each sonication, after which cell debris was removed by centrifugation (15,000 rpm for 25 minutes at 4°C). The resulting supernatant was then quantified for peptide/aminotransferase/esterase activity.

Detection of specific peptidase activities was conducted by fluorescence using 7-amino-4-methyl coumarin (AMC) coupled peptidase substrates; H-Lys-AMC.acetate (Lys-AMC) PepN and PepC, H-Asp (AMC)-OH (Asp-AMC) PepA, H-Pro-AMC.HBr (Pro-AMC) Proline imino peptidase, H-Gly-Pro-AMC. HBr (Gly-Pro-AMC) PepX, N-Suc-Gly-Pro-Leu-Gly-Pro-AMC (Gly-Pro-Leu-Gly-Pro-AMC) Endopeptidase and CBZ-Gly-Pro-AMC (Z-Gly-Pro-AMC) Carboxypeptidase, sourced from Bachem AG through VWR Ireland. The protocol was performed as described by Kato and colleagues [55], with the exception of reduced volumes for high throughput screening in 96-well plates. Released fluorescence was measured on a SpectraMax M3 Multi-Mode Microplate Reader from Molecular Devices. Enzyme activity was calculated in RFU PPDA (1 RFU = the amount of uM of AMC released min^{-1} by 1 mg of protein).

Amino acid transferase activity was determined (for Phe and Met) as previously described by Cavanagh and colleagues [16]. The final absorbance was read at wavelength, 300 nm in triplicate on a DU Series 730 spectrophotometer from Beckman Coulter, blanking the machine between each measurement. Standard curves were prepared for phenylalanine and methionine using phenylpyruvate and α-ketomethylthiobutyrate, respectively. Amino acid transferase activity was then expressed as micromoles per minute per milligram of protein.

Detection of short chain esterase activity was conducted via a spectrophotometric assay as previously described [56], utilising *p*-nitrophenyl butyrate as a substrate. Absorbance was measured on a DU Series 730 spectrophotometer from Beckman Coulter. All activities measured were normalised for each strain based on cell count.

3.2.8 Nucleotide sequence accession numbers

L. lactis subsp. *lactis* Il1403 AE005176, *L. lactis* subsp. *lactis* KF147 CP001834, *L. lactis* subsp. *lactis* CV56 CP002365, *L. lactis* subsp. *lactis* IO-1 AP012281, *L. lactis* subsp. *lactis* KLDS 4.0325 CP006766, *L. lactis* subsp. *lactis* NCDO 2118 CP009054, *L. lactis* subsp. *lactis* SO CP010050, *L. lactis* subsp. *lactis* AI06 CP009472, *L. lactis* subsp. *lactis* 184 CP015895, *L. lactis* subsp. *lactis* 229 CP015896, *L. lactis* subsp. *lactis* 275 CP015897, *L. lactis* subsp. *lactis* UC06 CP015902, *L. lactis* subsp. *lactis* UC08 CP015903, *L. lactis* subsp. *lactis* UC11 CP015904, *L. lactis* subsp. *lactis* UC063 CP015905, *L. lactis* subsp. *lactis* UC77 CP015906, *L. lactis* subsp. *lactis* UL8 CP015908, *L. lactis* subsp. *lactis* C10 CP015898, *L. lactis* subsp. *cremoris* SK11 CP000425, *L. lactis* subsp. *cremoris* MG1363 AM406671, *L. lactis* subsp. *cremoris* NZ9000 CP002094, *L. lactis* subsp. *cremoris* A76 CP003132, *L. lactis* subsp. *cremoris* UC509.9 CP003157, *L. lactis* subsp. *cremoris* KW2 CP004884, *L. lactis* subsp. *cremoris* 158 CP015894, *L. lactis* subsp. *cremoris* UC109 CP015907, *L. lactis* subsp. *cremoris* JM1 CP015899, *L. lactis* subsp. *cremoris* JM2 CP015900, *L. lactis* subsp. *cremoris* JM3 CP015901, *L. lactis* subsp. *cremoris* JM4 CP015909 and *S. thermophilus* LMG 18311 CP000023.

3.3 Results

3.3.1 General genome features

In this study, the chromosomal features of thirty *L. lactis* strains were assessed, eighteen of which belong to subspecies *lactis* and a further twelve to subspecies *cremoris* based on phylogenetic analysis of 16S RNA. For all selected strains, complete genome assemblies were available, of which fourteen were obtained from the NCBI (National Centre for Biotechnology Information) database, while the remaining sixteen were sequenced as part of the current study using the SMRT sequencing approach (Table 3.1). Although the NCBI database contains in total eighty four *L. lactis* genome assemblies only those, which are fully finished (i.e. present in the data base as a single chromosomal contig), were selected for this project due to the inherent limitations of draft assemblies. Briefly, the order and orientation of contigs of such draft assemblies remains unresolved and the differentiation between traits, which are verified to be chromosomally-encoded versus plasmid-encoded, is not possible particularly when one considers plasmid integration events. Most notably, however it is the finite nature of a finished genome which facilitates the comparison of the full genetic content of a strain rather than most of the genetic content, whereas in the case of a draft genome the likelihood of error from missing genes or incorrect copy number is significantly higher [23, 24].

The thirty *L. lactis* strains included in this study encompass isolates from six different ecological niches; dairy, plant, meat, fermented foods, human isolate (this is a vaginal isolate of a healthy woman) and a strain isolated from a sink drain, with the vast majority obtained from the dairy environment, most notably for the production of cheese (Table 3.1). Comparison of the thirty lactococcal genomes

established an average chromosome length of 2.428 Mbp, where it should be noted that generally the genomes of subsp. *lactis* are larger than their subsp. *cremoris* counterparts (Table 3.2). Genomes belonging to the subsp. *cremoris* contain a higher proportion of pseudogenes and insertion sequence (IS) elements/transposons, indicative of transpositions and (associated) genome decay within the subsp. *cremoris* genome. A defining characteristic of both subspecies is evident in the number of plasmids within each strain. *L. lactis* carries many niche-specific adaptations within its plasmid complement, particularly for the dairy environment, such as lactose utilisation and casein utilisation, and this is evident in the larger plasmid complement observed for subsp. *cremoris* strains predominantly isolated from the dairy niche (a detailed functional and comparative analysis of the plasmid complement will be presented in Chapter IV). A substantial proportion of the observed genomic diversity is due to a variable number of integrated prophage elements (Table 3.2).

General feature extractions conducted on each of the chromosomes generated an overall average of 2344 predicted CDS (Coding Sequences) per chromosome of which 77.6 % can be functionally assigned using BLAST (Basic Local Alignment Search Tool) based on *in silico* predictions, while the remaining 22.4 % are assigned as hypothetical proteins (Table 3.2).

Table 3.2: General genome features of thirty representative *L. lactis* genomes

Strain	Genome length (Kbp)	CDS	tRNA features	rRNA features	Hypothetical proteins %	Assigned function %	Pseudo genes	IS elements/ transposases	Prophage	Plasmids	GC %
L. lactis* subsp. *lactis											
184	2343	2312	51	15	19.6	80.4	15	59	2 In* 6 Re#	3	35.16
229	2455	2541	56	15	20.2	79.8	15	94	4 In 3 Re	5	35.19
275	2496	2418	58	18	20.2	79.8	14	43	3 In 6 Re	4	35.49
UC06	2571	2472	61	18	21.7	78.3	8	35	2 In 3 Re	3	35.26
UC08	2382	2246	62	18	20.0	80.0	14	18	2 Re	3	35.00
UC11	2382	2237	60	19	20.0	80.0	16	17	2 Re	6	35.00
UC063	2393	2361	59	18	19.2	80.8	14	59	3 In 5 Re	5	35.32
UC77	2538	2541	66	21	19.0	81.0	12	96	5 In 3 Re	2	35.26
UL8	2422	2405	59	17	18.5	81.5	13	56	3 In 7 Re	3	35.29
C10	2336	2294	50	15	17.7	82.3	21	53	5 In 3 Re	1	35.30
IL1403	2366	2267	62	18	21.0	79.0	43	43	3 In 3 Re	-	35.33
KLDS 4.032!	2589	2587	64	19	34.0	66.0	56	39	4 In 7 Re	-	35.36
NCDO 2118	2555	2334	66	19	28.0	72.0	52	16	2 In 3 Re	1	34.91
KF147	2598	2537	68	19	19.5	80.5	93	29	2 In 4 Re	1	34.91
SO	2489	2281	64	19	21.5	78.5	126	45	3 In 3 Re	-	35.23
AI06	2398	2197	61	19	22.9	77.1	2	5	1 In 1 Re	-	35.04
CV56	2399	2301	62	19	23.7	76.3	51	31	2 In 4 Re	5	35.24

IO-1	2422	2233	65	18	23.1	76.9	8	13	1 In 1 Re	-	35.10
Average: (*lactis*)	**2451**	**2364**	**60**	**18**	**21.6**	**78.4**	**31**	**41**	**3 In 4 Re**	**2.3**	**35.18**
L. lactis* subsp. *cremoris											
158	2250	2078	60	19	17.9	81.1	106	150	2 Re	6	35.88
UC109	2248	2081	60	19	20.0	80.0	98	149	2 Re	6	35.91
JM1	2397	2308	60	19	20.5	79.5	74	243	1 In 6 Re	7	36.01
JM2	2374	2316	58	19	19.6	80.4	68	167	1 In 3 Re	4	35.80
JM3	2454	2411	59	19	23.7	76.3	60	163	2 In 3 Re	5	35.87
JM4	2380	2293	60	19	20.9	79.1	88	181	1 In 4 Re	5	35.83
UC509.9	2250	1947	60	19	18.5	81.5	182	125	1 Re	8	35.88
SK11	2439	2390	61	20	26.2	73.8	144	159	2 In 3 Re	5	35.86
A76	2453	2643	57	19	25.8	74.2	193	198	2 In 7 Re	4	35.88
KW2	2427	2268	61	19	20.8	79.2	-	3	1 In	-	35.74
MG1363	2530	2516	62	7	30.8	69.2	81	60	2 In 4 Re	1	35.75
NZ9000	2530	2514	65	19	35.3	64.7	99	66	2 In 5 Re	-	35.74
Average: (*cremoris*)	**2394**	**2323**	**60**	**18**	**23.3**	**76.6**	**100**	**138**	**1 In 3 Re**	**4.25**	**35.84**
Average: (*lactis* & *cremoris*)	**2428**	**2344**	**60**	**18**	**22.3**	**77.6**	**59**	**80**	**2 In 4 Re**	**3.1**	**35.45**

[*]In: Complete intact prophage [#]Re: Partial/remnant prophage

3.3.2 Phylogenetic analysis and genome synteny

To investigate the phylogenetic relationship between the selected lactococcal isolates, a multifaceted approach was employed. Firstly, the 30 genomes were aligned based on 16S rRNA sequences with *Streptococcus thermophilus* used as an out-group to root the phylogenetic tree, resulting in a clear division into two major clades that correspond to the subsp. *lactis* and subsp. *cremoris* division (Fig. 3.1A). In order to improve the phylogenetic resolution of the analysis, a second approach was employed by constructing a phylogenetic supertree of 596 conserved orthologous proteins using an approach that has previously also been applied to other species [22, 25]. The conserved orthologues were selected based on all-against-all reciprocal BLASTP analysis with an e-value cut-off of 0.0001 and MCL (Markov Clustering) in order to identify single-copy genes conserved across all 31 (30 *L. lactis* plus *S. thermophilus* out-group) genomes in the phylogenetic analysis. The generated supertree displays the same bifurcation observed for the 16S rRNA analysis, substantiating this clear genomic differentiation between the two subspecies. This is also indicative of a unique allelic type for genes from subsp. *lactis* isolates in comparison to those from subsp. *cremoris* isolates, and is in agreement with the described differences in average nucleotide identity and tetranucleotide frequency correlation coefficients between the two subspecies [16]. To investigate a subspecies-specific allelic type, a subset of individual housekeeping genes from each of the genomes were aligned (involving the following genes: *radC* (Supp. Fig S3.1), *groEL*, *grpE*, *recX*, *ssbA*, *recA*, *recQ*, *rimM*, *radA*, and *hsp10* (Data not shown)), again resulting in each instance in a clear divide between representatives of each subspecies thus adding further evidence for an evolution-driven speciation event.

The lactococcal supertree revealed also a number of subclades of which members seem to be ordered based on niche specificity. Dairy isolates of subsp. *cremoris* cluster together into one clade, distinct from *L. lactis* KW2 isolated from fermented corn, while *L. lactis* NZ9000 and its parent strain *L. lactis* MG1363, which originated from the dairy niche formed their own clade. Dairy isolates of subsp. *lactis* also grouped together, with the exception of *L. lactis* UC06 and *L. lactis* SO. Furthermore, subsp. *lactis* isolates from meat and fermented foods each formed separate clades (Fig. 3.1B)

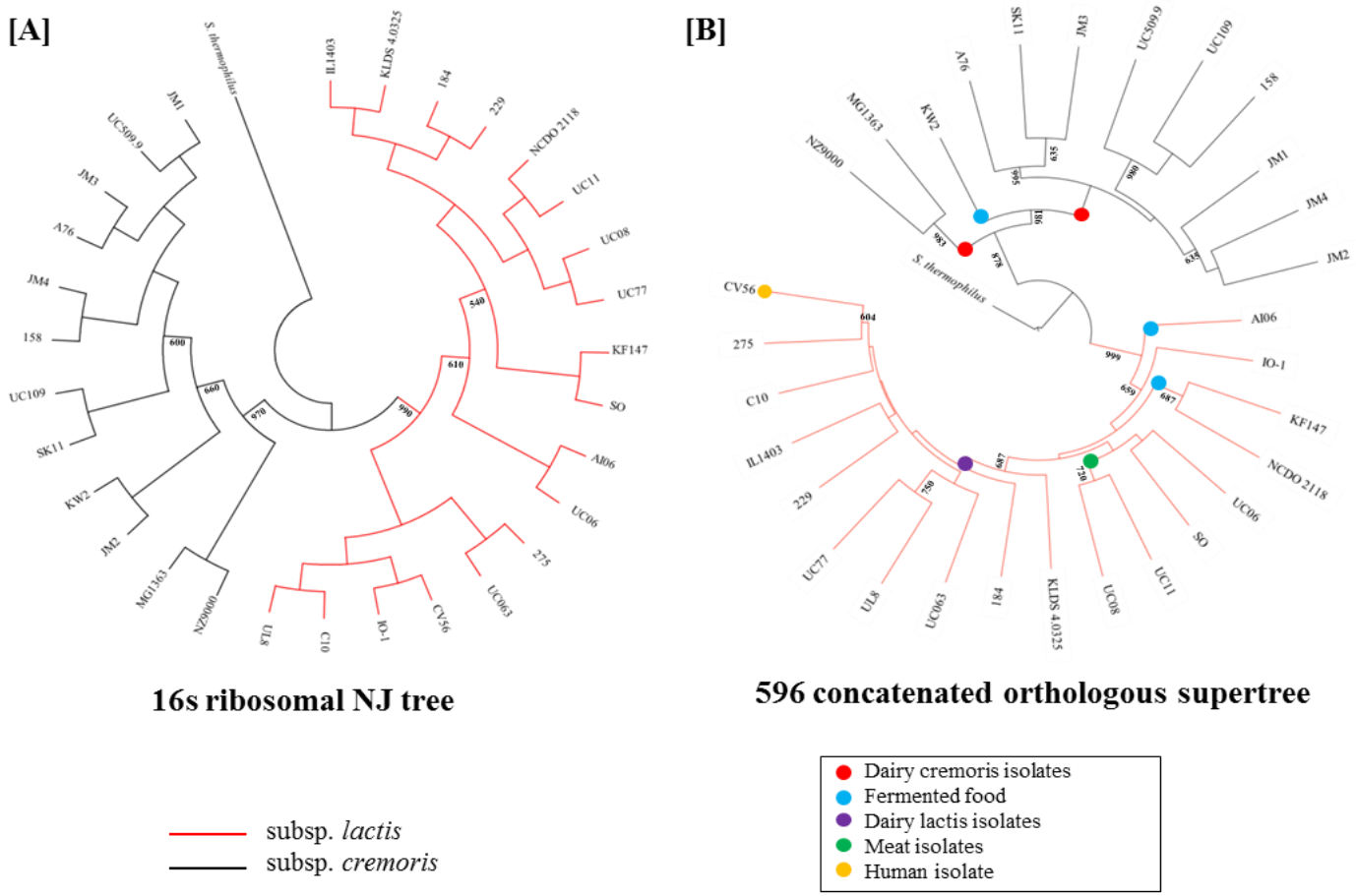


Figure 3.1: Phylogenetic analysis of *L. lactis* taxon

A) 16S neighbour-joining (NJ) tree, resulting from the alignment of the 16S rRNA-encoding genes of 30 *L. lactis* isolates. The corresponding 16S rRNA-specifying sequence of *S. thermophilus* LMG 18311 was used as to root the tree. B) Multilocus supertree resulting from the alignment of 596 orthologous genes selected from the core genome. Ecological niche of representative clades is also indicated.

To assess the synteny of the lactococcal genomes, whole genome nucleotide alignments were performed and represented as a dotplot matrix (Fig. 3.2). *L. lactis* subsp. *lactis* 184 was used as a representative strain for the subspecies, first aligned against itself and then against the remaining seventeen subsp. *lactis* genomes. This approach was also employed for the subsp. *cremoris* genomes using *L. lactis* subsp. *cremoris* 158 as the representative strain. Genome synteny was conserved in the *lactis* subspecies with the exception of the *L. lactis* subsp. *lactis* AI06 chromosome, which revealed a large inversion between coordinates 900 Kbp and 1633 Kbp as previously reported [26].

Genome synteny was significantly less conserved among the subsp. *cremoris* strains, with in particular *L. lactis* subsp. *cremoris* strains A76, JM1, JM2, MG1363 and NZ9000 presenting with multiple chromosomal inversions. In the case of genomes sequenced within the scope of this study (by SMRT sequencing, which generates long individual reads; average ~8 Kbp), these inversions are assumed to be genuine inversions rather than assembly errors. Visual inspection of the SMRT assembly at points intersecting these inversions indeed identified reads that in each case comfortably bridge the inversion points. The increased incidence of chromosomal inversions within these genomes is in accordance with the observed high number of transposons and other mobile elements (Table 3.2). The suspected role of mobile genetic elements in promoting chromosomal inversions was corroborated by sequence inspection of the borders of each of the identified inverted regions, which revealed in all incidences the presence of multiple transposable elements or integrated prophage(s).

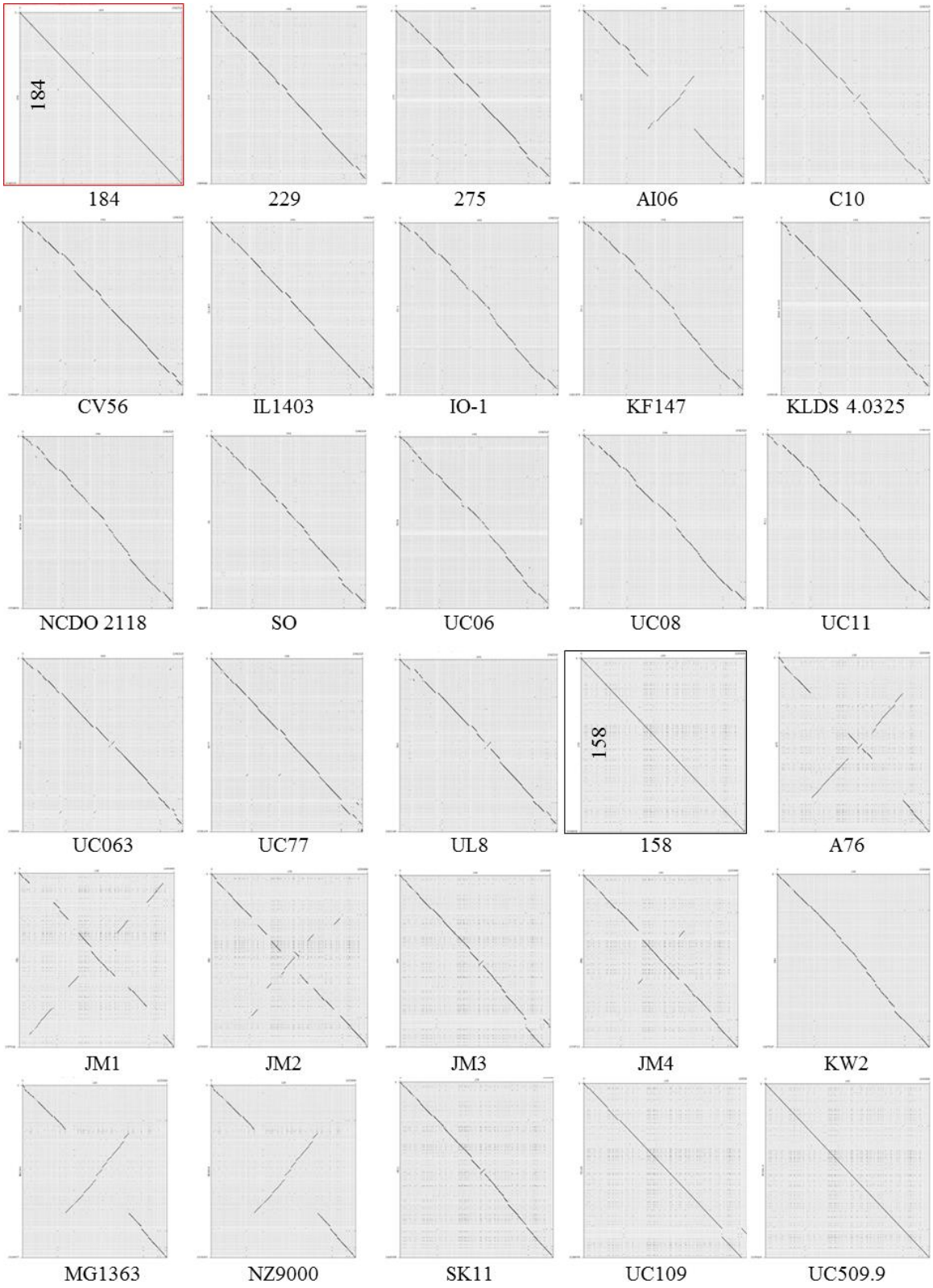


Figure 3.2: Whole genome nucleotide dotplots
Whole genome nucleotide alignments of thirty fully sequenced *L. lactis* genomes. Alignments 1(red)-18 represent subsp. *lactis* genomes. Alignments 19(black)-30 represent subsp. *cremoris* genomes.

3.3.3 Pan/core-genome analysis

To evaluate current sequencing efforts of the *L. lactis* taxon and to determine if additional genome sequencing is necessary to provide a complete overview of the chromosomal diversity of this taxon, pan-genome analysis was applied using the PGAP v1.0 pipeline [27]. The analysis was applied to the chromosomes of *L. lactis* only and excluded plasmid sequences. The resulting graph (Fig. 3.3) reveals an asymptotic curve increasing at an average rate of 209.44 genes for the first eleven chromosomes analysed. Beyond this point, the rate of pan-genome increase slows to an average of 86 genes per genome added for the remaining nineteen strains in the analysis resulting in a pan-genome constituted by 5906 genes. The majority of new genes added at this point in the analysis are short hypothetical CDSs which do not contribute greatly to our current understanding of the genetic diversity of these strains. The deduced mathematical function is also displayed (Fig. 3.3) and the exponential value (<0.5) indicates that the pan-genome is in a closed state [20].

Using the approach outlined above, it was also possible to deduce that the core genome of *L. lactis* consists of 1129 genes (Fig. 3.3). Conversely, when the subspecies are separated and the analysis repeated, the core genome size increases to 1406 genes for subsp. *cremoris* and 1413 genes for subsp. *lactis*, revealing that 277 and 284 core genes, respectively, are uniquely conserved for each subspecies. Overall, both analyses show that *L. lactis* contains an essentially closed pan-genome (excluding the plasmid complement) and that a sufficient number of strains have been included to describe the complete genetic repertoire of the taxon.

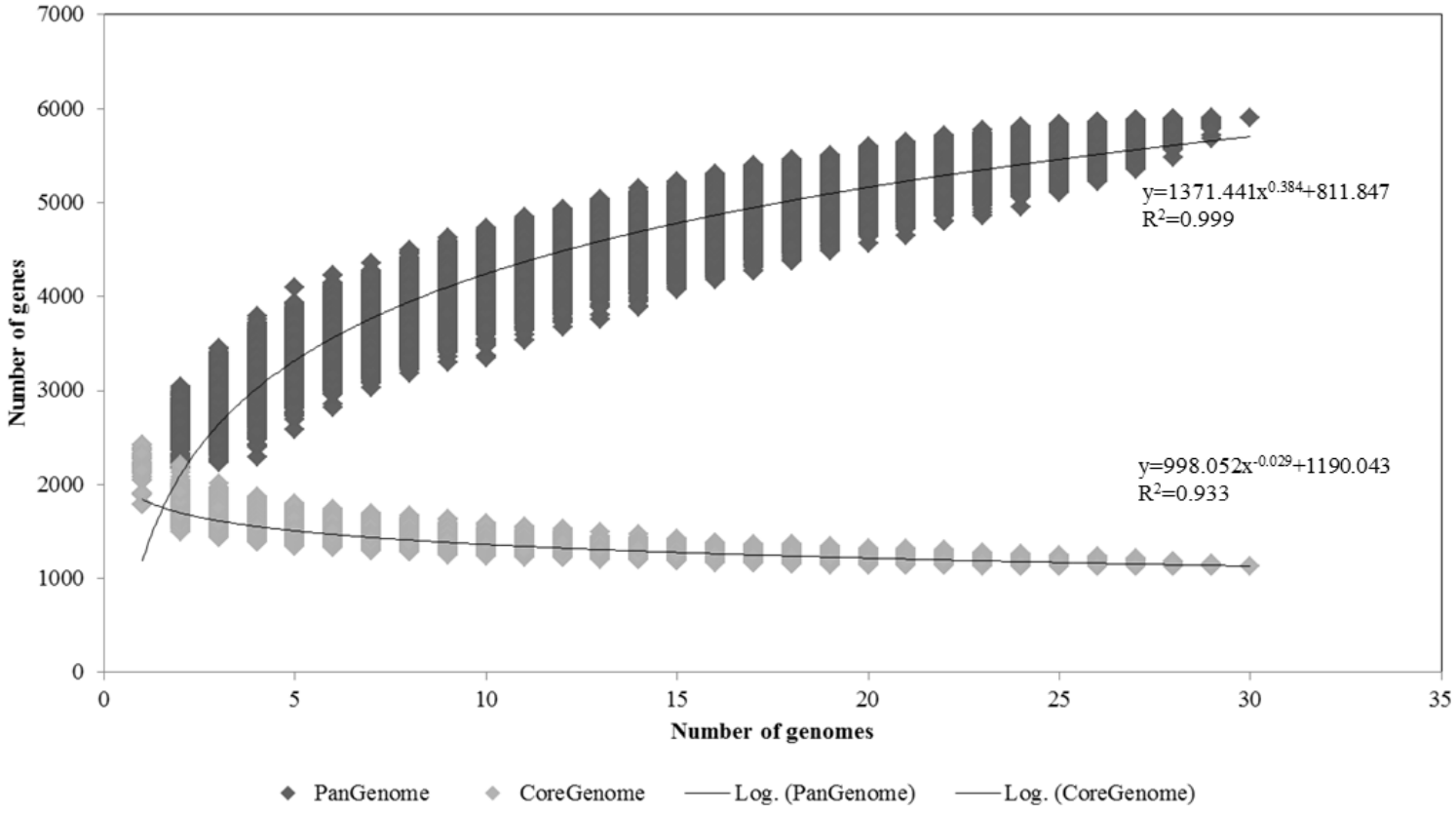


Figure 3.3: Pan-genome and core-genome analysis of *lactococcal chromosomes*
Pan-genome profile displays accumulated number of new genes in the *L. lactis* pan-genome plotted against the number of genomes added and the accumulated number of genes attributed to the core-genome plotted against the number of added genomes. The deduced mathematical functions are also indicated.

3.3.4 Comparative analysis of orthologous genes

To assess the level of (functional) diversity within the lactococcal core and dispensable genomes, comparative analysis was performed via all-against-all, bi-directional BLASTP alignment, and clustering implemented in the MCL pipeline [28, 29]. The core genome of 1129 genes, as defined above, was found to comprise 904 orthologous (single copy) gene families and 225 paralogous (multi-copy) gene families. Gene families unique to each chromosome were also calculated (Fig. 3.4A) and totalled 757 unique gene families across the 30 assessed *L. lactis* isolates. BLASTP analysis showed that 65 % of these unique or dispensable gene families encode proteins of unknown function, while a further 16 % encode phage proteins acquired through the integration of a particular prophage-like element. The remaining unique gene families were predominantly found to be representing plasmid integration events encoding proteins involved in mobilisation and conjugation, integrated mobile elements such as transposases and IS elements, or systems that provide specific benefit to the bacterium such as restriction-modification systems, bacteriocin production, and sugar transport and metabolism.

Cluster of Orthologous Group (COG) analysis was employed to further classify both the core and dispensable genome of *L. lactis*. The thirty lactococcal chromosomes analyzed in this study were classified using COG analysis. The core genome was predominantly composed of genes involved in housekeeping functions, fundamental to growth and survival, while 24 % of the genes contained in the core genome were assigned to COG groups [R] and [S] representing genes, for which a general function was predicted or which are of unknown function (Fig. 3.4B).

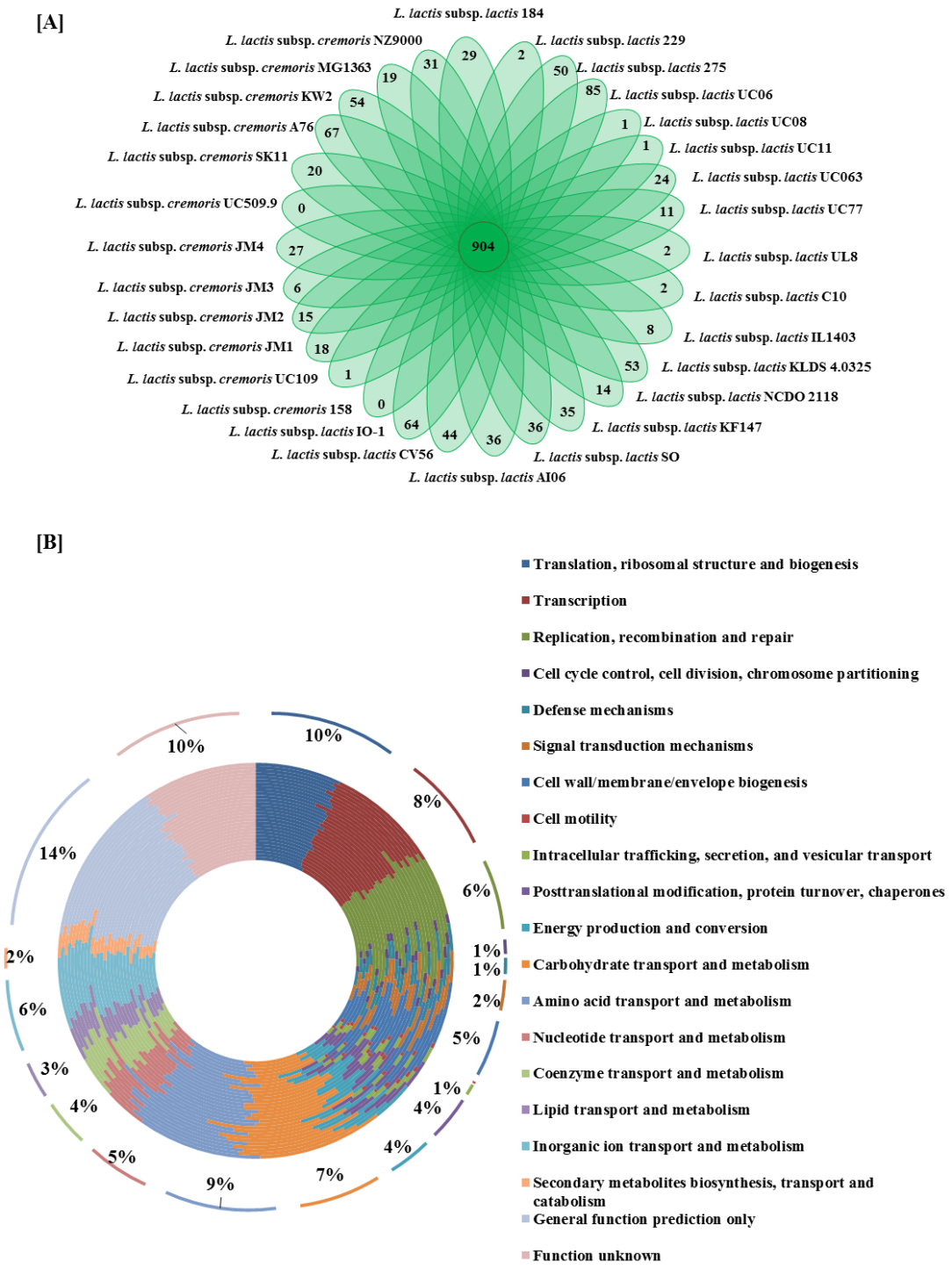


Figure 3.4: Comparative genomics of orthologous protein groups
A) Venn diagram displaying core gene families obtained by MCL clustering, and unique genes of 30 *L. lactis* isolates. B) Cluster of Orthologous Groups (COGs) classification of *L. lactis*. Circles from inner to outer represent: *L. lactis* 158, *L. lactis* 184, *L. lactis* 229, *L. lactis* 275, *L. lactis* C10, *L. lactis* JM1, *L. lactis* JM2, *L. lactis* JM3, *L. lactis* JM4, *L. lactis* KF147, *L. lactis* KLDS 4.0325, *L. lactis* KW2, *L. lactis* MG1363, *L. lactis* NCDO 2118, *L. lactis* NZ9000, *L. lactis* SK11, *L. lactis* SO, *L. lactis* UC06, *L. lactis* UC08, *L. lactis* UC11, *L. lactis* UC063, *L. lactis* UC77, *L. lactis* UC109, *L. lactis* UC509.9, *L. lactis* A76, *L. lactis* AI06, *L. lactis* CV56, *L. lactis* IL1403, *L. lactis* IO-1 and *L. lactis* core genome.

COG classification was also performed on the non-overlapping parts of the core genomes of subsp. *cremoris* and subsp. *lactis*, thus focusing on conserved features that differentiate the two subspecies (Table 3.3). This analysis identified CDSs predicted to be involved in metabolism, particularly carbohydrate transport and metabolism (Table 3.3) as the major discerning factor between the two subspecies. Further examination of these subspecies-specific, conserved gene set demonstrates that subsp. *lactis* conserved more unique genes than subsp. *cremoris*, particularly related to metabolism, 124 compared to 68, respectively. The reduced number of CDSs encoding products related to metabolism in subsp. *cremoris* strains is noteworthy as it is in agreement with the generally observed reduced metabolic capabilities of subsp. *cremoris* strains, and highlights the reductive pressure and genome decay imposed on these strains predominantly isolated from the dairy niche.

Table 3.3: COG classifications of the core genomes of *L. lactis*, *L. lactis* subsp. *lactis* and *L. lactis* subsp. *cremoris*

COG classification	*L. lactis* core genome	Unique core genomes	
		L. lactis subsp. *lactis*	*L. lactis* subsp. *cremoris*
Translation, ribosomal structure and biogenesis	10 %	**<1 %**	**5 %**
Transcription	8 %	**11 %**	**9 %**
Replication, recombination and repair	6 %	**3 %**	**6 %**
Cell cycle control, cell division, chromosome partitioning	1 %	<1 %	<1 %
Defence mechanisms	1 %	4 %	3 %
Signal transduction mechanisms	2 %	2 %	2 %
Cell wall/membrane/envelope biogenesis	5 %	4 %	4 %
Cell motility	<1 %	1 %	1 %
Intracellular trafficking, secretion, and vesicular transport	1 %	<1 %	2 %
Posttranslational modification, protein turnover, chaperones	4 %	1 %	<1 %
Energy production and conversion	4 %	**4 %**	**3 %**
Carbohydrate transport and metabolism	7 %	**14 %**	**10 %**
Amino acid transport and metabolism	9 %	**15 %**	**5 %**
Nucleotide transport and metabolism	5 %	1 %	1 %
Coenzyme transport and metabolism	4 %	2 %	2 %
Lipid transport and metabolism	3 %	3 %	3 %
Inorganic ion transport and metabolism	6 %	4 %	3 %
Secondary metabolites biosynthesis, transport and catabolism	2 %	1 %	2 %
General function prediction only	14 %	6 %	10 %
Function unknown	10 %	23 %	27 %

* Highlighted rows indicate those were a significant difference exists within the unique core genomes

3.3.5 Metabolism and niche adaptation

To explore the divide between the subspecies in terms of their metabolic capabilities and to highlight particular niche adaptations within the strains, MCL analysis was employed to compare the COG groupings based on function, i.e. [G] carbohydrate transport and metabolism, [E] amino acid transport and metabolism and [I] lipid transport and metabolism. These COG groups are fundamental to niche adaptation as they provide an overview of a strain's ability to metabolise different energy sources. They may also include key technological traits sought in strains utilised in the dairy niche where the majority of sequenced strains have been isolated. Until now, the focus of this study has been on chromosome-specific traits, however, in order to gain an overall view of the total metabolic capabilities of a strain it is necessary to also consider extra-chromosomal encoded traits. Therefore, both chromosomally- and plasmid-encoded features were considered for the remainder of the comparative analysis.

MCL analysis of COG [G] functions (genes involved in carbohydrate transport and metabolism) across all 30 isolates resulted in a gene presence/absence matrix displaying groupings specific to niche environments (Fig. 3.5). The majority of analysed lactococcal genome sequences are derived from isolates from the dairy niche, where the most important adaptation is the ability to ferment lactose, facilitated by the products of the plasmid-borne *lac* operon, which consists of the *lacABCDEFGX* genes [8, 9]. The complete *lac* operon was identified in all subsp. *cremoris* strains isolated from the dairy niche except for the plasmid-free strains MG1363 and its derivative NZ9000. However, MG1363 has previously been shown to metabolise lactose due to the activity of a cellobiose-specific phosphotransferase system (PTS), which can act as an alternative lactose utilisation pathway under

glucose starvation conditions [30]. The complete *lac* operon was also identified in six of the eleven subsp. *lactis* dairy isolates, yet not in the remaining five (strains 184, C10, UL8 and IL1403), of which *L. lactis* IL1403 is known to be a plasmid-cured strain [31]. When strains C10 and UL8 were inoculated in 10 % RSM (reconstituted skimmed milk), they displayed no signs of growth or acidification, which is consistent with the observed absence of the *lac* operon. However, in the case of strain 184, growth on lactose is still observed, which can be explained by the presence of the cellobiose-specific phosphotransferase system (PTS), similar to the situation in MG1363 [30]. Interestingly, while all dairy-derived *cremoris* strains form a single cluster based on genes involved in carbohydrate metabolism, all dairy-derived *lactis* strains with the exception of strains SO and UC06 form a single separate cluster to their *cremoris* counterparts based on carbohydrate utilisation. The only human isolate of *L. lactis* included in our analysis is also contained within this cluster. Differentiating factors, such as the clusters responsible for maltose utilisation found in all *lactis* strains and non-dairy *cremoris* strains, and for xylose metabolism as observed in all *cremoris* strains (with the exception of JM1), yet not present in *lactis* strains, contribute to this division.

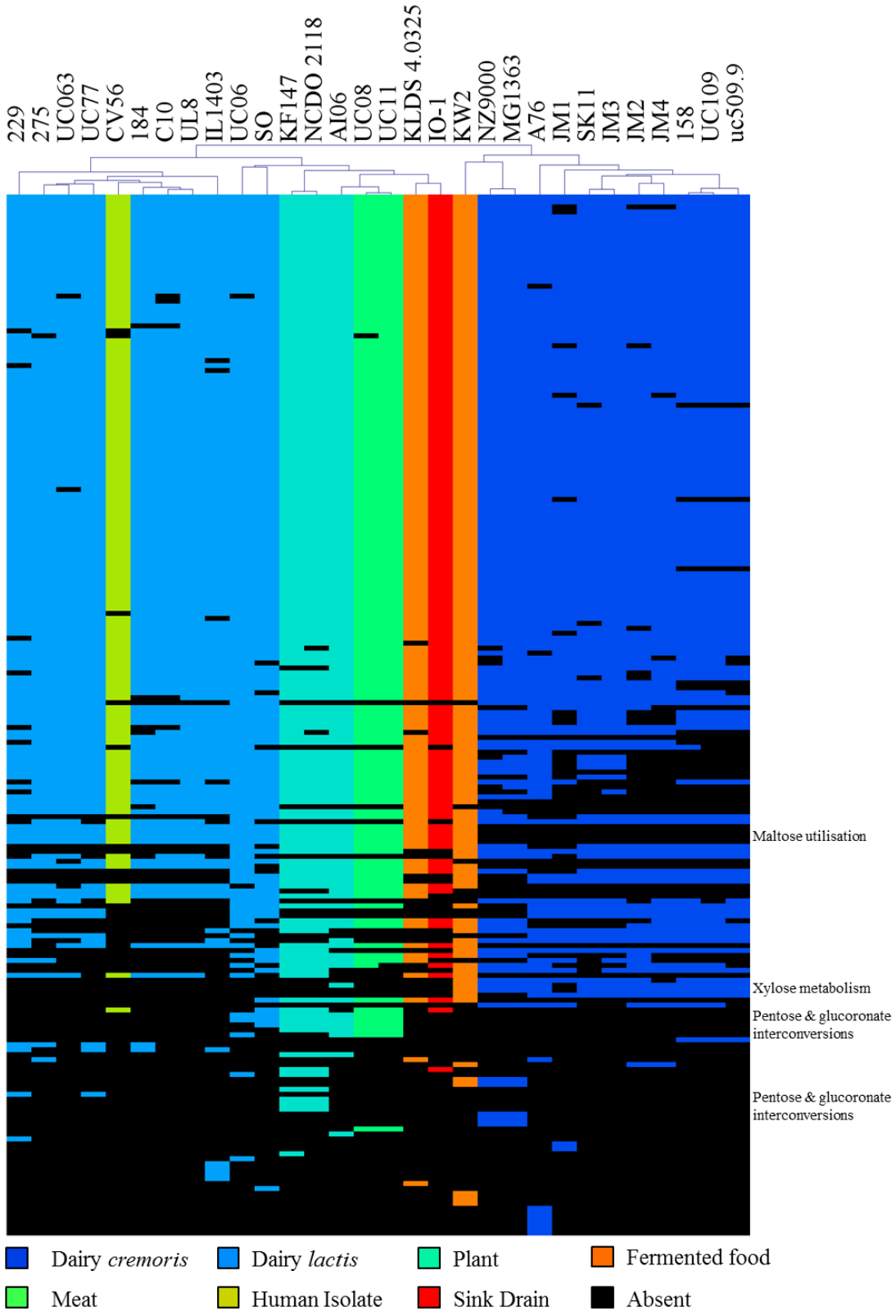


Figure 3.5: Carbohydrate utilisation and niche adaptation
Hierarchical clustering analysis representing the presence/absence of gene families from COG group [G], carbohydrate transport and metabolism. Colour indications refer to the particular niche from which the *L. lactis* strain had been isolated.

The genomes of *L. lactis* UC08 and UC11 represent the only two complete lactococcal genome sequences isolated from fermented meat. In this analysis, these strains clustered closely with those derived from non-dairy sources, particularly plant-derived strains based on carbohydrate metabolism. Genes encoding functions involved in pentose and glucuronate interconversions are found exclusively in strains isolated from plant and meat niches, and thus are not present in any other lactococcal strain. These sugars are generally not found in milk where the primary sugar source is lactose with only trace amounts of monosaccharides and oligosaccharides. The majority of strains examined in this study are dairy isolates and so it is plausible that these functions have been lost through reductive evolution in strains adapted to (the rich growth media provided by) the dairy environment.

Supplementing COG analysis with information obtained from KEGG (Kyoto encyclopaedia of genes and genomes) analysis, a full assessment of all major metabolic pathways present in *L. lactis* was undertaken. In this case complete pathways for D-galacturonate degradation (KEGG accession: M00631) and beta-oxidation, acyl-CoA synthesis (KEGG accession: M00086) were exclusively identified in the plant-derived strains NCDO2118 and KF147.

