Meta-analysis of cheese microbiomes highlights contributions to multiple aspects of quality

Aaron M. Walsh[a,b,+], Guerrino Macori[a,b,+], Kieran N. Kilcawley[a], and Paul D. Cotter[a,b,*].

[a]Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

[b]APC Microbiome Ireland, Cork, Ireland

[+]these authors contributed equally to this work

[*]corresponding author (email: paul.cotter@teagasc.ie)

ABSTRACT

A detailed understanding of the cheese microbiome is key to the optimization of flavour, appearance, quality and safety**.** Accordingly, we conducted a high resolution meta-analysis of cheese microbiomes and corresponding volatilomes. Using 77 new samples from 55 artisanal cheeses from 27 Irish producers combined with 107 publicly available cheese metagenomes, we recovered 328 metagenome assembled genomes, including 47 putative new species that could influence taste or colour through the secretion of volatiles or biosynthesis of pigments. Additionally, from a subset of samples, we found that differences in the abundances of strains corresponded with levels of volatiles. Genes encoding bacteriocins and other antimicrobials, such as pseudoalterin, were common, potentially contributing to the control of undesirable microorganisms. Although antibiotic resistance genes were detected, evidence suggested they are not of major concern with respect to dissemination to other microbiomes. Phage, a potential cause of fermentation failure, were abundant and evidence for phage-mediated gene transfer was detected. The anti-phage defence mechanism CRISPR was widespread and analysis thereof, and of anti-CRISPR proteins, revealed a complex interaction between phage and bacteria. Overall, our results provide new and significant technological and ecological insights into the cheese microbiome that can be applied to further improve cheese production.

INTRODUCTION

Given the essential role of microorganisms in the production of cheese, a detailed understanding of the microbiota of cheese is of considerable value with respect to the optimisation of cheese flavour, appearance, quality and safety. As cheese microbiotas are representative of low to medium complexity microbial communities, these populations are also of great fundamental interest, providing insights into microbial interactions that can translate to more complex communities. High-throughput sequencing has yielded invaluable insights into cheese microbiomes over the past decade [1], and recent studies have highlighted that this technology can further our understanding of the roles of microorganisms during the ripening of cheese [2-4].

Cheese also represents an environment in which microorganisms have evolved to adapt to abiotic stresses~~,~~ such as acidity or salinity, in addition to biotic stresses, such as competition or predation, and cheese microbiomes have also served as valuable model communities to study complex processes of relevance to more complex microbial communities, such as microbial community formation [5], microbial interaction [6], or horizontal gene transfer [7]. In this regard, shotgun metagenomic sequencing has proven to be a powerful tool for the characterisation of the microbiota of cheese and other fermented foods [8], providing taxonomic resolution to species and strain level, and functional information [9].

Importantly, shotgun metagenomics can also detect viruses, which is particularly relevant in studies of cheese, since phage infection is often deleterious from an industry perspective through its contribution to fermentation failure [10]. This technology, coupled with recent advances in bioinformatics, presents an unprecedented opportunity to further characterise the cheese microbiota in detail, including strain-level identification of microorganisms [11]. This is particularly important as strain-level variation among starters is already known to influence the flavour of cheese [12], and likely also contribute to other important features. In addition, it

has previously been reported that microorganisms that are not typically associated with cheeses might contain genes important for flavour [5], spoilage [13] and other attributes, and the ability to assemble the near complete new genomes of such microorganisms by metagenome binning has the potential to be of great value. However, to date, this approach has not been applied to cheese, or indeed other food, communities.

Here, a combination of tools is used to characterise new metagenomes corresponding to 55 cheeses from 27 Irish producers, in addition to 107 publicly available cheese metagenomes. We identify strain-level variations in cheeses that correspond with differences in the volatilome, conduct CRISPR analysis-based reconstruction of the history of phage infections, investigate the distribution and potential for transfer of antibiotic resistance genes, and provide evidence of the widespread distribution of bacteriocin- and other antimicrobial-associated genes in cheeses.

RESULTS

Taxonomic profiling highlights the prevalence of phage in the cheese microbiome

The taxonomic profiling of 184 cheese samples (Figure 1A) detected 23 species at ≥0.1% relative abundance in ≥10% of samples (the defined threshold of prevalence in this study). Overall, the mean abundance of bacteria, eukaryota, and viruses was 78.08%, 2.15%, and 19.76%, respectively (Figure 1B). Notably, given the importance of phage in cheese fermentations, particularly in relation to their potential to negatively impact on fermenting bacteria, five of the prevalent species were phage from the *Siphoviridae* family, including an unclassified C2-like virus (43.5%), *Lactococcus* phage ul36 (17.9%), *Lactococcus* phage P680 (14.7%), *Lactococcus* phage BM13 (11.4%) and *Streptococcus* phage ALQ13.2 (10.9%). The 3 most prevalent bacterial species were *Lactococcus lactis* (78%), *Streptococcus thermophilus* (43.5%), and *Lactobacillus helveticus* (37%), all being lactic acid bacteria (LAB) from the order Lactobacillales and well known cheese microorganims. Overall, LAB had a prevalence of 91.3%. The three most prevalent non-LAB bacterial species were *Brevibacterium linens* (28.9%), *Staphylococcus equorum* (28.9%), and an unclassified *Brachybacterium* species (22.8%). Notably, two other unclassified species, i.e., a *Halomonas* sp. (16.8%) and a *Brevibacterium* sp. (10.9%), were prevalent. The only prevalent eukaryotic species was *Debaryomyces hansenii* (35.9%).