It has previously been demonstrated that both *L. lactis* subsp. *cremoris* and subsp. *lactis* are capable of folate biosynthesis [32]. Interestingly, KEGG analysis showed all analysed subsp. *lactis* strains to lack a complete pathway for tetrahydrofolate biosynthesis (KEGG accession: M00126) which was found to be complete in all subsp. *cremoris* strains. In *cremoris* strains the pathway was found to consist of nine genes responsible for conversions from purine metabolism to folate, whereas in subsp. *lactis* strains, the *phoA* gene that encodes an alkaline phosphatase (E3.1.3.1), responsible for the conversion of 7,8-dihydroneopterin 3-triphosphate to

dihyroneopterin, appears to be absent. This may indicate that this step in tetrahydrofolate biosynthesis in subsp. *lactis* may be performed by an alternative and as yet unidentified enzyme (in comparison to their *cremoris* counterparts).

3.3.6 Amino acid transport and metabolism

Proteolysis (of casein) performed by *L. lactis* has been widely studied as it is considered to be an important technological trait in dairy lactococci due to its contribution to flavour in fermented dairy products such as cheese, as outlined by a number of reviews that detail this process [33-35]. The main categories of peptidases contributing to proteolysis in *L. lactis* are aminopeptidases, endopeptidases, di/tri-peptidases, proline peptidases, endopeptidases and carboxypeptidases. The majority of described peptidase-encoding genes represent monocistronic elements (e.g. *pepC*, *pepN* and *prtP*), while others are transcribed with genes apparently unrelated to proteolysis [36]. To assess the level of peptidase activity within *L. lactis*, both functional and genomic analyses were undertaken. Quantitative assays utilising fluorescently labelled substrates (see Materials and Methods section) were used to determine the activity levels of PepN/C, PepA, PepX, proline imino peptidase, carboxypeptidase and endopeptidase produced by each strain (Fig. 3.6A) and expressed as a percentage of their total proteolytic capability (Fig. 3.6B). The dominant peptidase activities expressed by each strain was that represented by the proline peptidase PepX and the aminopeptidases PepA and PepN/C. Interestingly, all of these peptidases are present in single-copy in each of the chromosomes, though the measured activity levels do vary considerably between strains. To ascertain a broader perspective on peptidase or amino acid digestion, an MCL analysis of COG group [E] amino acid transport and metabolism was performed (Fig. 3.6C).

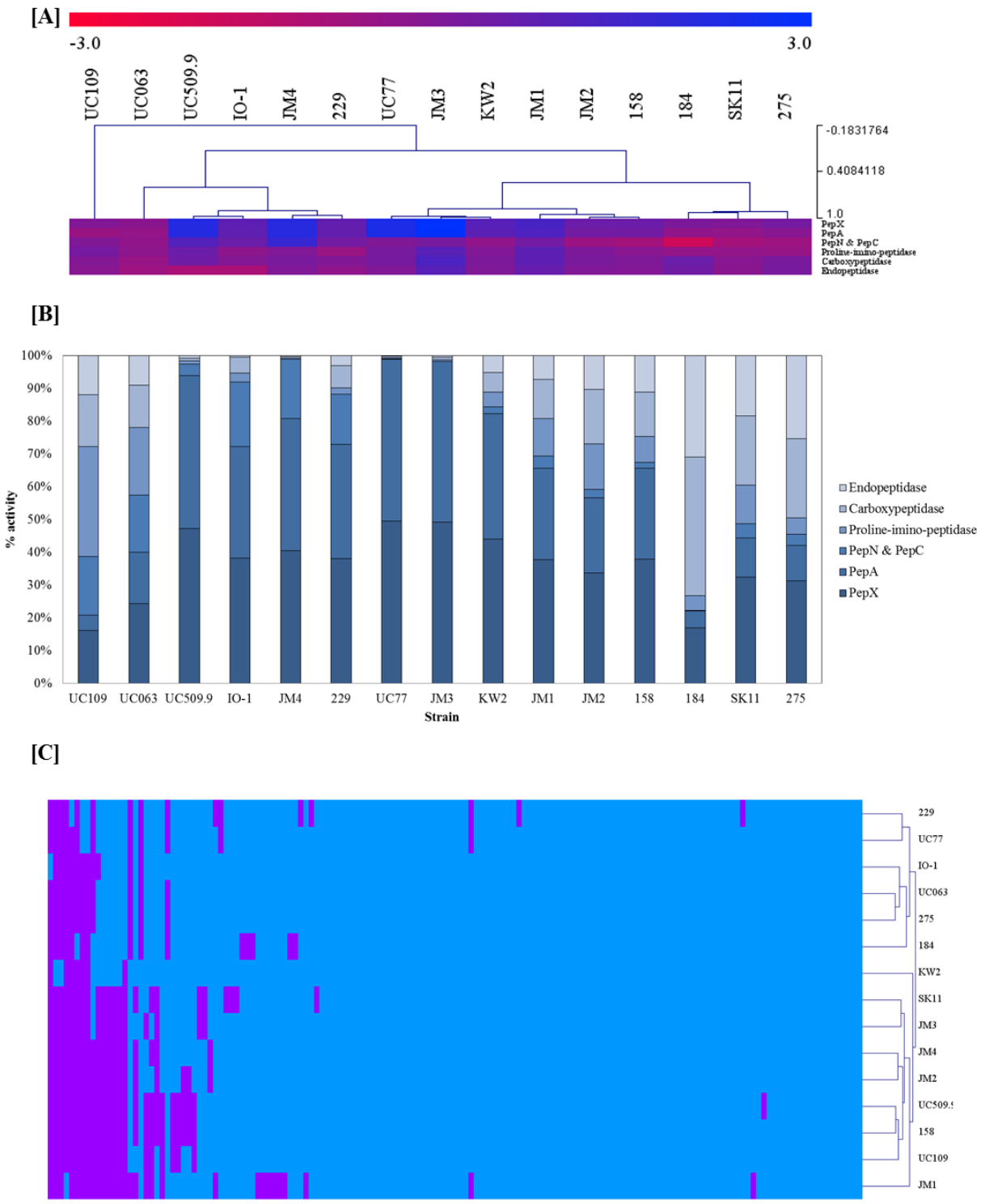


Figure 3.6: Peptide metabolism in *L. lactis*

A) Level of PepX, PepA, PepN/C, Proline imino peptidase, endopeptidase and carboxypeptidase activity, expressed by *L. lactis* in log(RFU PPDA) where (1 RFU = the amount of µM of AMC released min^{-1} by 1 mg of protein). Strains are clustered based on activity red-blue indicating increased activity. B) Histogram representing the percentage of total peptidase activity contributed by each peptidase for each strain. C) Hierarchical clustering analysis representing the presence/absence of gene families from COG group [E] amino acid transport and metabolism.

Clustering based on the presence or absence of genes involved in amino acid transport and metabolism resulted in two major groupings: the first composed of subsp. *lactis* strains and the second composed of *cremoris* strains indicating that the proteolytic system of these bacteria is distinct between and relatively well conserved within each subspecies.

Another important factor in assessing the proteolytic system of *Lactococcus* is the effect of amino acid transferases, which convert free amino acids to α-ketoacids. This is of particular importance when considering strains which may be used within the fermented food industry for the production of cheeses where aminotransferases contribute to flavour and aroma development [37]. As a high proportion of the available lactococcal dataset is currently composed of strains from the dairy niche, we assayed the strains for amino acid transferase activity against phenylalanine (aromatic amino acid) and methionine (sulphur amino acid), which are both common in milk and important in terms of cheese production. All strains demonstrated aminotransferase activity with phenylalanine as a substrate (Fig. 3.7A), and a considerably higher level of activity when methionine acted as a substrate (Fig. 3.7B). With the exception of *L. lactis* subsp. *cremoris* JM4, strains of the *cremoris* subspecies were shown to display significantly higher levels of aminotransferase activity compared to their *lactis* counterparts.

Markov clustering of aminotransferases in *L. lactis* strains was also carried out and resulted in clades, which closely resemble the level of activity expressed by the constituent strains (Fig. 3.7C). Interestingly, strains SK11, JM2, and KW2, which exhibited the highest level of aminotransferase activity using either phenylalanine or methionine as substrates, did not encode the highest number of aminotransferases, and none of these strains specify a histidinol-phosphate/aromatic

aminotransferase. Overall, both the peptidase and aminotransferase analyses revealed a very divergent proteolytic system between the two subspecies.

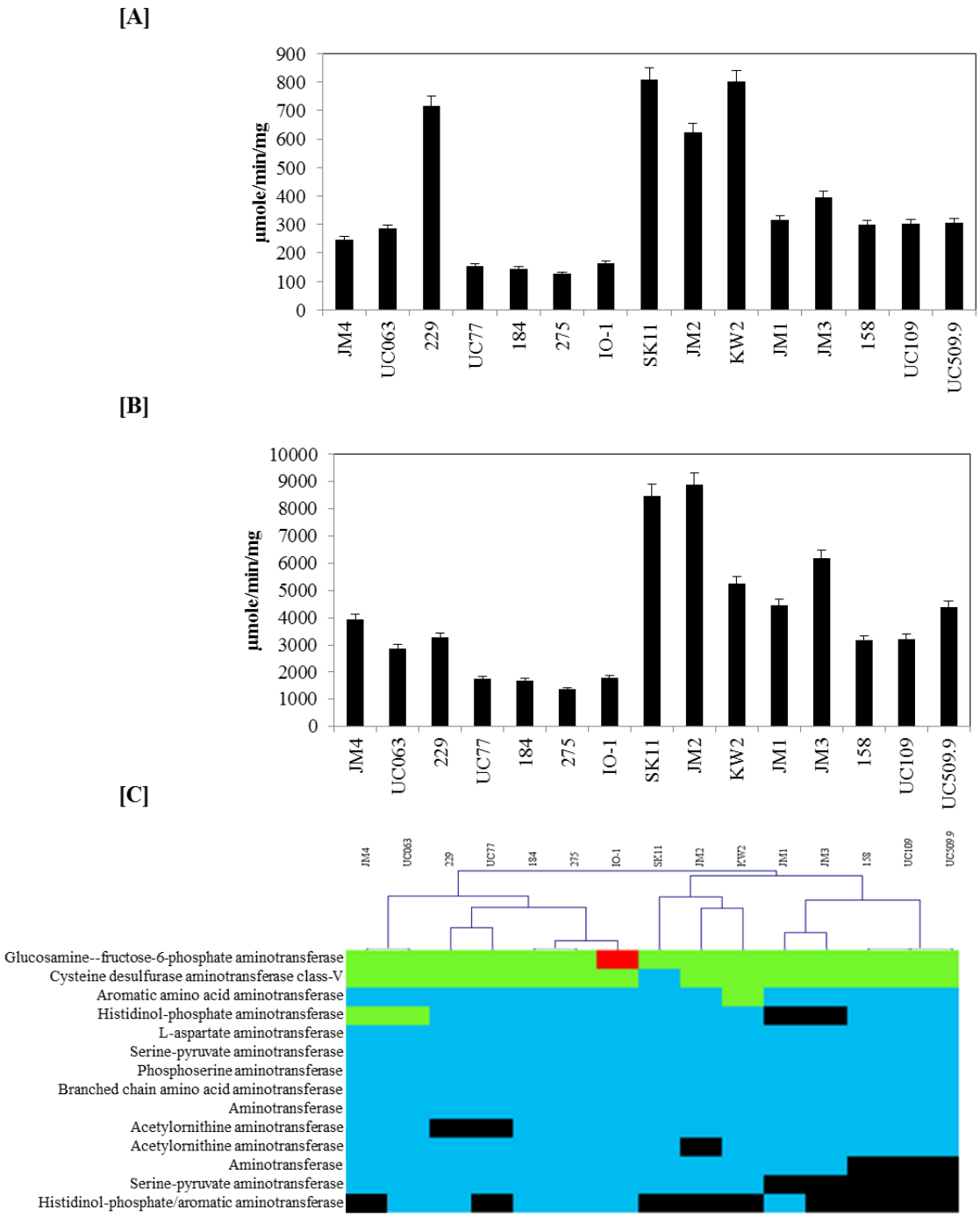


Figure 3.7: Aminotransferase activity in *L. lactis*
Amino acid transferase activities for (A) phenylalanine and (B) methionine. C) Hierarchical clustering analysis representing the presence/absence of genes involved in aminotransferase activities. Copy number is indicated by colour; red (x3), green (x2), blue (single-copy) and black (absent).

3.3.7 Lipid transport and metabolism

MCL analysis combined with hierarchical clustering of COG group [I] (lipid transport and metabolism) revealed two main groups based on predicted lipolytic activity; the first was composed of both subsp. *lactis* and *cremoris* strains from mixed sources, while the second was composed exclusively of dairy *cremoris* strains, namely strains JM1, JM2, JM3, JM4, 158, UC109 and UC509.9. These strains encode a well-conserved lipolytic system, while lipolytic assays utilizing *p*-nitrophenyl-butyrate for the detection of short chain esterase activity revealed a trend showing higher expression of esterase activity by these strains compared to their subsp. *lactis* and non-dairy subsp. *cremoris* counterparts (Fig. 3.8).

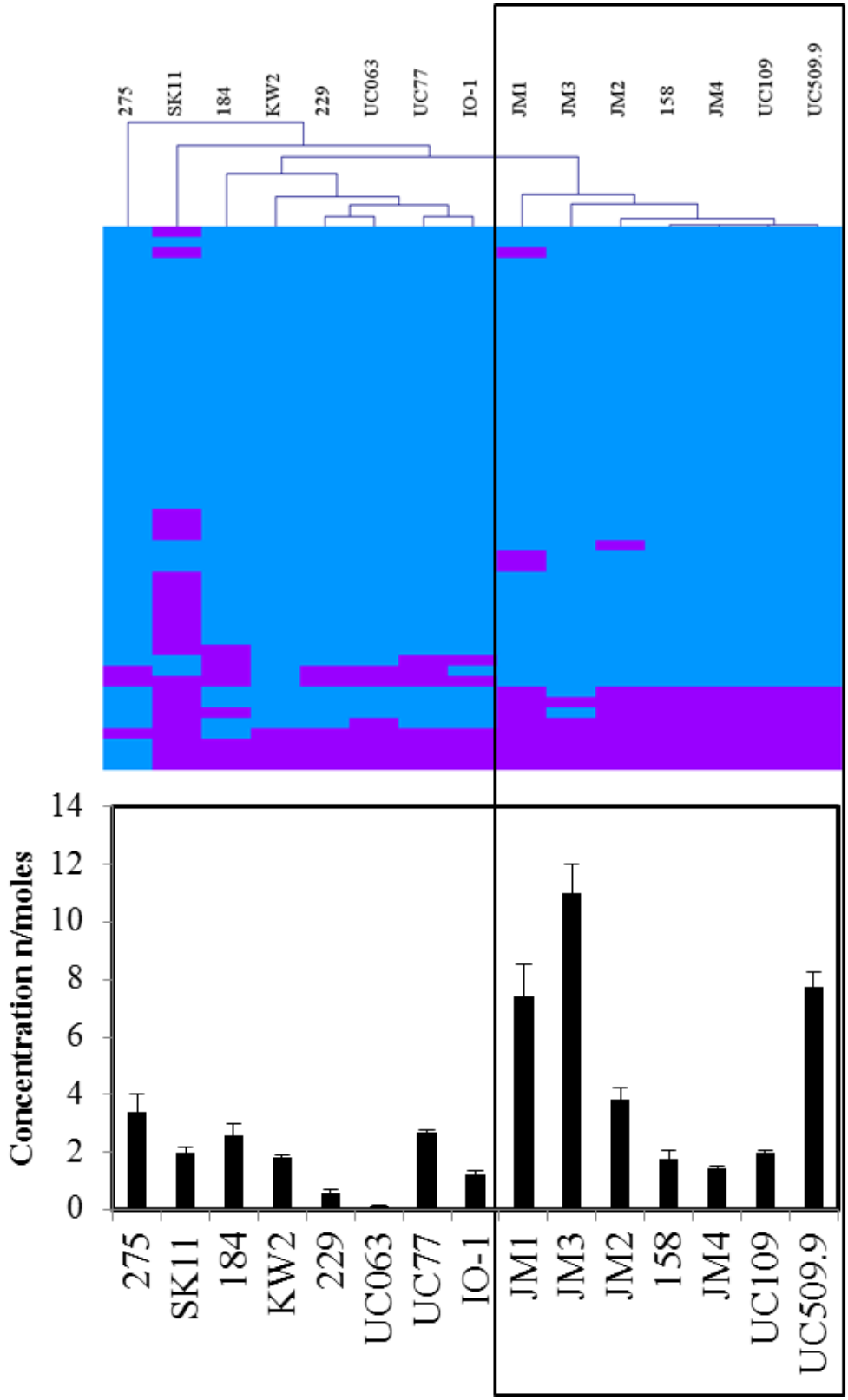


Figure 3.8: Lipid metabolism in *L. lactis*
Upper panel displays hierarchical clustering analysis representing the presence/absence of gene families from COG group [I] lipid transport and metabolism. Lower panel displays a histogram indicating level of short chain esterase activity of each constituent strain in n/moles.

3.3.8 Plasmid Integration

Bacterial adaptation relies heavily on the metabolic capabilities of the cell. In the case of *L. lactis* the most studied adaptations are those related to the dairy environment where reductive evolution or genome decay is observed among strains, a phenomenon which is believed to be due to repeated passaging in the nutrient-rich growth medium, milk [6, 7, 38]. As well as streamlining a bacterium's metabolic activities to reduce energy–demanding, unnecessary systems in such a niche, the acquisition of new genetic information encoding traits that are advantageous to the host (in the particular niche) is often necessary. In *L. lactis,* the most notable example is adaptation to the dairy environment via the plasmid-encoded *lac* operon, which allows for lactose utilisation as the primary sugar source, and the *prtP*-encoded protease and *opp* operon responsible for amino acid/nitrogen acquisition from the milk protein casein. However, in some instances integration of such genetic features into the host's chromosome may take place.

In silico based analysis of the chromosomes of thirty lactococcal isolates resulted in the identification of (1-6) integrated regions with significant (>90 %) nucleotide identity to previously sequenced lactococcal plasmids . The most notable of these putative integrations was the presence of the *opp* operon, originally identified as a plasmid-encoded trait in dairy *L. lactis* [39], conserved in the chromosomes of twenty-four out of thirty strains. The region shares (>90 %) nucleotide identity with lactococcal plasmids pIL4, pQA549, pCIS8, pSK11L /SK11 plasmid 4, pVF50 and pGdh442. *L. lactis* MG1363 and its derivative *L. lactis* NZ9000 also harbour *prtP* in the same integration site; however, it is integrated at approximately 680-690 Kbp on the chromosome. In one instance, for *L. lactis* SO, the associated *lac* operon, which controls lactose utilisation in the dairy niche, was

detected on the chromosome, 20 Kbp downstream of the integrated *opp* operon and sharing significant homology with plasmids pCV56B, pSK08, pKF147A and pNCDO2118.

A number of other (apparent) integrations were detected containing typical lactococcal plasmid features, such as genes encoding restriction-modification systems, conjugal transfer and mobilisation or *mob* genes, a partial lactococcin production gene cluster (four instances) and a partial (exo)polysaccharide biosynthesis gene cluster (nine instances). The frequency of these integrations suggests that the total genetic complement of *L. lactis* is in a state of flux, yet is also indicative of adaptations that are more permanent, particularly in the dairy niche where plasmid-encoded traits appear to become incorporated into the chromosomes of dairy strains.

3.4 Discussion

Recent advances in NGS technologies have made it easier to sequence a far greater number of high-quality bacterial genomes than ever before. In this study SMRT sequencing was applied for the complete sequencing of sixteen lactococcal genomes, more than doubling the existing number of publicly available, fully sequenced lactococcal genomes. The chromosomal features of *L. lactis* were assessed with particular emphasis on discerning the subspecies classification and niche adaptation of *L. lactis*.

Our analysis clearly identified a phylogenetic division between subspecies *lactis* and *cremoris*. This subspecies division was corroborated by hierarchical clustering based on both carbohydrate and amino acid metabolism, which indicates two main subgroups that correspond to each subspecies. Furthermore, for a number of conserved genes investigated in this study, a unique allelic type was observed for strains belonging to subsp. *lactis* and a separate allelic variant observed for strains belonging to subsp. *cremoris*. These observations support those made by Cavanagh and colleagues, who recently proposed a re-evaluation of the taxonomic group separating *L. lactis* into two distinct species *L. lactis* and *L. cremoris* based on ANI (average nucleotide identity) and TETRA (tetranucleotide frequency correlation coefficients) [16].

The genomes of *L. lactis* subsp. *cremoris* were found to contain a higher number of pseudogenes in comparison to their *L. lactis* subsp. *lactis* counterparts, on average 100 per strain compared to 31 per strain, respectively. The vast majority of these strains are isolated from the dairy niche where genome decay and redundancy is widely reported [6, 38, 40], and believed to be due to continuous growth in milk. These genomes were also shown to contain a high number of prophages and

transposable elements in agreement with Chopin and colleagues [5], and assumed to be the result of continued industrial pressures. Such prophages represent a risk factor, which warrants thorough assessment before introducing such strains into industrial fermentations. Conversely, the genomes of lactococcal strains isolated from both meat or plant environments displayed greater genetic variation and encode a higher number of metabolic pathways for the utilisation of a broader range of substrates compared to dairy-associated lactococci. The isolation of strains from these non-dairy sources may provide novel cultures for food fermentations and deliver desirable capabilities in terms of flavour and industrial robustness as dairy starter cultures.

COG analysis of *L. lactis* subsp. *cremoris* and subsp. *lactis* showed a higher proportion of genes involved in information processing and storage in *cremoris* strains, and in metabolism in *lactis* strains, in the specific portions of the core genome the two subspecies do not share. This is in agreement with the generally observed reduced metabolic capabilities of subsp. *cremoris* strains, and highlights the reductive pressure through genome decay imposed on these (mostly) dairy-derived strains. This may also be conducive to the observed faster growth rate of *lactis* strains compared to their *cremoris* counterparts under milk fermentation conditions. COG analysis was also utilised as a mechanism for functional genomic analysis in examining both peptide and lipid metabolism. It was determined that although strains can be genotypically clustered based on their subspecies and common niche, in agreement with a previous study [36], many of the peptidases for which functional assays are available exist in single copy in the majority of lactococcal genomes. Therefore, it may not always be possible to make genotype-phenotype links without the involvement of transcriptome and/or metabolome-based

studies to support the data. Interestingly, both peptidase and aminotransferase analyses indicated a very divergent proteolytic system between the two subspecies, yet being relatively well conserved within each subspecies.

Niche adaptation also relies heavily on the acquisition of new metabolic capabilities as well as the loss of unnecessary functions. The introduction of niche-specific adaptations via plasmid acquisition, such as lactose and citrate metabolism has been extensively studied in *L. lactis* in view of their role in dairy niche adaptation [4, 8-10, 41], however, chromosomal adaptations are largely under-represented by comparison. Interestingly, the division between plasmid- and chromosome-based traits is becoming less clear as multiple integration events within the lactococcal chromosome suggests a more fluid genome than previously thought [4].

3.5 Conclusions

In conclusion, the sequencing of 16 novel lactococcal isolates has doubled the number of complete finished quality lactococcal genomes available and allowed for large-scale comparative analysis of the complete metabolic systems of the taxon. Analysis of the two lactococcal subspecies revealed unique allelic subtypes for many of the conserved genes within each subspecies raising the question of their taxonomic placement and whether or not the two subspecies should be redefined as separate species. Niche adaptation appears to play a significant part in governing the genetic content of each constituent strain, while genome decay and redundancy in the dairy niche is also widely observed. The deduced pan-genome of *L. lactis* appears to be closed, indicating that the representatives of this analysis are sufficient to fully describe the genetic diversity of the taxon.

3.6 References

[1] Schleifer K, Kilpper-Bälz R: Molecular and chemotaxonomic approaches to the classification of streptococci, enterococci and lactococci: a review. Syst Appl Microbiol. 1987, 10:1-19.

[2] Latorre-Guzman BA, Kado CI, Kunkee RE: *Lactobacillus hordniae*, a New Species from the Leafhopper (*Hordnia circellata*). Int J Syst Evol Microbiol. 1977, 27:362-370.

[3] Perez T, Balcazar JL, Peix A, Valverde A, Velazquez E, de Blas I, Ruiz-Zarzuela I: *Lactococcus lactis* subsp. *tructae* subsp. nov. isolated from the intestinal mucus of brown trout (*Salmo trutta*) and rainbow trout (*Oncorhynchus mykiss*). Int J Syst Evol Microbiol. 2011, 61:1894-1898.

[4] Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D: The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. FEMS Microbiol Rev. 2014, 38:1066-1088.

[5] Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin MC: Analysis of six prophages in *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage populations. Nucleic Acids Res. 2001, 29:644-651.

[6] Ainsworth S, Zomer A, de Jager V, Bottacini F, van Hijum SA, Mahony J, van Sinderen D: Complete genome of *Lactococcus lactis* subsp. *cremoris* UC509. 9, host for a model lactococcal P335 bacteriophage. Genome announc. 2013, 1:e00119-00112.

[7] Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N: Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci. 2006, 103:15611-15616.

[8] van Rooijen RJ, De Vos W: Molecular cloning, transcriptional analysis, and nucleotide sequence of *lacR*, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. J Biol Chem. 1990, 265:18499-18503.

[9] Van Rooijen R, Gasson M, De Vos W: Characterization of the *Lactococcus lactis* lactose operon promoter: contribution of flanking sequences and LacR repressor to promoter activity. J Bacteriol. 1992, 174:2273-2280.

[10] Drider D, Bekal S, Prévost H: Genetic organization and expression of citrate permease in lactic acid bacteria. Genet Mol Res. 2004, 3:271-281.

[11] Kelly WJ, Ward LJH, Leahy SC: Chromosomal Diversity in *Lactococcus lactis* and the Origin of Dairy Starter Cultures. Genome Biol Evol. 2010, 2:729-744.

[12] Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A: The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. Genome Res. 2001, 11:731-753.

[13] Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP: Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. J Bacteriol. 2007, 189:3256-3270.

[14] Liu M, Nauta A, Francke C, Siezen RJ: Comparative genomics of enzymes in flavor-forming pathways from amino acids in lactic acid bacteria. Appl Environ Microbiol. 2008, 74:4590-4600.

[15] Liu M, Bayjanov J, Renckens B, Nauta A, Siezen R: The proteolytic system of lactic acid bacteria revisited: a genomic comparison. BMC Genomics. 2010, 11:36.

[16] Cavanagh D, Casey A, Altermann E, Cotter PD, Fitzgerald GF, McAuliffe O: Evaluation of *Lactococcus lactis* isolates from nondairy sources with potential dairy applications reveals extensive phenotype-genotype disparity and implications for a revised species. Appl Environ Microbiol. 2015, 81:3961-3972.

[17] Siezen RJ, Bayjanov JR, Felis GE, van der Sijde MR, Starrenburg M, Molenaar D, Wels M, van Hijum SA, van Hylckama Vlieg JE: Genome-scale diversity and niche adaptation analysis of *Lactococcus lactis* by comparative genome hybridization using multi-strain arrays. Microb Biotechnol. 2011, 4:383-402.

[18] Gupta PK: Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol. 2008, 26:602-611.

[19] McCarthy A: Third generation DNA sequencing: Pacific Biosciences single molecule real time technology. Chem Biol. 2010, 17: 675-676.

[20] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM: Genome analysis of

multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. Proc Natl Acad Sci USA. 2005, 102:13950-13955.

[21] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: The microbial pan-genome. Curr Opin Genet Dev. 2005, 15:589-594.

[22] Bottacini F, O’Connell-Motherway M, Kuczynski J, O’Connell KJ, Serafini F, Duranti S, Milani C, Turroni F, Lugli GA, Zomer A, Zhurina D, Riedel C, Ventura M, van Sinderen D: Comparative genomics of the *Bifidobacterium breve* taxon. BMC Genomics. 2014, 15:1-19.

[23] Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR: What is finished, and why does it matter. Genome Res. 2002, 12:669-671.

[24] Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW: Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014, 10:e1003998.

[25] De Angelis M, Bottacini F, Fosso B, Kelleher P, Calasso M, Di Cagno R, Ventura M, Picardi E, van Sinderen D, Gobbetti M: *Lactobacillus rossiae*, a vitamin B12 producer, represents a metabolically versatile species within the genus *Lactobacillus*. PloS one. 2014, 9:e107232.

[26] McCulloch JA, de Oliveira VM, de Almeida Pina AV, Perez-Chaparro PJ, de Almeida LM, de Vasconcelos JM, de Oliveira LF, da Silva DE, Rogez HL, Cretenet M, Mamizuka EM, Nunes MR: Complete genome sequence of *Lactococcus lactis* strain AI06, an endophyte of the Amazonian Acai Palm. Genome Announc. 2014, 2.

[27] Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J: PGAP: pan-genomes analysis pipeline, Bioinformatics. 2012, 28:416-418.

[28] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool, J Mol Biol. 1990, 215:403-410.

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

[30] Solopova A, Bachmann H, Teusink B, Kok J, Neves AR, Kuipers OP: A specific mutation in the promoter region of the silent *cel* cluster accounts for the appearance of lactose-utilizing *Lactococcus lactis* MG1363. Appl Environ Microbiol. 2012, 78:5612-5621.

[31] Chopin A, Chopin MC, Moillo-Batt A, Langella P: Two plasmid-determined restriction and modification systems in *Streptococcus lactis*: Plasmid. 1984, 11:260-263.

[32] Sybesma W, Starrenburg M, Kleerebezem M, Mierau I, de Vos WM, Hugenholtz J: Increased production of folate by metabolic engineering of *Lactococcus lactis*: Appl Environ Microbiol. 2003, 69:3069-3076.

[33] McSweeney PLH: Biochemistry of cheese ripening. Int J Dairy Technol. 2004, 57:127-144.

[34] Steele J, Broadbent J, Kok J: Perspectives on the contribution of lactic acid bacteria to cheese flavor development. Curr Opin Biotechnol. 2013, 24:135-141.

[35] Kelleher P, Murphy J, Mahony J, van Sinderen D: Next-generation sequencing as an approach to dairy starter selection. Dairy Sci Technol. 2015, 95:545-568.

[36] Guédon E, Renault P, Ehrlich SD, Delorme C: Transcriptional pattern of genes coding for the proteolytic system of *Lactococcus lactis* and evidence

for coordinated regulation of key enzymes by peptide supply. J Bacteriol. 2001, 183:3614-3622.

[37] Smit BA, Engels WJM, Wouters JTM, Smit G: Diversity of l-leucine catabolism in various microorganisms involved in dairy fermentations, and identification of the rate-controlling step in the formation of the potent flavour component 3-methylbutanal. Appl Microbiol Biotechnol. 2004, 64:396-402.

[38] Goh YJ, Goin C, O'Flaherty S, Altermann E, Hutkins R: Specialized adaptation of a lactic acid bacterium to the milk environment: the comparative genomics of *Streptococcus thermophilus* LMD-9. Microb Cell Fact. 2011, 10:Suppl 1S22.

[39] Yu W, Gillies K, Kondo JK, Broadbent JR, McKay LL: Loss of plasmid-mediated oligopeptide transport system in lactococci: Another reason for slow milk coagulation. Plasmid. 1996, 35:145-155.

[40] 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 JH, Diaz-Muniz 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: Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci USA. 2006, 103:15611-15616.

[41] Cavanagh D, Fitzgerald GF, McAuliffe O: From field to fermentation: the origins of *Lactococcus lactis* and its domestication to the dairy environment. Food Microbiol. 2015, 47:45-61.

[42] Sambrook J, Russell DW: Purification of nucleic acids by extraction with phenol:chloroform. Cold Spring Harb Protoc. 2006; doi:pdb.prot4455.

[43] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: The Pfam protein families database. Nucleic Acids Res. 2004, 32:D138-D141.

[44] Söding J, Biegert A, Lupas AN, The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005, 33:W244-W248.

[45] Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: PHAST: a fast phage search tool. Nucleic Acids Res. 2011, doi:10.1093/nar/gkr485.

[46] Darling AC, Mau B, Blattner FR, Perna NT, Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004,14:1394-1403.

[47] Krumsiek J, Arnold R, Rattei T: Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007, 23:1026-1028.

[48] Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000, 28:33-36.

[49] Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000, 28:27-30.

[50] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M, KEGG as a reference resource for gene and protein annotation. Nucleic acids research. 2016, 44:D457-D462.

[51] Crooks GE, Hon G, Chandonia JM, Brenner SE: WebLogo: a sequence logo generator. Genome Res. 2004, 14:1188-1190.

[52] Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004, 32:1792-1797.

[53] Guindon S, Gascuel O, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003, 52:696-704.

[54] Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987, 4:406-425.

[55] Kato T, Nagatsu T, Kimura T, Sakakibara S: Fluorescence assay of *X*-prolyl dipeptidyl-aminopeptidase activity with a new fluorogenic substrate. Biochem Med. 1978, 19:351-359.

[56] Vorderwülbecke T, Kieslich K, Erdmann H: Comparison of lipases by different assays. Enzyme Microb Technol. 1992, 14:631-639.

[57] Siezen RJ, Bayjanov J, Renckens B, Wels M, van Hijum SA, Molenaar D, van Hylckama Vlieg JE: Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium, J Bacteriol. 2010, 192:2649-2650.

[58] Gao Y, Lu Y, Teng KL, Chen ML, Zheng HJ, Zhu YQ, Zhong J: Complete genome sequence of *Lactococcus lactis* subsp. *lactis* CV56, a probiotic strain isolated from the vaginas of healthy women. J Bacteriol. 2011, 193:2886-2887.

[59] Kato H, Shiwa Y, Oshima K, Machii M, Araya-Kojima T, Zendo T, Shimizu-Kadota M, Hattori M, Sonomoto K, Yoshikawa H: 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. 2012, 194:2102-2103.

[60] Yang X, Wang Y, Huo G: Complete Genome Sequence of *Lactococcus lactis* subsp. *lactis* KLDS4. 0325: Genome Announc. 2013, 1:e00962-00913.

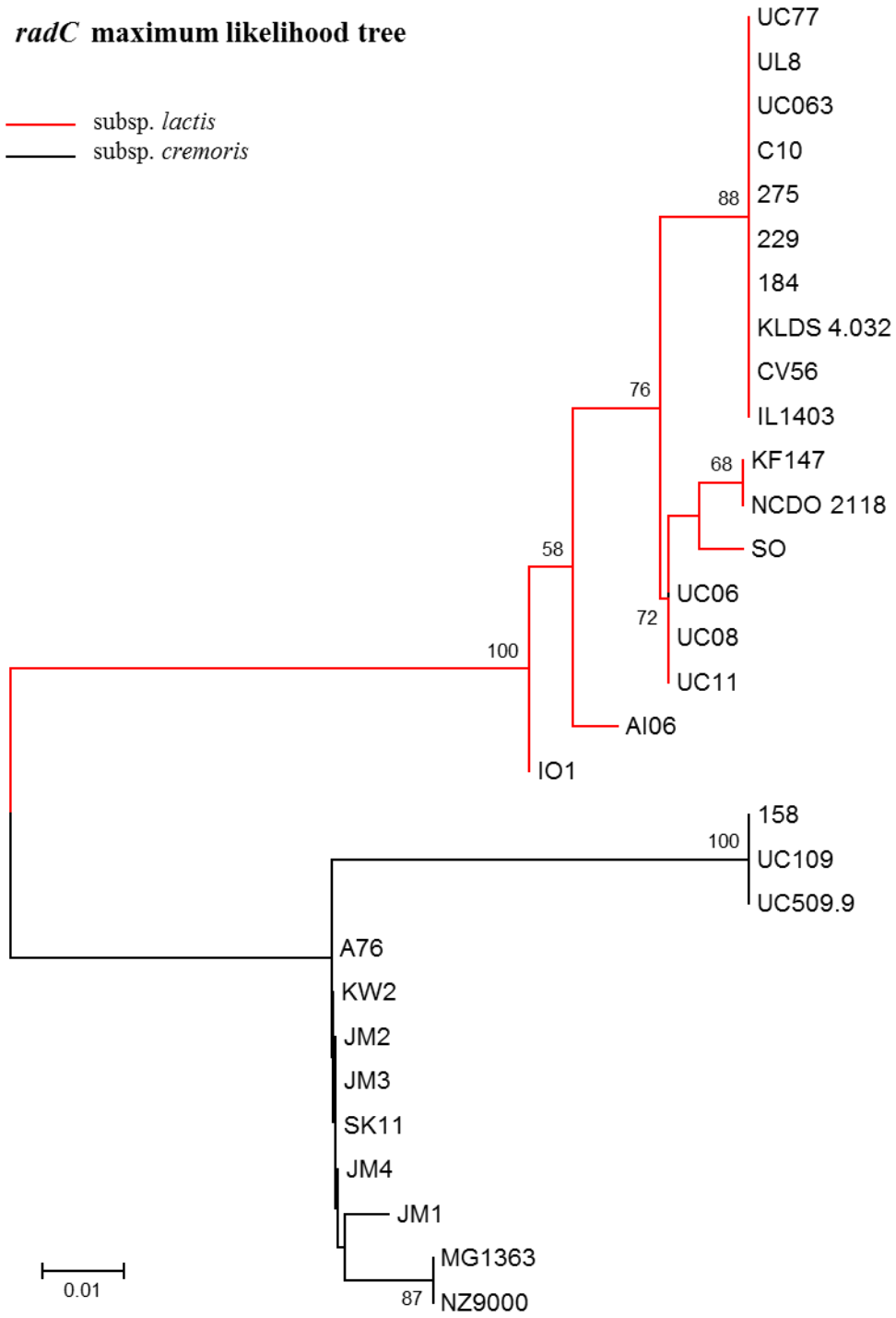
[61] Oliveira LC, Saraiva TD, Soares SC, Ramos RT, Sá PH, Carneiro AR, Miranda F, Freire M, Renan W, Júnior AF: Genome Sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-Producing Strain. Genome Announc. 2014, 2:e00980-00914.

[62] Zhao F, Ma H, Lu Y, Teng K, Kang X, Wang F, Yang X, Zhong J: Complete genome sequence of *Lactococcus lactis* S0, an efficient producer of nisin. J Biotechnol. 2015, 198:15-16.

[63] Linares DM, Kok J, Poolman B: Genome sequences of *Lactococcus lactis* MG1363 (revised) and NZ9000 and comparative physiological studies. J Bacteriol. 2010, 192:5806-5812.

[64] Bolotin A, Quinquis B, Ehrlich SD, Sorokin A, Complete genome sequence of *Lactococcus lactis* subsp. *cremoris* A76. J Bacteriol. 2012, 194:1241-1242.

[65] Kelly WJ, Altermann E, Lambie SC, Leahy SC: Interaction between the genomes of *Lactococcus lactis* and phages of the P335 species. Front Microbiol. 2013, 4.



Supplementary Figure S3.1: Phylogenetic analysis of *radC*
Maximum likelihood tree, resulting from the alignment of the *radC* genes of 30 *L. lactis* isolates resulting in a clear division between *L. lactis* subsp. *lactis* and subsp. *cremoris* strains.

Chapter IV

Comparative genomic analysis of the *Lactococcus lactis* plasmidome and assessment of its technological properties

Chapter IV contents

4 Abstract 185
4.1 Introduction 186
4.2 Materials & Methods 188
4.2.1 Sequencing 188
4.2.2 General feature predictions 188
4.2.3 Pan-plasmidome analysis 189
4.2.4 Comparative genomics 189
4.2.5 Pulse field gel electrophoresis (PFGE) 190
4.2.6 Bacteriocin Assays 190
4.2.7 Genbank accession numbers of applied strains 191
4.3 Results 192
4.3.1 General plasmid features 192
4.3.2 Pan-plasmidome calculation 199
4.3.3 MCL analysis of the lactococcal plasmidome 201
4.3.4 Lactococcal megaplasmids 201
4.3.5 Technological properties 204
4.3.6 Conjugation 206
4.3.7 Cell surface interactions (Adhesion & EPS) 210
4.3.7.1 Adhesion 210
4.3.7.2 EPS production 211
4.3.8 Bacteriocins 214
4.3.9 Phage resistance systems 215
4.3.10 Abortive infection systems 216
4.4 Discussion 218
4.5 References 222

4 Abstract

Plasmids are autonomous, self-replicating, extrachromosomal genetic elements that are typically not essential for growth of their host. They may encode additional metabolic capabilities which promote the maintenance of these genetic elements and may enhance the adaptation of bacterial strains to specific ecological niches. Genome sequencing of sixteen *Lactococcus lactis* strains revealed the presence of a total of sixty-seven plasmids, including two megaplasmids. Comparative genome analysis of these sequences combined with eighty one publicly available plasmids allowed the definition of the lactococcal plasmidome, and facilitated an investigation into technologically important plasmid-encoded traits such as conjugation, bacteriocin production, EPS production and (bacterio)phage resistance.

4.1 Introduction

Lactococcus lactis is globally applied as a starter culture for dairy-based food fermentations, such as those involved in the production of Cheddar, Colby, Gouda and blue cheeses, and from an economic and (food) biotechnological perspective represents one of the most important bacteria [1]. It is widely accepted that *L. lactis* originated from a plant-associated niche [2] and, whilst the majority of sequenced lactococcal representatives are isolated from the dairy environment, this is not representative of the presumed diversity of the taxon. It is evident from genome analyses of *L. lactis* strains isolated from the dairy niche that genome decay (due to functional redundancy) [3-6], in parallel with the acquisition of novel plasmid-encoded traits played a significant role in their adaptation to the nutrient-rich environment of milk. Analysis of the plasmid complement has revealed a relatively low abundance of plasmids among lactococcal strains isolated from non-dairy niches [1, 3, 5, 7]. Since various dairy-associated phenotypes are encoded by plasmids, horizontal acquisition to adapt to the dairy environment is likely to be one of the major drivers of plasmid transfer in *L. lactis* [1]. Plasmid transfer in *L. lactis* is believed to be predominantly governed by conjugation and transduction [1]. Plasmid transduction is a process in which plasmid transfer is carried out by a (bacterio) phage (i.e. a virus that infects a bacterium) due to accidental packaging of plasmid DNA, and has previously been observed in *L. lactis* [8, 9]. Conjugation involves the transfer of plasmid material via a conjugative apparatus [10] and is of particular importance as it represents a natural biological phenomenon that is suitable for the transfer of traits such as phage resistance systems in food grade processes [11].

Extensive research into the technological traits of *L. lactis* has been carried out in the past with a significant focus on lactose utilisation [12, 13], casein

metabolism [14], citrate metabolism [15], flavour formation [16, 17], and phage resistance mechanisms, all of which represent properties that are commonly plasmid-encoded. Lactose utilisation in *L. lactis* is governed by the *lac* operon, which provides dairy strains with the ability to rapidly ferment lactose and grow in milk. The *L. lactis lac* operon, which consists of the genes *lacABCDEFGX*, is generally plasmid-borne (see Chapter III) and is regulated by a repressor, encoded by the adjacent *lacR* gene [12, 18]. Citrate metabolism is conducted by citrate-positive (Cit^+) lactococci and is important as it leads to the production of a number of volatile flavour compounds [16]. Citrate uptake and subsequent diacetyl production is governed by the plasmid-encoded *citQRP* operon in lactococcal species [15]. Proteolysis also significantly contributes to flavour production in fermented dairy products, although high levels of proteolysis may cause bitterness in cheese [19]. The plasmid-encoded extracellular cell wall proteinase (lactocepin) has been shown to be directly involved in the bitter flavour defect in Cheddar cheese varieties, specifically involving starters which produce lactocepin of the so-called a, e, or h groups, and its characterisation is of particular importance when selecting novel starter cultures [19].

Lactococcal phages are recognized as the main cause of fermentation problems within the dairy industry with concomitant economic problems. Lactococcal strains possess an arsenal of phage defence mechanisms, such as restriction modification (R-M) systems and abortive infection (Abi) systems, many of which are plasmid-encoded.

In this study we assess the genetic content of lactococcal plasmids, define the current pan-plasmidome of *L. lactis*, and investigate corresponding plasmid-encoded (technological) traits.

4.2 Materials & Methods

4.2.1 Sequencing

In total, 67 plasmids (65 plasmids and 2 megaplasmids) were sequenced in the context of this study (Table 4.1). Sequencing was performed utilising the SMRT sequencing approach on a Pacific Biosciences RS II sequencing platform (executed by GATC Biotech Ltd., Germany). *De novo* assemblies were performed on the Pacific Biosciences SMRTPortal analysis platform (version 2.3.1), utilizing the RS_HGAP_Assembly.2 protocol. Assemblies were then repeated with a reduced minimum coverage threshold adjusted to 15X to ensure all plasmid-associated contigs had been detected. Remaining low quality regions and sequence conflicts were resolved by primer walking and Sanger sequencing of PCR products (performed by Eurofins MWG Operon, Germany).

4.2.2 General feature predictions

Open Reading Frame (ORF) prediction, defined as a continuous stretch of codons that do not contain a stop codon was performed with Prodigal v2.5 prediction software (http://prodigal.ornl.gov) and confirmed using BLASTX v2.2.26 alignments [20]. ORFs were automatically annotated using BLASTP v2.2.26 [20] analysis against the non-redundant protein databases curated by the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nig.gov/). Artemis v16 genome browser and annotation tool was used to manually curate identified ORFs (http://www.sanger.ac.uk/science/tools/artemis) and for the combination and inspection of ORF results. The final ORF annotations were refined where necessary

using additional software tools and database searches, such as Pfam [21], HHpred [22], Uniprot/EMBL (http://www.uniprot.org/) and Bagel3 [23].

4.2.3 Pan-plasmidome analysis

Pan-plasmidome analysis was performed utilising the PGAP v1.0 pipeline [24] according to Heaps law pan-genome model [25]. The ORF content of each plasmid was organised into functional gene clusters via the Gene Family method. ORFs which produced an alignment with a minimum of 50 % sequence identity across 50 % of the gene or protein length were clustered and a pan-plasmidome profile was subsequently generated [25].

4.2.4 Comparative genomics

All sequence comparisons at protein level were performed via all-against-all, bi-directional BLAST alignments [20]. An alignment cut-off value of E-value 0.0001, >50 % amino acid identity across 50 % of the sequence length was used. For analysis and clustering of these results, the Markov Clustering Algorithm (MCL) was implemented in the mclblastline pipeline v12-0678 [26]. BlastGraph v0.1 was used to visualise BLAST results (https://github.com/bigwiv/BlastGraph). Logo motifs were produced using WebLogo 3 [27]. TM4 MeV, MultiExperiment Viewer v4.9 was used to view MCL clustering data, conduct hierarchal clustering and to plot relevance networks (http://www.tm4.org/mev.html).

4.2.5 Pulse field gel electrophoresis (PFGE)

Lactococcal strains were cultured in M17 broth (Oxoid) supplemented with 0.5 % (w/v) lactose at 30 °C without agitation overnight. PFGE plugs were then prepared and restricted with SI nuclease as previously described [28].

A 1 % (wt/vol) PFGE agarose gel was prepared in 0.5X TBE (89 mM Tris-borate, 2 mM EDTA [pH 8.3]) buffer and the PFGE plugs were melted in and sealed with molten agarose in 0.5X TBE buffer. A CHEF-DR III pulsed-field system (Bio-Rad Laboratories, Hercules, CA) was used to resolve the DNA fragments at 6 V/cm for 18 h in 0.5X TBE running buffer maintained at 14 °C with a linear ramped pulse time of 3 to 50 seconds. DNA ladder (Chef DNA lambda) was included in each gel (number 170-3635; Bio-Rad Laboratories). The gels were stained in ethidium bromide (10 mg/ml) (25 µl/500 ml dH_2O) for 120 min under light-limited conditions and destained in distilled water for 60 min. Gels were visualised by UV transillumination.

4.2.6 Bacteriocin assays

Lactococcal strains were cultured in M17 broth (Oxoid) supplemented with 0.5 % (w/v) lactose or glucose (strain-dependent) at 30 °C without agitation overnight. 3 µl of overnight culture was spotted on M17 agar supplemented with 0.5 % (w/v) glucose and left at 30 °C overnight. Cells that had grown on the spotted areas were inactivated by exposure to UV light for 30 minutes. Plates were then overlaid with a semi-solid M17 agar (0.4 % agarose) containing indicator strain *L. lactis* HP. Zones of inhibition were visualised after 24 hours.

4.2.7 Genbank accession numbers of applied strains

L. lactis subsp. *lactis* Il1403: AE005176; *L. lactis* subsp. *lactis* IO-1: AP012281; *L. lactis* subsp. *lactis* 184: CP015895; *L. lactis* subsp. *lactis* 229: CP015896; *L. lactis* subsp. *lactis* 275: CP015897; *L. lactis* subsp. *lactis* UC06: CP015902; *L. lactis* subsp. *lactis* UC08: CP015903; *L. lactis* subsp. *lactis* UC11: CP015904; *L. lactis* subsp. *lactis* UC063: CP015905; *L. lactis* subsp. *lactis* UC77: CP015906; *L. lactis* subsp. *lactis* UL8: CP015908; *L. lactis* subsp. *lactis* C10: CP015898; *L. lactis* subsp. *cremoris* SK11: CP000425; *L. lactis* subsp. *cremoris* MG1363: AM406671; *L. lactis* subsp. *cremoris* NZ9000: CP002094; *L. lactis* subsp. *cremoris* A76: CP003132; *L. lactis* subsp. *cremoris* UC509.9: CP003157; *L. lactis* subsp. *cremoris* KW2: CP004884; *L. lactis* subsp. *cremoris* 158: CP015894; *L. lactis* subsp. *cremoris* UC109: CP015907; *L. lactis* subsp. *cremoris* JM1: CP015899; *L. lactis* subsp. *cremoris* JM2: CP015900; *L. lactis* subsp. *cremoris* JM3: CP015901; *L. lactis* subsp. *cremoris* JM4: CP015909; and *L. lactis* subsp. *cremoris* HP: JAUH00000000.1.

4.3 Results

4.3.1 General plasmid features

In this study the sequences of sixty seven plasmids were elucidated utilising the PacBio SMRT sequencing approach, and represent the detected plasmid complement of the sixteen genomes sequenced in Chapter III (Tables 3.1 & 3.2). These plasmids were combined with a further eighty one plasmids retrieved from the NCBI database (National Centre for Biotechnology Information) (Table 4.1). In total, the features of one hundred and forty eight plasmids derived from forty seven lactococcal strains in addition to seventeen lactococcal plasmids without an assigned strain were investigated. This extra-chromosomal DNA complement amounts to 4,005 Kb of DNA and is predicted to represent four thousand and four CDSs (coding sequences; ORFs which encode protein products), thus contributing very substantially to the diversification of *L. lactis*.

The vast majority of currently sequenced plasmids originate from strains that were isolated from the dairy niche (118 of 148). These dairy lactococci carry between one and nine plasmids (the latter in *L. lactis* UC509.9), which accounts for up to 355 Kbp of extra-chromosomal DNA in a given strain (as is the case for *L. lactis* JM1). The size of individual lactococcal plasmids varies widely from the smallest *L. lactis* KLDS4.0325 plasmid 2, with a size of 0.87 Kbp, to the two megaplasmids, each maintained by *L. lactis* JM1 and *L. lactis* JM2, with a size of 193 and 113 Kbp, respectively. The GC content of lactococcal plasmids ranges from ~30 - 38 %, whilst the average GC content of previously sequenced chromosomes is more constrained (34 – 36 %). Only three lactococcal plasmids deviate from this range; pWC1 29.48 %, pIL105 29.79 % and pHP003 40.05 %, where the latter is

closer in GC-content to *Streptococcus thermophilus* genomes, which ranges from 39 to 40 % [29].

Lactococcal plasmids are known to replicate via two alternative methods, rolling circle replication (RCR) or theta-type replication [1, 11]. Based on predicted plasmid replication proteins/origins it appears that the majority of lactococcal plasmids replicates via the theta-type mechanism, while only a small proportion appears to utilise RCR (twelve of the current data-set). The relatively small number of plasmids utilising RCR may be attributed to a number of factors, such as the fact that RCR plasmids can only support a limited replicon size (<10 Kb), incompatibility with other RCR type plasmids [30], and/or intrinsic structural and segregationally instability [1]. The analysis also identified in three instances, plasmids for which replication modes could not be clearly determined as the origin of replication of these plasmids did not conform to the typical origin of replication associated with RCR or theta replication.

Table 4.1: Characteristics of the plasmids analysed in this study

Name	Accession	Size (Kbp)	GC (%)	Genes	Niche	Replication mode
KLDS 4.0325 p1	CP006767	4.094	30.02	4	Fermented food	RCR
KLDS 4.0325 p2	CP007042	0.870	32.64	2	Fermented food	Undetermined
KLDS 4.0325 p3	CP007043	1.278	32.63	4	Fermented food	Undetermined
p158A *	CP016685	75.119	33.04	93	Dairy	Theta
p158B *	CP016686	57.981	33.56	22	Dairy	Theta
p158C *	CP016687	51.651	34.57	55	Dairy	Theta
p158D *	CP016688	33.287	37.39	32	Dairy	Theta
p158E *	CP016689	11.679	34.05	13	Dairy	Theta
p158F *	CP016690	6.164	35.84	4	Dairy	Theta
p184A *	CP016691	9.735	34.84	13	Dairy	Theta
p184B *	CP016692	5.929	34.51	6	Dairy	Theta
p184C *	CP016693	10.488	33.35	14	Dairy	Theta
p229A *	CP016694	56.368	34.81	59	Dairy	Theta
p229B *	CP016695	33.280	37.39	29	Dairy	Theta
p229C *	CP016696	30.272	35.15	29	Dairy	Theta
p229D *	CP016697	6.153	35.88	8	Dairy	Theta
p229E *	CP016698	39.612	32.40	51	Dairy	Theta
p275A *	CP016699	92.710	35.35	104	Dairy	Theta
p275B *	CP016700	56.332	33.36	65	Dairy	Theta
p275C *	CP016701	54.922	34.28	62	Dairy	Theta
p275D *	CP016702	54.046	31.77	60	Dairy	Theta
pAF04	JQ821353	3.801	32.02	4	Dairy	Theta
pAF07	JQ821354.1	7.435	36.44	6	Dairy	Theta
pAF12	JQ821355.1	12.067	33.30	11	Dairy	Theta
pAF14	JQ821356.1	14.419	34.07	11	Dairy	Theta
pAF22	JQ821357.1	22.388	34.95	23	Dairy	Theta
pAG6	AB198069	8.663	33.70	8	Unknown	Theta

pAH33	AF207855	6.159	35.85	7	Dairy	Theta
pAH82	AF243383	20.331	34.44	17	Dairy	Theta
pAR141	DQ288662	1.594	36.14	2	Dairy	RCR
pAW153	HQ646604.1	7.122	31.35	8	Unknown	Theta
pAW601	AJ132009.2	4.752	31.42	1	Unknown	Theta
pBL1	AF242367	10.899	32.62	8	Dairy	Theta
pBM02	AY026767	3.854	35.73	6	Dairy	RCR
pC10A *	CP016703	2.120	34.10	4	Dairy	RCR
pCD4	AF306799	6.094	33.43	5	Dairy	Theta
pCI305	AF179848	8.694	32.41	8	Dairy	Theta
pCIS1	CP003165	4.263	31.97	2	Dairy	Theta
pCIS2	CP003164	5.461	30.07	4	Dairy	Theta
pCIS3	CP003163	6.159	35.85	5	Dairy	Theta
pCIS4	CP003162	7.045	38.42	10	Dairy	Theta
pCIS5	CP003161	11.676	34.06	10	Dairy	Theta
pCIS6	CP003160	38.673	37.12	30	Dairy	Theta
pCIS7	CP003159	53.051	32.40	48	Dairy	Theta
pCIS8	CP003158	80.592	33.97	72	Dairy	Theta
pCL2.1	U26594	2.047	33.95	2	Unknown	RCR
pCRL1127	AF409136	8.278	34.82	7	Unknown	Theta
pCRL291.1	AF380336	4.640	33.51	3	Unknown	Theta
pCV56A	CP002366	44.098	32.08	41	Human	Theta
pCV56B	CP002367	35.934	34.54	31	Human	Theta
pCV56C	CP002368	31.442	32.49	27	Human	Theta
pCV56D	CP002369	5.543	32.24	6	Human	Theta
pCV56E	CP002370	2.262	33.82	4	Human	Theta
pDBORO	DQ089807	16.404	35.16	15	Unknown	Theta
pDR1-1	AB079381	7.412	33.70	6	Dairy	Theta
pDR1-1B	AB079380	7.344	33.74	6	Dairy	Theta
pFI430	DQ011112.1	59.474	34.63	57	Dairy	Theta
pGdh442	AY849557	68.319	35.11	63	Plant	Theta
pHP003	AF247159	13.433	40.05	6	Dairy	Theta

pIL1	HM021326	6.382	32.28	7	Dairy	Theta
pIL105	AF116286	8.506	29.79	7	Dairy	Theta
pIL2	HM021327	8.277	34.82	10	Dairy	Theta
pIL3	HM021328	19.244	35.11	20	Dairy	Theta
pIL4	HM021329	48.978	35.11	47	Dairy	Theta
pIL5	HM021330	23.395	34.49	22	Dairy	Theta
pIL6	HM021331	28.434	33.64	25	Dairy	Theta
pIL7	HM197723	28.546	34.10	26	Dairy	Theta
pJM1A *	CP016747	51.777	35.02	53	Dairy	Theta
pJM1B *	CP016748	48.280	33.94	63	Dairy	Theta
pJM1C *	CP016749	30.146	35.40	29	Dairy	Theta
pJM1D *	CP016750	15.360	35.25	12	Dairy	Theta
pJM1E *	CP016751	11.008	31.95	11	Dairy	Theta
pJM1F *	CP016752	5.329	34.28	6	Dairy	Theta
pJM2A *	CP016742	11.314	37.77	11	Dairy	Theta
pJM2B *	CP016743	13.334	34.48	13	Dairy	Theta
pJM2C *	CP016744	62.261	35.12	56	Dairy	Theta
pJM3A *	CP016737	75.814	35.44	80	Dairy	Theta
pJM3B *	CP016738	47.185	34.84	46	Dairy	Theta
pJM3C *	CP016739	45.257	33.11	59	Dairy	Theta
pJM3D *	CP016740	13.546	33.63	15	Dairy	Theta
pJM3E *	CP016741	3.729	32.90	5	Dairy	Theta
pJM4A *	CP016729	60.219	33.38	74	Dairy	Theta
pJM4B *	CP016730	2.239	33.50	5	Dairy	RCR
pJM4C *	CP016731	5.931	34.53	7	Dairy	Theta
pJM4D *	CP016732	6.207	35.98	8	Dairy	Theta
pJM4E *	CP016733	47.240	34.85	43	Dairy	Theta
pK214	X92946	29.871	32.45	29	Unknown	Theta
pKF147A	CP001835	37.510	32.38	32	Plant	Theta
pKL001	EU289287	6.068	32.86	4	Unknown	Theta
pKP1	FR872378	16.181	35.94	7	Dairy	Theta
pL2	DQ917780	5.299	32.46	5	Dairy	Theta

pLP712	FJ649478.1	55.395	37.39	44	Dairy	Theta
pMN5	AF056207	5.670	30.26	4	Dairy	RCR
pMPJM1 *	CP016746	193.245	33.83	186	Dairy	Theta
pMPJM2 *	CP016745	113.820	34.92	123	Dairy	Theta
pMRC01	AE001272	60.232	30.11	63	Dairy	Theta
pNCDO2118	CP009055	37.571	32.33	32	Plant	Theta
pND324	U44843	3.602	33.37	3	Unknown	Theta
pNP40	DQ534432	64.980	32.33	62	Dairy	Theta
pNZ4000	AF036485	42.810	33.31	45	Dairy	Theta
pQA504	CP003136	3.978	37.83	3	Dairy	Undetermined
pQA518	CP003135	17.661	37.40	13	Dairy	Theta
pQA549	CP003134	49.219	35.14	44	Dairy	Theta
pQA554	CP003133	53.630	34.86	54	Dairy	Theta
pS7a	AJ550509	7.302	33.43	5	Dairy	Theta
pS7b	AJ550510	7.264	33.65	5	Dairy	Theta
pSRQ700	U16027	7.784	34.19	9	Dairy	Theta
pSRQ800	U35629	7.858	31.33	7	Dairy	Theta
pSRQ900	AF001314	10.836	31.13	11	Dairy	Theta
pUC063A *	CP016715	75.962	35.31	79	Dairy	Theta
pUC063B *	CP016716	44.205	34.27	41	Dairy	Theta
pUC063C *	CP016717	11.663	32.55	15	Dairy	Theta
pUC063D *	CP016718	8.697	32.39	10	Dairy	Theta
pUC063E *	CP016719	8.551	31.53	11	Dairy	Theta
pUC06A *	CP016734	36.928	32.10	43	Dairy	Theta
pUC06B *	CP016735	48.632	34.82	55	Dairy	Theta
pUC06C *	CP016736	23.429	31.87	29	Dairy	Theta
pUC08A *	CP016726	89.015	34.19	102	Meat	Theta
pUC08B *	CP016727	49.037	34.22	52	Meat	Theta
pUC08C *	CP016728	15.396	30.83	21	Meat	Theta
pUC109A *	CP016707	64.175	33.17	83	Dairy	Theta
pUC109B *	CP016708	48.261	34.63	51	Dairy	Theta
pUC109C *	CP016709	11.868	32.20	14	Dairy	Theta

pUC109D *	CP016710	11.333	31.64	13	Dairy	Theta
pUC109E *	CP016711	4.237	33.35	5	Dairy	Theta
pUC109F *	CP016712	2.413	33.11	3	Dairy	RCR
pUC11A *	CP016720	59.284	33.91	65	Meat	Theta
pUC11B *	CP016721	49.307	34.22	53	Meat	Theta
pUC11C *	CP016722	19.351	35.19	18	Meat	Theta
pUC11D *	CP016723	15.393	30.82	17	Meat	Theta
pUC11F *	CP016725	5.238	30.99	4	Meat	RCR
pUC77A *	CP016713	6.083	35.75	7	Dairy	Theta
pUC77B *	CP016714	63.462	34.86	66	Dairy	Theta
pUL8A *	CP016704	7.652	33.95	6	Dairy	Theta
pUL8B *	CP016705	27.296	35.31	30	Dairy	Theta
pUL8C *	CP016706	2.119	34.07	3	Dairy	RCR
pVF18	JN172910	18.977	33.90	21	Dairy	Theta
pVF21	JN172911	21.728	33.59	14	Dairy	Theta
pVF22	JN172912	22.166	35.14	19	Dairy	Theta
pVF50	JN225497	53.876	34.50	41	Dairy	Theta
pWC1	L75827	2.846	29.48	1	Dairy	RCR
pWV01	X56954	2.178	33.43	4	Dairy	RCR
pWVO2	NC_002193.1	3.826	31.34	1	Unknown	Theta
SK11 p1	CP000426	14.041	34.37	13	Dairy	Theta
SK11 p2	CP000427	9.554	30.44	10	Dairy	Theta
SK11 p3	CP000428	74.750	35.41	69	Dairy	Theta
SK11 p4	CP000429	47.208	34.84	42	Dairy	Theta
SK11 p5	CP000430	14.206	33.55	10	Dairy	Theta

* Plasmids sequenced in the context of the current study

4.3.2 Pan-plasmidome calculation

The pan-plasmidome calculation provides an overview of the overall genetic diversity of the *L. lactis* plasmidome, the latter representing the total plasmid content harboured by members of the *L. lactis* taxon. To calculate the pan-plasmidome, a pan-genome analysis approach was applied using the PGAP v1.0 pipeline [24]. The resultant pan-plasmidome graph (Fig. 4.1) displays an asymptotic curve rising steadily as each of the one hundred and forty eight plasmids included in the analysis is added until a total pan-plasmidome size of one thousand one hundred and twenty nine coding sequences (CDSs) was reached. The trend observed in the pan-genome indicates that the pan-plasmidome remains in a fluid or open state, therefore, continued plasmid sequencing efforts are expected to further expand the observed genetic diversity among lactococcal plasmids. The PGAP pipeline was also used to determine the core genome of the lactococcal plasmid sequence data set. Interestingly, no single CDS is conserved across all plasmids therefore resulting in an empty core genome.

The *L. lactis* pan-genome, based on chromosomal sequences only, has previously been calculated to constitute 5906 CDSs (Chapter III). When compared with the calculated lactococcal plasmidome (1129 CDSs), it is obvious that the lactococcal plasmidome contributes very substantially to overall lactococcal genetic diversity.

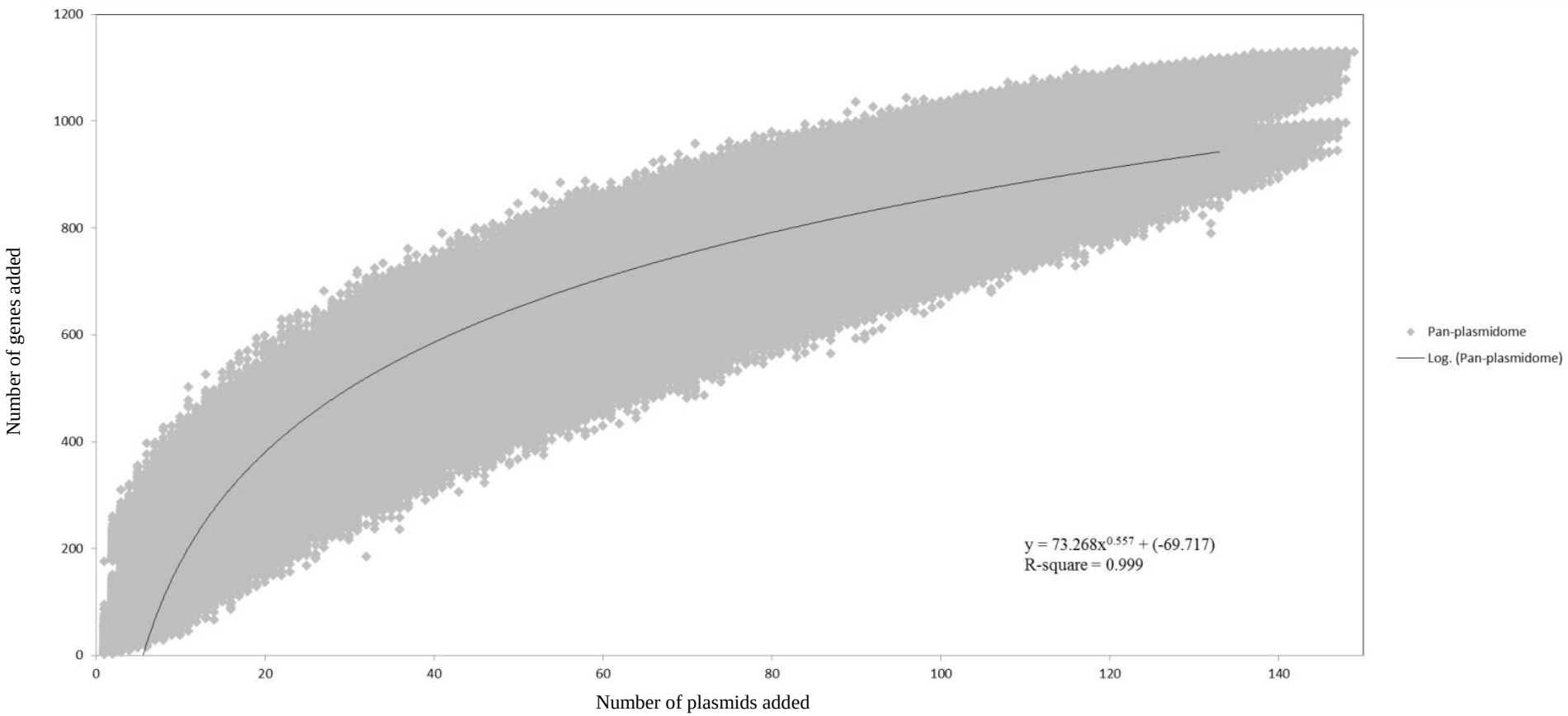


Figure 4.1: Pan-plasmidome of *L. lactis*
This represents accumulated number of new genes in the *L. lactis* pan-plasmidome plotted against the number of plasmids added. The deduced mathematical function is also indicated.

4.3.3 MCL analysis of the lactococcal plasmidome

To explore the genetic content of the one hundred and forty eight plasmids in this study, all-against-all reciprocal BLASTP (Basic local alignment search tool) analysis and MCL (Markov clustering) was conducted [31, 32]. The plasmidome was determined to comprise seven hundred and forty protein families, of which three hundred and forty nine represented unique proteins, evidence of the divergent nature of the plasmid sequences. At present, three hundred and five of these families constitute hypothetical protein families, representing a total of eight hundred and seventy seven individual proteins. These hypothetical proteins encompass 21.9 % of the total CDSs in the lactococcal plasmidome.

The largest constituent of the lactococcal plasmidome is that represented by transposable elements. Transposable elements encompass eight hundred and ninety two CDSs, or 22 % of the plasmidome, with members of the IS6, IS30, IS982 and ISL3 insertion families being among the most dominant genetic elements. These mobile elements are responsible for the transfer and recombination of DNA [33-35], and are likely to contribute to a fluid lactococcal plasmidome.

4.3.4 Lactococcal megaplasmids

Typically *L. lactis* plasmids range in size from 1-50 Kbp, and prior to this study the largest plasmid identified in *L. lactis* was pCIS8 (80.59 Kbp) from *L. lactis* UC509.9 [36]. In the current study, whole genome sequencing efforts resulted in the identification of two plasmids that were larger than 100 Kbp, namely pMPJM1 (193 Kbp) and pMPJM2 (113 Kbp) from *L. lactis* JM1 and *L. lactis* JM2, respectively, and owing to their size are defined as megaplasmids (Fig. 4.2A & B). Pulse field gel electrophoresis also identified bands which would be consistent with plasmids of that

size, although unambiguous validation will require Southern hybridization (Fig. 4.2C).

The larger of the two megaplasmids, pMPJM1, encompasses 186 CDSs and is presumed to replicate (as expected for such a large replicon) via the theta-type replication mechanism [based on the identification of the origin of replication (*ori*), comprised of an AT-rich region plus three and a half iterons of 22 bp in length] [37]. pMPJM1 encompasses, among others, gene clusters predicted to be responsible for (exo)polysaccharide biosynthesis, conjugation and nisin resistance, while it also specifies an apparently novel type I RM shufflon system (as well as a high proportion of unique/hypothetical CDSs). The overall sequence of the plasmid shows little homology to previously sequenced plasmids in the NCBI databases, however, it shares 24 % sequence coverage with 99 % nucleotide identity to the other identified megaplasmid pMPJM2, which indicates that they share a common ancestor. pMPJM2 encodes 123 CDSs and BLAST analysis identified sequence identity to a number of different lactococcal plasmids indicating a mosaic genetic structure commonly seen in large lactococcal plasmids [1]. pMPJM2 also encodes a putative conjugation operon and a very close homolog of the type I RM shufflon system of pMPJM1.

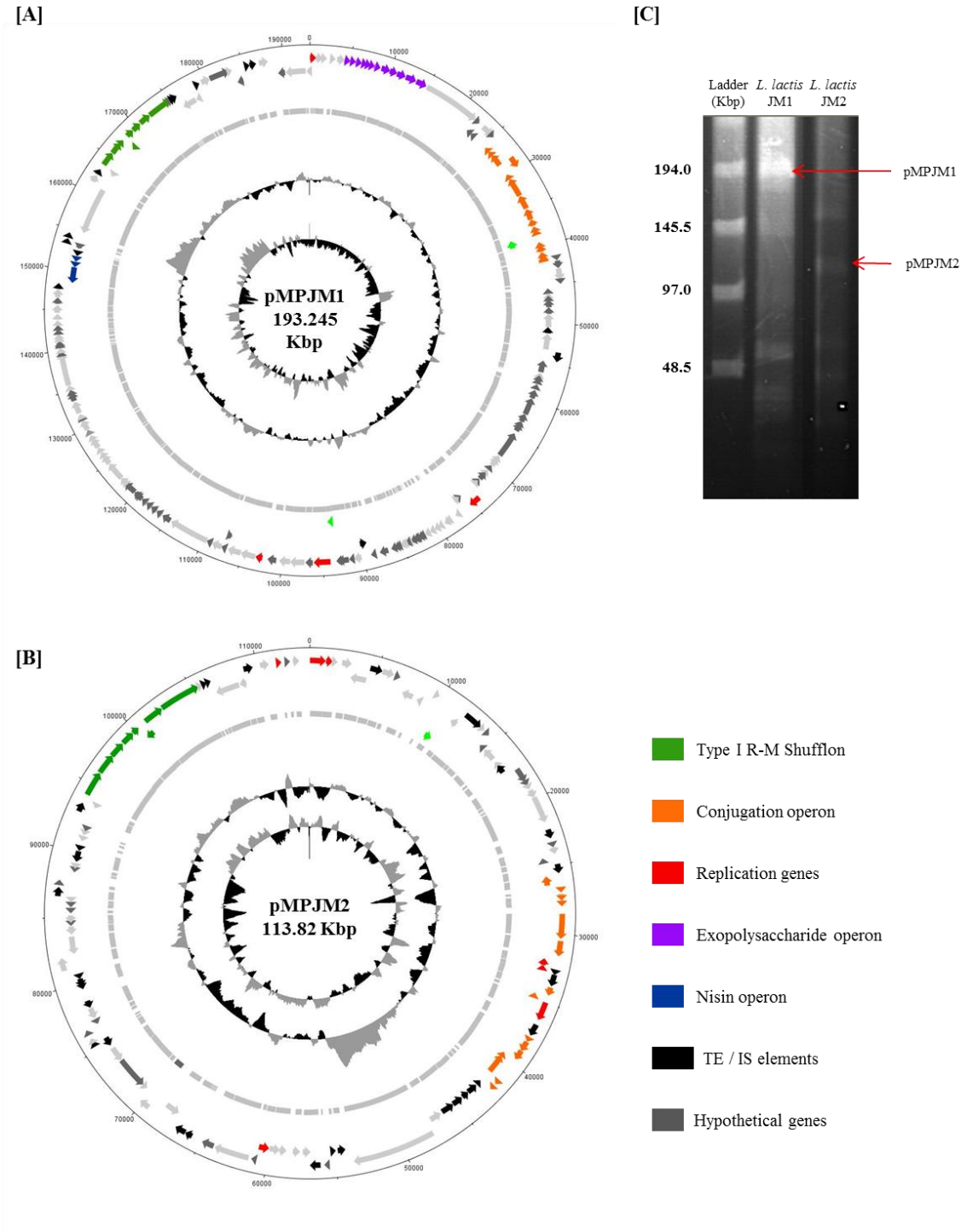


Figure 4.2: Megaplasmids pMPJM1 and pMPJM2 general features

[A] Circular maps of pMPJM1 and [B] pMPJM2. CDS of interest are highlighted in colour. [C] PFGE image of pMPJM1 (lane 2) and pMPJM2 (lane 3), the possible position of each of the two megaplasmids is indicated by a red arrow. CHEF lambda (Bio-Rad Laboratories, Hercules, CA) DNA ladder is also indicated (lane 1).

4.3.5 Technological properties

Strains of *L. lactis* are commonly used as starter cultures employed by the dairy industry [38], and their dairy adaptations such as citrate metabolism and lactose utilisation are frequently plasmid-encoded. In *L. lactis*, citrate uptake and subsequent diacetyl production is governed by the plasmid-encoded *citQRP* operon [15]. In the current data set, only two plasmids encompass the *citQRP* operon, *L. lactis* CRL1127 plasmid pCRL1127 and *L. lactis* IL594 plasmid pIL2 [39]. Lactose metabolism is controlled by the *lac* operon consisting of the genes *lacABCDEFGX* and is regulated by a repressor, encoded by the adjacent *lacR* gene [40], both citrate and lactose utilisation have previously been described in detail [15, 40].

In this study the *lac* operon was found to be present on twenty plasmids (in twenty different strains) (Table 4.2). The plasmids analysed were derived from forty seven lactococcal strains in addition to seventeen lactococcal plasmids unassigned to a particular strain, and represented the total plasmid complement of twenty five such strains. In all cases bar one, the strains were isolated from the dairy environment with the exception of *L. lactis* NCDO1867 isolated from peas (Table 4.1). Alternative lactose metabolism methods have previously been observed in *L. lactis* (Chapter 1). For example, *L. lactis* MG1363 does not harbour the *lac* operon, yet is capable of growth on lactose-supplemented media due to the activity of a cellobiose-specific phosphotransferase system (PTS), which can act as an alternative lactose utilisation pathway [41]. Another example of an alternative lactose metabolic pathway is found in the slow lactose fermenter *L. lactis* NCDO2054 which metabolises lactose via the Leloir pathway [42]. Plasmid integration events discussed in Chapter III have also resulted in the integration of the *lac* operon in the chromosome of *L. lactis* SO, where it is located 20 Kbp downstream of an integrated

opp operon, sharing significant homology with (the lac operons of) plasmids pCV56B, pSK08, pKF147A and pNCDO2118. Due to the lack of complete sequencing projects, defining the true frequency of lactose utilisation is problematic. However of those strains for which complete genome sequencing projects have been described (thirty strains in Chapter III) twenty two were found to be capable of metabolizing lactose based on growth in lactose supplemented broth, nineteen via plasmid-encoded *lac* operons, one via a chromosomally-encoded *lac* operon and two by an alternative pathway. This analysis included twelve subsp. *cremoris* strains, of which all but one possessed genes for a lactose utilisation mechanism, the exception being strain KW2, which lacks a plasmid complement.

Table 4.2: Overview of presence of plasmid-encoded *lac*/*opp* operons

Stain	Subspecies	Origin	Plasmid
SK11	*cremoris*	Dairy	pSK114
158	*cremoris*	Dairy	p158C
229	*lactis*	Dairy	p229A
275	*lactis*	Dairy	p275C
A76	*cremoris*	Dairy	pQA549
JM1	*cremoris*	Dairy	pJM1A
JM2	*cremoris*	Dairy	pJM2C
JM3	*cremoris*	Dairy	pJM3B
JM4	*cremoris*	Dairy	pJM4E
UC063	*lactis*	Dairy	pUC063A
UC06	*lactis*	Dairy	pUC06B
UC109	*cremoris*	Dairy	pUC109B
UC77	*lactis*	Dairy	pUC77B
UC509.9	*cremoris*	Dairy	pCIS8
DPC3901	*lactis* bv. diacetylactis	Dairy	pVF50
IL594	*lactis*	Dairy	pIL4
NCDO712	*cremoris*	Dairy	pLP712
UC08	*lactis*	Dairy	pUC08A
UC11	*lactis*	Dairy	pUC11A
NCDO1867	*lactis*	Plant	pGdh442

4.3.6 Conjugation

Conjugation and transduction are believed to be the dominant mechanisms of plasmid transfer in *L. lactis* [1]. Recently, particular emphasis has been placed on conjugation as it is considered a naturally occurring DNA transfer process and for this reason may be used in food-grade applications to confer beneficial traits to industrial strains [11]. Generally, during conjugation the AT-rich, so-called 'origin of transfer' or *oriT* of the conjugative plasmid is nicked by a nickase, and the resulting ssDNA strand is passed on to a recipient cell [10], though the precise mechanistic details of the conjugation process in *L. lactis* remain unclear.

The *tra* (transfer) locus is believed to be responsible for the donor-to-recipient DNA transfer process of conjugation. Previous studies have identified the role of *traF* as encoding a membrane-spanning protein involved in channel formation and membrane fusion. In addition, the *traE* and *traG* genes have been proposed to encode proteins involved in the formation of the conjugal pilus similar to type IV secretion systems [43, 44]. Typically, the three *tra* genes (i.e. *traE*, *traF* and *traG*) are part of a larger gene cluster (consisting of up to fifteen genes; Fig. 4.3), including *traA*, which encodes a relaxase. However, precise functions for the remainder of the genes in the *tra* gene cluster have yet to be elucidated, though additional predicted *tra* genes were identified in a small number of cases, the majority based on homology to the *trs* operon in *Staphylococcus* [45]. For example, *traJ* and *traL* were identified on plasmids pAF22 and pMRC01, and *traB*, *traC*, *traD*, *traF* (mating channel formation) and *traK* (P-loop NTPase) on plasmids pUC08B, pUC11B, pAF22 and pMRC01. Plasmids pAF22, pMRC01 and pNP40 have all previously been demonstrated to be capable of conjugation [43, 46-48], however, the annotation of the operons involved is not well defined and they are

currently poorly characterised. This is also amplified by both a lack of sequence conservation and synteny within these operons (Fig 4.3).

While the *tra* operon is thought to be responsible for the formation of conjugal pilus, previous studies have identified a number of genes believed to play a role in the mobilisation of other (non-conjugatable) plasmids in *L. lactis* [11, 43, 49]; principal among these are the *mob* (mobilisation) genes. Mobilisation genes are responsible for nicking the plasmid's dsDNA at a particular site and forming a relaxome which allows the transfer of a single stranded template to a recipient cell. Variants of four main *mob* genes are distributed throughout the lactococcal plasmidome; *mobA* and *mobD* encode nickases, and *mobB* and *mobC*, whose protein products are thought to form a relaxosome with an associated nickase (either *mobA* or *mobD*) are typically present in the genetic configuration *mobABC* or *mobDC*. Comparative analysis identified 372 occurrences of *mob* genes (and *mob*-associated genes) distributed on the 148 plasmids in this study, including thirteen occurrences of a predicted retron-type reverse transcriptase or maturase (located between *mobD* and *mobC*) believed to play a role in DNA recombination. The results indicate that 69.6 % of plasmids in the lactococcal plasmidome carry at least one or more genes encoding mobilisation proteins.

The lactococcal megaplasmids pMPJM1 and pMPJM2 harbour two (16 Kb) regions putatively involved in conjugation and/or mobilisation. In the case of pMPJM2 the predicted region was found to contain homologues of *mobC* and *mobD*, encoding a nickase and associated relaxase near a `possible secondary replication origin, although the presence of five transposase-encoding genes and the lack of predicted *tra* genes with conserved functions suggest that this putative conjugation system is unlikely to be functional.

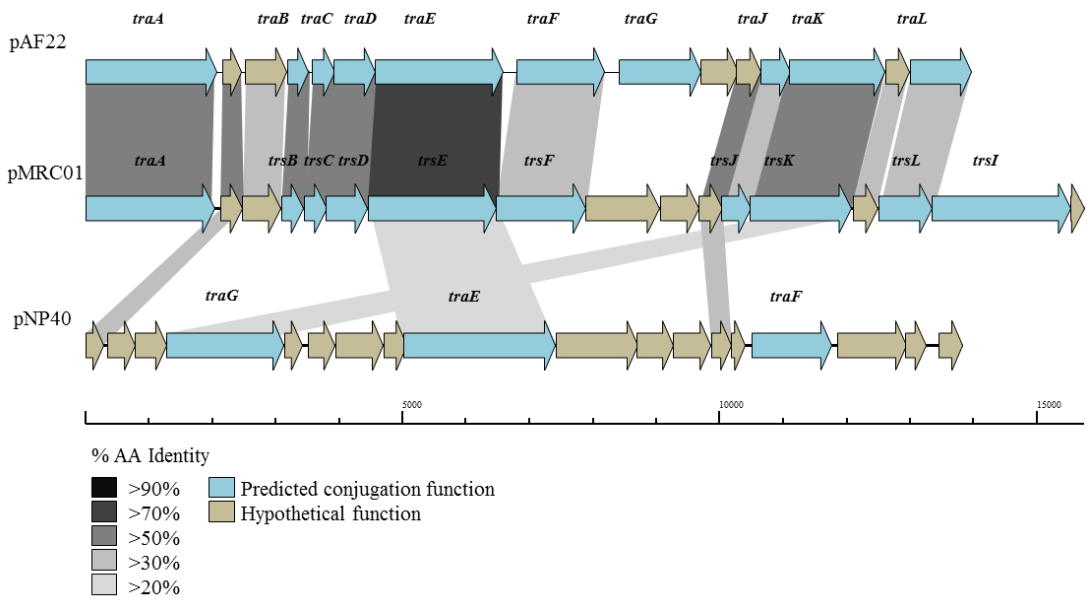


Figure 4.3: BLAST map of active lactococcal conjugation operons
The image describes the genetic organisation of the conjugation operons from plasmids; pAF22, pMRC01 and pNP40. All three plasmids have previously been shown to be conjugatable. Gene synteny is highly conserved between pAF22 and pMRC01, but amino acid identity is not, while pNP40 represents a more divergent system. Amino acid identity is indicated by the shaded boxes. Arrows coloured blue indicate predicted conjugative function, while arrows shaded mustard indicate hypothetical functions.

Conversely, analysis of pMPJM1 identified a more divergent system to that typically found in lactococcal plasmids. Three hypothetical proteins were found to contain the PFAM domain (pfam12846) usually conserved in conjugation proteins, in addition to a homolog of *virB11*, whose deduced product acts as a type IV secretory pathway ATPase (pfam00437). Cellular localisation analysis of the operon using PsortB was also indicative of a transmembrane complex (Fig. 4.4). The divergence of both operons from typical lactococcal conjugative operons suggests that these two megaplasmids have lost their conjugative ability.

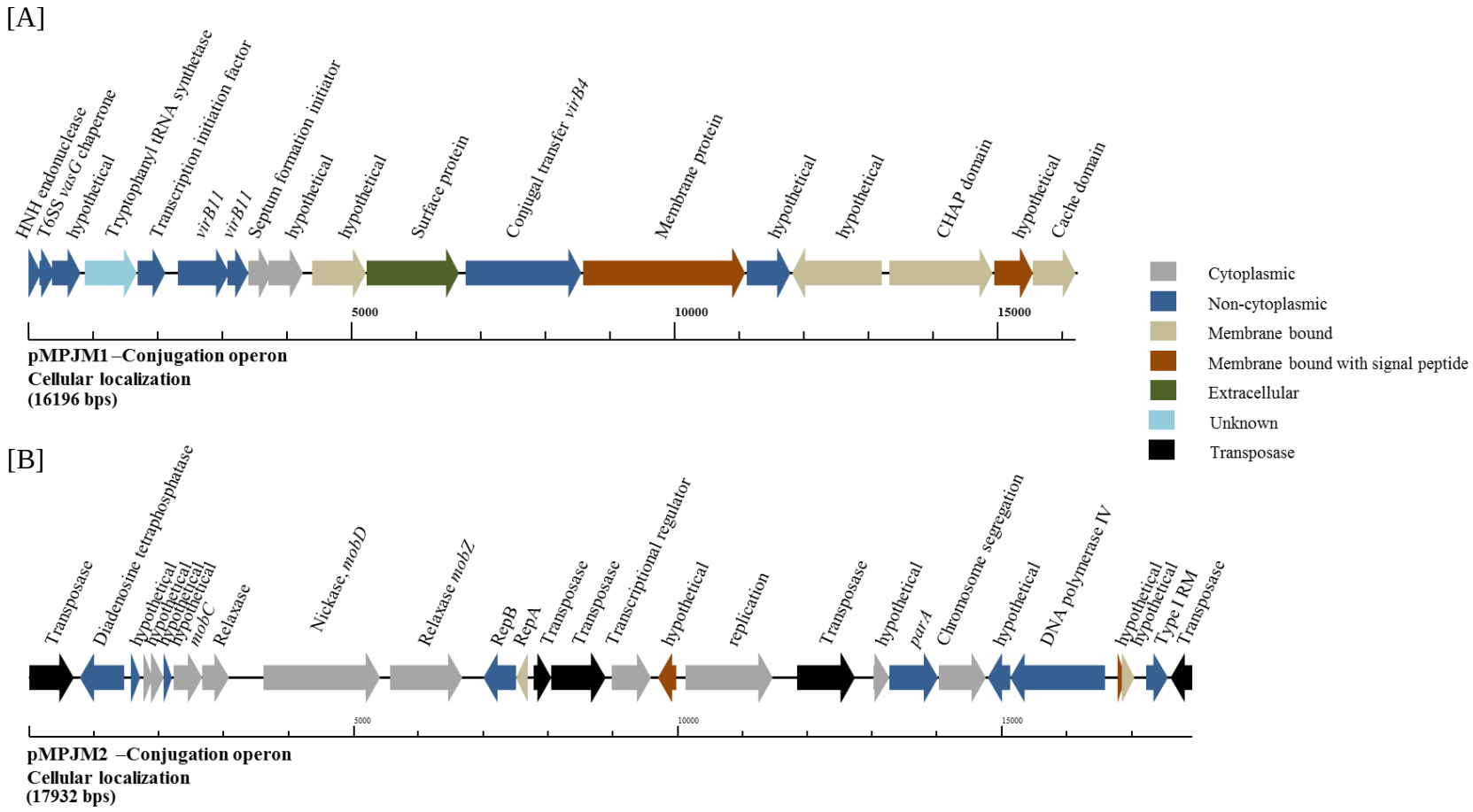


Figure 4.4: Genetic organisation of the putative conjugation operons in pMPJM1 and pMPJM2
[A] Represents the putative conjugation locus in pMPJM1. [B] Represents the putative conjugation locus in pMPJM2. Colours indicate the predicted cellular localization of each product. The system in pMPJM1 appears to encode proteins involved in conjugal transfer, while the cellular localisation data is predictive of a transmembrane complex. Conversely, the conjugation locus in pMPJM2 appears to be involved in mobilisation rather than conjugation, and the presence of a number of insertion elements suggest it is unlikely to be functional.