Here, we identify a nonlinear correlation ($p=0.007$, $R=0.523$) between the relative abundance of *Siphoviridae* and *Streptococcaceae* (Figure 1C). Specifically, when *Siphoviridae* were present below 15% relative abundance, they positively correlated with *Streptococcaceae*, but when *Siphoviridae* were present above this level, a negative correlation existed, thereby highlighting a threshold above which phage are likely to inhibit these important LAB. Further

details relating to the taxonomy of the newly sequenced Irish cheese microbiome can be found in the Supplementary notes, which include Supplementary Figs. 1-5, and Supplementary Tables 1-5.

Strain-level variation corresponds to differences in the volatilome

Across 48 of the newly studied cheese samples, a total of 63 volatiles that may contribute to flavour were detected by GC-MS (Supplementary Table 6). This was combined with data from a previous study [3] to identify correlations between strain abundance and volatile levels. We detected a total of 7 species that were present in both datasets at >1% relative abundance in ≥12 samples and the corresponding strains were identified in each case. Only strains that were present in ≥6 samples were considered for subsequent correlation analysis. These strains were *B. linens* strains GCF_001729525 (n=9) and GCF_002332445 (n=7), *L. casei paracasei* strains GCF_000194765 (n=6) and GCF_003957435 (n=17), *Lactobacillus plantarum* strains GCF_00469115 (n=7) and GCF_00473935 (n=7), *L. lactis* strains GCF_000006865 (n=57) and GCF_900240895 (n=18), and *S. thermophilus* strains GCF_000253395 (n=8) and GCF_000836675 (n=39). The abundances of these strains were inferred from the abundances of the corresponding species. Overall, 32 volatile compounds were shared between both datasets (Supplementary Table 7). Analysis of the effect of strain level variation on the volatilome revealed that *B. linens* strain GCF_001729525 (p=0.002, R=0.586), *L. lactis* strain GCF_900240895 (p=0.002, R=0.158), and the *S. thermophilus* strains GCF_000253395 (p=0.008, R=0.201), and GCF_000836675 (p=0.001, R=0.3) all significantly correlated with variance in the volatilome (Figure 2A). Correlation coefficient analysis revealed that 9 volatiles that were positively correlated with *B. linens* GCF_001729525 were negatively correlated with *B. linens* GCF_002332445 (Figure 2B). Additionally, significant differences in the associations between strains of the same species to volatiles were identified (Figure 2C).

Newly characterized bacteria may contribute to cheese quality

A total of 924 metagenome assembled genomes (MAGs), which represent actual individual genomes based on the binned metagenomes, were identified. We focused on 328 high quality MAGs (over 80% complete with less than 10% contamination) for subsequent analysis (Figure 3A; Supplementary Table 8). Overall, taxonomic identification revealed that 186 MAGs were Firmicutes, 89 were Actinobacteria, and 53 were Proteobacteria (Figure 3B, 3C). 105 of the MAGs were not assigned to a species, with 47 MAGs having less than 95% average nucleotide identity (ANI) to reference genomes and fell into 22 primary clusters (PCs), that represent up to 22 putative novel species. Based on their ANI to reference genomes, they may belong to the genera *Psychrobacter* (21 MAGs belonging to 9 PCs), *Brachybacterium* (8 MAGs belonging to 2 PCs), *Corynebacterium* (6 MAGs belonging to 3 PCs), *Brevibacterium* and *Halomonas* (each with 3 MAGs belonging to 2 PCs), *Advenella*, *Arthrobacter*, *Idiomarina*, *Proteus*, *Streptomyces* and *Vibrio* (1 MAG each).

Genome-scale metabolic models for the 47 MAGs from putatively novel species predicted that the following metabolites that may influence flavour, i.e., acetate, succinate, and lactate and ammonium, (Figure 3D), were secreted by ≥10 of these MAGs. We also assess the prevalence of genes encoding carotenoids, i.e., pigments that contribute to the appearance of cheese, across MAGs and identified 246 hits to carotenoids genes across 58 MAGs (BLAST *E*-value ≤1e-05) (Figure 3E). The majority of hits to carotenoid genes were found in *Brevibacterium* MAGs (52.4% of hits), followed by *Glutamicibacter* (32.9% of hits). Overall, genes associated with the biosynthesis of 14 types of carotenoid were detected. The most widely distributed carotenoid genes were involved in the biosynthesis of lycopene, neurosporene, phytofluene, and zeta-carotene; these genes were detected across 5 genera.

Previously, *Psychrobacter* has been reported to cause purpling of cheeses, and it was proposed that this defect was through the conversion of indole to indigo [14]. Therefore, we assessed the presence of genes associated with the biosynthesis of indigo (Supplementary Table 9) and found that genes encoding indole-3-acetate monooxygenase (EC 1.14.13.235) were present on 13 of the 21 *Pyschrobacter* MAGs recovered in this study (Figure 3F). Indole-3-acetate monooxygenase catalyses the formation of indoxyl from indole, which then forms indigo upon reaction with oxygen and, thus, our study supports the previously provided hypothesis.