4.3.7 Cell surface interactions (Adhesion & EPS)

4.3.7.1 Adhesion

Mucin-binding proteins, i.e. those which allow adhesion to the mucin layer of the gastrointestinal tract, are considered essential for stable and extended gut colonisation by LAB [50]. While lactococci are typically not associated with the human gut, instances of such proteins encoded by lactococcal plasmids have been reported [51-53]. Muco-adhesive proteins are considered of paramount importance for the efficacy of probiotic bacteria [50] and the presence of such elements in *L. lactis* may have significant commercial impact for their role in functional foods.

Analysis of the plasmids in this study identified a number of strains with predicted novel muco-adhesive elements, similar to those found in pKP1 [52]. Plasmid pKP1 encodes two proteins, a mucin-binding domain-containing protein and an aggregation-promoting protein AggL, which promotes its binding to colonic mucosa [53]. While no direct homologue of AggL was detected, mucus-binding protein-encoding genes were identified on plasmids p275A, p275B, pUC08B and pUC11B, perhaps reflecting a potential for gastrointestinal persistence conferred to the strains that carry these plasmids. A number of additional proteins predicted to be host cell surface-associated, were detected during the analysis. For example, pUC11C encodes two class C sortases, which are commonly involved in pilus biosynthesis [50, 54], while p275A encodes a LPXTG anchor domain, cell surface-associated protein. Interestingly each of these strains belongs to subspecies *lactis* and is capable of growth at 37 °C, which would impede the growth of their *cremoris* counterparts, which are generally less thermo-tolerant. *L. lactis* JM1 is the sole *cremoris* strain that is predicted to encode proteins directly involved in host cell surface alterations. This plasmid encodes five putative proteins containing a 26-

residue repeat domain found in predicted surface proteins (often lipoproteins) and one collagen-binding domain protein.

4.3.7.2 EPS production

EPS production by *L. lactis* is a characteristic trait of strains isolated from viscous Scandinavian fermented milk products and is widely reported as a plasmid-encoded trait [55-58]. EPS production by *L. lactis* strains is of particular importance for functional foods, as the EPS produced by these strains is considered to be a food-grade additive that significantly contributes to properties such as mouth-feel and texture in fermented dairy products [59]. The *L. lactis* EPS biosynthesis gene cluster (*eps*) contained on pNZ4000 has previously been characterised [55] and consists of 14 genes *epsRXABCDEFGHIJK*. Comparison of the *eps* gene cluster from pNZ4000 with all sequenced plasmids in the current dataset identified a further four plasmids which harbour *eps* clusters, namely p229E, pJM3C, p275B and pMPJM1 (Fig. 4.5). In pNZ4000, EPS production is regulated by *epsRX*, EPS subunit polymerisation and export is believed to be executed by the encoded products of *epsABIK*, while the proteins encoded by *epsDEFGH* are responsible for the biosynthesis of the EPS subunit [55]. Homology-based analysis with the four newly identified gene clusters shows that in all cases *epsRXABCD* are conserved (except in pMPJM1 where *epsR* is absent), while the remainder of the gene cluster in each case consists of variable genes. These *eps* gene clusters consist of a highly conserved region at the proximal end of the cluster and a variable distal region, which is similar to other lactococcal polysaccharide biosynthesis clusters [60-62]. The conserved *epsRX* genes are responsible for transcriptional regulation, the products of *epsAB* are required for EPS export, while the deduced proteins of *epsCD* are putative glycosyltransferases of

which EpsD (priming glycosyltransferase) has previously been demonstrated to be essential for EPS subunit biosynthesis [55]. The variable region, *epsEFGHIJKLP* in pNZ4000, encodes the enzymatic machinery responsible for EPS subunit biosynthesis.

In the case of p229E, the variable *eps* region is composed of CDSs predicted to encode products with functions similar to the CWPS operon in strain 229. Plasmid pJM3C contains genes predicted to encode a rhamnosyltransferase, UDP-glucose dehydrogenase, capsular biosynthesis protein and five glycosyltransferases. The p275B variable region is heavily rearranged due to the presence of nine transposase-encoding genes. The megaplasmid pMPJM1 encodes a 9 Kb EPS region with well conserved synteny to pNZ4000, although with relatively low homology (Fig. 4.5). Further analysis of these plasmid-borne *eps* gene clusters revealed that in all cases *mob* elements are also present indicating that they may be mobilisable via conjugation. To assess if these plasmids had a common lineage, nucleotide homology based analysis was conducted utilising BLASTN [31]. This analysis however did not identify significant homology or common hits between the plasmids outside of the conserved region of the EPS gene cluster.

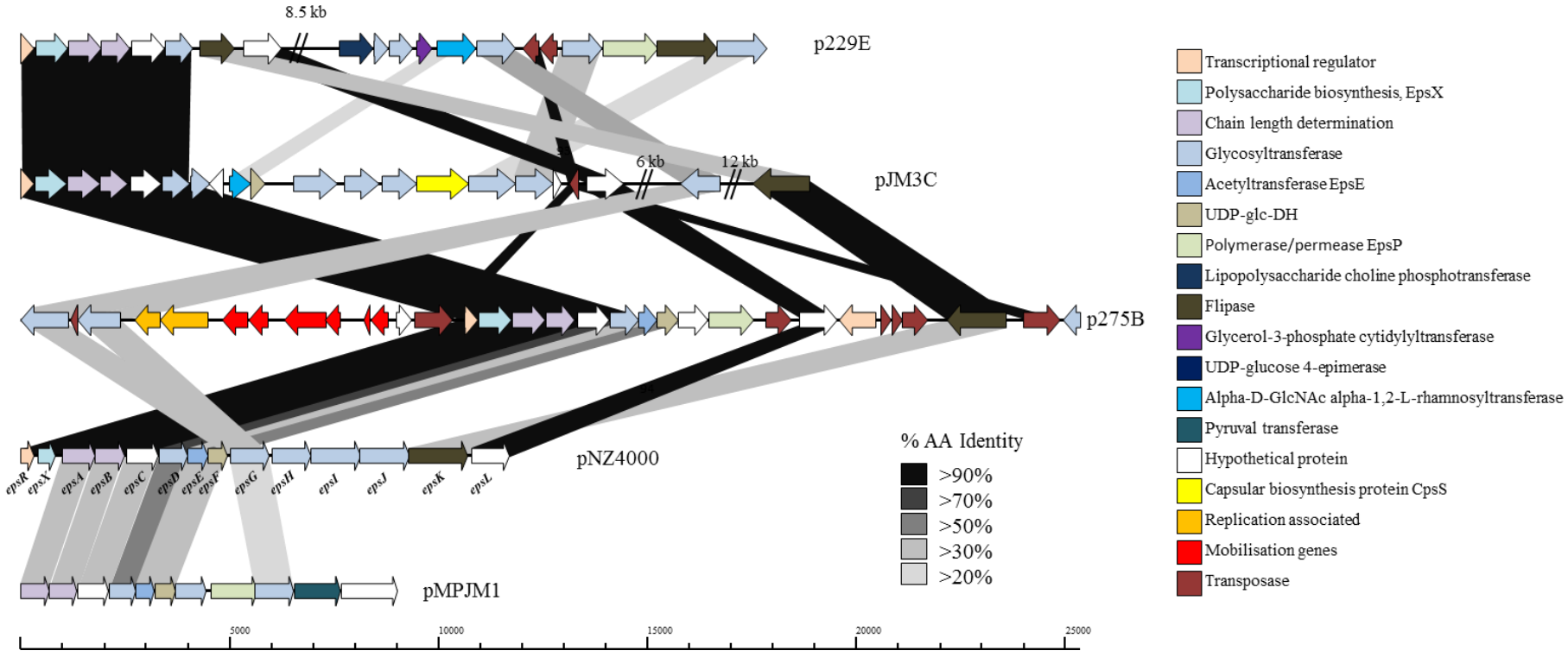


Figure 4.5: Linear BLAST map of the lactococcal EPS gene clusters
Linear BLAST map of *eps* gene clusters from [1] p229E, [2] pJM3C, [3] p275B, [4] pNZ4000, [5] pMPJM1. Arrow colour indicates predicted product, while shaded region indicated percentage amino acid identity between BLAST hits. The highly conserved region of the gene cluster is apparent from EpsR to EpsD while the variable region is strain specific.

4.3.8 Bacteriocins

Bacteriocins are a diverse group of ribosomally synthesized bacterial peptides, which when secreted inhibit growth of other bacteria by interfering with cell wall biosynthesis or disrupting membrane integrity [63]. To investigate bacteriocin production in the lactococcal plasmidome, available strains were screened for bacteriocin production against an indicator strain *L. lactis* subsp. *cremoris* HP. In total six strains were found to produce clearly defined zones of inhibition, indicating bacteriocin production, namely *L. lactis* subsp. *lactis* IO-1, 184, UC06, UC08, UC11 and *L. lactis* subsp. *cremoris* 158. Analysis of the plasmid complement of each of these strains indicated that strains 158, UC06 and UC08 each possess a plasmid-borne bacteriocin gene cluster, while IO-1, 184 and UC11 contain a bacteriocin gene cluster of chromosomal origin. In each case these were identified as lactococcin producers: p158A is predicted to be responsible for lactococcin A & B production, pUC08A for lactococcin A production, and pUC06C for lactococcin B biosynthesis. Lactococcin has a narrow spectrum of activity, targeting predominantly closely related lactococcal species [64] and as such is an important consideration when selecting strains for use in mixed starter cultures.

Sequence analysis of the remaining plasmids in the current study (for which strains were not available for phenotypic analysis) identified additional putative bacteriocin-encoding gene clusters (Table 4.3), which were found to be responsible for the production of lactococcin A or B, and in one case (pMRC01) for the lantibiotic lacticin 3147 (Table 4.3) [65].

Table 4.3: Plasmid-encoded antimicrobial peptides

Plasmid	Bacteriocin	Activity detected
pBL1	Lactococcin 972	N/A[$]
pCIS7	Lactococcin A	N/A
pMN5	Predicted/uncharacterised bacteriocin	N/A
pMRC01	Lacticin 3147	N/A
SK11 plasmid 1	Lactococcin A	No
p158A	Lactococcin A and B	Yes
pUC08C	Lactococcin A	Yes
pUC06C	Lactococcin B	Yes

[$]N/A, host strain unavailable to screen phenotypically

4.3.9 Phage resistance systems

Lactococcal strains possess an arsenal of phage defence mechanisms including Restriction Modification (R-M) systems, Superinfection exclusion systems (Sie) (encoded by integrated prophages) and Abortive infection systems (Abi). Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR associated (*cas*) genes form an acquired adaptive immunity system against foreign DNA in bacteria [66]. To date only one such system has been characterised in *Lactococcus* on a conjugation-transmissible plasmid, pKLM which encodes a novel type III CRISPR-Cas system (though it is unable to incorporate new spacers) [49]. Analysis of plasmid sequences in this study did not detect any further instances of CRISPR systems in lactococci.

R-M systems are extremely diverse and widespread and are encoded by approximately 90 % of all currently available bacterial and archaeal genomes [67]

(NB. An in-depth analysis of lactococcal R-M systems is conducted in Chapter V). Similarly, Sie systems are a prophage-encoded defence mechanism [68, 69] and will be discussed in detail as part of an investigation into lactococcal prophages in Chapter VI.

4.3.10 Abortive infection systems

Abortive infection systems (Abi) represent an abundant phage defence mechanism in *L. lactis* [70] and are frequently plasmid-encoded [11]. To date, twenty three Abi systems have been identified in *L. lactis* of which, twenty one are plasmid-encoded [1]. They are single gene systems, with the exception of three multigene systems, AbiE [71], AbiR [72] and AbiT [73]. Analysis of the plasmids in this study identified fourteen Abi occurrences based on homology, namely AbiF, AbiC, AbiK, AbiQ and the two component system AbiEi/AbiEii, in addition to one predicted uncategorised Abi (Table 4.4).

Table 4.4: Lactococcal Abi systems detected

Similar to Abi system	Plasmid	Locus tag
AbiF	p158B	LL158_pB41
AbiF	pCIS8	UC509_RS11675
AbiF	pIL105	pIL105p7
AbiF	pNP40	pNP40_p16
AbiC	p275A	LL275_pA087
AbiEi-Eii	p275A	LL275_pA051-052
AbiEi-Eii	pNP40	pNP40_p19-20
AbiK	pSRQ800	pSRQ800_04
AbiQ	pCV56A	CVCAS_RS12180
AbiQ	pSRQ900	pSRQ900_04
Uncharacterised Abi *	p158E	LL158_pE13
Uncharacterised Abi	pUC063B	LLUC063_pB07
Uncharacterised Abi	pCIS8	UC509_RS11625
Uncharacterised Abi	pCIS5	UC509_RS12350

* Uncharacterised Abi, based on amino acid homology to unclassified Abi's in the NCBI database

4.4 Discussion

The advent of next generation sequencing technologies has made genome sequencing more accessible and has led to a dramatic rise in the number of available genome sequences. In this study one such technology, SMRT sequencing was applied for the elucidation of sixty seven novel lactococcal plasmids. The main advantage of SMRT technology is the long read length it achieves, which is particularly useful when assembling lactococcal plasmids due to the high frequency of repetitive transposable elements which can lead to incorrectly assemblies. This is also beneficial for assembling larger lactococcal plasmids which are frequently composed of a mosaic type structure and may encode multiple identical IS elements which may complicate assemblies with shorter read lengths [1]. However, during the course of the current study some cautionary notes also emerged. These were predominantly related to smaller plasmids and plasmids with lower average consensus coverage which could potentially be filtered out under standard assembly parameters. It was found that by repeating the assembly with a reduced minimum coverage cut-off to 15-fold coverage permitted the detection of these plasmids, although it may well be that some plasmids may still have been missed, particularly if they are very small (<3 Kbp).

In the course of this study, the pan-plasmidome of *L. lactis* was calculated and found to be in a fluid state, making it likely that continued sequencing efforts will expand the diversity of this data set and lead to an increase in the identification of novel plasmid features. At present, the lactococcal plasmidome was found to consist of over 4000 Kbp of extra-chromosomal DNA encoding an arsenal of diverse features. Significantly, the current open plasmidome contributes the equivalent of 19.11 % of the CDSs contained in the pan-genome of the *L. lactis* chromosomes

which is in a closed state (see Chapter III). BLAST-based analysis of these features identified 742 protein families, of which 393 represented unique families, evidence of the divergent nature of the plasmid sequences. There is, however, a skew in the data set towards the dairy niche which has arisen due to a number of factors. Primarily, the majority of strains sequenced to date have been sequenced due to their commercial value in the production of fermented dairy products. The impact of these strains on the overall data set is then further amplified as these strains generally carry a larger plasmid complement than their non-dairy counterparts (Chapter III, Table 3.2) as many desirable dairy-associated traits are typically plasmid-encoded (e.g. *lac* operon). As such, these features account for a large proportion of the plasmidome. However, as efforts to isolate new diverse starter cultures for the dairy industry continue, screening of more diverse cultures particularly from the plant niche should lead to increased novelty in the lactococcal plasmidome.

Megaplasmids have been found in LAB previously, in particular in members of the *Lactobacillus genus* [74-77]. In this study sequencing efforts resulted in the identification of the first examples of lactococcal megaplasmids (> 100 Kbp), substantially surpassing the size of any previously sequenced plasmids in this taxon, and providing further diversity within the plasmidome. While megaplasmids are not expected to be essential for the growth of their host, they can encode additional metabolic capabilities. The lactococcal megaplasmids were also examined for the presence of conjugation machinery. A novel gene cluster encoding a number of conjugation-related proteins located in pMPJM1 was predicted to be involved in the conjugal transfer of the plasmid based on the presence of conserved structural domains involved in conjugation. Further analysis of *mob* and *tra* genes across the plasmidome identified a number of genes predicted to encode proteins involved in

conjugal transfer. The frequency (484 genes across 148 plasmids) of these genes is indicative of the mobilisable nature of lactococcal plasmids.

There has been limited research performed to date in the area of lactococcal gut adhesion as *L. lactis* is not commonly associated with the human gut. In this study, potential gut adhesion factors were identified within the lactococcal plasmidome, a key trait for persistence in the gastrointestinal tract. Similarly, this may offer further insights for the use of *L. lactis* as a vector for vaccine and biomolecule delivery, a rapidly growing area of research [78, 79]. Further technological properties of *L. lactis* were also investigated including EPS production. Analysis of a large dataset of newly sequenced plasmids facilitated the identification and comparison of a number of novel EPS gene clusters. The major outcome of this work was the definition of "conserved" and "variable" regions within these EPS clusters. The conserved region encodes the transcriptional regulation, export and biosynthesis initiation machinery, while the variable region contains various genes that are predicted to encode glycosyltransferases, which are believed to be responsible for the production of a diverse set of EPS subunits.

Finally, phage resistance mechanisms were assessed with particular emphasis on Abi systems. Abi systems confer defence against phage infection and are commonly found in lactococcal strains where they are frequently plasmid encoded [11]. Analysis of the plasmids sequences identified fourteen plasmid-encoded Abi systems, while further analysis also identified frequent occurrences of these systems within the lactococcal chromosomes [70]. The presence of these systems and a range of R-M systems is evidence of the adaptation of these strains towards phage resistance. Discovery of the first lactococcal megaplasmids along with a host of novel features is evidence that the diversity of the lactococcal plasmidome represents

a relatively untapped resource, and suggests that continued future sequencing will increase the observed diversity carried by these elements, potentially leading to new avenues of research and applications.

4.5 References

1. Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D: The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. FEMS Microbiol Rev. 2014, 38(5):1066-1088.
2. Price CE, Zeyniyev A, Kuipers OP, Kok J: From meadows to milk to mucosa – adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. FEMS Microbiol Rev. 2012, 36(5):949-971.
3. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N et al: Comparative genomics of the lactic acid bacteria. PNAS US. 2006, 103(42):15611-15616.
4. Goh YJ, Goin C, O'Flaherty S, Altermann E, Hutkins R: Specialized adaptation of a lactic acid bacterium to the milk environment: the comparative genomics of *Streptococcus thermophilus* LMD-9. Microb Cell Fact. 2011, 10:Suppl 1S22.
5. Ainsworth S, Zomer A, de Jager V, Bottacini F, van Hijum SA, Mahony J, van Sinderen D: Complete genome of *Lactococcus lactis* subsp. *cremoris* UC509. 9, host for a model lactococcal P335 bacteriophage. Genome announc. 2013, 1:e00119-00112.
6. Kelleher P, Murphy J, Mahony J, Van Sinderen D: Next-generation sequencing as an approach to dairy starter selection. Dairy Sci & Technol. 2015, 95(5):545-568.
7. Kelly WJ, Ward LJH, Leahy SC: Chromosomal Diversity in *Lactococcus lactis* and the Origin of Dairy Starter Cultures. Genome Biol Evol. 2010, 2:729-744.

8. Wegmann U, Overweg K, Jeanson S, Gasson M, Shearman C: Molecular characterization and structural instability of the industrially important composite metabolic plasmid pLP712. Microbiol. 2012, 158(12):2936-2945.

9. Ammann A, Neve H, Geis A, Heller KJ: Plasmid transfer via transduction from *Streptococcus thermophilus* to *Lactococcus lactis*. J Bacteriol. 2008, 190(8):3083-3087.

10. Grohmann E, Muth G, Espinosa M: Conjugative Plasmid Transfer in Gram-Positive Bacteria. Microbiology Mol Biol Rev. 2003, 67(2):277-301.

11. Mills S, McAuliffe OE, Coffey A, Fitzgerald GF, Ross RP: Plasmids of lactococci – genetic accessories or genetic necessities? FEMS Microbiol Rev. 2006, 30(2):243-273.

12. Van Rooijen R, Gasson M, De Vos W: Characterization of the *Lactococcus lactis* lactose operon promoter: contribution of flanking sequences and LacR repressor to promoter activity. J Bacteriol. 1992, 174(7):2273-2280.

13. van Rooijen RJ, De Vos W: Molecular cloning, transcriptional analysis, and nucleotide sequence of *lacR*, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. J Biol Chem. 1990, 265(30):18499-18503.

14. Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM: Complete sequences of four plasmids of *Lactococcus lactis* subsp. *cremoris* SK11 reveal extensive adaptation to the dairy environment. Appl Environ Microbiol. 2005, 71(12):8371-8382.

15. Drider D, Bekal S, Prévost H: Genetic organization and expression of citrate permease in lactic acid bacteria. Genet Mol Res. 2004, 3(2):271-281.

16. McSweeney PLH, Sousa MJ: Biochemical pathways for the production of flavour compounds in cheeses during ripening: A review. Lait. 2000, 80(3):293-324.

17. McSweeney PLH: Biochemistry of cheese ripening. Int J Dairy Technol. 2004, 57(2-3):127-144.

18. van Rooijen RJ, de Vos WM: Molecular cloning, transcriptional analysis, and nucleotide sequence of *lacR*, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. J Biol Chem. 1990, 265(30):18499-18503.

19. Broadbent JR, Barnes M, Brennand C, Strickland M, Houck K, Johnson ME, Steele JL: Contribution of *Lactococcus lactis* cell envelope proteinase specificity to peptide accumulation and bitterness in reduced-fat Cheddar cheese. Appl Environ Microbiol. 2002, 68(4):1778-1785.

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990, 215(3):403-410.

21. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: The Pfam protein families database. Nuc Acids Res. 2004, 32(suppl 1):D138-D141.

22. Söding J, Biegert A, Lupas AN: The HHpred interactive server for protein homology detection and structure prediction. Nuc Acids Res. 2005, 33(Web Server issue):W244-W248.

23. van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP: BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucl Acids Res. 2013, 41(W1):W448-W453.

24. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J: PGAP: pan-genomes analysis pipeline. Bioinformatics. 2012, 28(3):416-418.

25. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al: Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". PNAS (US). 2005, 102(39):13950-13955.

26. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. Nucl Acids Res. 2002, 30(7):1575-1584.

27. Crooks GE, Hon G, Chandonia J-M, Brenner SE: WebLogo: a sequence logo generator. Genome Res. 2004, 14(6):1188-1190.

28. Bottacini F, O'Connell Motherway M, Casey E, McDonnell B, Mahony J, Ventura M, van Sinderen D: Discovery of a Conjugative Megaplasmid in *Bifidobacterium breve*. Appl Environ Microbiol. 2015, 81(1):166-176.

29. Fernández E, Alegría Á, Delgado S, Martín MC, Mayo B: Comparative Phenotypic and Molecular Genetic Profiling of Wild *Lactococcus lactis* subsp. *lactis* Strains of the *L. lactis* subsp. *lactis* and *L. lactis* subsp. *cremoris* Genotypes, Isolated from Starter-Free Cheeses Made of Raw Milk. Appl Environ Microbiol. 2011, 77(15):5324-5335.

30. Leenhouts KJ, Tolner B, Bron S, Kok J, Venema G, Seegers JF: Nucleotide sequence and characterization of the broad-host-range lactococcal plasmid pWVO1. Plasmid. 1991, 26(1):55-66.

31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990, 215.

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

33. Machielsen R, Siezen RJ, van Hijum SA, van Hylckama Vlieg JE: Molecular description and industrial potential of Tn6098 conjugative transfer conferring alpha-galactoside metabolism in *Lactococcus lactis*. Appl Environ Microbiol. 2011, 77(2):555-563.

34. Nicolas P, Bessières P, Ehrlich SD, Maguin E, Van De Guchte M: Extensive horizontal transfer of core genome genes between two *Lactobacillus* species found in the gastrointestinal tract. BMC Evol Biol. 2007, 7(1):1.

35. Alkema W, Boekhorst J, Wels M, van Hijum SAFT: Microbial bioinformatics for food safety and production. Brief Bioinform. 2016, 17(2):283-292.

36. Kao SM, Olmsted SB, Viksnins AS, Gallo JC, Dunny GM: Molecular and genetic analysis of a region of plasmid pCF10 containing positive control genes and structural genes encoding surface proteins involved in pheromone-inducible conjugation in *Enterococcus faecalis*. J Bacteriol. 1991, 173.

37. Terzaghi BE, Sandine WE: Improved medium for lactic streptococci. Curr Microbiol. 1975, 7.

38. Beresford TP, Fitzsimons NA, Brennan NL, Cogan TM: Recent advances in cheese microbiology. Int Dairy J. 2001, 11(4–7):259-274.

39. Górecki RK, Koryszewska-Bagińska A, Gołębiewski M, Żylińska J, Grynberg M, Bardowski JK: Adaptative Potential of the *Lactococcus lactis* IL594 Strain Encoded in Its 7 Plasmids. PLOS ONE. 2011, 6(7):e22238.

40. Cords BR, McKay LL, Guerry P: Extrachromosomal elements in group N streptococci. J Bacteriol. 1974, 117(3):1149-1152.

41. Solopova A, Bachmann H, Teusink B, Kok J, Neves AR, Kuipers OP: A Specific Mutation in the Promoter Region of the Silent *cel* Cluster Accounts

for the Appearance of Lactose-Utilizing *Lactococcus lactis* MG1363. Appl Envir Microbiol. 2012, 78(16):5612-5621.

42. Bissett DL, Anderson RL: Lactose and d-Galactose Metabolism in Group N Streptococci: Presence of Enzymes for Both the d-Galactose 1-Phosphate and d-Tagatose 6-Phosphate Pathways. J Bacteriol. 1974, 117(1):318.

43. O'Driscoll J, Glynn F, Fitzgerald GF, Sinderen Dv: Sequence Analysis of the lactococcal Plasmid pNP40: a Mobile Replicon for Coping with Environmental Hazards. J Bacteriol.2006, 188(18):6629-6639.

44. Górecki RK, Koryszewska-Bagińska A, Gołębiewski M, Żylińska J, Grynberg M, Bardowski JK: Adaptative Potential of the *Lactococcus lactis* IL594 Strain Encoded in Its 7 Plasmids. PLoS ONE. 2011, 6(7):e22238.

45. Sharma VK, Johnston JL, Morton TM, Archer GL: Transcriptional regulation by TrsN of conjugative transfer genes on staphylococcal plasmid pGO1. J Bacteriol.1994, 176(12):3445-3454.

46. Harrington A, Hill C: Construction of a bacteriophage-resistant derivative of *Lactococcus lactis* subsp. *lactis* 425A by using the conjugal plasmid pNP40. Appl Envir Microbiol. 1991, 57(12):3405-3409.

47. Fallico V, Ross R, Fitzgerald G, McAuliffe O: Novel conjugative plasmids from the natural isolate *Lactococcus lactis* subspecies *cremoris* DPC3758: a repository of genes for the potential improvement of dairy starters. J Dairy Sci. 2012, 95(7):3593-3608.

48. Coakley M, Fitzgerald G, Ros R: Application and evaluation of the phage resistance-and bacteriocin-encoding plasmid pMRC01 for the improvement of dairy starter cultures. Appl Envir Microbiol. 1997, 63(4):1434-1440.

49. Millen AM, Horvath P, Boyaval P, Romero DA: Mobile CRISPR/Cas-Mediated Bacteriophage Resistance in *Lactococcus lactis*. PLoS ONE. 2012, 7(12):e51663.

50. von Ossowski I, Reunanen J, Satokari R, Vesterlund S, Kankainen M, Huhtinen H, Tynkkynen S, Salminen S, de Vos WM, Palva A: Mucosal Adhesion Properties of the Probiotic *Lactobacillus rhamnosus* GG SpaCBA and SpaFED Pilin Subunits. Appl Envir Microbiol. 2010, 76(7):2049-2057.

51. Le DT, Tran TL, Duviau MP, Meyrand M, Guerardel Y, Castelain M, Loubiere P, Chapot-Chartier MP, Dague E, Mercier-Bonin M: Unraveling the role of surface mucus-binding protein and pili in muco-adhesion of *Lactococcus lactis*. PLoS One. 2013, 8(11):e79850.

52. Kojic M, Jovcic B, Strahinic I, Begovic J, Lozo J, Veljovic K, Topisirovic L: Cloning and expression of a novel lactococcal aggregation factor from *Lactococcus lactis* subsp. *lactis* BGKP1. BMC Microbiol. 2011, 11(1):265.

53. Lukić J, Strahinić I, Jovčić B, Filipić B, Topisirović L, Kojić M, Begović J: Different Roles for Lactococcal Aggregation Factor and Mucin Binding Protein in Adhesion to Gastrointestinal Mucosa. Appl Envir Microbiol. 2012, 78(22):7993-8000.

54. Lebeer S, Claes I, Tytgat HLP, Verhoeven TLA, Marien E, von Ossowski I, Reunanen J, Palva A, de Vos WM, De Keersmaecker SCJ et al: Functional Analysis of *Lactobacillus rhamnosus* GG Pili in Relation to Adhesion and Immunomodulatory Interactions with Intestinal Epithelial Cells. Appl Envir Microbiol. 2012, 78(1):185-193.

55. Kranenburg Rv, Marugg JD, Van Swam II, Willem NJ, De Vos WM: Molecular characterization of the plasmid-encoded eps gene cluster essential

for exopolysaccharide biosynthesis in *Lactococcus lactis*. Mol Microbiol. 1997, 24(2):387-397.

56. Vedamuthu ER, Neville JM: Involvement of a plasmid in production of ropiness (mucoidness) in milk cultures by *Streptococcus cremoris* MS. Appl Envir Microbiol. 1986, 51(4):677-682.

57. von Wright A, Tynkkynen S: Construction of *Streptococcus lactis* subsp. *lactis* strains with a single plasmid associated with mucoid phenotype. Appl Envir Microbiol. 1987, 53(6):1385-1386.

58. Neve H, Geis A, Teuber M: Plasmid-encoded functions of ropy lactic acid streptococcal strains from Scandinavian fermented milk. Biochimie. 1988, 70(3):437-442.

59. Kleerebezem M, van Kranenburg R, Tuinier R, Boels IC, Zoon P, Looijesteijn E, Hugenholtz J, de Vos WM: Exopolysaccharides produced by *Lactococcus lactis*: from genetic engineering to improved rheological properties? In: Lactic Acid Bacteria: Genetics, Metabolism and Applications. Springer; 1999: 357-365.

60. Ainsworth S, Sadovskaya I, Vinogradov E, Courtin P, Guerardel Y, Mahony J, Grard T, Cambillau C, Chapot-Chartier M-P, Van Sinderen D: Differences in lactococcal cell wall polysaccharide structure are major determining factors in bacteriophage sensitivity. MBio. 2014, 5(3):e00880-00814.

61. Mahony J, Kot W, Murphy J, Ainsworth S, Neve H, Hansen LH, Heller KJ, Sørensen SJ, Hammer K, Cambillau C: Investigation of the relationship between lactococcal host cell wall polysaccharide genotype and 936 phage receptor binding protein phylogeny. Appl Envir Microbiol. 2013, 79(14):4385-4392.

62. Mahony J, Randazzo W, Neve H, Settanni L, van Sinderen D: Lactococcal 949 group phages recognize a carbohydrate receptor on the host cell surface. Appl Envir Microbiol. 2015, 81(10):3299-3305.

63. Dobson A, Cotter PD, Ross RP, Hill C: Bacteriocin production: a probiotic trait? Appl Envir Microbiol. 2012, 78(1):1-6.

64. Geis A, Singh J, Teuber M: Potential of lactic *streptococci* to produce bacteriocin. Appl Envir Microbiol. 1983, 45(1):205-211.

65. Valenzuela AS, ben Omar N, Abriouel H, López RL, Veljovic K, Caňamero MM, Kojic M, Topisirovic L, Gálvez A: Virulence factors, antibiotic resistance, and bacteriocins in enterococci from artisan foods of animal origin. Food Control 2009, 20.

66. Horvath P, Barrangou R: CRISPR/Cas, the immune system of bacteria and archaea. Science 2010, 327(5962):167-170.

67. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SK, Dryden DTF, Dybvig K et al: A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Research 2003, 31(7):1805-1812.

68. Mahony J, McGrath S, Fitzgerald GF, van Sinderen D: Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. Appl Envir Microbiol. 2008, 74(20):6206-6215.

69. McGrath S, Fitzgerald GF, Sinderen Dv: Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Molecular microbiology 2002, 43(2):509-520.

70. Chopin M-C, Chopin A, Bidnenko E: Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol. 2005, 8(4):473-479.

71. Garvey P, Fitzgerald GF, Hill C: Cloning and DNA sequence analysis of two abortive infection phage resistance determinants from the lactococcal plasmid pNP40. Appl Envir Microbiol. 1995, 61(12):4321-4328.

72. Twomey DP, De Urraza PJ, McKay LL, O'Sullivan DJ: Characterization of AbiR, a Novel Multicomponent Abortive Infection Mechanism Encoded by Plasmid pKR223 of *Lactococcus lactis* subsp. *lactis* KR2. Appl Envir Microbiol. 2000, 66(6):2647-2651.

73. Bouchard JD, Dion E, Bissonnette F, Moineau S: Characterization of the Two-Component Abortive Phage Infection Mechanism AbiT from *Lactococcus lactis*. J Bacteriol.2002, 184(22):6325-6332.

74. Li Y, Canchaya C, Fang F, Raftis E, Ryan KA, van Pijkeren J-P, van Sinderen D, O'Toole PW: Distribution of Megaplasmids in *Lactobacillus salivarius* and Other Lactobacilli. J Bacteriol.2007, 189(17):6128-6139.

75. Roussel Y, Colmin C, Simonet JM, Decaris B: Strain characterization, genome size and plasmid content in the *Lactobacillus acidophilus* group (Hansen and Mocquot). J Appl Bacteriol. 1993, 74(5):549-556.

76. Muriana PM, Klaenhammer TR: Conjugal Transfer of Plasmid-Encoded Determinants for Bacteriocin Production and Immunity in *Lactobacillus acidophilus* 88. Appl Envir Microbiol. 1987, 53(3):553-560.

77. Fang F, Flynn S, Li Y, Claesson MJ, van Pijkeren J-P, Collins JK, van Sinderen D, O'Toole PW: Characterization of endogenous plasmids from *Lactobacillus salivarius* UCC118. Appl Envir Microbiol. 2008, 74(10):3216-3228.

78. Bermúdez-Humarán LG, Aubry C, Motta J-P, Deraison C, Steidler L, Vergnolle N, Chatel J-M, Langella P: Engineering lactococci and lactobacilli for human health. Curr Opin Microbiol. 2013, 16(3):278-283.

79. Bermúdez-Humarán LG: *Lactococcus lactis* as a live vector for mucosal delivery of therapeutic proteins. Hum Vaccines. 2009, 5(4):264-267.

Chapter V

Base modification analysis of *Lactococcus lactis* strains and their corresponding restriction-modification systems

Note: REBASE analysis and assignment of methylation motifs and enzyme nomenclature was performed by Dr Richard J. Roberts of New England Biolabs, USA.

Chapter V contents

5 Abstract 235

5.1 Introduction 236

5.2 Materials & Methods 239

5.2.1 Strain growth conditions and media 239

5.2.2 Sequencing 239

5.2.3 Base modification analysis 239

5.2.4 Comparative genomics 241

5.2.5 DNA restrictions 241

5.2.6 Quantitative polymerase chain reaction, qPCR 244

5.3 Results 244

5.3.1 Motif analysis 244

5.3.2 Analysis of encoded MTases and REases 247

5.3.3 Type I R-M systems 250

5.3.4 Type II R-M systems 254

5.3.5 Type III/IV R-M systems 259

5.3.6 Type I R-M Shufflon system 261

5.4 Discussion 267

5.5 References 269

5 Abstract

SMRT sequencing is the first and so far only sequencing technology to allow concomitant detection of base modifications with primary sequence analysis. In the present study, this technology was applied to determine the methylome of sixteen *Lactococcus lactis* strains, which revealed fifty two methylation motifs consisting of N^6-methyladenine and N^4-methylcytosine base modifications. Five of these motifs were validated as they prevented site-specific cleavage by commercially available restriction enzymes. The sixteen strains were predicted to encode a number of unique Type I, II, III and IV restriction-modification (R-M) systems, including a novel Type I R-M shufflon system, represented by multiple *hsdS* subunits arranged around a recombinase gene. The presumed genetic rearrangement activity of this system was corroborated by the presence of different *hsdS* subunit combinations in the raw sequence data and their subsequent confirmation within a heterogeneous population via qPCR.

5.1 Introduction

Methylation of prokaryotic genomes by DNA methyltransferases (MTases) plays an important role in expanding the functionality of the four DNA bases [1]. MTases encoded by prokaryotes (and the base modifications they confer) are known to be involved in a variety of processes, such as cell cycle regulation, DNA repair and pathogenesis. MTases may also be involved in preventing invasion of foreign DNA, in which case the MTase is linked to a cognate restriction endonuclease (REase) activity to form a restriction-modification (R-M) system (where the MTase methylates 'self' DNA and the REase targets the invading, unmethylated DNA) [2-5]. Three main MTase classes are identified in prokaryotes which function by methyl transfer from S-adenosyl-L-methionine (SAM) to a target nucleotide base [1]. Class I and II MTases target exocyclic nitrogens at position *N6* in adenine and position *N4* in cytosine, to form N^6-methyladenine (^{6m}A) and N^4-methylcytosine (^{4m}C) modifications, respectively, while class III MTases target cytosine at position C^5 to form C^5-methylcytosine (^{5m}C) [6].

R-M systems are generally classified in Types I-IV based on sub-unit composition, ATP (GTP) requirements and cleavage mechanisms [6]. Type I R-M systems are multi-subunit proteins that function as a single protein complex, usually composed of one or two REase subunits (HsdR), one or two MTase subunits (HsdM) and one specificity (S) subunit (HsdS) [6]. Type I R-M systems recognize long, non-palindromic motifs, typically composed of two components, the first of 3 or 4 bp and the second of 4 or 5 bp (the sequence of each specified by particular HsdS domains), separated by a non-specific spacer of 6 to 8 bp [7]. Type II R-M systems are composed of separate REase and MTase activities. Type II REases act as homodimers to target specific DNA sequences, usually represented by short (4-8 bp)

palindromic sequences, cleave at a specific position within the recognition sequence, and act independently of their cognate MTase [6]. Type II MTases act as monomers and transfer a methyl group from the donor SAM directly to double-stranded DNA forming ^{4m}C, ^{5m}C or ^{6m}A modifications. Type II R-M systems are among the most thoroughly studied due to their importance in molecular biology [8, 9]. Type III R-M systems are composed of two subunits that function either in DNA recognition and modification (Mod) or restriction (Res) [10]. These systems target a non-palindromic recognition sequence, present on both strands in inverse orientation, and cut at a defined location (25 -27 bp) downstream of the associated recognition site [11]. Type III systems require ATP hydrolysis to function [6] and are frequently found in prokaryotic genomes [8, 11]. Type IV R-M systems are those which, unlike Types I-III, only target methylated DNA. Type IV systems are composed of two genes and their target motifs are not well defined [6].

The development of single molecule real time (SMRT) sequencing by Pacific Biosciences has, for the first time, allowed the detection of DNA base modifications concomitantly with primary sequence analysis [12]. SMRT technology utilises a single polymerase molecule bound to a zero-mode waveguide (ZMW) nanostructure to incorporate fluorescently labelled nucleotides complementary to a DNA template strand [13, 14]. The incorporation of a nucleotide generates a specific fluorescent signal called a ‘pulse’ [15]. Distinct variations in pulse width (PW), which reflects the length of time the polymerase is bound to a particular base, and interpulse duration (IPD), representing the time it takes for the polymerase to move from one base to the next, are observed when the polymerase encounters a modified base in the DNA template. This signature allows SMRT sequencing to differentiate between unmodified bases and those with ^{6m}A, ^{4m}C or ^{5m}C base modifications, allowing for

the determination of specific methylation motifs, which may be paired to specific R-M systems [16]. However, the m5C kinetic signature is difficult to detect accurately and accurate detection of such modifications requires treatment of template DNA with Tet1 enzyme prior to sequencing [17].

In the current study, methylome analysis was performed on sixteen lactococcal strains sequenced utilising the SMRT approach. Comparative analysis of their predicted R-M systems was used in conjunction with the generated SMRT data in order to identify active R-M systems and resolve their target methylation motifs.

5.2 Methods

5.2.1 Strain growth conditions and media

Bacterial strains used in this study are detailed in Table 5.1. *L. lactis* strains were routinely cultured at 30 °C in M17 broth (Oxoid) supplemented with 0.5 % glucose/lactose without agitation.

5.2.2 Sequencing

In total, 16 *Lactococcus lactis* strains were sequenced (Table 5.1) and employed here for the purpose of methylome analysis. Sequencing was performed utilising the SMRT sequencing approach on a Pacific Biosciences RS II sequencing platform (executed by GATC Biotech Ltd., Germany). *De novo* assemblies were performed using the Pacific Biosciences SMRT Portal analysis platform (version 2.3.1), utilizing the RS_HGAP_Assembly.2 protocol.

5.2.3 Base modification analysis

Identification of DNA base modifications was performed by means of SMRT sequencing, utilising the RS_Modification_and_Motif_Analysis.1 protocol and the finished genome assemblies as reference files. The identified methylation motifs were refined based on three criteria: (i) a mean modification QV cut-off of 40 %, equivalent to a P-value of <0.0005 was applied; (ii) secondly motifs of unknown type were removed; (iii) motifs methylated at less than 50 % of possible positions were removed.

Table 5.1: Strains used in this study from the UCC strain collection

Strain	Accession Number	Origin	Sequencing technology	Average coverage
subsp. *lactis*				
L. lactis 184	CP015895	Dairy product	PacBio SMRT	72.56
L. lactis 229	CP015896	Dairy product	PacBio SMRT	107.27
L. lactis 275	CP015897	Dairy product	PacBio SMRT	60.88
L. lactis UC06	CP015902	Dairy product	PacBio SMRT	66.25
L. lactis UC08	CP015903	Fermented meat product	PacBio SMRT	159.42
L. lactis UC11	CP015904	Fermented meat product	PacBio SMRT	113.79
L. lactis UC063	CP015905	Dairy product	PacBio SMRT	95.48
L. lactis UC77	CP015906	Dairy product	PacBio SMRT	97.73
L. lactis UL8	CP015908	Dairy product	PacBio SMRT	45.42
L. lactis C10	CP015898	Dairy product	PacBio SMRT	81.29
subsp. *cremoris*				
L. lactis 158	CP015894	Dairy product	PacBio SMRT	113.98
L. lactis UC109	CP015907	Dairy product	PacBio SMRT	134.74
L. lactis JM1	CP015899	Dairy product	PacBio SMRT	49.26
L. lactis JM2	CP015900	Dairy product	PacBio SMRT	99.08
L. lactis JM3	CP015901	Dairy product	PacBio SMRT	72.01
L. lactis JM4	CP015909	Dairy product	PacBio SMRT	206.24

5.2.4 Comparative genomics

ORFs encoding putative MTases and REases were identified by homology-based BLASTP v2.2.26 [18] analysis against the non-redundant protein databases curated by the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nig.gov/) and REBASE [8]. The Artemis (v16) genome browser and annotation tool was used to inspect and (where necessary) manually curate ORFs (http://www.sanger.ac.uk/science/tools/artemis). ORF annotations were refined where necessary using alternative databases; Pfam [19], HHpred [20] and Uniprot/EMBL (http://www.uniprot.org/). All sequence comparisons at protein level were performed via all-against-all, bi-directional BLAST alignments [18]. Alignment cut-off was: E-value <0.0001, with >30 % amino acid identity across 80 % of the sequence length. For analysis and clustering of results, the Markov Clustering algorithm (MCL) was implemented in the mclblastline pipeline v12-0678 [21]. TM4 MeV, MultiExperiment Viewer v4.9 was used to view MCL clustering data and conduct hierarchal clustering (http://www.tm4.org/mev.html).

5.2.5 DNA restrictions

Chromosomal DNA from *L. lactis* strains was isolated as previously described [22]. DNA restrictions were performed on genomic DNA in a 50 µl reaction volume, containing: 1 µl restriction enzyme, 5 µl reaction buffer, 10 µl DNA and 34 µl dH_2O. Restriction enzymes were sourced from New England Biolabs, USA (BmtI, NsiI, SfaNI and ScrFI) and Roche, USA (DpnI). Restrictions were performed at 37 °C for 15 mins (NsiI), 60 mins (BmtI, SfaNI and ScrFI) and 3 hours (DpnI) according to manufacturer's instructions. Electrophoresis of DNA was conducted at 100 V for 30 mins on a 1 % agarose gel.

5.2.6 Quantitative polymerase chain reaction, qPCR

Detection and quantification of *hsdS* domain configuration was performed via qPCR on a LightCycler 480 qPCR instrument (Roche Life Science) utilising LightCycler 480 SYBR Green I Master mix (Roche Life Science). DNA samples were prepared by phenol-chloroform extraction as described previously [22] and the quantity was estimated on a Nanodrop 2000 (Thermo Scientific), with the final DNA concentration adjusted to 10 ng/µl for each sample. Serial dilution of standard DNA was used to prepare a standard curve. Primers used are described in Table 5.2 and were synthesized by Eurofins MWG Operon (Germany).

PCR reaction mixtures contained: 3 µl of ultrapure water, 2 µl 10X primers, 10 µl 2X master mix and 5 µl of DNA template (template DNA was replaced by dH_20 for negative controls). Quantitative PCR reactions were carried out with a 5 min pre-incubation at 95 °C followed by 45 cycles of denaturation at 95 °C for 10 s, annealing/extension at 50 °C for 10 s/ 72 °C for 10 s. All samples were tested at least in triplicate. Absolute quantification analysis was used to calculate the crossing point (Cp, the point at which the fluorescence of a sample rises above the background fluorescence) for each sample in the analysis using the Fit points analysis method in LightCycler 480 qPCR software (Roche Life Science).

Table 5.2: qPCR primers for pMPJM2 shufflon used in this study

Oligo name	Sequence (5' -3')	Target/Comment
*hsdS*1A_1B_F	TCATGCAGTATCAGATTCCAGA	Targets *hsdS* sub-unit combination 1A – 1B
*hsdS*1A_1B_R	GCAAGAGATCAAACTGAGCATC	
*hsdS*1A_2B_F	CCGCGTGGAGATAAATCAG	Targets *hsdS* sub-unit combination 1A – 2B
*hsdS*1A_2B_R	GCCAATCATTTGGCATAACA	
*hsdS*2A_1B_F	AAAGGGTTCAACTTGATGTGC	Targets *hsdS* sub-unit combination 2A – 1B
*hsdS*2A_1B_R	GCAAGAGATCAAACTGAGCATC	
*hsdS*2A_2B_F	AGGGTTCAACTTGATGTGCTT	Targets *hsdS* sub-unit combination 2A – 2B
*hsdS*2A_2B_R	GGCATAACACCATCATAGGG	
res1A_2B_F	GCCTTAGATGATAGAATTGCTGAA	Targets reservoir combination 1A – 2B
res1A_2B_R	AATACCAGTTAAGTTTGATAATTGCC	
res1A_1B_F	CCACTTGAGGATCAACGAAC	Targets reservoir combination 1A – 1B
res1A_1B_R	ATGCTATTGCCAAAGCTAATGT	
res2A_1B_F	TCCATCGTTGGAAGAACAGA	Targets reservoir combination 2A – 1B
res2A_1B_R	ATGCTATTGCCAAAGCTAATGT	
res2A_2B_F	AGGGTTCAACTTGATGTGCTT	Targets reservoir combination 2A – 2B
res2A_2B_R	GGCATAACACCATCATAGGG	

5.3 Results

5.3.1 Motif analysis

SMRT sequencing technology was applied here to determine the DNA modifications of sixteen lactococcal genomes (described in detail in Chapter III) to assess their encoded MTases, with particular emphasis on MTases linked to cognate REases to form functional R-M systems. In total 51 ^{6m}A type and 1 ^{4m}C type methylation motifs were detected (Table 5.3). Initial analysis of detected methylation motifs identified isoschizomers of four motifs, namely; 5′-ATGC^{6m}AT-3′, 5′-GCTAG^{6m}C-3′, 5′-GC^{6m}ATC-3′ and 5′-G^{6m}ATC-3′. Analysis of these motifs indicated that they represent Type II R-M motifs based on their short (4-8 bp) palindromic recognition sites. Type I methylation motifs were the most frequently encountered (34/52 detected motifs), indicating a high level of diversity of Type I systems in *L. lactis* (Table 5.3).

Table 5.3: Methylated motifs detected in strains sequenced by Pacific Biosciences SMRT sequencing

Strain	Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
158	*TAAANNNNNNTTYG	3	m6A	100.00%	643	643	95.53	60.4	CRAANNNNNNTTTA
184	VTACNNNNNGGT	3	m6A	97.05%	263	271	65.19	39.61	ACCNNNNNGTAB
	GGCTNA	6	m6A	96.24%	3429	3563	64.67	37.71	
	ACCNNNNNGGT	1	m6A	95.33%	429	450	59.85	35.4	ACCNNNNNGGT
	TTAMNNNNNGGT	3	m6A	94.59%	630	666	62.97	37.89	ACCNNNNNKTAA
	GGAGA	5	m6A	94.57%	3222	3407	64.11	38.13	
229	*GATGNNNNNTTTA	2	m6A	86.51%	218	252	54.44	30.7	TAAANNNNNCATC
	GAYNNNNNTTTA	2	m6A	81.06%	1211	1494	50.97	30.69	TAAANNNNNRTC
	TAAANNNNNRTC	3	m6A	70.15%	1048	1494	46.15	31.62	GAYNNNNNTTTA
	TAAANNNNNNTTYG	3	m6A	79.33%	572	721	49.56	30.62	CRAANNNNNNTTTA
JM1	GAGNNNNNTGA	2	m6A	99.84%	1227	1229	92.21	55.9	TCANNNNNCTC
	*AGCYAC	5	m6A	99.77%	1768	1772	91.49	57.92	
	*ATGCAT	5	m6A	98.91%	635	642	93.83	58.88	ATGCAT
	*CCAAT	4	m6A	98.86%	8203	8298	87.63	54.51	
	*GAAYNDNNNNTARC	3	m6A	18.87%	50	265	53.48	57.54	
	GYTANNNNNDRTTC	4	m6A	21.68%	49	226	52.06	60.04	
JM2	CCANNNNNGTC	3	m6A	99.37%	629	633	82.74	50.71	GACNNNNNTGG
	AGYNNNNNCGT	1	m6A	99.19%	853	860	84.9	49.41	ACGNNNNNRCT
	TCACNNNNNNATGA	3	m6A	98.84%	85	86	83.96	54.92	
	TCAYNNNNNNATGB	3	m6A	98.53%	401	407	82.08	52.22	
	ACANNNNNNRTAA	3	m6A	98.40%	981	997	81.8	50.2	TTAYNNNNNNTGT
	*AGAAG	4	m6A	98.10%	8814	8985	69.59	49.33	
	*CATNNNNNNRTGA	2	m6A	97.68%	632	647	81.67	52.47	
JM3	GMAGG	3	m6A	97.22%	5939	6109	68.53	38	
	GRTAAAT	6	m6A	94.59%	1817	1921	62.81	36.81	
JM4	GRTANAG	6	m6A	92.18%	2498	2710	112.98	69.55	
	AGAAGC	4	m6A	91.10%	1862	2044	110.18	65.8	
	*YTCANNNNNNRTTA	4	m6A	90.59%	549	606	106.02	71.64	TAAYNNNNNNTGAR

	*GCTAGC	6	m4C	86.20%	281	326	65.68	66.14	GCTAGC
	*TTAANNNNNNVTTG	3	m6A	85.98%	423	492	101.67	70.32	CAABNNNNNNTTAA
UC06	CAGNNNNNNTAYC	2	m6A	94.50%	584	618	62.08	33.11	GRTANNNNNNCTG
	CACNNNNNNTTYG	2	m6A	92.43%	476	515	58.44	33.93	CRAANNNNNNGTG
	*GATC	2	m6A	92.00%	4510	4902	61.28	34.19	GATC
	ACTNNNNNNTYTC	1	m6A	90.42%	774	856	58.43	33.75	GARANNNNNNAGT
	*GCDGCAGC	2	m4C	31.70%	71	224	40.24	37.03	Actual motif = GCNGC
UC08	CNACNNNNNNNTGG	3	m6A	90.30%	549	608	56.53	32.26	CCANNNNNNNGTNG
	GGANNNNNNNTTCA	3	m6A	86.35%	329	381	58.51	33.24	TGAANNNNNNNTCC
	*GATGC	2	m6A	52.96%	1751	3306	45.07	37.16	GCATC
UC11	*GATGC	2	m6A	99.91%	3280	3283	99.62	59.01	GCATC
	CNACNNNNNNNTGG	3	m6A	99.84%	606	607	88.99	58.66	CCANNNNNNNGTNG
	GGANNNNNNNTTCA	3	m6A	99.48%	381	383	91.03	60.16	TGAANNNNNNNTCC
UC063	GACNNNNNNTTYG	2	m6A	99.26%	675	680	82.21	49.76	CRAANNNNNNGTC
	*YTCANNNNNNRTTC	4	m6A	98.34%	534	543	83.73	49.78	GAAYNNNNNNTGAR
	AGCNNNNNCCT	1	m6A	98.28%	573	583	88.37	49.02	AGGNNNNNGCT
UC77	GATGNNNNNTTTA	2	m6A	99.60%	246	247	95.77	54.94	TAAANNNNNCATC
	TAAANNNNNNTTYG	3	m6A	99.17%	713	719	77.63	50.84	CRAANNNNNNTTTA
	GAYNNNNNTTTA	2	m6A	91.86%	1366	1487	59.07	52.27	TAAANNNNNRTC
UC109	ACCNNNNNNTTAA	1	m6A	100.00%	306	306	96.72	63.2	TTAANNNNNNGGT
	GRTCNAG	6	m6A	99.80%	994	996	98.09	61.38	
	*GAATC	3	m6A	99.51%	5103	5128	95.34	59.91	
	GARANNNNNNNTTTA	4	m6A	99.17%	718	724	95.65	60.9	TAAANNNNNNNTYTC
	GCANNNNNNATTA	3	m6A	98.81%	415	420	97.75	64.97	TAATNNNNNNTGC

* Indicates a motif which has been resolved to its associated restriction modification system

5.3.2 Analysis of encoded MTases and REases

In order to link the identified methylated DNA modifications to methylases of R-M systems, a bioinformatics-based search was undertaken to identify predicted chromosomally- and plasmid-encoded MTases and REases. The plasmid complement of the sixteen strains in this study was found to encode sixty eight proteins involved in DNA methylation and restriction, while the chromosomes were predicted to encode a further seventy four such proteins. Hierarchical clustering utilising all against all bidirectional BLASTP analysis was used to cluster the constituent strains based on the presence or absence of specific encoded MTases, REases and specificity subunits identified above which revealed a high degree of divergence between the constituent strains in the analysis (Fig. 5.1). Encoded MTases, REases and specificity subunits were further categorized into predicted R-M systems based on homology to previously identified systems and their genetic organisation. This resulted in the identification of putative, complete and incomplete Type I (24), Type II (19), Type III (45) and Type IV (1) systems, which will be discussed individually (Table 5.4) (Fig 5.2).

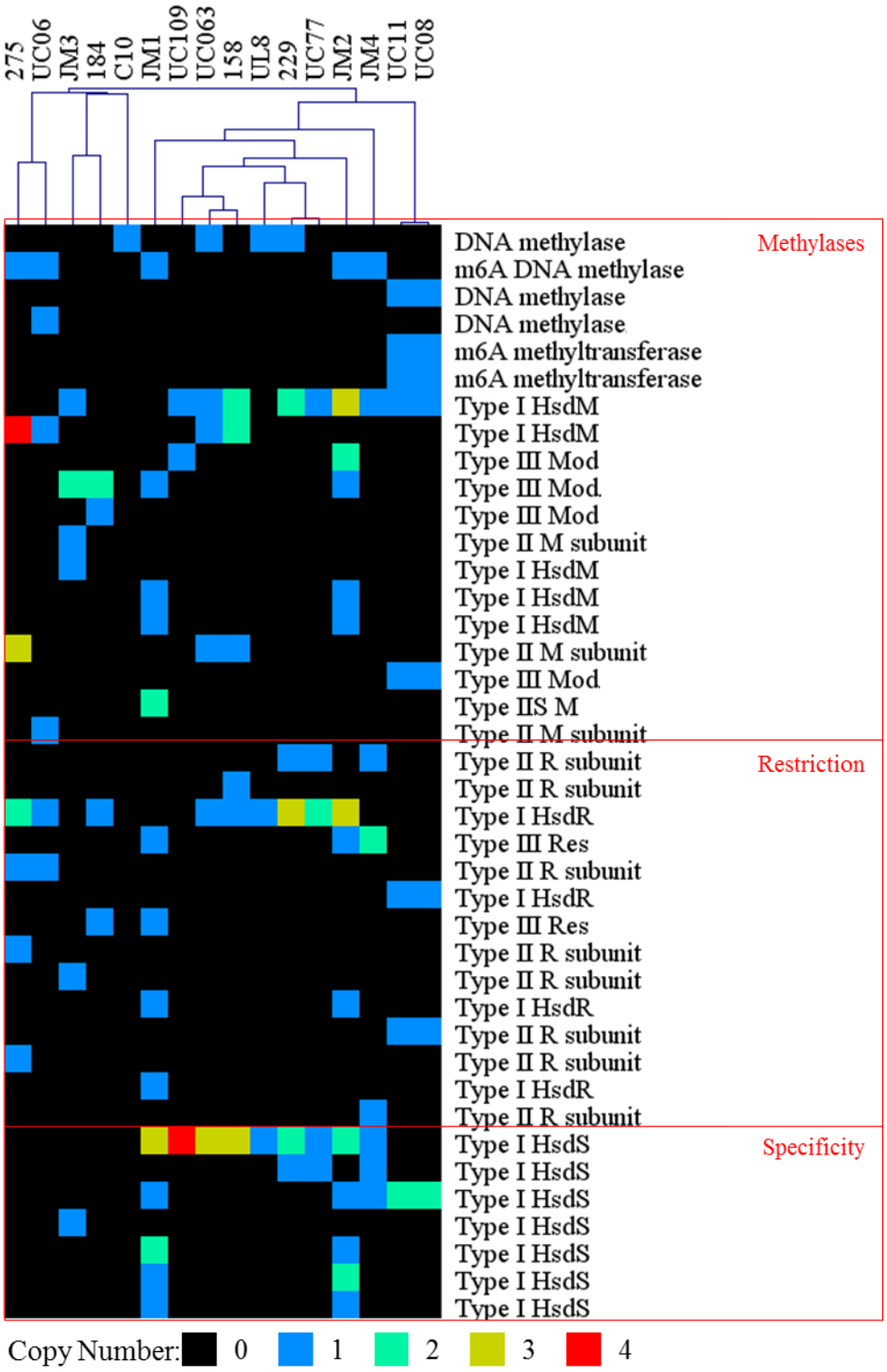


Figure 5.1: MCL analysis of encoded MTases and REases
MCL analysis of DNA MTases, REases and S subunits grouped by hierarchical clustering based on presence/absence of CDS. Colour indicates copy number.

Table 5.4: Overview of predicted chromosomally- and plasmid-encoded R-M Systems

	158	184	229	275	C10	JM1	JM2	JM3	JM4	UC06	UC08	UC11	UC77	UC063	UC109	UL8
Chromosomally encoded																
Type I	2	1	1	-	-	1	1	-	1	-	1	1	1	1	1	-
Type II	1	-	-	1	-	1	-	3	1	4	1	1	-	1	-	-
Type III	-	1	-	-	-	-	1	-	-	-	-	-	-	-	1	-
Type IV	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
Plasmid encoded																
Type I	1	-	2	-	-	2	3	-	-	-	-	-	1	2	1	-
Type II	-	-	-	1	-	1	-	-	1	-	1	1	-	-	-	-
Type III	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
Type IV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

5.3.3 Type I R-M systems

Type I R-M systems are usually comprised of three sub-units: REase (designated HsdR), MTase (designated HsdM) and a specificity determinant (designated HsdS) [7]. Based on their deduced recognition sequence, thirty five of the fifty two identified methylation motifs were assigned to Type I systems (see Table 5.3). BLAST analysis identified fifteen Type I R-M systems which appear to be complete, while also nine additional, orphan specificity subunits were detected (Fig 5.2). However, due to their divergent nature most of these thirty five motifs could not be attributed to specific type I R-M systems. Only eight methylation motifs (Table 5.3) were attributed to their putative specificity subunits based on homology to the target recognition domains (TRD) of previously identified Type I systems.

S.Lla158ORFAP, located on plasmid p158F, is a reasonable candidate for CRAANNNNNNTTTA based on sequence homology between its TRD1 and that of S.Spy743I (CRAANNNNNNNTGC) (REBASE Enz. Num.113363). However, since S.Lla158ORFAP represents a solitary specificity subunit and both alternate subunits in strain 158 are associated with methylase genes, this specificity appears to be the result of a more complicated interaction.

The motifs GAAYNDNNNNTARC and GYTANNNNNDRTTC detected in strain JM1 are most likely complementary strands and the probable motif is GAAYNNNNNNTARC. S.LlaJM1ORFDP located on plasmid pJM1E is a reasonable candidate based on sequence homology of its TRD1 to that of S.Sth9I (GAAYNNNNNNTAYG) (REBASE Enz. Num. 137120). However, as was the case for S.Lla158ORFAP, S.LlaJM1ORFDP represents a solitary specificity subunit and indicates that there may be more complex interactions occurring in this strain.

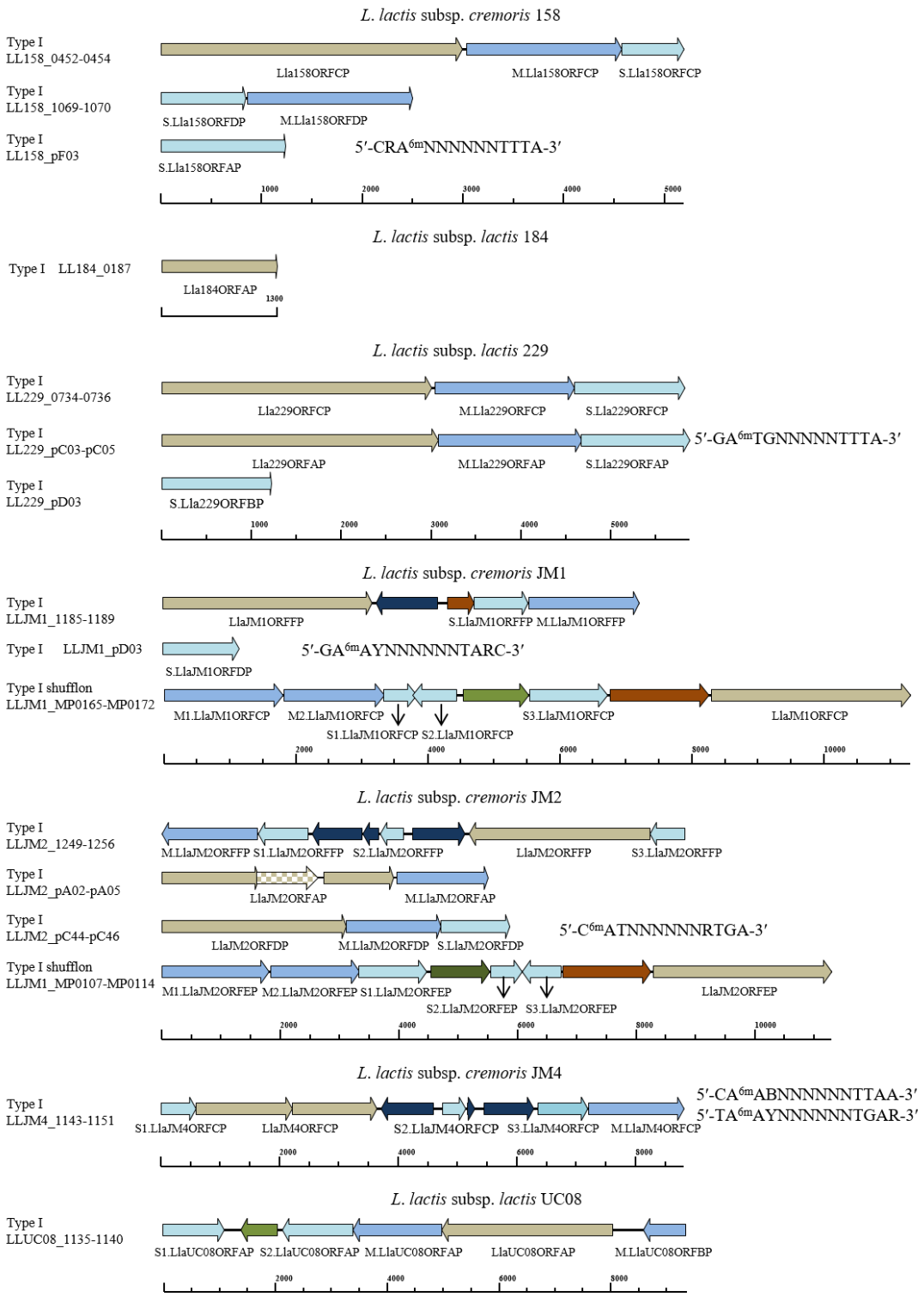


Figure 5.2: Overview of identified Type I R-M systems
Genetic organisation of encoded Type I R-M systems in; *L. lactis* 158, *L. lactis* 184, *L. lactis* 229, *L. lactis* JM1, *L. lactis* JM2, *L. lactis* JM4, *L. lactis* UC08, *L. lactis* UC11, *L. lactis* UC063, *L. lactis* UC77 and *L.* lactis UC109. Methylation motifs are indicated where resolved.

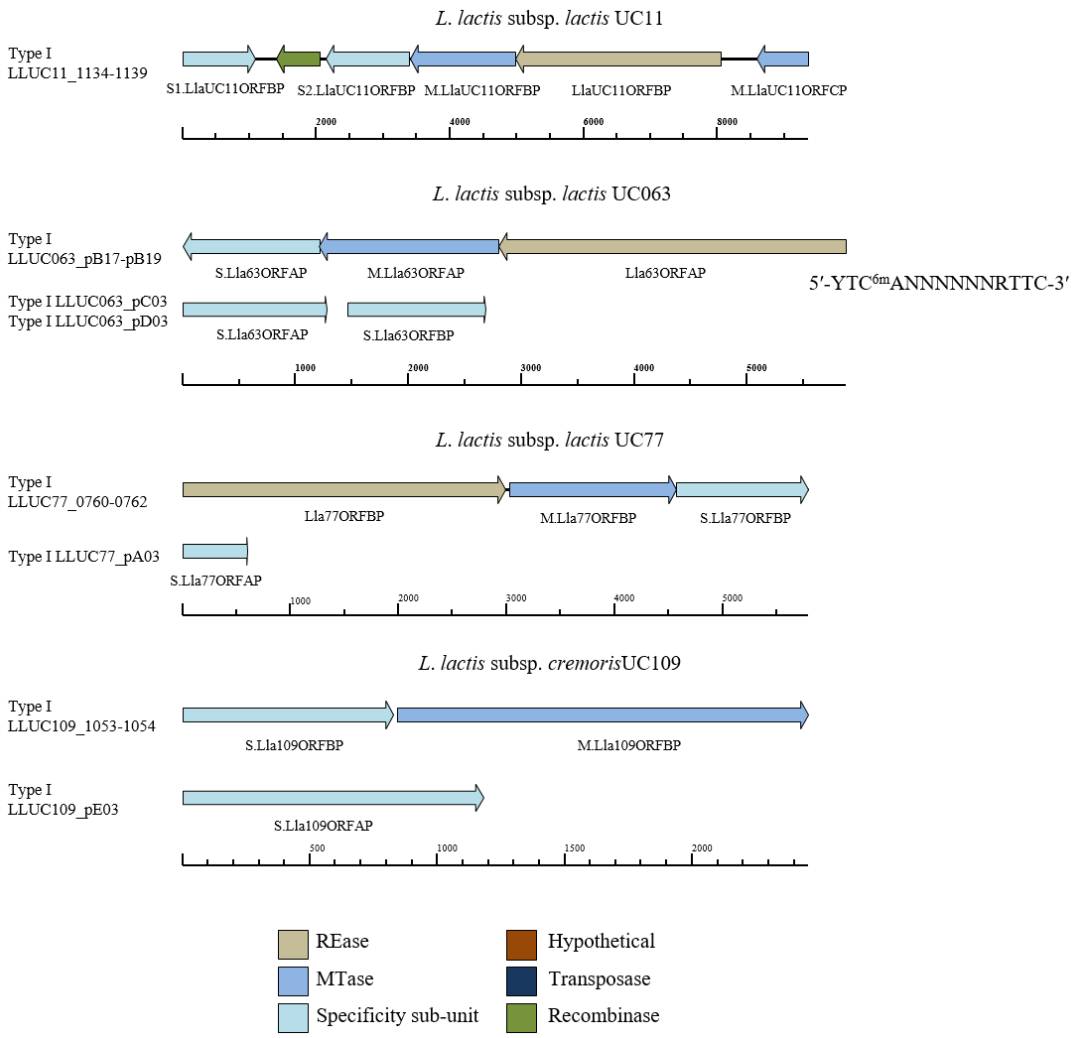


Figure 5.2 continued:
Genetic organisation of encoded Type I R-M systems in; *L. lactis* 158, *L. lactis* 184, *L. lactis* 229, *L. lactis* JM1, *L. lactis* JM2, *L. lactis* JM4, *L. lactis* UC08, *L. lactis* UC11, *L. lactis* UC063, *L. lactis* UC77 and *L.* lactis UC109. Methylation motifs are indicated where resolved.

S.LlaJM2ORFDP located on plasmid pJM2C is the most likely to be responsible for the recognition of motif CATNNNNNNRTGA based on sequence similarity of its TRD2 with that of S.SauSTORF499P (ACCNNNNNRTGA) (REBASE Enz. Num. 23368). LlaJM2ORFDP represents a complete Type I R-M system composed of a *hsdR*, *hsdM* and *hsdS*.

In strain JM4, the two Type I motifs CAABNNNNNNTTAA and TAAYNNNNNNTGAR are presumably the results of odd combinations of the specificity subunits from the ORFC system located on the chromosome. This system is composed of a *hsdR*, *hsdM* and three partial *hsdS*, and appears to be interrupted by two transposon-elements, suggestive of a recombination event.

S.Lla229ORFAP encoded on plasmid p229C in strain 229 is a good candidate for GATGNNNNNTTTA based on sequence homology of its TRD1 to that of S.Awo1030III (GATGNNNNNNTGC) (REBASE Enz. Num. 4579). Lla229ORFAP is one of two complete Type I R-M systems in strain 229 composed of a *hsdR*, *hsdM* and a single complete *hsdS*.

S.Lla63ORFAP is part of a complete R-M system encoded on plasmid pUC063B in strain UC063 and represents a good candidate for the recognition of motif YTCANNNNNNRTTC based on the similarity of its TRD1 to that of S.Bsp3003III (YTCANNNNNNNTCNNC) (REBASE Enz. Num. 70536). However, there appears to be a specificity subunit missing unless there is an unknown interaction occurring with S.Lla63ORFBP on plasmid pUC063D.

The remaining Type I motifs remain unassigned and therefore an increased dataset appears to be necessary to resolve more of these motifs. Furthermore, resolving these target specificities is highly complicated by the genetic make-up of some of the Type I systems, which appear to be subject to genetic rearrangements,

and multiple Type I specificity subunits which may be combining in the cell leading to unusual specificities.

5.3.4 Type II R-M systems

Analysis of the methylation motifs detected in this study resulted in the identification of eight Type II recognition motifs which were associated with their respective Type II R-M systems; LlaJM1IP, LlaJM1ORFAP, LlaJM4I, LlaUC06IP M.LlaUC06ORFDP, LlaUC08IP, LlaUC11IP and Lla158ORFBP. Complementary homology based analysis of the CDS in the respective genomes utilising BLASTP, MCL and REBASE allowed the identification of unique Type II R-M systems proposed to carry out this methylation function (Fig 5.3).

Analysis of the detected R-M systems in strain JM1 indicated that the methylated motif CCAAT was a possible product of M.LlaJM1ORFEP, while the methylated motif AGCYAC was found to be a potential product of M.LlaJM1ORFAP. However, without cloning (and subsequent characterisation) the corresponding genes it is difficult to annotate such functions accurately. In strain JM3 it was found that the motif AAGGAAGWNNNR represents an inaccurate assignment for a simpler sequence, perhaps AGGAAG. However, while it and the other two motifs are the products of the three Type IIG enzymes encoded in JM3, none of the enzymes could be assigned motifs unambiguously. In strain UC06, it was found that the motif GCDGCAGC probably represents an inaccurate assignment for GCNGC, the product of the Type II methylase; M.LlaUC06ORFDP (Fig 5.3).

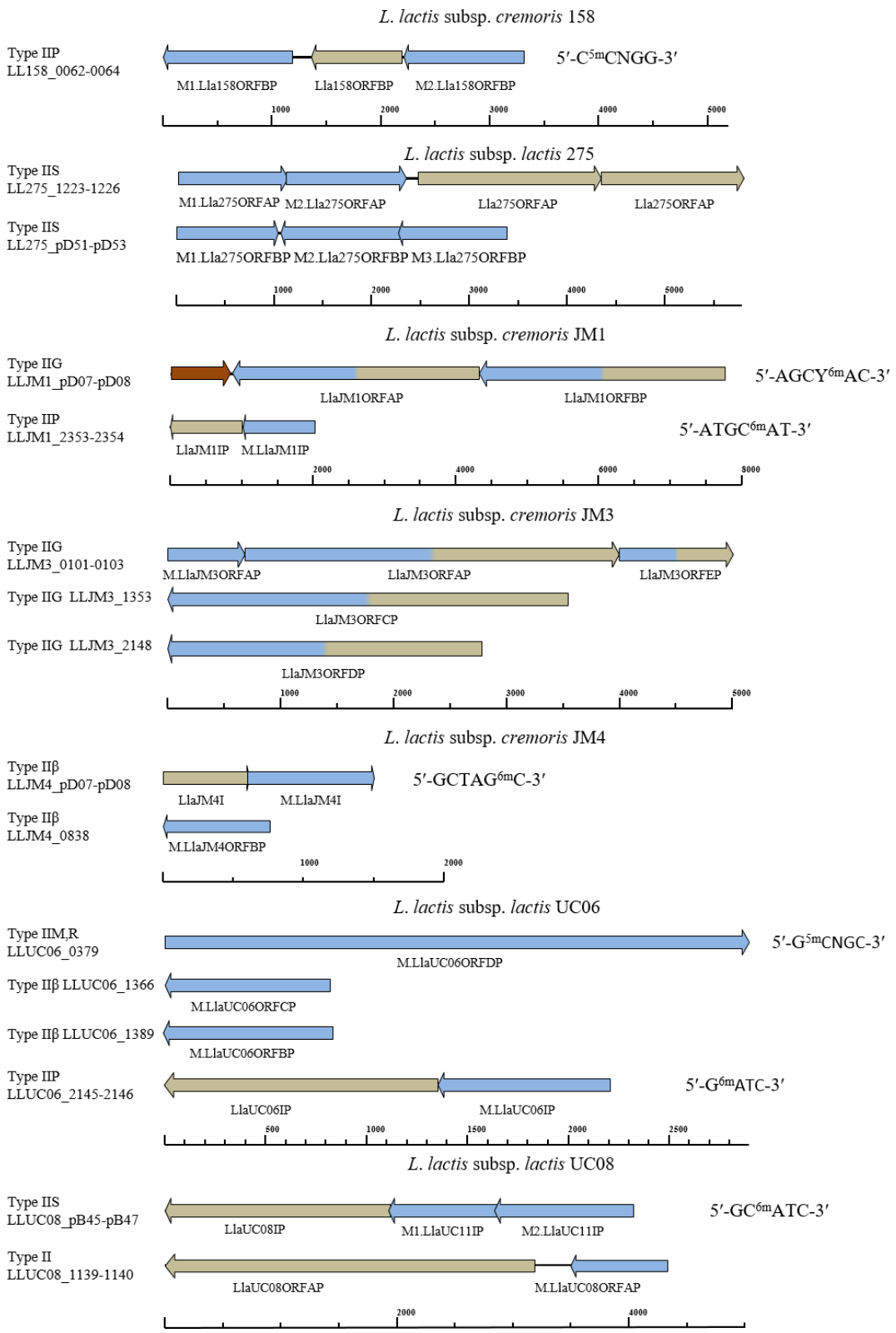


Figure 5.3: Overview of identified Type II R-M systems
Genetic organisation of encoded Type II R-M systems in; *L. lactis* 158, *L. lactis* 275, *L. lactis* JM1, *L. lactis* JM3, *L. lactis* JM4, *L. lactis* UC06, *L. lactis* UC08, *L. lactis* UC11, and *L. lactis* UC063. Methylation motifs are indicated where resolved.

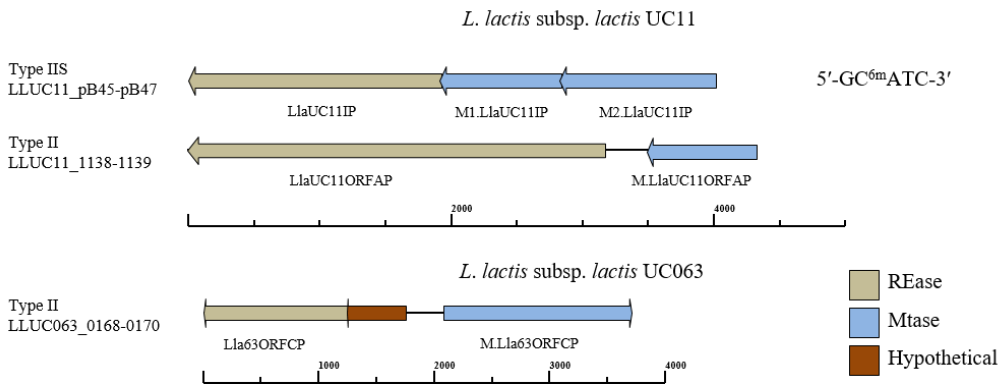


Figure 5.3 continued:
Genetic organisation of encoded Type II R-M systems in; *L. lactis* 158, *L. lactis* 275, *L. lactis* JM1, *L. lactis* JM3, *L. lactis* JM4, *L. lactis* UC06, *L. lactis* UC08, *L. lactis* UC11, and *L. lactis* UC063. Methylation motifs are indicated where resolved.

To assess if certain identified Type II methylated motifs prevent restriction, the total DNA complement of particular strains was treated with a restriction enzyme which targeted its respective base modification motif (Fig. 5.4). Treatment of strain *L. lactis* JM1 with NsiI did not result in DNA fragmentation confirming that the detected methylation motif is correct, while similarly BmtI did not restrict *L. lactis* JM4 (Fig. 5.4A). Both *L. lactis* UC08 and UC11 presented the methylation pattern corresponding to SfaNI, however only UC11 was protected from restriction (Fig. 5.4B). Analysis of these motifs showed that while 99.84 % of these motifs were methylated in UC11, only 53.84 % of detected motifs in UC08 were methylated allowing the remaining motifs (1526 on the leading strand, 1555 on the lagging strand) to be restricted. This would indicate that the *hsdR* subunit in UC08 is not functional, but as the sequence of both systems is well conserved this may also be due to differential expression levels of the encoding plasmids.

The Lla158ORFBP system detected on the genome of *L. lactis* 158 was also found to be active in methylation as treatment with the R.ScrFI enzyme did not result

in restriction (Fig. 5.4B). The Lla158ORFBP target site is an ^{5m}C modification in the recognition motif 5′-C^{5m}CNAGG-3′. This methylation motif was not detected by SMRT sequencing, most likely because such ^{5m}C modifications are difficult to detect using SMRT [17]. Restriction with DpnII was also found to be blocked in the *L. lactis* UC06 genome (Fig. 5.4C), while *L. lactis* UC063 which did not contain any Type II methylation motifs was restricted by all of the enzymes employed in this analysis (Fig. 5.4A,B,C).

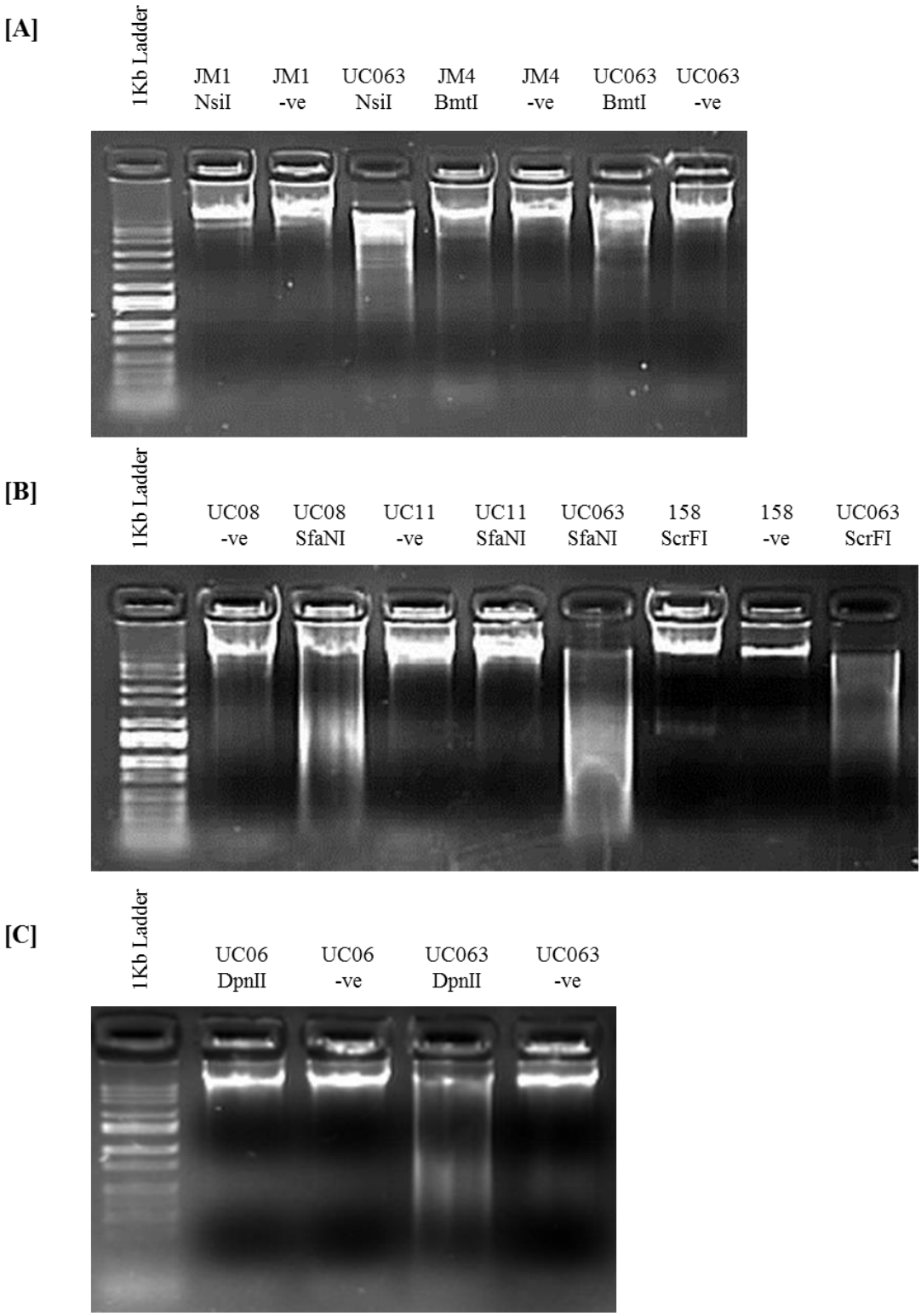


Figure 5.4: DNA restriction analyses
Genomic DNA restrictions of [A] *L. lactis* JM1 and JM4, [B] *L. lactis* UC08, UC11 and 158, [C] *L. lactis* UC06. In each case unrestricted DNA from the strain tested is used as a negative control, while *L. lactis* UC063 is used as a positive control. The enzymes used are indicated on each lane.

5.3.5 Type III/IV R-M systems

Four Type III R-M systems were identified, three of which appear to be complete: Lla184ORFBP in *L. lactis* 184, LlaJM1ORFEP in *L. lactis* JM1 and Lla109ORFCP in *L. lactis* UC109. The fourth, M.LlaJM2ORFGP in *L. lactis* JM2 appears to be responsible for the motif AGAAG as it is the only candidate in JM2. However, the gene currently has a frameshift, which may indicate a sequencing error. Alternatively, several active Type II methylases are known that also contain frameshifts [28]. The Type III methylase encoded by UC109, M.Lla109ORFCP, is probably the best candidate for the methylation of motif GAATC in that strain, while in strain JM1, the methylated CCAAT motif is a possible product of M.LlaJM1ORFEP. However, verification of these assignments will require cloning and characterisation of such methylases. In strain 184, the Type III motifs remain unassigned; GGAGA and GGCTNA both look as though they could be Type III motifs, but while M1.Lla184ORFBP appears functional, M2.Lla184ORFBP looks as though it is inactive.

A single Type IV system (LlaUC06McrCP and LlaUC06McrBP) was identified in *L. lactis* UC06 and appears to be the only complete system. Type III and IV systems are both significantly under-represented in *L. lactis* compared to Type I and II systems and no motifs were assigned to Type IV systems in this analysis which may be a result of difficulty in detecting cytosine modifications. A complete overview of all identified Type III/IV systems with predicted R-M activities is presented in Figure 5.5.

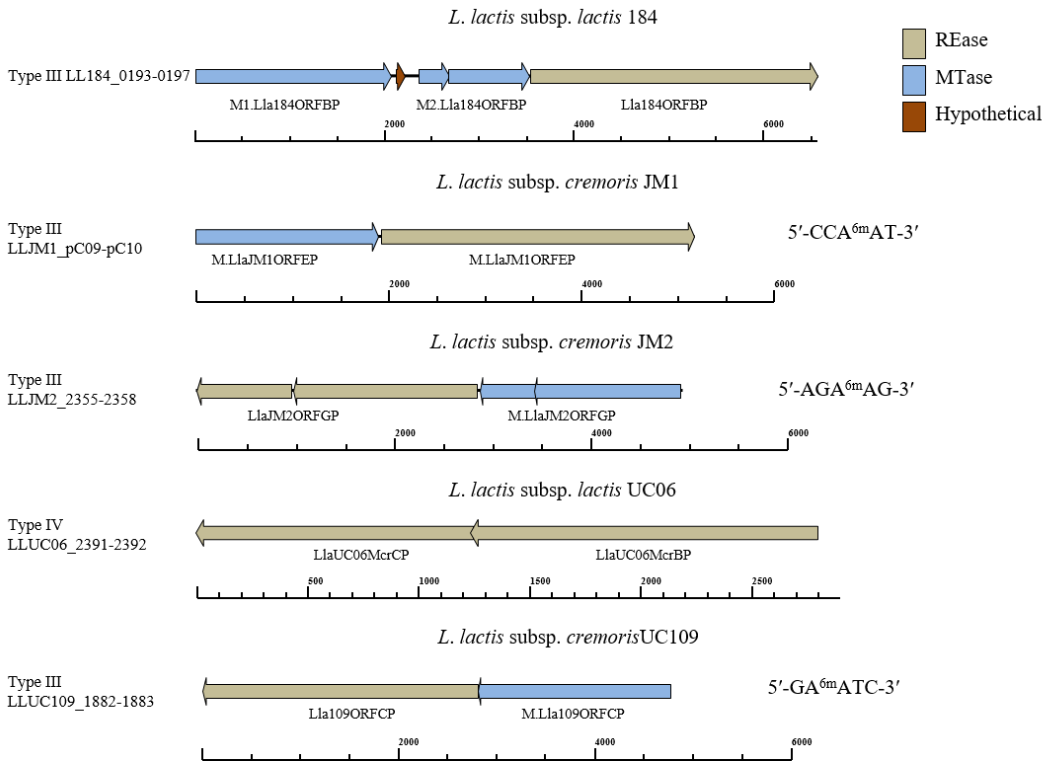


Figure 5.5: Overview of identified Type III & IV R-M systems
Genetic organisation of encoded Type III and Type IV R-M systems in; *L. lactis* 184, *L. lactis* JM1, *L. lactis* JM2, *L. lactis* UC06, and *L. lactis* UC109. Methylation motifs are indicated where resolved.