The frequency of antibiotic resistance genes is comparable to that ~~of~~ the human gut microbiome

We identified 40 ARGs (antibiotic resistance genes) across 35 MAGs in the cheese microbiome. These included lincosamide ARGs in *Streptococcaceae,* multidrug ARGs in *Moraxellaceae*, and fosfomycin ARGs in *Staphylococcaceae* (Figure 4A). We then investigated ARGs on plasmids to assess the potential for the transfer of ARGs to other microorganisms via conjugation. A total of 74 ARGs were detected on plasmids from 66 samples, and these ARGs were predicted to confer resistance to fosfomycin, phenicol, sulphonamide, diaminopyramidine, tetracycline or multidrug resistance (Figure 4A). These plasmid-associated contigs were assigned to the families *Enterobacteriaceae*, *Moraxellaceae*, *Staphylococcaceae,* and *Vibrionaceae* (Figure 4B). No ARGs were detected on LAB-associated plasmids, suggesting a limited potential for plasmid-mediated dissemination of ARGs by LAB. We also noted that four Proteobacteria-associated MAGs had integrons containing ARGs, including one multidrug ARG, and three phenicol ARGs.

Next, the frequency of ARGs among MAGs recovered from cheeses was compared to the publicly available MAGs recovered from human, ocean, and rumen samples. Following dereplication, the number of representative MAGs from each environment was 99 cheese MAGs; 4,929 human MAGs; 2,139 ocean MAGs; and 859 rumen MAGs. The percentage of representatives containing ARGs was as follows: 2.02% of cheese representatives; 2.25% of human representatives; 0.09% of ocean representatives; and 0.23% of rumen representatives (Figure 4E). We found no significant differences in ARG frequency between cheese and human representatives ($p=1$). However, ARGs were enriched among cheese representatives relative to ocean ($p=0.011$) and rumen ($p=0.055$) representatives.

Antimicrobial peptide genes are common among cheese microbes

Bacteriocins are ribosomally synthesised antimicrobial peptides produced by bacteria that can be regarded as a means of providing an innate immunity to foods against pathogens[15] (Cotter et al 2005). Here, we assessed the prevalence of bacteriocin genes within cheese-derived MAGs (Figure 4C). Overall, 210 sequences on 106 MAGs that were recovered from 71 samples were predicted to encode 25 Class II/III bacteriocins, and these MAGs belonged to 9 genera (Supplementary Table 10). Class II bacteriocins detected in ≥10 MAGs were lactococcin A-like, lactococcin B-like, BlpK-like bacteriocins and a putative bacteriocin in *Staphylococcus*. Class III bacteriocins detected in ≥10 MAGs were helveticin J-like, enterolysin A-like and linocin CFP29-like bacteriocins. It was found that 13 Class II/III bacteriocins had ≥2 variants, based on their amino acid sequences.

The percentage of representative MAGs from each environment that were found to contain Class II/III bacteriocin genes was as follows: 20.20% among cheese representatives; 12.94% of human representatives; 5.70% of ocean representatives; and 5.36% of rumen

representatives (Figure 4E). Bacteriocin genes were found to be enriched among cheese representatives relative to human (p=0.048), ocean (p=1.74e-06) and rumen (p=2.43e-06) representatives.

With respect to other antimicrobials, we found that variants of the metalloprotease, pseudoalterin, which has recently been shown to contribute to predator-prey interactions between marine Gram-negative bacteria and Gram positive bacteria [16], were present in 15 MAGs from the genera *Alcaligenes*, *Corynebacterium*, *Halomonas*, *Idiomarina*, *Marinomonas*, *Streptomyces*, and *Vibrio* (Figure 4D).

Evidence for phage- and transposon-mediated lateral gene transfer in the cheese microbiome

We identified 624 occurrences of LGT (lateral gene transfer) across 75 of the 184 samples analysed. Although LGTs were detected at each taxonomic rank, most occurred between members of the same order (Figure 5A). LGTs were most frequent between members of the order Lactobacillales, especially among *Streptococcaceae*. The mean LGT frequency varied across the datasets (Figure 5B), ranging from 2.44-e05 to 4.32e-04–LGTs/kb. Interestingly, the frequency of LGTs was lowest in those datasets that included rind samples, which may reflect the lower overall relative abundance of Lactobacillales in these samples. Of the 132 instances where the LGT direction could be determined, including 75 genes that had UniRef90 annotations, 21 were identified as phage replication gene A protein (GPA) (UniRef90_I6TNE5), while 2 were transposases (UniRef90_C8Q1E5 and UniRef90_W9EFK6). Phage-related and transposase genes were identified in 9% and 19% of loci, respectively, where the direction of LGT was not determined.

Elucidating patterns of phage infections in cheeses

The level of viruses detected was consistent with phage infection being common across the cheeses. Overall, viral signals in 88 MAGs were–hypothetically lysogenic, with 74% of *Lactococcus* MAGs and 6% of *Streptococcus* MAGs containing prophage (Figure 6A). The history of phage infections among bacteria in the cheeses was explored by analysing CRISPR spacers within the metagenome. Overall, 1894 putative CRISPRs were identified in 181 samples, and 102,407 putative spacers were found across these loci. 2,633 spacers from 381 CRISPRs had significant matches with viral genomes. Spacers homologous to sequences from viruses that infect the following genera were detected: *Acinetobacter* (20 spacers); *Aeromonas* (1); *Geobacillus* (2); *Halomonas* (2); *Lactobacillus* (149); *Lactococcus* (60); *Propionibacterium* (42); *Pseudoalteromonas* (1); *Psychrobacter* (3); *Streptococcus* (2,337); and *Thermus* (14) (Figure 6B). Additionally, 2 spacers were found to have ambiguous hits, meaning that they aligned to the genomes of two types of phage (i.e., *Lactococcus* ul36 and 3 *Streptococcus* phage). We established that 270 CRISPRs contained spacers that can be aligned to multiple phage genomes (Figure 6C), indicating that bacteria were exposed to multiple infections.