5.3.6 Type I R-M Shufflon system

Modulation of type I R-M recognition specificity by intergenic shuffling of HsdS-encoding genes has previously been reported [24-26]. Sequence analysis and annotation revealed the presence of a conserved Type I R-M shufflon system in the megaplasmids pMPJM1 (LlaJM1ORFCP) and pMPJM2 (LlaJM2ORFEP) (discussed in Chapter IV), consisting of multiple (apparently complete and incomplete) *hsdS* genes arranged around a recombinase-encoding gene (Fig. 5.4A). The presumed activity of this system was corroborated by the presence of a number of unassigned Type I methylase motifs in *L. lactis* JM2 containing m6A base modifications (Table 5.3). The six detected motifs were on average methylated in 99.87 % of positions present in the genomes. Strain JM2 also encodes two additional complete Type I R-M systems of which LlaJM2ORFDP was assigned the motif CATNNNNNNRTGA. However, it is not possible to assign additional motif(s) to the remaining individual systems.

Comparative analysis of the pMPJM2 *hsdS* subunits identified four conserved TRDs (target recognition domain) termed ARD (amino-proximal recognition domain) and CRD (carboxy-proximal recognition domain) in subunits 1A/B, and TRD in subunits 2A and 2B (Fig. 5.4B). Sequence analysis of the pMPJM2 *hsdS* genes and their respective recognition domains indicated the presence of a putative recombinatorial sequence (AATCATCATTTA) termed ‘vipareetus’ by Sitaraman and colleagues (from the Sanskrit ‘vipareet’ meaning inverted or opposite) thought to behave as inversion sites (Fig. 5.5) [27]. Computational analysis of these vipareetus sequences indicate four possible functional *hsdS* combinations, with the ‘unused’ subunits stored as a so-called reservoir (Fig. 5.4C).

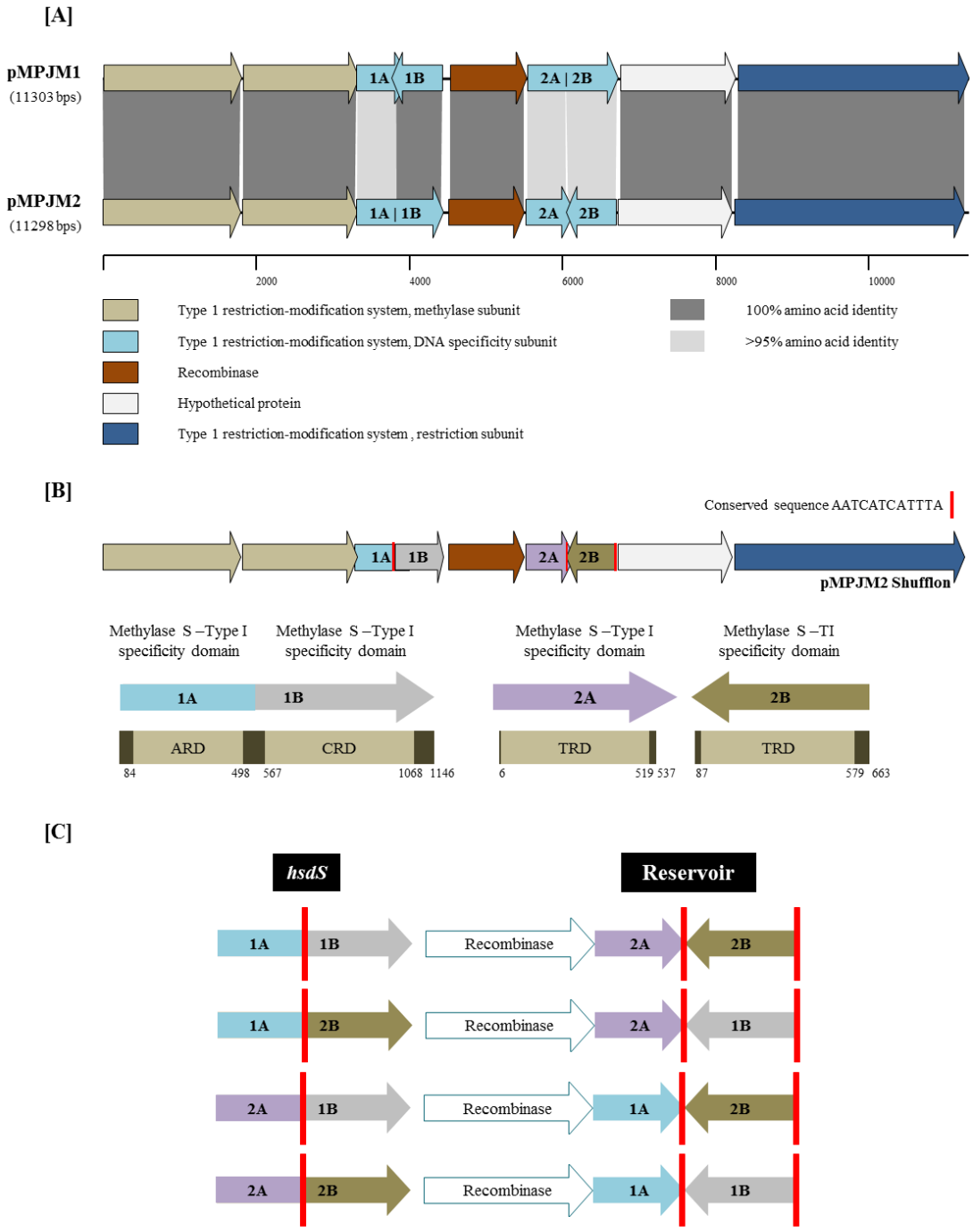


Figure 5.4: Analysis of the pMPJM2 shufflon system
[A] Arrangement of the LlaJM1ORFCP and LlaJM2ORFEP shufflon systems, shaded areas indicate BLAST amino acid identity. [B] Gene maps of conserved specificity domains and predicted combinations of the pMPJM2 LlaJM2ORFEP shufflon system. Locations of conserved recombination sequences are indicated by red lines.

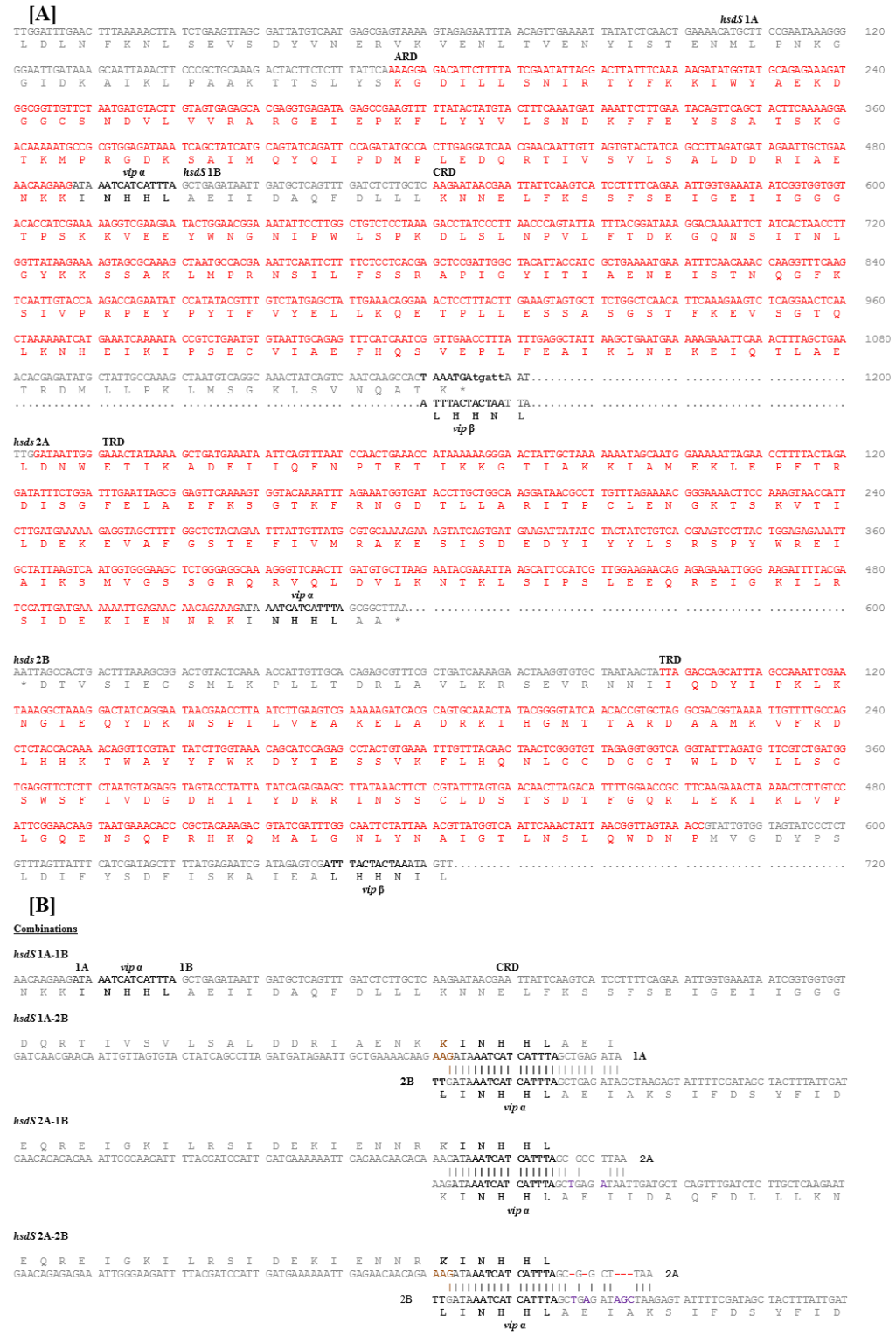


Figure 5.5: Specificity sub-units of shufflon system
[A] Nucleotide sequence analysis of the *hsdS* specificity subunits of the pMPJM2 shufflon system. Nucleotide sequence is displayed (grey) with each specificity sub-unit highlighted (red) and the predicted recombination sequences 'vip' (Black). [B] The mechanism of recombination of each sub-unit sequence is also displayed.

To assess this model, qPCR was employed as a method of detection and quantification of both the active subunit and the reservoir for each of the predicted combinations. Primers (Table 5.2) were designed based on sequences that represent each of the theoretical combinations (Subunits; 1A-1B, 1A-2B, 2A-1B, 2A-2B) and their corresponding respective reservoirs (Subunits; 2A-2B, 2A-1B, 1A-2B, 1A-1B). The analysis indicated the presence of each of these possible *hsdS* combinations with that of the sequenced conformation (*hsdS* 1A-1B) representing the dominant sub-unit (Fig. 5.6A & B) and confirming subunits are organised as predicted based on the vipareetus sequences.

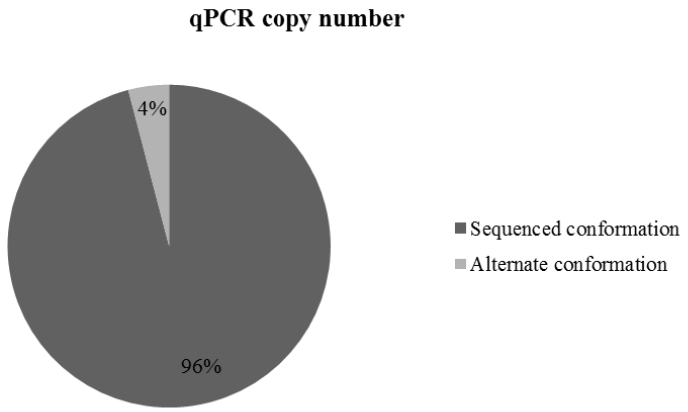
To further corroborate the proposed shuffling scenario, analysis of the raw sequencing reads which overlapped the shufflon region was undertaken. In accordance with the qPCR results, the sequenced conformation was found to be the dominant arrangement of the *hsdS* subunits (91 % of reads) (Fig 5.6C). Both the initial sequencing run and the qPCR analysis were conducted on a culture isolated from a single colony and it is believed this may have impacted on the heterogeneity of the population. An interesting observation from the raw read data was the conformation of the “reservoir” sub-units which in some cases were identified in alternate arrangements to those proposed in (Fig. 5.4C). In some instances they occurred in a different orientation or alternate strands to those described, while in a small number of cases a sub-unit was missing completely from the sequence, indicating that the shuffling of unused *hsdS* sub-units is less well conserved than previously thought. Interestingly, the homologous system sequenced on pMPJM1 (Fig 5.4A), was sequenced in an alternative combination, indicating the proposed shuffling scenario is likely to result from selective pressure within a population.

Finally, analysis of the ^{6m}A modified motifs detected by SMRT base modification analysis was conducted. It is expected that, in the case of an active type I R-M system, the ARD would contribute the first part of the target motif while the CRD would contribute the second part separated by $N^{[x]}$. In the case of the shufflon system, this would indicate four distinct methylation motifs, composed of two ARD patterns and two CRD patterns, replicating each possible combination of the *hsdS* subunits. While the detection of several type I motifs in *L. lactis* JM2 would appear to corroborate this hypothesis (Table 5.3), at sequencing coverage of 99X, it was not possible to definitively confirm the sequences associated with each domain. The identification of two (one chromosomal- and one plasmid-encoded) additional Type I R-M systems in *L. lactis* JM2, further complicates the determination of which motifs should be associated with each system. *L. lactis* JM1, which encodes one additional type I R-M system on its chromosome, was found to produce one Type I motif with high confidence, while a further two Type I motifs were found to be complementary strands of the motif GAAYNNNNNNTARC which was associated with a solitary specificity subunit S.LlaJM1ORFDP located on a separate plasmid.

[A]

hsdS conformation	Replicate	Cp	Copies	(+/-) SD
1A-1B	1	11.58	60829.88179	
1A-1B	2	11.67	57321.06096	
1A-1B	3	11.61	59637.03661	
Average			59263	1457
1A-2B	1	17.63	1120.981961	
1A-2B	2	17.58	1158.599899	
1A-2B	3	17.31	1384.655526	
Average			1221	116
2A-1B	1	17.92	925.6704968	
2A-1B	2	17.79	1008.618812	
2A-1B	3	17.78	1015.299183	
Average			983	41
2A-2B	1	19.52	321.9147125	
2A-2B	2	19.59	307.3774606	
2A-2B	3	19.46	334.9212169	
Average			321	11

[B]



[C]

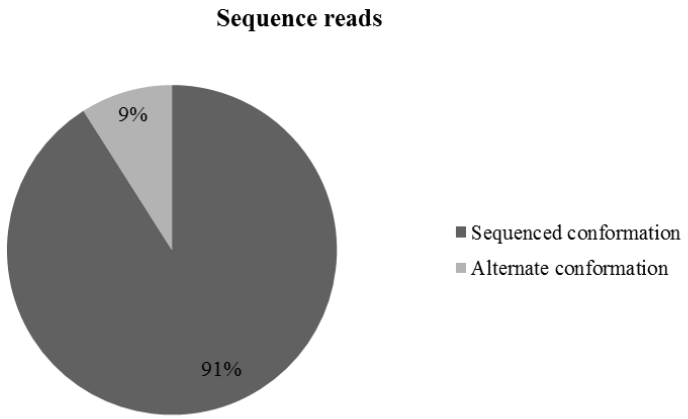


Figure 5.6: qPCR analysis of the pMPJM2 shufflon system
[A] The levels of each identified *hsdS* sub-unit combination as detected by qPCR. [B] Pie chart representing the percentage of occurrence of the sequenced *hsdS* conformation versus alternate conformations detected by qPCR. [C] Pie chart representing the percentage of occurrence of the sequenced *hsdS* conformation versus alternate conformations found in the raw sequencing reads.

5.4 Discussion

SMRT sequencing may be used for the identification of methylated DNA bases and their associated motifs. Methylome analysis of the lactococcal strains sequenced in the framework of this study was applied to identify methylation motifs that are linked to Type I and Type II R-M systems. The sequencing of a larger number of strains using this technology would permit an expansion and refinement of our knowledge of these systems in the future. A limitation of this technology is a difficulty in detecting ^{5m}C base modifications. While not performed in this study, detection of such modifications can be improved by treating DNA with Tet1 enzyme prior to sequencing [17]. A further dependent factor is the fold-coverage of the sequencing data with higher fold coverage resulting in more accurate base modification detection. For this reason a minimum of 250X coverage is recommended by Pacific Biosciences (https://github.com/PacificBiosciences/ Bioinformatics-Training/).

Comparative analysis of the lactococcal isolates in this study indicates a large degree of divergence in the encoded R-M systems present in each of the strains. This is also indicative of their phage defence abilities. *L. lactis* C10 and UL8 which encode no R-M systems contain five and three complete integrated prophages respectively, while strains *L. lactis* JM1 and JM2 which encode significantly more R-M systems present with one complete integrated prophage each. These two strains also present an adaptive phage response in the form of a plasmid-encoded Type I shufflon system. This system, the first of its type in *L. lactis* was composed of multiple *hsdS* subunits arranged around a recombinase-encoding gene allowing for the intergenic shuffling of specificity subunits, resulting in an effective adaptive defence mechanism against phage infection. Sequence analysis indicates the

recombination events rely on conserved overlapping ‘vipareetus’ sequences and analysis by qPCR confirms the proposed shuffling scenario.

Detection of a number of Type II R-M methylation motifs allowed for functional analysis with commercial restriction preparations to test the functionality of these systems. In each case it was found that the detected methylation motif blocked restriction by the associated enzyme, with the exception of *L. lactis* UC08. Both *L. lactis* UC08 and *L. lactis* UC11were found to contain the methylation motif 5′-GC^{6m}ATC-3′, but restriction with SfaNI was blocked in the case of UC11 only. Analysis of these motifs showed that while 99.84 % of these motifs were methylated in UC11, only 53.84 % of detected motifs in UC08 were methylated allowing the remaining motifs (1526 on the leading strand, 1555 on the lagging strand) to be restricted.

The major advantages of these predictions are the ease with which data can now be mined for the detection of novel restriction enzymes. The technology also presents the ability to assess a strain rapidly and efficiently in terms of its abilities to withstand foreign DNA, particularly valuable in lactococcal strains which are frequently used in industrial fermentations, or in contrast to assess the ease with which a strain may be transformed in a laboratory setting. In this study overlapping motifs and clustering of REases and MTases allowed for the identification of the systems responsible for seventeen of the detected motifs from a total of fifty two; however, with increased data sets and continued improvements in sequencing coverage, it is envisioned that significantly more of these systems will be elucidated in the future.

5.5 References

1. Korlach J, Turner SW: Going beyond five bases in DNA sequencing. Cur Opin Struc Biol. 2012, 22(3):251-261.
2. Boye E, Lobner-Olesen A: The role of dam methyltransferase in the control of DNA replication in *E. coli*. Cell 1990, 62(5):981-989.
3. Julio SM, Heithoff DM, Provenzano D, Klose KE, Sinsheimer RL, Low DA, Mahan MJ: DNA adenine methylase is essential for viability and plays a role in the pathogenesis of *Yersinia pseudotuberculosis* and *Vibrio cholerae*. Infect Immun. 2001, 69(12):7610-7615.
4. Mohapatra SS, Fioravanti A, Biondi EG: DNA methylation in *Caulobacter* and other Alphaproteobacteria during cell cycle progression. Trend Microbiol. 2014, 22(9):528-535.
5. Marinus MG, Casadesus J: Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEMS Microbiol Rev. 2009, 33(3):488-503.
6. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev S, Dryden DT, Dybvig K et al: A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nuc Acids Res. 2003, 31(7):1805-1812.
7. Galli D, Lottspeich F, Wirth R: Sequence analysis of *Enterococcus faecalis* aggregation substance encoded by the sex pheromone plasmid pAD1. Mol Microbiol. 1990, 4.
8. Roberts RJ, Vincze T, Posfai J, Macelis D: REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nuc Acid Res. 2015, 43(Database issue):D298-D299.

9. Roberts RJ, Vincze T, Posfai J, Macelis D: REBASE—enzymes and genes for DNA restriction and modification. Nuc Acid Res. 2007, 35(suppl 1):D269-D270.

10. Källström H, Blackmer Gill D, Albiger B, Liszewski MK, Atkinson JP, Jonsson AB: Attachment of *Neisseria gonorrhoeae* to the cellular pilus receptor CD46: identification of domains important for bacterial adherence. Cell Microbiol. 2001, 430.

11. Bae T, Schnewind O: The YSIRK-G/S motif of staphylococcal protein A and its role in efficiency of signal peptide processing. J Bacteriol. 2003, 185.

12. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Meth. 2010, 7(6):461-465.

13. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: Real-time DNA sequencing from single polymerase molecules. Meth Enzymol. 2010, 472:431-455.

14. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al: Real-time DNA sequencing from single polymerase molecules. Science. 2009, 323(5910):133-138.

15. Timp W, Mirsaidov UM, Wang D, Comer J, Aksimentiev A, Timp G: Nanopore Sequencing: Electrical Measurements of the Code of Life. IEEE Trans Nanotechnol. 2010, 9(3):281-294.

16. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J: Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nuc Acid Res. 2012, 40(4):e29.

17. Clark T, Lu X, Luong K, Dai Q, Boitano M, Turner S, He C, Korlach J: Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biol. 2013, 11(1):1-10.

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990, 215(3):403-410.

19. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: The Pfam protein families database. Nuc Acid Res. 2004, 32(suppl 1):D138-D141.

20. Söding J, Biegert A, Lupas AN: The HHpred interactive server for protein homology detection and structure prediction. Nucl Acid Res. 2005, 33(Web Server issue):W244-W248.

21. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. Nuc Acid Res. 2002, 30(7):1575-1584.

22. Sambrook J, Russell DW: Purification of nucleic acids by extraction with phenol:chloroform. CSH protocols. 2006, 2006(1).

23. Galli D, Friesenegger A, Wirth R: Transcriptional control of sex-pheromone-inducible genes on plasmid pAD1 of *Enterococcus faecalis* and sequence analysis of a third structural gene for (pPD1-encoded) aggregation substance. Mol Microbiol. 1992, 6.

24. O'sullivan D, Twomey DP, Coffey A, Hill C, Fitzgerald GF, Ross RP: Novel type I restriction specificities through domain shuffling of HsdS subunits in *Lactococcus lactis*. Mol Microbiol. 2000, 36(4):866-875.

25. Cerdeño-Tárraga AM, Patrick S, Crossman LC, Blakely G, Abratt V, Lennard N, Poxton I, Duerden B, Harris B, Quail MA et al: Extensive DNA

Inversions in the *B. fragilis* Genome Control Variable Gene Expression Science. 2005, 307(5714):1463-1465.

26. Claesson MJ, Li Y, Leahy S, Canchaya C, van Pijkeren JP, Cerdeño-Tárraga AM, Parkhill J, Flynn S, O'Sullivan GC, Collins JK et al: Multireplicon genome architecture of *Lactobacillus salivarius*. PNAS. 2006, 103(17):6718-6723.

27. Sitaraman R, Dybvig K: The *hsd* loci of *Mycoplasma pulmonis*: organization, rearrangements and expression of genes. Mol. Microbiol. 1997, 26(1):109-120.

28. Ershova AS, Karyagina A S, Vasiliev, MO, Lyashchuk AM, Lunin VG, Spirin SA, Alexeevski AV: Solitary restriction endonucleases in prokaryotic genomes. Nuc Acid Res. (2012), 40(20), 10107–10115. http://doi.org/10.1093/nar/gks853.

Chapter VI

Assessing functionality and genetic diversity of lactococcal prophages

Note: Prophage inductions and DNA restriction profiling were performed by Dr Jennifer Mahony.

Chapter VI contents

6 Abstract 275
6.1 Introduction 276
6.2 Materials & Methods 279
6.2.1 Bacterial Strains and growth conditions 279
6.2.2 Genome sequencing and data assembly 279
6.2.3 General feature predictions 279
6.2.4 Prophage identification 280
6.2.5 Identification of phage-encoded phage-resistance systems 280
6.2.6 Comparative genomics 281
6.2.7 Phylogenetic analysis 281
6.2.8 Pan- and core-virome analysis 281
6.2.9 Prophage induction 282
6.2.10 Validation of prophage induction by DNA restriction profiling 282
6.2.11 Nucleotide sequence accession numbers 284
6.3 Results 286
6.3.1 Prophage identification 286
6.3.2 MCL analysis of structural regions 290
6.3.3 Analysis of the adhesion module of the Group V members 294
6.3.4 Pan- and core-virome analysis 295
6.3.5 Prophage induction trials 297
6.3.6 Validation of the presence of induced prophages 299
6.3.7 Prophage-encoded phage-resistance systems 301
6.4 Discussion 303
6.5 References 307
Supplementary Table S6.1 313
Supplementary Table S6.2 319

6 Abstract

Lactococcus lactis strains are the most extensively exploited lactic acid bacteria (LAB) in commercial dairy fermentations. Though the presence of prophages in lactococcal genomes is widely reported, only a small number of studies pertaining to the stability of the prophages in the genomes have been performed. The current study explored the genomes of thirty lactococcal strains for the presence of potentially intact prophages, so as to assess their genomic diversity and the associated risk (or benefit) of harbouring such prophages. Genomic predictions partnered with mitomycin C-induction tests revealed that only four strains consistently produced intact phage particles. Interestingly, our analysis revealed the widespread presence of phage-resistance systems encoded by lactococcal prophages highlighting the potential benefits for host fitness. Most of the identified lactococcal prophages are shown to belong to the so-called P335 phage group, while various (presumed) phage remnants bear similarity to members of the 936 phage group. The P335 phage group was recently shown to encompass four distinct genetic lineages. Our study identified an additional lineage, thus expanding the diversity of this industrially significant phage group.

6.1 Introduction

Consistent cheese production relies on the application of technologically robust starter cultures, which in many cases consist of *Lactococcus lactis* strains. One of the key characteristics of technologically robust strains is resistance to virulent (bacterio)phages. However, many lactococcal chromosomes are known to harbour one or more integrated prophage genomes, which may excise following induction, culminating in starter cell lysis and release of intact phage particles. Prophage induction represents a double-edged sword phenomenon since on the one hand it may cause unwanted or premature lysis leading to poor quality or loss of product, while on the other hand phage-mediated cell lysis in maturing cheese is considered favourable because of the release of intracellular enzymes involved in flavour development [1].

In the context of phage therapy, the presence of prophages may equally be considered both beneficial and problematic. For example, the development of unusually virulent derivatives of *Streptococcus pyogenes* was linked to prophage acquisition highlighting the role of prophages in the evolutionary fitness of the host [2]. This example is mirrored across a spectrum of bacterial pathogens, which include, among others, *Bacillus anthracis*, *Staphylococcus aureus* and *Vibrio cholera* [3-5]. Therefore, while integrated prophage genomes are observed to suffer from considerable genome decay, with a majority believed to become functionally defective, it is important to assess their presence, diversity and functionality.

The genomes of *L. lactis* MG1363 and IL1403 were the first lactococcal genomes to be fully sequenced [6, 7], with each chromosome containing six predicted prophage-encoding regions, of which two and three, respectively, appear to represent intact prophages [6-8]. Various studies have applied UV, mitomycin C

(MitC), acid and thermal treatments with varying success to study prophage-mediated lytic potential of lactococcal strains [9-11]. Furthermore, bacteriocin-provoked prophage induction has been reported [12]. It is crucial to assess the induction ability of putative lactococcal lysogens in order to establish the risk they pose to dairy fermentations. To date, the majority of studies relating to prophage induction have focused on one or a limited number of lactococcal strains, encumbering our ability to generally appreciate the risk presented by such prophages.

Lactococcal prophages are typically classified as members of the polythetic P335 phage group, which are a genetically diverse group of phages. Excluding integrated prophages, the genome sequences of ten P335 phages are currently publicly available and have recently been divided into four subgroups (designated as subgroup I, II, III and IV), based on their overall nucleotide similarity and associated virion morphology [13]. The structural elements of phages determine the morphology of the phage including features that comprise the adhesion module. These modules dictate the initial interactions of the phage with its host and consequently, are an essential factor in phage classification. The adhesion modules in lactococcal P335 phage genomes encode elements of the so-called “initiator” and “baseplate” complexes of the distal tail region. The adhesion module is comprised of (the C-terminus of) the tail tape measure protein (TMP), the distal tail protein (Dit), the the tail-associated lysin (Tal) or tail fibre, and the receptor binding protein (RBP) and in some cases additional baseplate proteins (Bpp’s). Since the RBP (which typically makes up [part of] the baseplate) is the primary determinant of host range, the baseplate-encoding region is explored in further detail here. Sub-group I phages typically possess tails with a long tail fibre, which is thought to consist of a long Tal

fused with an RBP. Sub-group II phages possess a single large RBP-encoding gene or a multi-component baseplate structure with a double-disc morphology [13]. P335 phages assigned to Sub-groups III or IV typically display stubby tail tips that are reminiscent of the 936 phages of *L. lactis* and presumed to be solely composed of a homo-oligomeric RBP [13]. However, this subgrouping of P335 phages has not been applied to the extended analysis of lactococcal prophages and this constitutes a major knowledge gap in terms of the genetic diversity and interactions of these phages and their hosts.

Significant advances in genome sequencing technologies in recent years have facilitated an increasing availability of high quality complete genome sequences and improved our ability to predict and assess technologically appropriate and advantageous strains, including the presence of prophage-associated DNA [14]. Complete genome sequences of fourteen lactococcal strains are currently available in the public data bases and represent a useful resource for the analysis of the genetic diversity and identification of strains carrying prophages that are fully functional. In the current study, complete genome sequences of a further sixteen lactococcal strains were assessed in order to derive information on the presence and diversity of lactococcal prophages. A survey of all thirty genomes was undertaken in this study to assess the genetic diversity of, and potential risk and/or benefit associated with prophages in the dairy industry.

6.2 Materials & Methods

6.2.1 Bacterial Strains and growth conditions

Bacterial strains used in this study are detailed in Table 6.1. *L. lactis* strains were routinely cultured at 30 ^{o}C in M17 broth (Oxoid) supplemented with 0.5 % glucose without agitation.

6.2.2 Genome sequencing and data assembly

Sequencing and data assembly for the genomes of newly sequenced *L. lactis* strains employed in this study is detailed in Chapter III.

6.2.3 General feature predictions

Open Reading Frame (ORF) prediction was performed using a combinatorial approach of Prodigal v2.5 prediction software (http://prodigal.ornl.gov) and BLASTX v2.2.26 alignments [15]. Automatic annotation of ORFs was performed using BLASTP v2.2.26 [15] analysis against the non-redundant protein databases curated by the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nig.gov/). Manual curation of predicted ORFs was conducted using Artemis v16 genome browser and annotation tool (http://www.sanger.ac.uk/science/tools/artemis), which was used to combine and inspect ORF results, to adjust where necessary start codons of predicted genes, and to aid in the identification of pseudogenes. Further refinement of annotations was performed where required using alternative databases; Pfam [16] and Uniprot/EMBL (http://www.uniprot.org/). Ribosomal RNA (rRNA) and transfer RNA (tRNA) genes were predicted using RNAmmer v1.2 (http://www.cbs.dtu.dk/services/RNAmmer/)

and tRNA-scan-SE v1.4 (http://lowelab.ucsc.edu/tRNAscan-SE/), respectively. Predicted RNA encoding genes were manually added using Artemis.

6.2.4 Prophage identification

PHAST (PHAge Search Tool) [17] was used to screen genome sequences for the presence of integrated prophage genomes, and classifies its outputs in three categories: intact, incomplete and questionable. In order to further validate the presence of predicted complete or remnant prophages, relevant genomic regions where manually annotated as described above for bacterial genomes, in order to ascertain if all genes required to produce a functional phage particle were present. A complete phage particle was defined as one which contained genes necessary for lysogeny (integrase and repressor), replication/transcription/packaging (e.g. topoisomerase, replisome organiser, DNA-binding proteins, small & large terminases), morphogenesis (capsid, and tail, whiskers and other decorations) and lysis (holin[s] and lysin). Regions containing components of all the above-mentioned functional modules were predicted as intact, all others were predicted as incomplete phage (Supplemental Table S6.1).

6.2.5 Identification of phage-encoded phage-resistance systems

Potential abortive infection systems (Abi) were detected by constructing a database of the amino acid sequences of all currently known Abi systems (Supplemental Table S6.2) and performing an all-against-all reciprocal BLASTP [18] of the phage-encoding regions against the database using an alignment cut-off value; E-value 0.0001, and >50 % amino acid identity across 50 % of the sequence length. Sie (Superinfection exclusion) proteins were manually identified using the

following criteria: a small protein (~160 amino acids in length), possessing an N-terminal transmembrane domain detected with TMHMM Server, v. 2.0, and encoded by a gene situated between the integrase- and repressor-encoding gene within the lysogeny module. Phage-encoded methylases were detected as described above for general feature predictions.

6.2.6 Comparative genomics

Sequence comparisons at the protein level were performed by all-against-all, bi-directional BLAST alignment [15] using the following alignment cut-off criteria: E-value < 0.0001, and > 50 % amino acid identity across at least 50 % of the sequence length. The Markov Clustering Algorithm (MCL) was implemented in the mclblastline pipeline v12-0678 [9], as previously described [19].

6.2.7 Phylogenetic analysis

Whole phage-genome nucleotide alignments were performed using MUSCLE v3.8.31 [20]. Phylogenetic trees were computed by the maximum-likelihood method in PhyML v3.0 and bootstrapped x1000 replicates [21]. Tree files were visualised using ITOL (Interactive Tree of Life) (http://itol.embl.de/index.shtml).

6.2.8 Pan- and core-virome analysis

PGAP v1.0 [22] was used to perform the pan-genome analysis according to Heaps law pan-genome model [23]; the ORF content of each genome is organised in functional gene clusters using the Gene Family method where ORFs produce an alignment with a minimum of 50 % sequence identity across 50 % of their length

and a pan-genome profile was subsequently generated by plotting the number of genes in the pan-virome against the number of genomes added.

6.2.9 Prophage induction

To assess the functionality and lytic capability of the (predicted) resident prophages of the lactococcal strains employed in this study, induction trials with the DNA intercalating agent mitomycin C (MitC) were undertaken. Initial screening of prophage induction was performed in 96-well microtitre plate assays. 10 ml of GM17 broth was inoculated with 2 % of a fresh overnight of the bacterial strains to be analysed. The cultures were incubated at 30 °C until an OD at 600 nm of approximately 0.2 was reached at which point either 0.2 or 2 $\mu g.ml^{-1}$ MitC (final concentration) was added. 0.2 ml of the treated cultures was transferred in triplicate (three independent cultures) to a 96-well microtitre plate. A negative control of uninduced culture of each strain was included, as well as a positive control of *L. lactis* NZ9000 carrying the inducible prophage TP901-1*erm* [24]. The microtitre plate was incubated at 30 °C for 16 hours in a microtitre plate reader (MWG Sirius HT plate reader, BIO-TEK® Instruments, USA) and OD_{600} readings recorded at 30 minute intervals. 0.2 $\mu g.ml^{-1}$ MitC is a relatively low level of this prophage-inducing agent, and where induction was observed at this level, it is considered to represent genuine prophage-induction mediated cell lysis as opposed to growth arrest or cell death due to toxicity as may be observed at the higher MitC level (2 $\mu g.ml^{-1}$).

6.2.10 Validation of prophage induction by DNA restriction profiling

To validate the induction of prophage, DNA was isolated from representative induced and uninduced samples. Since strains C10 and IL1403 yielded positive

induction profiles according to the MitC induction trials described above they were employed as presumed prophage-positive samples. Conversely, SK11 displayed a negative prophage induction profile and was thus employed as a phage-negative candidate. Four 50 ml cultures of each strain were grown to an OD_{600} of 0.2 and three of the four samples were induced by 0.2 $\mu g.ml^{-1}$ MitC (final concentration) as described above, while the remaining sample acted as an uninduced control. After overnight incubation, two of the four samples for each strain were DNase treated (Roche, Ireland) according to the manufacturer's instructions and all samples were PEG-precipitated. The resulting pellets were resuspended in 0.4 ml TE and treated with 40 µl of 20 $mg.ml^{-1}$ proteinase K for 20 min at 56 ˚C, followed by treatment with SDS at a final concentration of 2 % at 65 ˚C for 20 minutes. Potassium acetate was added to a final concentration of 1 M followed by incubation on ice for 20 min before centrifugation at 13,200 g for 10 min. The supernatant was then phenol/chloroform (25:24:1 phenol:chloroform:isoamyl alcohol, Sigma Aldrich, MO, USA) treated at least twice and the aqueous phase precipitated with 2.5 volumes of ice cold 96 % ethanol and 0.1 volume of sodium acetate (pH 4.8). Subsequent to centrifugation, the pellet was washed in 70 % ethanol and resuspended in 100 µl of TE buffer (pH 8.0). The extracted DNA was subsequently restricted with EcoRV (Roche diagnostics, Ireland) according to the manufacturer's instructions. For each strain, two induced samples were DNase-treated prior to EcoRV restriction, the third was not treated with DNase to allow residual host chromosomal to remain; while the fourth sample was uninduced and DNase-treated to account for spontaneously induced prophage (if any).

6.2.11 Nucleotide sequence accession numbers

Sequences used in the analysis were retrieved from the GenBank database under the following accession numbers: *L. lactis* IL1403 [GenBank: NC_002662], *L. lactis* MG1363 [GenBank: NC_009004], *L. lactis* SK11 [GenBank: NC_008527], *L. lactis* KF147 [GenBank: NC_013656], *L. lactis* NZ9000 [GenBank: NC_017949], *L. lactis* CV56 [GenBank: NC_017486], *L. lactis* A76 [GenBank: NC_017492], *L. lactis* UC509.9 [GenBank: NC_019435], *L. lactis* IO-1 [GenBank: NC_020450], *L. lactis* KW2 [GenBank: NC_022369], *L. lactis* NCDO 2118 [GenBank: NZ_CP009054], *L. lactis* KLDS 4.0325 [GenBank: NC_022593], *L. lactis* AI06 [GenBank: NZ_CP009472], *L. lactis* SO [GenBank: NZ_CP010050], *Lactococcus* phage 4268 [GenBank: NC_004746], *Lactococcus* phage BK5-T [GenBank: NC_002796], *Lactococcus* phage phiLC3 [GenBank: NC_005822], *Lactococcus* phage P335 [GenBank: DQ838728], *Lactococcus* phage r1t [GenBank: NC_004302], *Lactococcus* phage TP901-1 [GenBank: NC_002747], *Lactococcus* phage Tuc2009 [GenBank: NC_002703], *Lactococcus* phage ul36 [GenBank: NC_004066], *Lactococcus* phage 28201 [GenBank: KX456206], *Lactococcus* phage 50101 [GenBank: KX456207], *Lactococcus* phage 50901 [GenBank: KX456208], *Lactococcus* phage 56701 [GenBank: KX456209], *Lactococcus* phage 62501 [GenBank: KX456210], *Lactococcus* phage 63301 [GenBank: KX456211], *Lactococcus* phage 86501 [GenBank: KX456212], *Lactococcus* phage 98201 [GenBank: KX456213], *L. lactis* 184 [GenBank: CP015895], *L. lactis* 229 [GenBank: CP015896], *L. lactis* 275 [GenBank: CP015897], *L. lactis* UC06 [GenBank: CP015902], *L. lactis* UC08 [GenBank: CP015903], *L. lactis* UC11 [GenBank: CP015904], *L. lactis* UC063 [GenBank: CP015905], *L. lactis* UC77 [GenBank: CP015906], *L. lactis* UL8 [GenBank: CP015908], *L. lactis* C10

[GenBank: CP015898], *L. lactis* 158 [GenBank: CP015894], *L. lactis* UC109 [GenBank: CP015907], *L. lactis* JM1 [GenBank: CP015899], *L. lactis* JM2 [GenBank: CP015900], *L. lactis* JM3 [GenBank: CP015901] and *L. lactis* JM4 [GenBank: CP015909].

6.3 Results

6.3.1 Prophage identification

The genomes of thirty fully sequenced lactococcal strains were analysed for the presence of prophages using PHAST as an initial screen for prophage-encoding regions, followed by manual validation and curating of putative prophage-encompassing regions, resulting in the predicted presence of 59 intact and 106 incomplete prophages. A summary of potential prophage-encoding regions by PHAST and manual examination in individual strains is provided in Table 6.1.

Regions specifying predicted intact and incomplete prophages were extracted and a phylogenetic analysis was performed based on the nucleotide sequences of all prophage elements identified in previously and newly sequenced lactococcal genomes combined with representatives of sequenced P335 phages (temperate: Tuc2009, TP901-1, LC3, and BK5-T; and lytic: P335, ul36, r1t, 4268, Q33 and BM13). This analysis resulted in a tripartite grouping of the analysed phage genomes. A clear bifurcation of the major clade revealed two distinct genetic lineages, designated here as Cluster A and Cluster B, in addition to a minor clade, designated here as Cluster C. Cluster A is composed of 15 (predicted) intact prophages belonging to the previously recognized P335 sub-groups I-III and 82 incomplete prophages, while Cluster B includes the ten sequenced P335 phage isolates (Tuc2009, TP901-1, LC3, BK5-T, P335, ul36, r1t, 4268, Q33 and BM13) (sub-groups I-IV) and 41 (predicted) intact prophages (Fig. 6.1) [13].

Table 6.1: Prophage regions predicted by PHAST and manual curation in *Lactococcus lactis* genomes.