Overall, 138 putative CRISPRs were identified in 90 MAGs, and 1965 putative spacers were found across these loci. A total of 159 spacers from 57 CRISPRs had significant matches in the database to phage that infect *Streptococcus* (118 spacers), *Propionibacterium* (24 spacers), *Lactobacillus* (13 spacers), *Acinetobacter* (3 spacers), and *Staphylococcus* (1 spacer) (Figure 6B). A total of 32 of these MAG-associated CRISPRs contained spacers that aligned to multiple phage genomes (Figure 6C). All of the spacers from the CRISPRs on *Lactococcus* MAGs aligned to *Streptococcus* phage (Figure 6D). Similarly, a spacer from a

CRISPR on a *Staphylococcus* MAG aligned to a *Streptococcus* phage. These may reflect a horizontal gene transfer event, the existence of phage that have a broad target range (i.e., can target multiple taxa) or of closely related phage that, while resembling one another, target distinct genera.

CRISPRs on contigs that were predicted to be plasmids were also identified. 277 putative CRISPRs were identified on plasmids-with 967 spacers from 195 of these CRISPRs matching phage that infect *Streptococcus* (884 spacers), *Lactobacillus* (45), *Lactococcus* (30), *Thermus* (5), and *Propionibacterium* (3). A total of 155 CRISPRs identified on plasmids contained spacers that aligned to multiple phage genomes (Figure 6C). Interestingly, in each of the CRISPRs that contained spacers from *Lactococococcus* phage, the other spacers were from *Streptococcus* phage. Overall, spacers were found to have homology to 10 lactococcal phage genomes (Figure 6E). The presence of sequences from *Streptococcus* phage alongside sequences from *Lactococcus* phage in the same CRISPRs may represent horizontal gene transfer of CRISPRs from *Streptococcus* to *Lactococcus* via plasmids.

Metagenome assemblies were aligned against a database of 45 anti-CRISPRs (Acrs) to determine if phage in cheese possessed Acrs (Figure 6 F). Homologs to Acrs were detected in 54.9% of samples, and 4Acrs were detected in >10% of samples: AcrIIA6, AcrIIA7, AcrIIA3, and AcrVA2. These and 3 other Acrs, AcrIIA1 and AcrIF12, AcrIF3, were found in ≥2 datasets.

DISCUSSION

Studying the microbiota of cheese offers valuable insights into biotechnologically important processes, such as flavour formation, in addition to fundamentally important processes that shape microbiomes. Recent advances in bioinformatic tools facilitate the strain-level identification of microorganisms and the recovery of genomes from metagenomes [17]. Here, these approaches provide the opportunity to characterise the cheese microbiome in a manner equivalent to that of human gut, rumen and ocean microbiomes [18-22]. In this study, 328 MAGs were recovered from the 184 cheese metagenomes, including 47 MAGs that represent putatively novel species. Notably, the majority of these species were inferred to belong to halophilic genera (e.g., *Psychrobacter*, *Halomonas*) that have been detected in cheeses previously [5,23]. It is likely that these halophiles are introduced during cheese making and thus their presence is not unexpected. Aside from halophiles, a number of the putative novel species were inferred to belong to genera that are associated with the rind (i.e., *Brevibacterium*, *Corynebacterium*, and *Arthrobacter*; SI1). The recovery of MAGs of putative novel species from cheeses provides an additional opportunity to investigate if such species influence the flavour of cheeses. Metabolic modelling of the genomes predicted that these species secrete compounds that influence flavour (e.g., ammonium, acetate), but require validation. The distribution of genes associated with pigment production across different taxa was assessed, and *Psychrobacter* were found to encode an enzyme that converts indole to indoxyl, which then oxidises to form indigo. *Psychrobacter* had previously been isolated from cheeses that were discoloured purple [14] and the authors had proposed that this might be due to the production of indigo. Here, we identify a metabolic pathway that is likely responsible for this phenomenon.

Previous studies were conducted to identify the correlation between food microbiomes and metabolomes to examine the ways in which species might influence flavour [3,24]. However,

correlation between species and metabolites may mask the effects of strain level variations. Here, the integration of strain-level metagenomics with metabolomics indicated that differences in the abundances of strains did correspond to differences in the levels of volatiles. Specifically, strains from 3 species (i.e., *B. linens*, *L. lactis*, and *S. thermophilus*) demonstrated measurable differences in their associations with metabolites. This is consistent with recent *in situ* experiments demonstrating that manipulating the composition of strains in cheese influenced its metabolome [25]. We propose that the future combined use of strain-level metagenomics with metabolomics has the potential to expand our knowledge of the effects of strains on flavour, and may guide the selection and/or development of starters for cheese and other fermented foods.

A high abundance of phage-associated sequences were identified across the cheeses, accounting for a predicted 19.76% of the population, although their abundance did vary between samples. It is worth noting that the DNA extraction method used in this study was not tailored specifically for the isolation of phage, as was done elsewhere [26], and so the level of phage detected here might be an underrepresentation of the actual virome. Predation by phage is a factor that shapes the formation of the cheese microbiota [27], and infection of starters is the principal cause of fermentation failures during cheese production [28]. We observed that when the abundance of *Siphoviridae* was present above ~15%, these phage negatively correlated with the abundance of *Streptococcaceae*, which suggests that infection was ongoing. However, the bacteria in the cheeses also possessed defences against phage, with CRISPRs identified in the metagenome. We analysed the spacers from the detected CRISPRs to reconstruct the history of infections in the cheeses and found spacers that were homologous to a combination of different phage that infect 12 genera of bacteria, but the majority of spacers were homologous to streptococcal phage. Many CRISPRs were found to contain multiple spacers that corresponded to different phage, which suggested that strains

within cheeses had been exposed to multiple infections. This, combined with evidence of the importance of phage in LGT, further emphasised the role of phage in shaping the cheese microbiome.

We found some evidence that CRISPRs were transferred between members of the microbiota. Firstly, *Lactococcus* and *Staphylococcus* MAGs contained CRISPRs whose spacers aligned to *Streptococcus* phage. Secondly, numerous spacers were homologous to lactococcal phage. This is notable as CRISPRs associated with lactococci are rare, and may represent the acquisition of CRISPRs by lactococci through LGT, after which point exposure to lactococcal phage occurred. Indeed, there has only been one report of a *L. lactis* strain with a CRISPR, and in that case the locus, which was predicted to be inactive, was identified on a plasmid [29]. Notably, CRISPRs containing lactococcal phage were only identified on plasmid contigs and, interestingly, these loci also contained spacers from streptococcal phage, which is indicative of possible acquisition of the locus by a *Lactococcus* species from a *Streptococcus* species. However, it should be noted that some lactococcal phage, such as *Lactococcus* phage ul36, share similarity to streptococcal phage belonging to the 987 group [30], so this data must be interpreted with caution. The transfer of CRISPRs between species is not without precedent [29], but it is interesting in the context of cheeses. We have already noted the abundance of phage in cheese, and this might select for resistance among bacteria, especially if these bacteria co-inhabit the cheeses over generations.

The fact that phage were abundant despite the prevalence of CRISPRs, suggests that phage can counteract the host defence, including through anti-CRISPR proteins (Acrs) [31]. Indeed, homologs to Acrs were identified in over half of the samples analysed here. The most prevalent Acr detected here was AcrIIA6, found on 33% of *S. thermophilus* phage genomes[32]. Other prevalent forms were AcrIIA3, associated with *Listeria* and *Streptococcus* phage[33], and AcrIIA7 and AcrVA2, which are widespread[34]. Our results point to coevolution

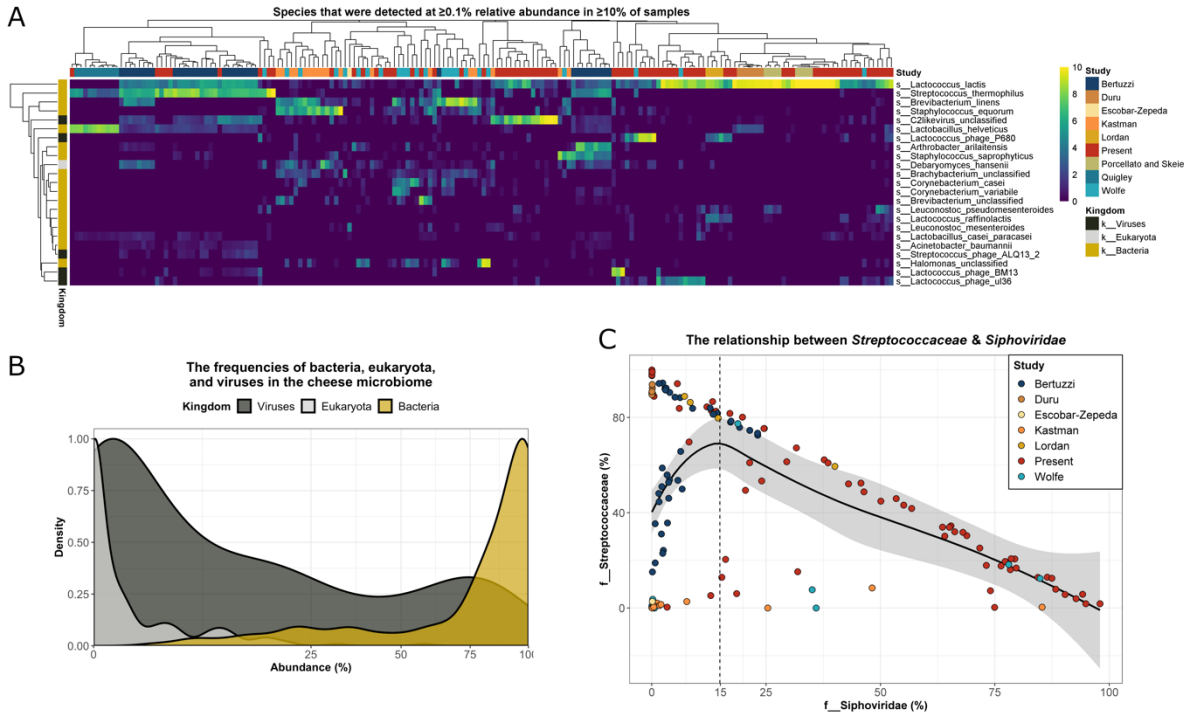
of microorganisms within cheese, wherein bacteria evolve defence against phage who subsequently evolve to overcome this defence.