Strain (ref/source)	No. prophage regions detected by PHAST			No. prophage regions identified manually	
	Intact	Questionable	Incomplete	Intact	Incomplete
158 $	-	-	2	-	2
JM1 $	1	2	4	1	6
JM2 $	2	-	2	1	3
JM3 $	1	4	-	2	3
JM4 $	2	-	3	1	4
UC109 $	-	-	2	-	2
MG1363 [25, 26]	3	-	3	2	4
SK11 [27]	2	3	-	2	3
NZ9000 [25]	4	1	2	2	4
A76 [28]	4	3	2	2	7
UC509.9 [29]	-	-	1	-	1
KW2 [30]	1	-	-	1	-
184 $	4	2	2	2	6
229 $	5	2	-	4	3
275 $	1	2	6	3	6
C10 $	7	-	1	5	3
UC06 $	4	1	-	2	3
UC08 $	-	-	2	-	2
UC11 $	-	-	2	-	2
UC063 $	5	1	2	3	5
UC77 $	7	1	-	5	3
UL8 $	6	1	3	3	7
IL1403 [31]	6	-	-	3	3
KF147 [32]	2	2	2	2	4
CV56 [33]	3	2	1	2	4
IO-1 [34]	1	-	1	1	1
NCDO 2118 [35]	2	2	1	2	3
KLDS4.0325 [36]	6	1	4	4	7
AI06 [37]	2	-	-	1	1
S0 [38]	2	1	3	3	3

$Strains sequenced in Chapter III

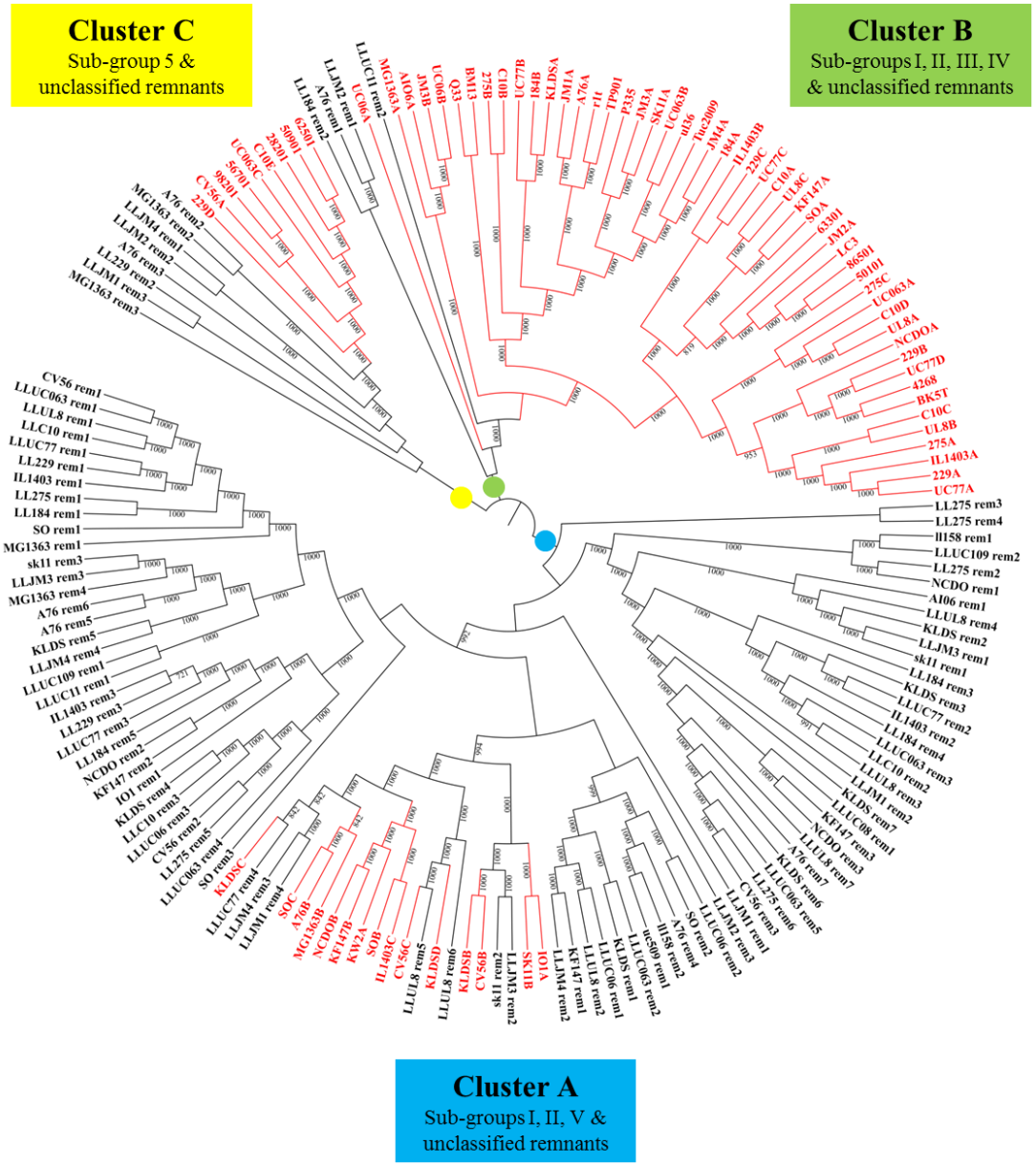


Figure 6.1: Whole genome phylogenetic tree of the P335 prophage.
Whole genome nucleotide alignment of the P335 type lactococcal prophage shows 3 distinct genetic groups. The (predicted) intact prophages are coloured red while incomplete prophages are coloured black. Cluster A, marked blue contains predominantly incomplete prophage. Cluster B, marked green contains the experimentally proven active P355 prophage and (predicted) intact prophages. Cluster C, coloured yellow represents a novel sub-group of P335 prophages.

In the majority of instances the phage remnants in Cluster A lack morphogenesis and lysis modules, the groupings were therefore made on the modules that are present. Furthermore, since the grouping of P335 phages is based on overall identity and morphology (as defined by the structural module), the prophage remnants in Cluster A were not considered to represent a new sub-group, but, rather, a group of incomplete prophages which cannot be classified due to the absence of group-determining genetic elements. The remaining prophage-encoding regions (Cluster C), which appear genetically distinct from Clusters A and B, encompasses eight incomplete prophage elements and nine intact prophage genomes (Fig. 6.1). The three clusters encompass the previously described P335 sub-groups I-IV [39]. Subgroups I-IV are contained within Cluster B, while a new sub-group (V) is contained in Cluster C as an evidently distinct genetic lineage. The overall tree is thought to be in agreement with previously described groupings of the P335 phage [39], yet contains more variation as a result of a vastly increased dataset and the inclusion of a large number of incomplete phage.

6.3.2 MCL analysis of structural regions

Due to the genetic variation and lack of conservation observed within the P335 prophages, the Markov Clustering Algorithm (MCL) was employed to classify the protein complement of those prophages that had been predicted to be complete. Prophage sequences that were presumed to be incomplete were excluded from the analysis at this point as partial phage regions would heavily skew the analysis and prevent accurate clustering. All-against-all reciprocal BLASTP analysis of the total protein complement of the analysed (presumed complete) prophages was initially performed and sorted using MCL, revealing a number of (distinct) clusters of conserved protein families. Further investigation highlighted that all of these conserved protein families were within the structural module of the prophages, allowing refinement of the analysis to this region. The amino acid sequences of all predicted ORFs corresponding to the predicted small terminase subunit (TerS) to the lysin-encoding gene of each prophage were taken and the analysis repeated. The analysed group of proteins includes those that are specified by the lysis cassette, which was used as a genetic marker to indicate the end of the structural module (although their encoded proteins are not components of the mature virion). The results of the MCL were formatted into a presence/absence matrix and hierarchical clustering (HCL) was applied to organise the prophages into groupings based on the variable content of their structural modules (Fig. 6.2). This analysis resulted in nine distinct and highly conserved structural classes each belonging to one of the four previously defined classical P335 sub-groups (defined as sub-group IA/B/C, sub-group IIA/B, sub-group IIIA/B and sub-group IV) plus one additional sub-group (sub-group V) (Fig. 6.2). Two representatives from each group were employed in further comparative analysis, which revealed that within each group a high level of

amino acid similarity was observed across the entire structural and lysis modules, yet that very little amino acid sequence relatedness was evident between groupings, indicating clear divisions between the groups (Fig. 6.3). Interestingly, while amino acid sequence similarities were not conserved between groupings, predicted functional synteny was preserved across the structural regions of all prophages in this study.

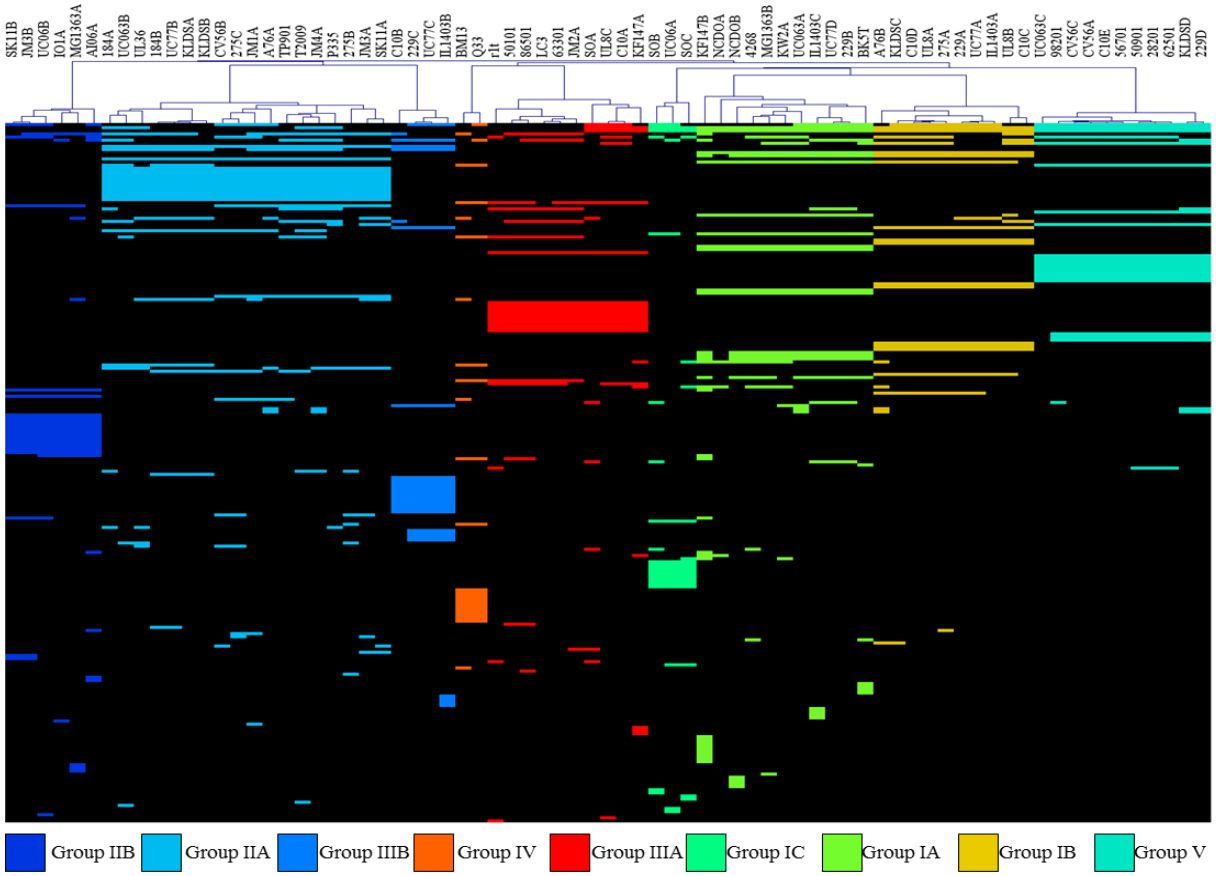


Figure 6.2: HCL presence/absence matrix of the P335 prophage structural regions.
HCL was performed on the basis of presence (coloured squares) or absence (black squares) of proteins from the constituent protein families of the P335 phage structural region. The structural regions of the P335 group phages form nine distinct clusters. The nine clusters shown include; subgroup IA/B/C; subgroup IIA/B; subgroup IIIA/B and subgroup IV, along with the newly discovered subgroup V phage.

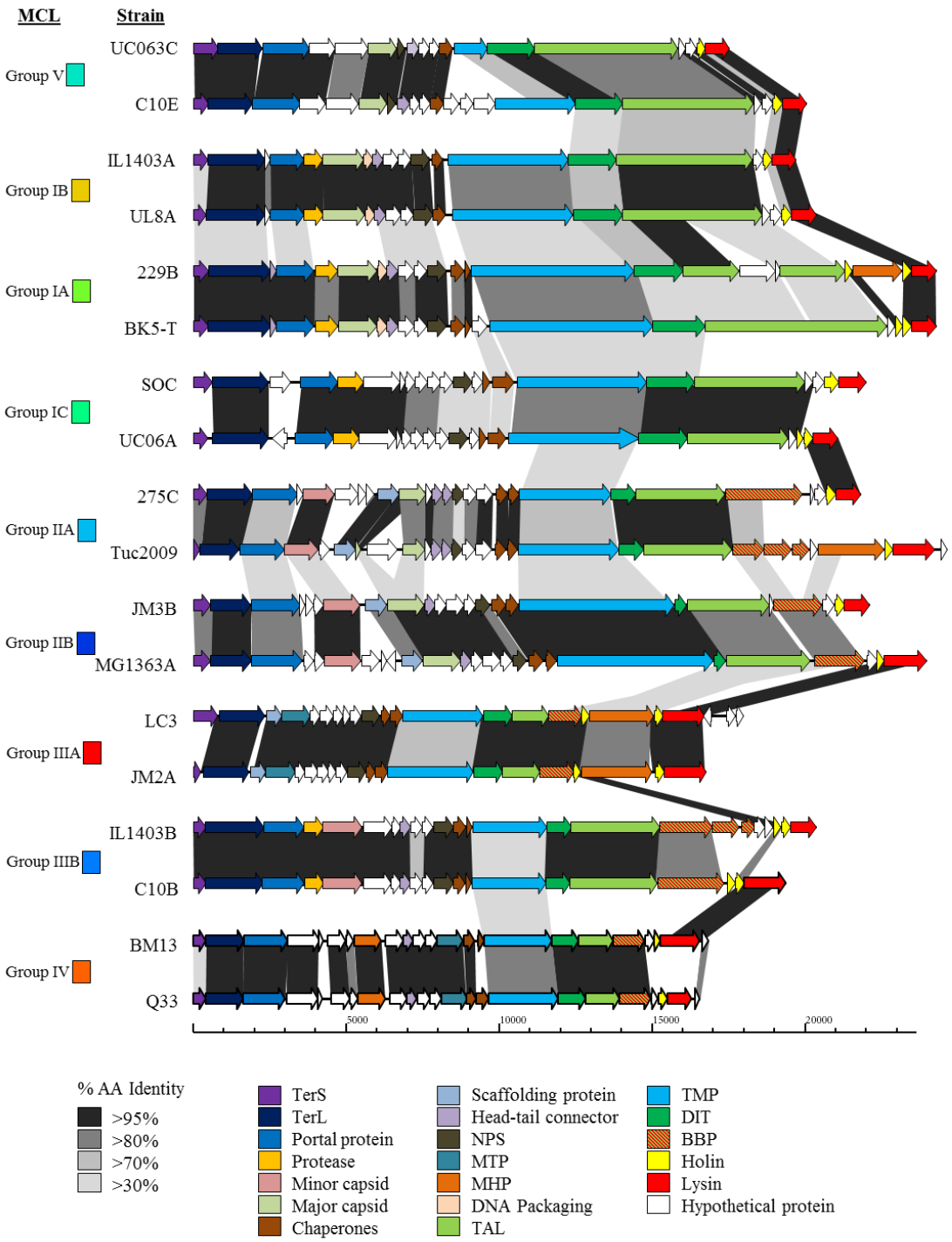


Figure 6.3: Gene synteny in the P335 prophage structural regions.
Representative strains form conserved groupings of the P335 phage structural regions are shown with arrows representing ORF coloured according to predicted function. Shaded boxes correspond to percentage amino acid identity between ORF.

6.3.3 Analysis of the adhesion module of the Group V members

In this study, phylogenetic analysis of the prophages of all sequenced lactococcal genomes and the P335 phages that have been sequenced revealed the presence of a fifth subgroup of P335 phages, termed subgroup V (represented by Cluster C in Fig. 6.1 and subgroup V in Fig. 6.3). Members of subgroup V possess TMPs of varying lengths (350 – 900 AA in length), well conserved Dit elements and large (~ 1500 AA in length) RBPs (Fig. 6.3). The RBPs of subgroup V phages are comparable in size to that of the Group I phage BK5-T and conserved domain search results for the RBP of C10E highlights the presence of a RBP N-terminal domain (residues 147-386) and a collagen triple helix (20 copies of a G-X-Y motif) at the protein's carboxy terminus (residues 1116-1144). A distinct gene encoding the Tal component is not observed in C10E or UC063C (Fig. 6.3), an observation that is consistent with BK5-T, which encodes a protein with an apparently fused Tal and RBP. Interestingly, the BK5-T virion contains a long tail fibre extending from the tail tip region [13], while recently it has been shown that the induced prophage 98201, a member of the newly identified subgroup V, also possesses such a long tail fibre [40]. Therefore, while the sequences of the RBPs of the Group I and V phages may be disparate, it is tempting to speculate that similar roles and structural features are conserved between these two phage subgroups.

6.3.4 Pan- and core-virome analysis

To ascertain genetic diversity levels and the extent to which additional genome sequencing will enhance current knowledge on this group of phages, pan- and core-virome analysis of all 74 predicted intact prophage-encoding regions of currently available lactococcal genome sequences was undertaken. Pan-virome analysis of prophage-encoding regions revealed an asymptotic curve indicating that the pan-virome is reaching a plateau and as additional genome sequences are added to this analysis, very limited new genetic information is expected to be added to the dataset (Fig. 6.4A). The resulting deduced mathematical function displays an exponential value < 0.5 confirming the closed state of the pan-virome. Conversely, core-virome analysis of the P335 prophage reveals the extent of genetic diversity and lack of conservation within the P335 group phages. Effectively no single gene is conserved among all of the P335 group phages (Fig. 6.4B). Effectively the discrepancy between the pan- and core- virome of these prophage is a result of conserved blocks of genes in the morphogenesis modules of the phage. These regions are highly conserved within each of the 5 P335 sub-groups and account for a large proportion of the pan-virome, thus reducing the perceived genetic variance in a large data-set. Conversely the morphogenesis regions do not share significant amino acid homology between the sub-groups resulting in an empty core-virome, whereas a core-virome could be described for each of the five distinct P335 sub-groups.

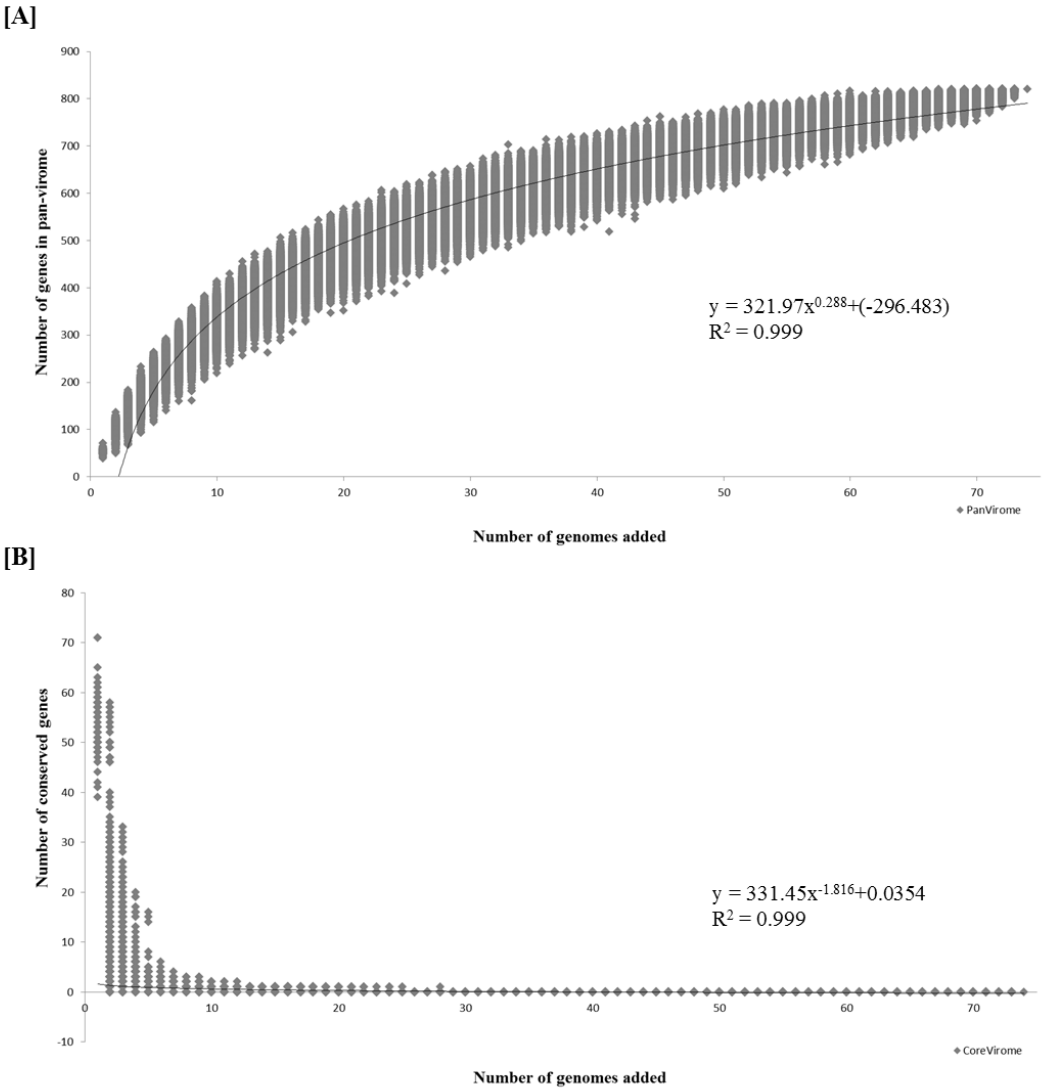


Figure 6.4: Pan-virome analysis of the P335 prophage genomes.
[A] Plot of accumulated number of genes in the P335 pan-virome (y-axis) versus the number of genomes added (x-axis), with deduced mathematical function. **[B]** Core-virome analysis of the P335 prophage. Plot of accumulated number of genes in the P335 core-virome (y-axis) versus the number of genomes added (x-axis), with deduced mathematical function

6.3.5 Prophage induction trials

Small-scale prophage induction trials were performed to assess if prophages could be induced from the thirty sequenced strains, while it was also used to ascertain if the predictions of the presence of intact (and thus functional) prophages is consistent with inducibility of such phages. Prophage inductions were implemented by the use of a sub-lethal and a relatively high dose of MitC in order to distinguish between genuine prophage-induction mediated cell lysis on the one hand, or growth arrest and/or cell death mediated by a lethal dose of MitC on the other. This yielded three distinct growth/cell lysis profiles: (i) both levels of MitC induced cell lysis, thus indicating prophage induction (Fig. 6.5A); (ii) only the addition of 2 $\mu g.ml^{-1}$ MitC induced cell death as delineated by a reduction in optical density at 600 nm (Fig. 6.5B); and (iii) lysis (as an indication of induction) is not observed at either level of MitC (Fig. 6.5C). A representative of each profile is presented in Fig. 6.5. Strains 184, 158, KF147, 275, A76, UC77, NZ9000, UC06, IO-1 and UC109 all exhibited growth profile (iii) and do not appear to contain inducible prophages (under the assessed conditions), while UC063, SK11, UC08, JM1, JM2, JM4, UC509.9 and UL8 are observed to lyse upon the addition of 2 $\mu g.ml^{-1}$ but not in the presence of 0.2 $\mu g.ml^{-1}$ MitC (growth profile ii), indicating cell death rather than prophage induction. Conversely, IL1403, C10, 229 and JM3 were observed to lyse upon the addition of 0.2 $\mu g.ml^{-1}$ MitC (growth profile i), thus indicative of prophage induction.

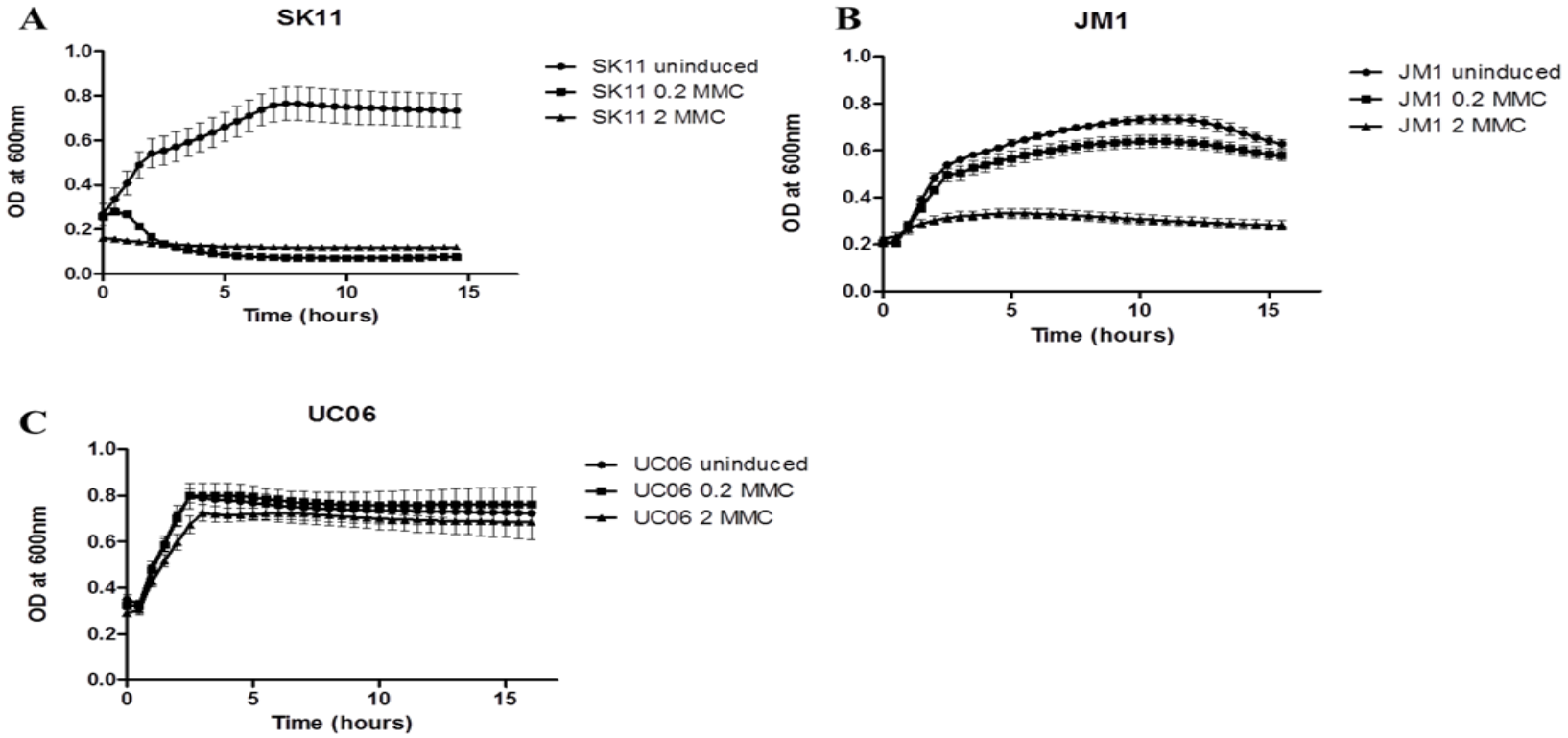


Figure 6.5: MitC induction profiles of representative lactococcal strains.
[A] Graph indicating the induction profile of *L. lactis* SK11 in the presence of 0.2 µg.ml^{-1} or 2 µg.ml^{-1}, or in the absence of MitC. Culture lysis was observed at both levels of MitC. [B] Graph indicating the induction profile of *L. lactis* JM1 in the presence of 0.2 µg.ml^{-1} or 2 µg.ml^{-1}, or in the absence of MitC. Culture lysis was observed only in the presence of 2 µg.ml^{-1} of MitC indicating that the culture is killed in the presence of this high level of inducing agent. [C] Graph indicating the induction profile of *L. lactis* UC06 which displays similar growth profiles in the presence or absence of MitC. These graphs present the data for representative strains and all strains analysed in this study exhibited one of the three profile types. The results are representative of at least three independent assays.

6.3.6 Validation of the presence of induced prophages

Four of the 24 strains assessed were identified as inducible by MitC treatment (IL1403, C10, 229 and UC063), with a further four strains yielding variable induction profiles (UC77, NZ9000, A76 and kw2). To ensure that the observed lysis corresponds to DNA-filled phage particle release and to further validate the induction profiles of the lactococcal strains, DNA was extracted from induced and uninduced culture supernatants from *L. lactis* C10 and IL1403. The recovered DNA was restricted with EcoRV and the restriction products separated by agarose gel electrophoresis (Fig. 6.6). The cell-free supernatants of the induced lysates of IL1403 and C10 exhibited clear profiles while the supernatant of the uninduced cultures did not indicate the presence of DNA, ruling out the possibility of (substantial) spontaneous prophage induction or background host DNA contributing to the restriction profiles since the supernatants of the uninduced strains were treated with DNase to remove background host chromosomal DNA.

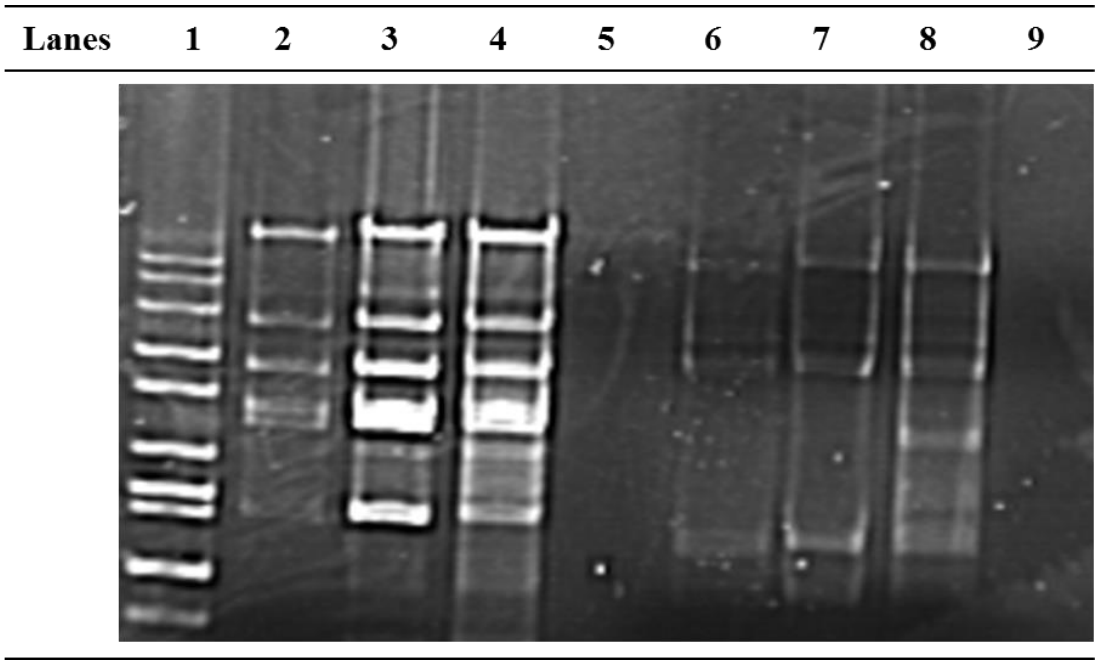


Fig 6.6: Restriction profiling of induced prophage
Lane 1; DNA ladder, lanes 2 and 3; IL1403 induced, lane 4; IL1403 induced & without DNase and lane 5; IL1403 uninduced. Lanes 6 and 7; C10 induced, lane 8; C10 induced & without DNase and lane 9; C10 uninduced.

6.3.7 Prophage-encoded phage-resistance systems

While prophages are considered a threat to production processes, it must also be considered that their presence may confer some advantages on the host. One such advantage is the potential for the provision of phage-resistance systems. An analysis of the prophages predicted to be intact in this study were assessed for the presence of potential phage-resistance systems based on previously established criteria [41, 42], or based on BLASTP analysis. On this basis, 14 and 9 out of 29 strains assessed were predicted to harbour prophages that encode at least one superinfection exclusion system or abortive infection system, respectively. In many cases multiple predicted systems were observed to be “stacked” in the strains owing to the presence of multiple prophages within a given strain (Table 6.2). The presence of such phage-resistance systems is expected to confer protection against a variety of phages thus providing a fitness benefit upon the host. It is also noteworthy that this is a conservative number since in many cases genes encoding hypothetical proteins are observed in the lysogeny modules that may possess Sie activity with characteristics that are beyond those that are currently proven to be active against the 936 phages. Similarly, additional, but as yet, unknown Abi systems may be encoded by prophages and it is therefore plausible that a much higher number of prophage-encoded phage-resistance systems is present. Conversely, twelve phage-encoded methylases were detected in this analysis, which can aid phage in overcoming host encoded R-M systems (discussed in Chapter V).

Table 6.2: Identification of phage-resistance systems on lactococcal prophages (Sie & Abi systems) and methylases to overcome R-M systems.

Strain	No. prophage-encoded predicted potential Sie systems	No. prophage-encoded predicted Abi systems	No. predicted prophage-encoded methylases
C10	1 (C10A)	1 (C10D)	2 (C10A, C10E)
229	-	-	1 (229D)
UC77	-	1 (UC77B)	1 (UC77B)
IL1403	1 (IL1403A)	-	-
UC063	-	1 (UC063A)	1 (UC063B)
UL8	-	1 (UL8A)	-
275	1 (275B)	-	2 (275A, 275C)
NZ9000	1 (NZ9000A)	1 (NZ9000A)	-
A76	1 (A76A)	-	-
SK11	1 (SK11A)	-	-
UC06	2 (UC06A, UC06B)	1 (UC06_rem2)	-
KF147	2 (KF147A, KF1477B)	-	1 (KF147A)
184	2 (184A, 184B)	-	-
JM1	1 (JM1A)	-	-
JM2	1 (JM2B)	-	-
JM3	1 (JM3A)	-	-
JM4	-	1 (pJM4A)	-
KW2	-	-	1 (KW2A)
IO-1	1 (IO1A)	-	-
UC509.9	-	-	1 (UC509_rem1)
UC08	-	-	-
UC11	-	-	-
158	-	-	-
UC109	-	-	-
AI06	-	1 (AI06A)	-
SO	1 (SOC)	-	-
KLDS 4.0325	-	2 (KLDSA, KLDSB)	1 (KLDSD)
NCDO2118	-	-	-
CV56	-	-	1 (CV56A)

6.4 Discussion

Lactococcal phages persist as a major threat to commercial fermentation processes required for the manufacture of dairy products, particularly cheese. While lactococci are prone to infection by lytic phages, the threat of prophage induction and concomitant cell lysis presents an equally challenging risk factor. Recently, the stability of active lactococcal prophages under dairy processing conditions was assessed for three lactococcal strains and it was observed that the prophages were not induced in media incorporating acids and osmotic stressors or through thermal treatments that would typically be encountered during dairy fermentation processes [43]. In contrast, MitC treatment was shown to be effective in inducing prophage elements from each of these three strains. However, the limited number of strains employed in this study constrains the assertions that can be applied to dairy strains in general as each strain will behave uniquely. To counter this issue the current study investigated the incidence of prophage induction in a larger set of strains so as to assess the genetic diversity of and risk factor presented by lactococcal prophages.

Thirty lactococcal genomes were explored for potential prophage-encoding regions using the PHAST software and followed by manual assessment of this analysis. This resulted in the identification of 84 potentially intact prophages; 31 questionable (and likely non-functional) prophage regions and 51 incomplete prophages. Phylogenetic analysis of the nucleotide sequence of all identified prophage regions combined with the sequences of previously sequenced P335 phages revealed two major groups of lactococcal prophages with a third minor group composed of a newly identified genetic lineage of prophage. The two major groups specify two distinct genetic lineages with the P335 phages (as distinct from prophages sequenced as part of bacterial chromosomal sequences) aligning within

Cluster B (Fig. 6.1). This suggests that lytically active P335 phages may all have derived from the genetic lineage constituted by Cluster B. Indeed, all but four of the phage sequences represented in Cluster B are P335 phages or predicted intact prophage regions (Fig. 6.1), while Cluster A contains predominantly phage remnants. Therefore, while Cluster A prophages largely appear to be permanent residents within their host bacterial genomes, Cluster B prophages present a much higher degree of likelihood of presenting with lytically active geno/phenotypes. Furthermore, while there are also a significant number of incomplete/non-functional prophage remnants in this group, it is possible that these prophages may contribute to the overall genetic diversity of incoming virulent or temperate phages since the lytically active P335 phages are contained within this overall genotypic group.

The lactococcal strains MG1363 and IL1403 were among the first to be analysed with respect to their prophage-encoding regions, each possessing six prophage regions [6-8]. Prophage induction of *L. lactis* MG1363 and IL1403 has been reported to result in variable inducibility profiles for MG1363 [8, 10, 43] and positive induction of two prophage elements of IL1403 [11]. Induction of the lactococcal strains ASCC890310 and ASCC890049 resulted in the release of phages detected using DNA sequencing before and after exposure to heat, acid, osmotic, oxidative and antibiotic stressors, with similarity to P335 subgroup I (BK5-T-like) and subgroup II (TP901-1-like) phages, respectively, among others [43]. This is also reflected in the current analysis since several strains including UC77, 229, NCDO 2118, UL8, 275, UC063, IL1403 and C10 possess at least one prophage with similarity to the sub-group I phages BK5-T and 4268 (Fig. 6.1), while a smaller number of strains possess prophages with similarity to subgroup II phages.

To assess the overall inducibility of lactococcal prophages resident within the host chromosomes, small-scale induction profiles were undertaken. Induction profiling under the assessed conditions determined that just four out of the 24 assessed strains contained inducible prophage. This is a relatively low number given the significant presence of seemingly intact phage genomes within lactococcal chromosomes and, as such, appears to represent a containable risk factor for the dairy industry. However, this entails a cautionary note as the appropriate conditions for induction may not have be achievable using MitC based inductions and induction under industrial fermentation conditions may still pose a valid threat. There appears to be a significant discrepancy between the number of predicted "intact" prophage genomic elements and the number of genuinely inducible prophage particles. There also appears to be a discrepancy between the number of intact prophages by PHAST and those by manual curation indicating that this tool should be used only as a guideline or indicator for the potential presence of intact prophage. However, ultimately manual checking of these prophage-encoding regions is essential for accurate determination of potential prophage-encoding regions.

From this study, it is clear that the majority of identified lactococcal prophage genomic regions are stable residents within their lactococcal host chromosome. Their replication *in situ* with the host is favourable to their continued existence, and induction of seemingly intact prophages appears at a relatively low frequency, approximately one in six strains are likely to be inducible under harsh conditions with a lower risk of induction expected in the dairy environment. However, while these lactococcal prophages are seemingly silent, they represent a vast genetic pool with the potential to increase the genetic diversity and adaptability of virulent phages. This is illustrated by the P335 phage ul36, which was previously

observed to circumvent two Abi systems, AbiK and AbiT, resident on the genome of *L. lactis* SMQ86 (UL8) by recombining with a resident prophage to produce progeny with altered receptor binding proteins and baseplate components [44]. Furthermore, through the acquisition of DNA replication functions, phage-resistance associated genes such as superinfection exclusion and abortive infection functions, it is clear that prophages may positively contribute to the overall fitness of the host.

In conclusion, prophages may represent a relatively low direct risk to cheese production processes, but their potential to expedite the evolution of virulent phages and the fitness of the host are key features that should be considered when selecting starter cultures. It is expected that rapid turnaround time on modern genome sequencing methods combined with the reduced costs will endorse the continued and vastly increased availability of lactococcal genomes permitting advanced assessments of prophage distribution, diversity and evolution, information that wil be crucial for the selection of genome-informed next generation starter cultures.

6.5 Referneces

1. O'Sullivan D, Ross RP, Fitzgerald GF, Coffey A: Investigation of the relationship between lysogeny and lysis of *Lactococcus lactis* in cheese using prophage-targeted PCR. Appl Environ Microbiol. 2000, 66(5):2192-2198.
2. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H: Prophage genomics. Microbiol Mol Biol Rev. 2003, 67(2):238-276.
3. Faruque SM, Albert MJ, Mekalanos JJ: Epidemiology, genetics, and ecology of toxigenic *Vibrio cholerae*. Microbiol Mol Biol Rev. 1998, 62(4):1301-1314.
4. Casas V, Sobrepena G, Rodriguez-Mueller B, Ahtye J, Maloy SR: Bacteriophage-encoded shiga toxin gene in atypical bacterial host. Gut Pathog. 2011, 3(1):10.
5. Coleman DC, Sullivan DJ, Russell RJ, Arbuthnott JP, Carey BF, Pomeroy HM: *Staphylococcus aureus* bacteriophages mediating the simultaneous lysogenic conversion of beta-lysin, staphylokinase and enterotoxin A: molecular mechanism of triple conversion. J Gen Microbiol. 1989, 135(6):1679-1697.
6. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A: The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. Genome Res. 2001, 11(5):731-753.
7. Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP et al: Complete genome sequence of the prototype lactic acid bacterium

Lactococcus lactis subsp. *cremoris* MG1363. J Bacteriol. 2007, 189(8):3256-3270.

8. Ventura M, Zomer A, Canchaya C, O'Connell-Motherway M, Kuipers O, Turroni F, Ribbera A, Foroni E, Buist G, Wegmann U et al: Comparative analyses of prophage-like elements present in two *Lactococcus lactis* strains. Appl Environ Microbiol. 2007, 73(23):7771-7780.

9. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. Nuc Acids Res. 2002, 30(7):1575-1584.

10. Wegmann U, Overweg K, Jeanson S, Gasson M, Shearman C: Molecular characterization and structural instability of the industrially important composite metabolic plasmid pLP712. Microbiol. 2012, 158(Pt 12):2936-2945.

11. Chopin MC, Chopin A, Rouault A, Galleron N: Insertion and amplification of foreign genes in the *Lactococcus lactis* subsp. *lactis* chromosome. Appl Environ Microbiol. 1989, 55(7):1769-1774.

12. Madera C, Garcia P, Rodriguez A, Suarez JE, Martinez B: Prophage induction in *Lactococcus lactis* by the bacteriocin Lactococcin 972. Int J Food Microbiol. 2009, 129(1):99-102.

13. Mahony J, Martel B, Tremblay DM, Neve H, Heller KJ, Moineau S, van Sinderen D: Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. Appl Environ Microbiol. 2013, 79(14):4401-4409.

14. Kelleher P, Murphy J, Mahony J, van Sinderen D: Next-generation sequencing as an approach to dairy starter selection. Dairy Sci Technol. 2015, 95:545-568.

15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990, 215(3):403-410.

16. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: The Pfam protein families database. Nuc Acids Res. 2004, 32(suppl 1):D138-D141.

17. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: PHAST: A Fast Phage Search Tool. Nuc Acids Res. 2011.

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990, 215.

19. Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D: Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. Sci Rep. 2016, 6:21345.

20. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nuc Acids Res. 2004, 32(5):1792-1797.

21. Guindon S, Gascuel O: A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Sys Biol. 2003, 52(5):696-704.

22. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J: PGAP: pan-genomes analysis pipeline. Bioinformatics. 2012, 28(3):416-418.

23. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al: Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. PNAS US. 2005, 102(39):13950-13955.

24. Stockdale SR, Mahony J, Courtin P, Chapot-Chartier MP, van Pijkeren JP, Britton RA, Neve H, Heller KJ, Aideh B, Vogensen FK et al: The lactococcal phages Tuc2009 and TP901-1 incorporate two alternate forms of their tail

fiber into their virions for infection specialization. J Biol Chem. 2013, 288(8):5581-5590.

25. Linares DM, Kok J, Poolman B: Genome sequences of *Lactococcus lactis* MG1363 (revised) and NZ9000 and comparative physiological studies. J Bacteriol. 2010, 192(21):5806-5812.

26. Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP: Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. J Bacteriol. 2007, 189(8):3256-3270.

27. Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos WM: Complete sequences of four plasmids of *Lactococcus lactis* subsp. *cremoris* SK11 reveal extensive adaptation to the dairy environment. Appl Environ Microbiol. 2005, 71(12):8371-8382.

28. Bolotin A, Quinquis B, Ehrlich SD, Sorokin A: Complete genome sequence of *Lactococcus lactis* subsp. *cremoris* A76. J Bacteriol. 2012, 194(5):1241-1242.

29. Ainsworth S, Zomer A, de Jager V, Bottacini F, van Hijum SA, Mahony J, van Sinderen D: Complete genome of *Lactococcus lactis* subsp. *cremoris* UC509. 9, host for a model lactococcal P335 bacteriophage. Gen Announc. 2013, 1(1):e00119-00112.

30. Kelly WJ, Altermann E, Lambie SC, Leahy SC: Interaction between the genomes of *Lactococcus lactis* and phages of the P335 species. Front Microbiol. 2013, 4:257.

31. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A: The complete genome sequence of the lactic acid

bacterium *Lactococcus lactis* ssp. *lactis* IL1403. Gen Res. 2001, 11(5):731-753.

32. Siezen RJ, Bayjanov J, Renckens B, Wels M, van Hijum SA, Molenaar D, van Hylckama Vlieg JE: Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium. J Bacteriol. 2010, 192(10):2649-2650.

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

34. Kato H, Shiwa Y, Oshima K, Machii M, Araya-Kojima T, Zendo T, Shimizu-Kadota M, Hattori M, Sonomoto K, Yoshikawa H: 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. 2012, 194(8):2102-2103.

35. Phalip V, Monnet C, Schmitt P, Renault P, Godon J-J, Diviès C: Purification and properties of the α-acetolactate decarboxylase from *Lactococcus lactis* subsp. *lactis* NCDO 2118. FEBS letters 1994, 351(1):95-99.

36. Yang X, Wang Y, Huo G: Complete Genome Sequence of *Lactococcus lactis* subsp. *lactis* KLDS4. 0325. Gen Announc. 2013, 1(6):e00962-00913.