Numerous studies have reported the isolation of antibiotic resistant bacteria from cheeses [35], and we found that ARGs were present on 35 MAGs. The occurrence of ARGs in cheeses is not problematic necessarily if the genes occur on the chromosome and not easily transferred to gut microbes after consumption. We found that the frequency of ARGs among representative cheese MAGs was no different to that among representative human MAGs. Furthermore, ARGs did not occur on the plasmids of LAB, being most commonly found on plasmids from *Enterobacteriaceae* and *Staphylococcaceae*. Similarly, ARGs were only detected in integrons of Proteobacteria MAGs. Overall, our results highlight that there is potential for transfer of ARGs between microorganisms in cheese, but this is only among bacteria that are more likely to be introduced from the environment. Therefore, measures to optimise hygiene during the manufacture of cheese might minimise the chance of cheese serving as a reservoir of transmissible ARGs. It is also important to note that future laboratory based investigations are required to determine the extent to which these ARG-based finding correspond to an associated resistance phenotype.

In addition to the mechanisms available to bacteria to protect against biotic stresses in the form of antibiotics and phage, bacteria can also kill other competing bacteria via the production of bacteriocins [36], producers of which have frequently been isolated from cheese [37]. Here, we found that Class II/III bacteriocin genes were present in 32.32% of the MAGs recovered from the samples. As expected, most of the bacteriocins were associated with species frequently employed in cheese manufacture, such as BlpK from *S. thermophilus*, helveticin from *L. helveticus*, lactococcin from *L. lactis*, or linocin from *B. linens*. Indeed, 68.87% of MAGs on which bacteriocins were detected were classified as *Brevibacteriaceae*, *Lactobacillaceae* or *Streptococcaceae*, though bacteriocins were also detected on MAGs

from other families, including *Enterobacteriaceae* and *Staphylococcaceae*. The frequency of bacteriocin genes across the cheese samples suggests that bacteriocins are enriched in cheese microbiomes relative to other environments (i.e., human, ocean, rumen), supporting the view that bacteriocin production is an important trait in cheese. It was also notable that homologs of the metalloprotease pseudoalterin gene were identified in several halophiles (e.g., *Marinomonas* and *Halomonas*). Pseudoalterin was first identified in *Pseudoalteromonas*, but homologs are widespread among Proteobacteria. It has been demonstrated that pseudoalterin is active against many Gram-positive bacteria. The presence of these genes in cheese-associated Proteobacteria may also provide these bacteria with a competitive advantage when colonising cheeses.

In conclusion, the present study highlights the heterogeneity of the microbiota of cheese. Our findings providing further genetic evidence that abiotic and biotic stresses shape cheese microbiomes. Notably, our results confirm that cheeses contain microbial communities in flux, wherein bacteria compete amongst themselves by producing bacteriocins and other antimicrobials, while also protecting themselves from phage by employing CRISPR. Additionally, the putatively novel species detected may influence the qualities of the cheeses, which showcases the potential for shotgun metagenomics to further our understanding of even relatively well characterised environments such as fermented foods.

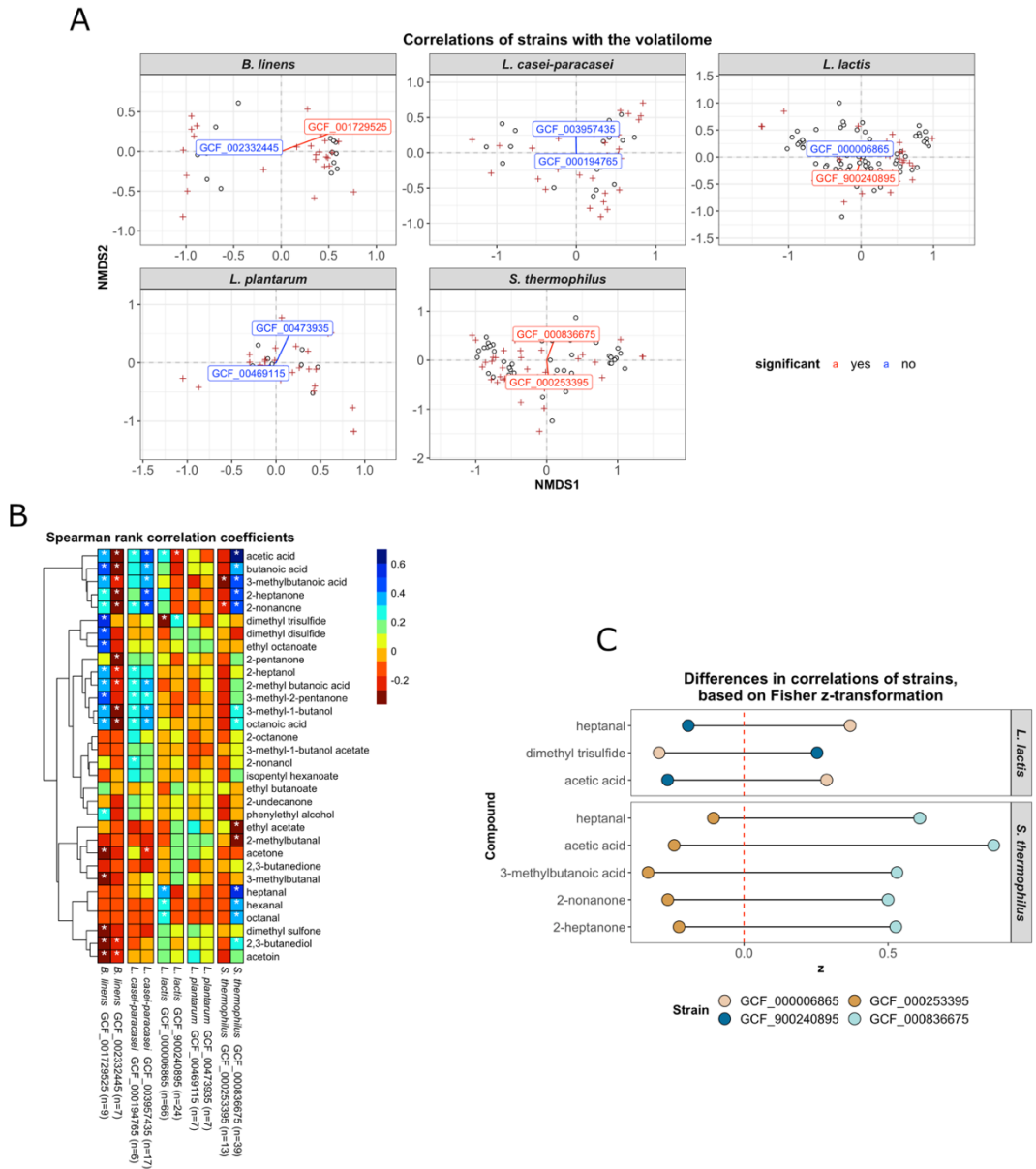


368

369 **Figure 1.** The microbial composition of cheeses. (A) Species that were detected at ≥0.1% relative abundance in ≥10% of all

370 samples (n=184 biologially independent samples). (B) The frequencies of bacteria, eukaryota, and viruses across all samples

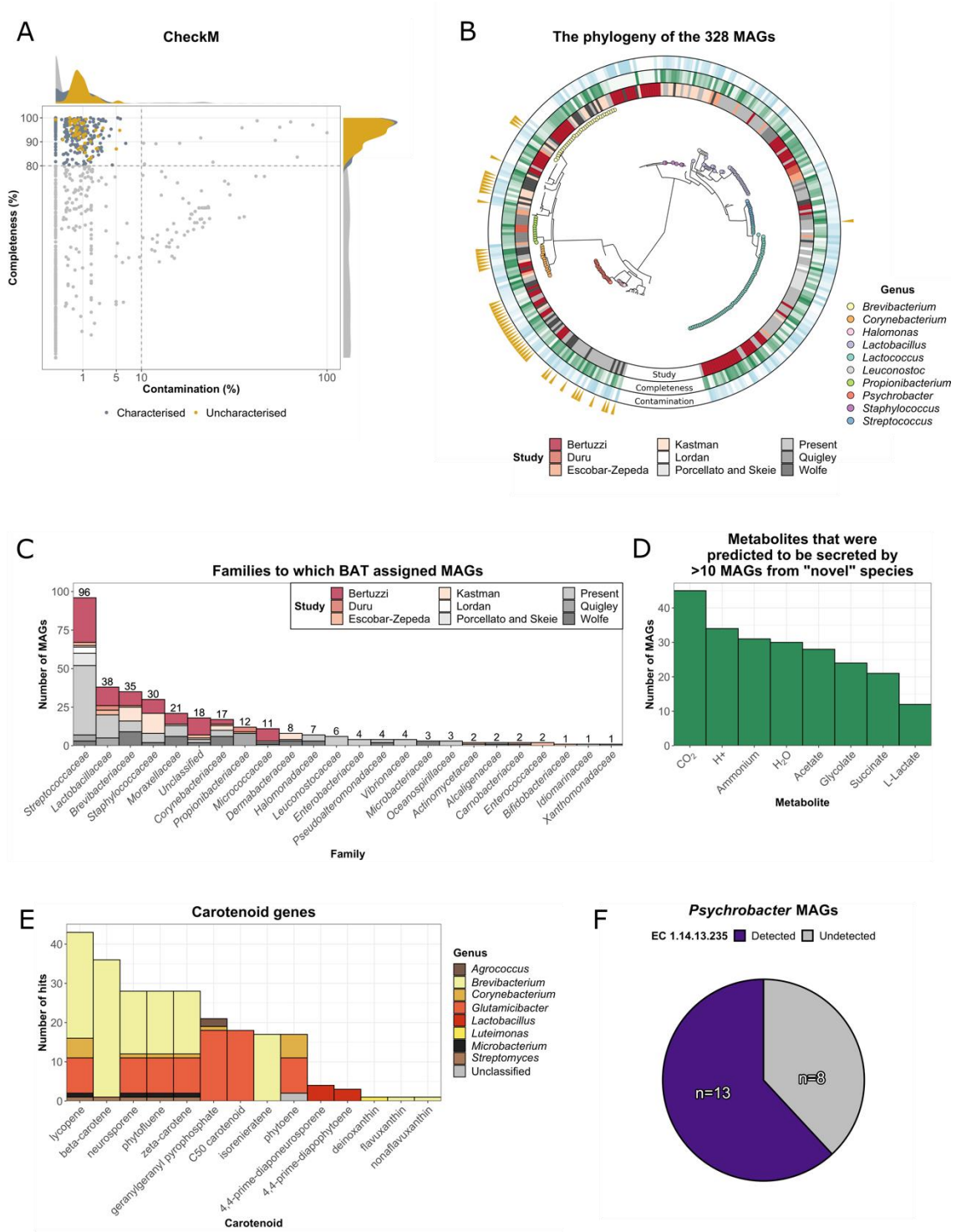
371 (n=184). (C) The relationship between *Streptococcaceae* and *Siphoviridae* across all samples (n=184).