37. McCulloch JA, de Oliveira VM, de Almeida Pina AV, Perez-Chaparro PJ, de Almeida LM, de Vasconcelos JM, de Oliveira LF, da Silva DE, Rogez HL, Cretenet M et al: Complete Genome Sequence of *Lactococcus lactis* Strain AI06, an Endophyte of the Amazonian Acai Palm. Gen Announc. 2014, 2(6).

38. Zhao F, Ma H, Lu Y, Teng K, Kang X, Wang F, Yang X, Zhong J: Complete genome sequence of *Lactococcus lactis* S0, an efficient producer of nisin. J Biotechn. 2015, 198:15-16.

39. Mahony J, Martel B, Tremblay DM, Neve H, Heller KJ, Moineau S, van Sinderen D: Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. Appl Environ Microbiol. 2013, 79(14):4401-4409.

40. Oliveira J, Mahony J, Lugli GA, Hanemaaijer L, Kouwen T, Ventura M, van Sinderen D: Genome Sequences of Eight Prophages Isolated from Lactococcus lactis Dairy Strains. Gen Announc. 2016, 4(6).

41. Mahony J, McGrath S, Fitzgerald GF, van Sinderen D: Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. Appl Environ Microbiol. 2008, 74(20):6206-6215.

42. McGrath S, Fitzgerald GF, van Sinderen D: Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Mol Microbiol. 2002, 43(2):509-520.

43. Ho CH, Stanton-Cook M, Beatson SA, Bansal N, Turner MS: Stability of active prophages in industrial *Lactococcus lactis* strains in the presence of heat, acid, osmotic, oxidative and antibiotic stressors. Int J Food Microbiol. 2016, 220:26-32.

44. Labrie SJ, Moineau S: Abortive infection mechanisms and prophage sequences significantly influence the genetic makeup of emerging lytic lactococcal phages. J Bacteriol. 2007, 189(4):1482-1487.

Supplementary Table S6.1: *In silico* detected P335 type (pro)phage fragments

Lactococcal subsp.	Strain name	Genbank accession no.	PHAST prediction	Phage co-ordinates	Genome size (bp)	Re-annotation
cremoris	158		Incomplete	1126714-1140695	13982	Incomplete
			Incomplete	1371155-1377902	6748	Incomplete
	JM1		Questionable	217275-234210	16936	Incomplete
			Incomplete	277856-297428	19573	Incomplete
			Incomplete	862688-884009	21322	Incomplete
			Questionable	1007488-1066452	58965	Intact
			Incomplete	1046019-1073805	27787	Incomplete
			Incomplete	1345807-1355395	9589	Incomplete
			Intact	1843663-1864000	20338	Incomplete
	JM2		Incomplete	486030-502980	16951	Incomplete
			Intact	581889-608595	26707	Incomplete
			Intact	1302471-1347786	45316	Intact
			Incomplete	1922969-1957200	34232	Incomplete
	JM3		Intact	291444-304882	13439	Incomplete
			Questionable	1043875-1105910	62036	Intact
			Questionable	1673961-1710094	36134	Incomplete
			Questionable	1989769-2029364	39596	Intact
			Questionable	2235012-2264102	29091	Incomplete
	JM4		Incomplete	540075-559569	19495	Incomplete
			Intact	843211-892526	49316	Intact
			Incomplete	1611464-1630709	19246	Incomplete
			Intact	1936722-1985616	48895	Incomplete
			Incomplete	2056181-2085124	28944	Incomplete
	UC109		Incomplete	312684-329292	16069	Incomplete
			Incomplete	1109821-1123802	13982	Incomplete
	MG1363	NC_009004	Intact	25908-60678	34771	Incomplete
			Intact	778852-821910	43059	Intact
			Incomplete	861704-872284	10581	Incomplete
			Incomplete	1310020-1335809	25790	Incomplete
			Intact	2061037-2110526	49490	Intact

			Incomplete	2203214-2237668	34455	Incomplete
	SK11	NC_008527	Intact	276137-289575	13439	Incomplete
			Questionable	1033838-1076036	42199	Intact
			Questionable	1660466-1696599	36134	Incomplete
			Intact	1976301-2015895	39595	Intact
			Questionable	2109819-2138908	29090	Incomplete
	NZ9000	NC_017949	Intact	25908-60678	34771	Incomplete
			Intact	583073-616375	33303	Incomplete
			Intact	776265-822742	46478	Intact
			Incomplete	862536-873116	10581	Incomplete
			Incomplete	1310854-1336625	25772	Incomplete
			Intact	2061853-2108338	46486	Intact
			Questionable	2204031-2238484	34454	Incomplete
	A76	NC_017492	Questionable	485882-509826	23945	Incomplete
			Questionable	622605-662110	39506	Intact
			Incomplete	854181-864183	10003	Incomplete
			Intact	958351-993297	34947	Incomplete
			Intact	1396426-1408335	11910	Incomplete
			Intact	1971765-2014239	42475	Intact
			Questionable	2111391-2125958	14568	Incomplete
			Intact	2111920-2190374	78455	Incomplete
			Incomplete	2352929-2381466	28538	Incomplete
	UC509.9	NC_019435	Incomplete	1372116-1378862	6747	Incomplete
	KW2	NC_022369	Intact	1878426-1919139	40714	Intact
lactis	184		Questionable	28101-54060	25960	Incomplete
			Questionable	154591-175458	20868	Incomplete
			Incomplete	347415-363059	15645	Incomplete
			Intact	524039-567880	43842	Intact
			Intact	586126-607576	21451	Incomplete
			Intact	728038-767971	39934	Intact
			Intact	2080049-2093308	13260	Incomplete
			Incomplete	2169282-2199842	30545	Incomplete
	229		Intact	28102-56029	27928	Incomplete

		Intact	521533-563579	42047	Intact
		Questionable	753845-802465	48621	Intact
		Questionable	1163201-1190289	27089	Incomplete
		Intact	1328423-1377207	48785	Intact
		Intact	1902578-1932903	30326	Incomplete
		Intact	2089335-2129151	39817	Intact
lactis	275	Incomplete	28095-55010	26916	Incomplete
		Intact	471423-513897	42475	Intact
		Incomplete	1197985-1211911	13927	Incomplete
		Incomplete	1266386-1288094	21709	Incomplete
		Incomplete	1291245-1306284	15039	Incomplete
		Questionable	1849809-1871796	21988	Incomplete
		Incomplete	1912611-1956567	43957	Intact
		Questionable	2063557-2107092	43536	Intact
		Incomplete	2164503-2183534	19032	Incomplete
	C10	Intact	28092-53276	25185	Incomplete
		Intact	454052-472631	18580	Incomplete
		Incomplete	970790-1027408	56619	Intact
		Intact	1317704-1360928	43225	Intact
		Intact	1397295-1440769	43475	Intact
		Intact	1762798-1800375	37578	Incomplete
		Intact	1944983-1988623	43641	Intact
		Intact	2179908-2238998	59091	Intact
	UC06	Questionable	361041-414377	53337	Incomplete
		Intact	1080213-1120371	40159	Intact
		Intact	1129338-1159934	30597	Incomplete
		Intact	1884301-1927291	42991	Intact
		Intact	2083183-2105865	22683	Incomplete
	UC08	Incomplete	1690095-1719553	29459	Incomplete
		Incomplete	2279979-2300415	20437	Incomplete
	UC11	Incomplete	558102-578503	20402	Incomplete
		Incomplete	1149693-1164391	14699	Incomplete
	UC063	Intact	28095-53289	25195	Incomplete

		Incomplete	138756-162176	23421	Incomplete
		Intact	493605-513193	19589	Incomplete
		Incomplete	1676175-1718240	42066	Incomplete
		Intact	1861779-1912694	50916	Intact
		Intact	2097464-2139720	42257	Intact
		Intact	2161841-2206154	44314	Intact
		Questionable	2296161-2315914	19754	Incomplete
UC77		Intact	28111-56038	27982	Incomplete
		Intact	521544-563590	42027	Intact
		Intact	581834-603842	22009	Incomplete
		Intact	1055162-1097440	42279	Intact
		Intact	1623401-1672185	48785	Intact
		Questionable	1794408-1844579	50172	Intact
		Intact	1946942-1977267	30326	Intact
		Intact	2120957-2177555	56599	Incomplete
UL8		Intact	28101-53285	25185	Incomplete
		Incomplete	293250-312083	18834	Incomplete
		Intact	485553-504150	18598	Incomplete
		Intact	547240-587879	40640	Intact
		Intact	735539-773116	37578	Incomplete
		Intact	1096224-1139671	43448	Intact
		Incomplete	1487223-1538320	51098	Intact
		Intact	2008049-2052163	44115	Incomplete
		Incomplete	2259191-2268474	9248	Incomplete
		Questionable	2288790-2309454	20665	Incomplete
IL1403	NC_002662	Intact	28459-56386	27928	Incomplete
		Intact	442048-484094	42047	Intact
		Intact	502338-520485	18148	Incomplete
		Intact	1030421-1075411	44991	Intact
		Intact	1414112-1460426	46315	Intact
		Intact	1997699-2028705	31007	Incomplete
KF147	NC_013656	Incomplete	311989-324003	12015	Incomplete
		Questionable	1055159-1110009	54851	Intact

			Intact	1534073-1593606	59534	Intact
			Intact	2052627-2073606	20980	Incomplete
			Incomplete	2278255-2308949	30695	Incomplete
			Questionable	2501570-2524177	22608	Incomplete
	CV56	NC_017486	Intact	28450-53643	25194	Incomplete
			Intact	1013597-1061046	47450	Intact
			Questionable	1722092-1743035	20944	Incomplete
			Questionable	1883260-1937861	54602	Intact
			Intact	2145779-2187692	41914	Intact
			Incomplete	2270197-2287233	17037	Incomplete
	IO-1	NC_019435	Intact	1706646-1759771	53126	Intact
			Incomplete	1951310-1976355	25046	Incomplete
	NCDO 2118	NZ_CP009054	Intact	995741-1039125	43385	Intact
			Incomplete	1179864-1214211	34348	Incomplete
lactis			Questionable	1773540-1821085	47546	Intact
			Intact	2060517-2081495	20979	Incomplete
			Questionable	2458115-2480723	22609	Incomplete
	KLDS 4.0325	NC_022593	Incomplete	153649-177366	23718	Incomplete
			Questionable	320922-333276	12355	Incomplete
			Intact	506771-548782	42012	Incomplete
			Intact	955456-1007952	52500	Intact
			Intact	1717014-1754981	37968	Intact
			Intact	1906998-1920058	13061	Incomplete
			Intact	2072376-2121522	49147	Intact
			Incomplete	2203268-2230575	27308	Incomplete
			Intact	2336619-2393566	56948	Intact
			Incomplete	2464584-2483381	18798	Incomplete
			Incomplete	2527610-2552014	24405	Incomplete
	AI06	NZ_CP009472	Intact	285344-297987	12644	Incomplete
			Intact	1042997-1090892	47896	Intact
	SO	NZ_CP010050	Incomplete	28625-50595	21971	Incomplete
			Incomplete	901021-916134	15114	Incomplete
			Questionable	1041956-1088069	46114	Intact

Intact	1432907-1477592	44686	Intact
Incomplete	1900585-1921034	20450	Incomplete
Intact	2084506-2122279	37774	Intact

Supplementary Table S6.2: Abi sequences used to create Abi database for screening of lactococcal prophages.

Genbank accession	Product
gi|60461909|	abi (plasmid) [*Lactococcus lactis*]
gi|695269642|	abi [*Lactococcus lactis*]
gi|695261980|	abi [*Lactococcus lactis*]
gi|695261751|	abi [*Lactococcus lactis*]
gi|501454300|	abi [*Lactococcus lactis*]
gi|499170988|	abi [*Lactococcus lactis*]
gi|695262046|	abi [*Lactococcus lactis*]
gi|695261979|	abi [*Lactococcus lactis*]
gi|695262149|	abi [*Lactococcus lactis*]
gi|691500870|	abi [*Lactococcus lactis*]
gi|504894708|	abi [*Lactococcus lactis*]
gi|504894644|	abi [*Lactococcus lactis*]
gi|500161265|	abi [*Lactococcus lactis*]
gi|500159963|	abi [*Lactococcus lactis*]
gi|499429749|	abi [*Lactococcus lactis*]
gi|499429738|	abi [*Lactococcus lactis*]
gi|552525936|	Abi [*Lactococcus lactis* subsp. *lactis* Dephy 1]
gi|695198230|	phage abi (plasmid) [*Lactococcus lactis*]
gi|2865246|	phage abi (plasmid) [*Lactococcus lactis*]
gi|457140|	abi mechanism-related protein [*Lactococcus lactis*]
gi|499994905|	abortive phage infection protein [*Lactococcus lactis*]
gi|695197890|	abortive phage resistance protein (plasmid) [*Lactococcus lactis*]
gi|2072188|	abortive phage resistance protein (plasmid) [*Lactococcus lactis*]
gi|1304597|	abortive phage resistance protein (plasmid) [*Lactococcus lactis*]
gi|695197889|	abortive phage resistance protein (plasmid) [*Lactococcus lactis*]
gi|2072187|	abortive phage resistance protein (plasmid) [*Lactococcus lactis*]
gi|2765135|	abiN (abi gene) [*Lactococcus lactis*]
gi|10441471|	abi phage resistance protein abiU [*Lactococcus lactis* subsp. *lactis*]
gi|578496740|	abi bacteriophage resistance protein [*Lactococcus lactis* subsp. *cremoris* HP]
gi|578495886|	abi mechanism-related protein [*Lactococcus lactis* subsp. *cremoris* HP]
gi|413975337|	abi mechanism-related protein (plasmid) [*Lactococcus lactis* subsp. *cremoris* UC509.9]
gi|413975227|	abi mechanism-related protein (plasmid) [*Lactococcus lactis* subsp. cremoris UC509.9]
gi|525227584|	Putative phage abi [*Lactococcus lactis* subsp. lactis A12]
gi|695209020|	abort lactococcal phage infection AbiTii (plasmid) [*Lactococcus lactis*]
gi|695209019|	abort lactococcal phage infection AbiTi (plasmid) [*Lactococcus lactis*]
gi|24421167|	abort lactococcal phage infection AbiTii (plasmid) [*Lactococcus lactis*]
gi|24421166|	abort lactococcal phage infection AbiTi (plasmid) [*Lactococcus lactis*]
gi|32455447|	AbiK (plasmid) [*Lactococcus lactis*]
gi|312831083|	unnamed protein product; ORF24 similar to abi K of *Lactococcus lactis* domain protein (plasmid) [*Staphylococcus aureus* subsp. *aureus* ECT-R 2]
gi|312831082|	unnamed protein product; ORF24 similar to abi K of *Lactococcus lactis*

	domain protein (plasmid) [*Staphylococcus aureus* subsp. *aureus* ECT-R 2]
gi|14251228|	AbiK (plasmid) [*Lactococcus lactis*]
gi|32455435|	AbiQ (plasmid) [*Lactococcus lactis*]
gi|695212062|	AbiA (plasmid) [*Lactococcus lactis*]
gi|4079668|	AbiQ (plasmid) [*Lactococcus lactis*]
gi|639891|	AbiA (plasmid) [*Lactococcus lactis*]
gi|695197413|	AbiF from pNP40 (plasmid) [*Lactococcus lactis*]
gi|60392783|	AbiD1 (plasmid) [*Lactococcus lactis*]
gi|1039480|	AbiF from pNP40 (plasmid) [*Lactococcus lactis*]
gi|695197410|	AbiEii (plasmid) [*Lactococcus lactis*]
gi|695197409|	AbiEi (plasmid) [*Lactococcus lactis*]
gi|149360|	abiC, partial [*Lactococcus lactis*]
gi|1039477|	AbiEii (plasmid) [*Lactococcus lactis*]
gi|1039476|	AbiEi (plasmid) [*Lactococcus lactis*]
gi|149358|	abi829 [*Lactococcus lactis*]
gi|1616605|	abiH [*Lactococcus lactis*]
gi|695198026|	abiI (plasmid) [*Lactococcus lactis*]
gi|2304799|	abiI (plasmid) [*Lactococcus lactis*]
gi|190571770|	AbiF (plasmid) [*Lactococcus lactis*]
gi|108736169|	AbiF (plasmid) [*Lactococcus lactis*]
gi|190571774|	AbiEi (plasmid) [*Lactococcus lactis*]
gi|190571773|	AbiEii (plasmid) [*Lactococcus lactis*]
gi|501454304|	AbiEi [*Lactococcus lactis*]
gi|501454303|	AbiEii [*Lactococcus lactis*]
gi|108736173|	AbiEi (plasmid) [*Lactococcus lactis*]
gi|108736172|	AbiEii (plasmid) [*Lactococcus lactis*]
gi|15674277|	abi phage resistance [*Streptococcus pyogenes* M1 GAS]
gi|33575906|	abi phage resistance protein [*Bordetella bronchiseptica* RB50]
gi|13621356|	abi phage resistance [*Streptococcus pyogenes* M1 GAS]
gi|17366546|	RecName: Full=Abortive phage resistance protein AbiGii
gi|416568|	RecName: Full=Abortive phage resistance protein AbiC
gi|1405404|	AbiGi [*Lactococcus lactis* subsp. cremoris]
gi|1405405|	AbiGii [*Lactococcus lactis* subsp. cremoris]
gi|695197296|	AbiD (plasmid) [*Lactococcus lactis* subsp. lactis]
gi|705395|	AbiD (plasmid) [*Lactococcus lactis* subsp. lactis]
gi|288547034|	CAAX amino terminal protease family protein
gi|17366543|	RecName: Full=Abortive phage resistance protein AbiGi
tr|O06042|	Abortive phage resistance protein
gb|AAB53711.1|	abortive phage resistance protein [*Lactococcus lactis*]
gb|AAC15900.1|	phage abi [*Lactococcus lactis*]
gb|AAN60762.1|	abort lactococcal phage infection AbiTi [*Lactococcus lactis*]
gb|AAN60763.1|	abort lactococcal phage infection AbiTii [*Lactococcus lactis*]
ref|WP_032398699.1|	AbiZ [*Lactococcus lactis*]
ref|WP_058206056.1|	hypothetical protein [*Lactococcus lactis*]

Chapter VII

General Discussion

The overall goal of the work described in this thesis was to assess the potential application of comparative and functional genomics in the selection of starter cultures, for example for optimum flavour production in particular cheeses (such as half-fat and/or low-salt cheese). This was approached via a focused genomic analysis of four strains of *Lactococcus lactis* that are known to produce good quality half-fat reduced salt cheese.

The research described in Chapter II provides an in-depth functional analysis of twenty *L. lactis* strains with particular emphasis on performance in terms of growth and autolysis coupled to cheese flavour development characteristics of lactococcal starter cultures. The functional characteristics of these strains generated selection criteria to screen candidate strains for whole genome sequencing. Chapter III describes the whole genome sequencing of sixteen *L. lactis* isolates; doubling the number of finished quality lactococcal genomes currently available in public databases. A comparative genomic investigation of the chromosomes of the sixteen strains sequenced in the context of this study and a further fourteen finished quality genomes available from the NCBI database was conducted with particular emphasis on dairy niche adaptations. Chapter IV describes the current lactococcal plasmidome and the discovery of the first lactococcal megaplasmids. In chapter V the restriction modification systems and associated methylome of sixteen *L. lactis* strains are investigated with the aid of single molecule real time sequencing, identifying a novel Type I shufflon RM system. Chapter VI represents the largest analysis to date of integrated lactococcal prophages, resulting in the identification of fifty nine intact and one hundred and six incomplete prophage regions within the thirty genomes assessed. This work also aided in the identification of an additional P335 phage

lineage, thus expanding knowledge on the diversity of this industrially significant phage group.

Lactococcal starter strains are a fundamental element of the dairy industry and consequently have been the focus of significant research interest. Commercial suppliers and producers are constantly looking to expand their product portfolios and overcome issues of phage sensitivity to meet both economic, production and consumer demands. Therefore, there is an ever-increasing demand to improve technologies for the selection of novel starter culture blends. Single molecule real time sequencing presents a promising new approach through whole genome sequencing and functional genome analyses; for the rapid identification and selection of such strains.

The contribution of lactococcal starter strains to cheese flavour development is predominantly through the major flavour pathways of lactose, lactate and citrate metabolism, lipolysis, proteolysis and the catabolism of free amino acids [1]. Functional analysis of the lactococcal starter strains in this study focused on assessing their performance in terms of these flavour-associated pathways. The subspecies divide between *lactis* and *cremoris* was found to be fundamentally important in terms of Cheddar production. The typical cooking temperatures used in Cheddar cheese fermentations is suitable for inducing temperature-controlled autolysis in *cremoris* strains but not *lactis* due to their higher thermal tolerance [2]. It suggests that *cremoris* strains are very suitable for Cheddar production, whereas the typical representatives of subspecies *lactis* strains are less appropriate for this purpose.

It was found in chapter II that subspecies *cremoris* strains elicit the highest overall enzymatic activity levels (in terms of aminotransferase and peptidases; pepX,

pepA, pepN/C), but the variations in enzyme profiles suggest that a number of subspecies *lactis* strains contribute a wider variety of flavour characteristics. It is widely accepted that the original niche environment of *L. lactis* is plant-based [3-5] and that the majority of dairy strains in use today are derived from a small number of closely related lineages. Therefore, plant-based *lactis* strains appear to offer the greatest possibilities for the expansion of flavours and novel types of dairy products. Since their performance is not ideally suited to current Cheddar type fermentations, an interesting avenue of investigation would be the transfer of peptidases or unique carbohydrate/lipolytic characteristics from *lactis* to *cremoris* strains utilising "natural" food-grade transfer mechanisms such as conjugation or transduction

Phenotypic analysis of four lactococcal starter cultures used in the Irish dairy industry for the production of low-fat Cheddar cheese allowed for the selection of potential novel starter cultures from the UCC starter culture collection (12 subsp. *cremoris* and 8 subsp. *lactis* strains were assessed) which may be useful for this type of dairy fermentation. The *L. lactis* subsp. *cremoris* strains were found to perform in a similar manner to the industrial isolates JM1-JM4. In particular strains 158 and UC109 presented with very similar growth and enzymatic profiles. Consequently these strains were selected for whole genome sequencing in addition to the industrial strains JM1-JM4, to further investigate their genetic composition.

The genome sequencing of sixteen novel lactococcal isolates has doubled the number of complete finished quality lactococcal genomes available and allowed for large-scale comparative analysis of the complete metabolic systems of the taxon. Our analysis clearly identifies a phylogenetic division between subspecies *lactis* and *cremoris*. This subspecies division is corroborated by hierarchical clustering based on both carbohydrate and amino acid metabolism, which indicates

two main subgroups that correspond to each subspecies. These observations support those of Cavanagh and colleagues, who recently proposed a re-evaluation of the taxonomic group separating *L. lactis* into two distinct species *L. lactis* and *L. cremoris* based on ANI (average nucleotide identity) and TETRA (tetranucleotide frequency correlation coefficients) [6]. The genomes of *L. lactis* subsp. *cremoris* were found to contain a higher number of pseudogenes in comparison to their *L. lactis* subsp. *lactis* counterparts, on average 100 per strain compared to 31 per strain, respectively. The vast majority of these strains are isolated from the dairy niche where genome decay and redundancy is widely reported [6, 38, 40], and believed to be due to continuous growth in milk.

To evaluate current sequencing efforts of the *L. lactis* taxon and to determine if additional genome sequencing is necessary to provide a complete overview of the chromosomal diversity of this taxon, the pan-genome of *L. lactis* was calculated and found to constitute 5906 genes. The deduced pan-genome of *L. lactis* was found to be closed, indicating that the representative data sets employed for this analysis are sufficient to fully describe the genetic diversity of the taxon. The core genome was also calculated, indicating a core genome size of 1129 genes.

It was determined that although strains can be clustered genotypically based on their subspecies and common niche, in agreement with a previous study [7], many of the flavour-related peptidases for which functional data are available, exist in single copy in the majority of lactococcal genomes. Therefore, it may not always be possible to make the genotype-phenotype link without the involvement of transcriptome and/or metabolome-based studies. It is therefore expected that continued work in this area would focus on complementing the genomics data with microarray or RNA-seq based analysis to gain a deeper understanding of these links.

The current study has provided a greater number of genome sequences and targets on which such platforms could be readily designed.

Niche adaptation also relies heavily on the acquisition of new metabolic capabilities as well as the loss of unnecessary functions. The introduction of niche-specific adaptations via plasmid acquisition, such as lactose and citrate metabolism has been extensively studied in *L. lactis* in view of their role in dairy niche adaptation [6, 8-11]; however, chromosomal adaptations are largely under-represented by comparison. Interestingly, the division between plasmid- and chromosome-based traits is becoming less clear as multiple integration events within the lactococcal chromosome suggests a more fluid genome than previously thought [8].

Genome sequencing of sixteen *L. lactis* strains revealed the presence of a total of sixty-seven plasmids, including two megaplasmids. Comparative genomic analysis of these sequences combined with those of publicly available plasmids (eighty one publicly available) allowed the definition of the lactococcal plasmidome. The lactococcal pan-plasmidome calculation constituted 1129 CDSs and indicated that the pan-plasmidome remains in a fluid or open state, and continued plasmid sequencing efforts are therefore expected to further expand the observed genetic diversity among lactococcal plasmids.

There has been limited research performed to date in the area of lactococcal gut adhesion as *L. lactis* is not commonly associated with the human gut. In Chapter IV, potential gut adhesion factors were identified within the lactococcal plasmidome, a key trait for persistence in the gastrointestinal tract. This may offer further insights into the potential application of *L. lactis* as a vector for vaccine and biomolecule delivery, a rapidly growing area of research [12, 13] or indicate probiotic potential

and/or functional food applications. It is envisioned that evolving nutritional preferences will further influence a trend towards such products and the identification of such traits within industrial strains may have beneficial implications for both industry and human health.

Discovery of the first lactococcal megaplasmids along with a host of novel features is evidence that the diversity of the lactococcal plasmidome is a relatively untapped resource, coinciding with evidence of an open or fluid plasmidome suggests that continued future sequencing will increase the observed diversity carried by these elements, leading to new avenues of research. The previously calculated lactococcal pan-plasmidome by Ainsworth et al. constituted the complete plasmid complement if eight strains and also resulted in an open pan-plasmidome [8].

Abi systems confer defence against phage infection and are commonly found in lactococcal strains where they are frequently plasmid-encoded [14]. Analysis of the plasmid sequences identified fourteen plasmid-encoded Abi systems while further analysis also identified frequent occurrences of these systems within the lactococcal chromosomes [15]. The presence of these systems combined with a host of R/M systems is evidence of the adaptation of these strains towards phage-resistance. The study of phage-resistance mechanisms will continue to be a valuable avenue of investigation as phages constitute one of the single greatest threats to dairy fermentations.

SMRT sequencing may be employed for the identification of methylated DNA bases and their associated motifs. Methylome analysis of the lactococcal strains sequenced in the framework of this study was applied to identify methylation motifs that are linked to Type I and Type II R-M systems. Comparative analysis of the lactococcal isolates in this study indicates a large degree of divergence in the

encoded R-M systems present in each of the strains. This is also indicative of their phage defence capabilities. *L. lactis* C10 and UL8 which apparently do not encode R-M systems contain five and three complete integrated prophages, respectively, while strains *L. lactis* JM1 and JM2 which encode significantly more R/M systems present with one complete integrated prophage each.

L. lactis JM1 and JM2 present an adaptive phage response in the form of a plasmid-encoded shufflon system. Analysis of the megaplasmids pMPJM1 and pMPJM2 resulted in the identification of a novel Type I shufflon R/M system. This system, the first of its type in *L. lactis,* is composed of multiple *hsdS* subunits arranged around a recombinase-encoding gene allowing for the intergenic shuffling of specificity subunits, resulting in an effective adaptive defence mechanism against phage infection. This system appears to be novel to these strains, though other systems with similar predicted functions have previously been reported in LAB [16].

SMRT technology allows a rapid assessment of a strain's abilities to withstand foreign DNA, or in contrast to assess the ease with which a strain may be transformed in a laboratory setting. In Chapter V, overlapping motifs and clustering of REases and MTases identified the systems responsible for ten of the 49 detected motifs; however, with increased data sets and continued improvements in sequencing coverage, it is envisioned that significantly more of these systems will be elucidated in the future.

Lactococcal phages persist as a major threat to commercial fermentation processes required for the manufacture of dairy products, particularly cheese. While lactococci are prone to infection by lytic phages, the threat of prophage induction and concomitant cell lysis presents an equally challenging risk factor. Thirty lactococcal genomes were explored for potential prophage-encoding regions using

the PHAST software followed by manual assessment and improvement. This resulted in the identification of fifty nine possibly intact prophages and one hundred and six incomplete prophage regions in total. Phylogenetic analysis of the nucleotide sequence of all such prophage regions combined with the sequences of previously sequenced P335 phages revealed two major groups of lactococcal prophages with a third minor group composed of a newly identified genetic lineage of prophages. The phylogenetic classification of the complete lactococcal prophage in this study is in agreement with a previous study [17] and previous P335 type phage sub-groupings [18] with the addition of one new sub-group V.

Prophages in this study were found to have limited inducibility which represents a relatively low direct risk to cheese production processes but their potential to expedite the evolution of virulent phages and the fitness of the host are key features that should be considered when selecting starter cultures. It is expected that rapid turnaround time on modern genome sequencing methods combined with the reduced costs will endorse the continued and vastly increased availability of lactococcal genomes permitting advanced assessments of prophage distribution, diversity and evolution, information that will be crucial for the selection of genome-informed next generation starter cultures.

Finally, from the inception of this project, the primary goal was to establish a methodology for the selection of novel dairy starter cultures applicable to low fat Cheddar cheese fermentations. Functional and comparative genomic analysis with four industrial isolates, *L. lactis* JM1-JM4, permitted the selection of similarly performing strains. The closest performing strain *L. lactis* subsp. *cremoris* 158 was selected for large scale cheese trials, with professional cheese grading later applied. The results of these trials indicated a Cheddar cheese with smooth texture reported as

“good” overall and most notably it would not be apparent to a consumer that it represented Cheddar which was low in fat and salt. The results of the cheese trial are encouraging in the context of the current work and offer validation to genome sequencing as a useful tool for assessing dairy strain collections. A secondary, perhaps more useful advantage of this system is the ability to predict divergent traits within strains which may lead to expanded starter diversity in the future.

The research presented in this thesis provides a solid foundation for further investigations into the comparative and functional genomics of *L. lactis*. The availability of a significantly enlarged *L. lactis* genomic data base will allow detailed studies into all aspects of *L. lactis* genomics. This work has also highlighted the potential of next generation sequencing technologies for applications such as strain screening/selection and the investigation of phage-host interactions; in terms of a so-called arms race between lytic phages and host defence mechanisms and the double-edged association of integrated prophages.

The significant proportion of unassigned and hypothetical plasmid-encoded proteins presents a wealth of available avenues for further studies to explore, which will undoubtedly result in the elucidation of novel traits in the future. Further methylome studies with an increased strain pool is likely to reveal more novel R/M systems and aid in the identification of their associated motifs, which has far-reaching potential implications within molecular biology.

In conclusion, the work presented in this thesis significantly increases our knowledge of the *L. lactis* taxon and is expected to lead to the development of strategies to expand and diversify lactococcal starter stains used for dairy fermentations.

References:

1. McSweeney PLH: Biochemistry of cheese ripening. Int J Dairy Technol. 2004, 57(2-3):127-144.
2. Harrigan WF: Laboratory methods in food microbiology (3rd ed). Academic Press, San Diego, CA 1998.
3. Siezen RJ, Bayjanov JR, Felis GE, van der Sijde MR, Starrenburg M, Molenaar D, Wels M, van Hijum SA, van Hylckama Vlieg JE: Genome-scale diversity and niche adaptation analysis of *Lactococcus lactis* by comparative genome hybridization using multi-strain arrays. Microbial Biotechnol. 2011, 4(3):383-402.
4. Price CE, Zeyniyev A, Kuipers OP, Kok J: From meadows to milk to mucosa – adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. FEMS Microbiol Rev. 2012, 36(5):949-971.
5. Cavanagh D, Fitzgerald GF, McAuliffe O: From field to fermentation: the origins of *Lactococcus lactis* and its domestication to the dairy environment. Food Microbiol. 2015, 47:45-61.
6. Cavanagh D, Casey A, Altermann E, Cotter PD, Fitzgerald GF, McAuliffe O: Evaluation of *Lactococcus lactis* Isolates from Nondairy Sources with Potential Dairy Applications Reveals Extensive Phenotype-Genotype Disparity and Implications for a Revised Species. Appl Environ Microbiol. 2015, 81(12):3961-3972.
7. Guédon E, Renault P, Ehrlich SD, Delorme C: Transcriptional Pattern of Genes Coding for the Proteolytic System of *Lactococcus lactis* and Evidence for Coordinated Regulation of Key Enzymes by Peptide Supply. J Bacteriol. 2001, 183(12):3614-3622.

8. Ainsworth S, Stockdale S, Bottacini F, Mahony J, van Sinderen D: The *Lactococcus lactis* plasmidome: much learnt, yet still lots to discover. FEMS Microbiol Rev. 2014, 38(5):1066-1088.

9. van Rooijen RJ, De Vos W: Molecular cloning, transcriptional analysis, and nucleotide sequence of *lacR*, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. J Biol Chem. 1990, 265(30):18499-18503.

10. Van Rooijen R, Gasson M, De Vos W: Characterization of the *Lactococcus lactis* lactose operon promoter: contribution of flanking sequences and LacR repressor to promoter activity. J Bacteriol. 1992, 174(7):2273-2280.

11. Drider D, Bekal S, Prévost H: Genetic organization and expression of citrate permease in lactic acid bacteria. Genet Mol Res. 2004, 3(2):271-281.

12. Bermúdez-Humarán LG, Aubry C, Motta J-P, Deraison C, Steidler L, Vergnolle N, Chatel J-M, Langella P: Engineering lactococci and *lactobacilli* for human health. Curr Opin Microbiol. 2013, 16(3):278-283.

13. Bermúdez-Humarán LG: *Lactococcus lactis* as a live vector for mucosal delivery of therapeutic proteins. Hum Vacc. 2009, 5(4):264-267.

14. Mills S, McAuliffe OE, Coffey A, Fitzgerald GF, Ross RP: Plasmids of lactococci – genetic accessories or genetic necessities? FEMS Microbiol Rev. 2006, 30(2):243-273.

15. Chopin M-C, Chopin A, Bidnenko E: Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol. 2005, 8(4):473-479.

16. Claesson MJ, Li Y, Leahy S, Canchaya C, van Pijkeren JP, Cerdeño-Tárraga AM, Parkhill J, Flynn S, O'Sullivan GC, Collins JK et al: Multireplicon

genome architecture of Lactobacillussalivarius. Proc Nat Acad Sci. 2006, 103(17):6718-6723.

17. Ventura M, Zomer A, Canchaya C, O'Connell-Motherway M, Kuipers O, Turroni F, Ribbera A, Foroni E, Buist G, Wegmann U: Comparative analyses of prophage-like elements present in two *Lactococcus lactis* strains. Appl Environ Microbiol. 2007, 73(23):7771-7780.

18. Mahony J, Martel B, Tremblay DM, Neve H, Heller KJ, Moineau S, van Sinderen D: Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. Appl Environ Microbiol. 2013, 79(14):4401-4409.

Appendix A

Large-scale cheese fermentation trial results

Large scale Cheddar cheese manufacturing trial was conducted by Prof. Tim Guinee and Catherine McCarthy at the MTL facility at Teagasc Rood Research Centre Moorepark. Cheese grading was performed by Enda Howley of Kerrygold.

To assess the functional and comparative genomic method for the selection of lactococcal starter strains for reduced fat/salt Cheddar cheese production, a candidate strain was used to perform a large scale cheese trial. *L. lactis* subsp. *cremoris* 158 was selected based on functional and genetic similarities to industrial comparators *L. lactis* JM1-JM4. The main functional similarities (Chapter II) are summarized in (Fig. A1) and the main genetic similarities (Chapter III & IV) are summarised in (Table. A1).

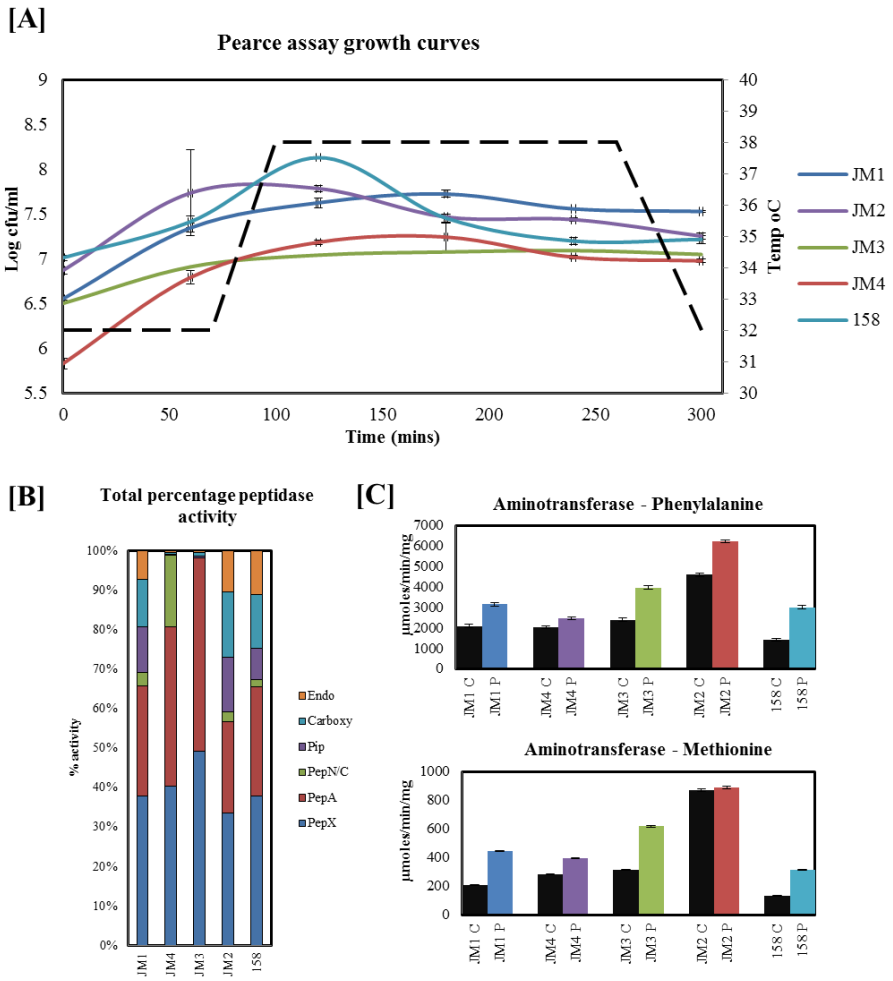


Figure A1: Overview of main performance and flavour based selection criteria

Summary of functional analysis (Chapter I); [A] Growth performance under simulated cheese fermentation conditions, temperature profile is indicated by black dashed-line [B] Overview of peptidase activities as normalised percentage of total strain activity, [C] Amino acid transferase activity utilising phenylalanine and methionine based substrates for four commercial starters and strain 158.

Table A1: Overview of general genome features of representative *L. lactis* genomes

Strain	158	JM1	JM2	JM3	JM4
Genome length (Mbp)	2250	2397	2374	2454	2380
CDS	2078	2308	2316	2411	2293
tRNA features	60	60	58	59	60
rRNA features	19	19	19	19	19
Hypothetical proteins %	17.9	20.5	19.6	23.7	20.9
Assigned function %	81.1	79.5	80.4	76.3	79.1
Pseudo genes	106	74	68	60	88
IS elements/ transposases	150	243	167	163	181
Prophage	2 Re	1 In 6 Re	1 In 3 Re	2 In 3 Re	1 In 4 Re
Plasmids	6	7	4	5	5
Plasmid complement (Kbp)	235.8	355.1	200.7	185.5	121.8
GC %	35.88	36.01	35.8	35.87	35.83

The cheese trial was performed at Teagasc Food Research Centre, Moorepark in 400 litre vats, utilising; semi-skimmed milk, camel trypsin, under lactate buffered conditions with a standard protocol optimised for reduced fat/salt Cheddar cheese. A commercially supplied Cheddar cheese starter was used as a control. The results of the starter performance during cheese manufacture are described in (Table A2).

Table A2: Overview of Cheese trial cooking temperatures, time and pH

	Vat 1 – Control	**Vat 2 – *L. lactis* 158**
Milk	454.1 kg / pH 6.57	456 kg / pH 6.57
Acidity to 6.1	Lactate buffered	Lactate buffered
Starter Added	0 mins	0 mins
Rennet	40 mins / pH 5.98	40 mins / pH 6.02
Cut / Finish Cut	52 mins	51 mins
Cooking 31-38.5ºC	88 mins / pH 6.01	90 mins / pH 5.99
Drain @ pH 7.5	142 mins / pH 5.95	139 mins / pH 5.88
Trench Slab Turn (x3)	232 mins / pH 5.66	184 mins / pH 5.69
Milk @ 5.3-5.35	345 mins / pH 5.38	301 mins /pH 5.36
Weight of Curd	39.94 kg	43.12 kg
Salt Addition (1%)	0.39 kg	0.43 kg
Mellow	365 mins	321 mins
Press	400 mins	366 mins

L. lactis 158 performed comparably with the commercial starter strain and delivered a high cheese yield. The manufactured Cheddar was matured at 8 °C for six months before grading. Cheeses were assessed at six months by a commercial grader from a local Cheddar factory, who was previously informed that the cheeses were half-fat [1]. The cheese grader reported that the Cheddar produced had "smooth texture", "good cheese", "tastes like a traditional table Cheddar in that it has acidic notes, unlike the 'new' Cheddar sold today which are considered quite sweet" and notably "a consumer wouldn't know it is low in fat and salt".

References:

1. Fenelon M, Beresford T, Guinee T: Comparison of different bacterial culture systems for the production of reduced-fat Cheddar cheese. Int J Dairy Technol. 2002, 55(4):194-203.

Acknowledgements

Firstly I'd like to thank Douwe for giving me the opportunity to undertake my PhD. Thank you for your encouragement, support, guidance and seemingly endless patience throughout my PhD. It has been a pleasure to work under your supervision.

To Jenn, I can't thank you enough for being both a mentor and a friend. For always being free to listen to my ideas and always being willing to help with anything and everything regardless of how busy you were, I wouldn't have gotten this far without you.

To all of the Cheeseboard collaborators who I had the pleasure of working with, in particular to Kieran, Tim and Catherine, thank you! To our international collaborator Richard, your experience and skills greatly added to the value of this thesis.

An enormous thank you to Francesca, who was fundamental to the completion of this project, teaching me almost everything I know about genomics (and becoming a lifelong friend), I can't thank you enough.

To lab friends old and new, hopefully I won't forget anyone! (I'll start with the oldies) Thank you all for making me feel so welcome, to James one of the funniest people I've ever had the pleasure of knowing, Eoghan who is the only other person I know who is almost as much of a cranky old man as me, Brian for the company on many a late nights struggle in 339, Joana my office and writing friend, Stuart, Stephen, Sarah, Barry, Anneleen, and anyone else I've forgotten, I'm sorry its old age! To our new lab friends many of whom feel like old, especially my favourite

two Americans Juan and Stephen! To Mona, Andrea, Tom, Nicolas, Angela, Joanna, Marine, and Ilias I hope I haven't been too torturous to be around for the last few months. To everyone in Douwe's other lab, who are a constant source of coffee, restriction enzymes, gossip and fun; Valentina, Amy, Muireann, Kieran, Ruth, Noreen, Emer, Sandra, and Rebecca, thank you all!

To everyone else I've had the pleasure of working with in the microbiology department, thank you, especially all the staff who make it all possible; Paddy, Dan, two Colettes, Aine, Maire, Julianna, Carmel, Hilda, Maurice, John, and Mary.

A big thank you to my closest friends Alan, Odile and Ruairi, who have been unlucky enough to endure all of my complaining and even worst my ramblings about sequencing in the pub, thank you so much! To all of my other friends for all of the cancelled plans, it is finally over!

Finally, to my family for all of your love and support throughout my life and in particular when I said I'm going back to college again, this thesis is for you. I really am finished in college this time!

Thanks!