372

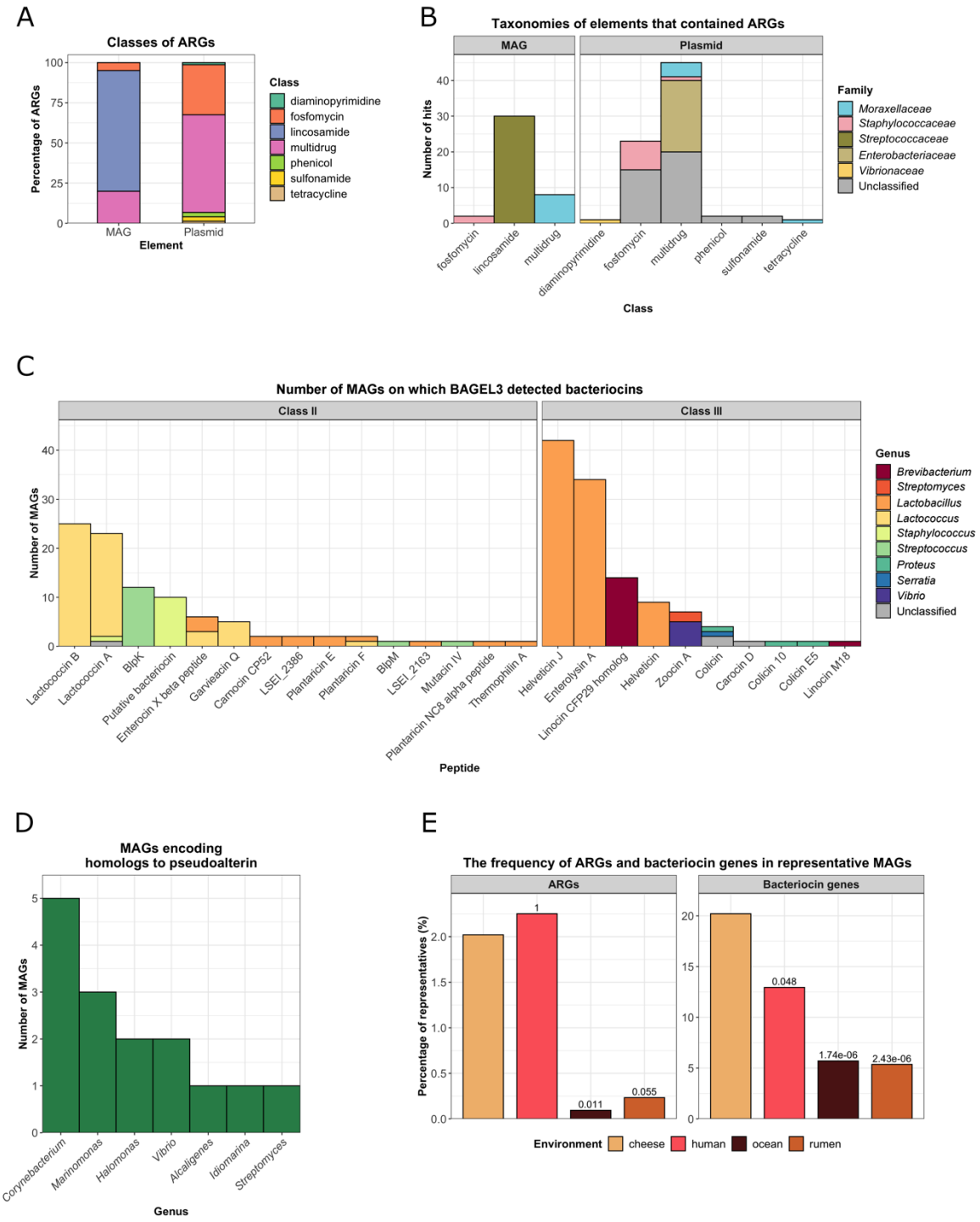
373 **Figure 2.** The relationship between strain-level variation and the metabolome. (A) Variance in the metabolome explained by
374 the abundances of strains. The colour of a strain indicates if it was significantly associated with variance in the metabolome,
375 as shown in the legend. Samples are represented by circles while the volatiles that were detected in those samples are
376 represented by crosses. (B) Spearman rank correlation coefficients of strains and volatiles. Significant correlations are
377 denoted by asterisks. (C) Differences in the correlations of strains, based on the comparison of Fisher z-transformed
378 Spearman correlation coefficients. Strains are labelled with the assembly accession number of their best match in the
379 pangenome database. The number of samples in which each strain was detected was as follows: GCF_001729525 (n=9),
380 GCF_002332445 (n=7), GCF_000194765 (n=6), GCF_003957435 (n=17), GCF_000006865 (n=66), GCF_900240895
381 (n=24), GCF_00469115 (n=7), GCF_00473935 (n=7), GCF_000253395 (n=13), and GCF_000836675 (n=39).

382



383

384 **Figure 3.** Assembly and characterisation of cheese MAGs. (A) The quality of metagenome assembled genomes (MAGs).
385 (B) The phylogeny of MAGs. The triangles on the edge of the tree indicate that a MAG could not be assigned to a species.
386 (C) The families to which MAGs were assigned. (D) Metabolites that were predicted to be secreted by >10 MAGs from
387 putative novel species. (E) The prevalence of carotenoid genes in MAGs. (F) The proportion of *Psychrobacter* MAGs
388 encoding indole-3-acetate monooxygenase (EC 1.14.13.235).



389

390 **Figure 4.** Analysis of antibiotic resistance genes (ARGs) on cheese metagenome assembled genomes (MAGs) and plasmids
391 (A) The classes of ARGs detected on MAGs and plasmids. (B) The taxonomic classification of MAGs and plasmids on
392 which ARGs were detected. (C) The number and taxonomic classification of MAGs harboring bacteriocin genes. (D) The
393 number and taxonomic classification of MAGs putatively encoding homologs to pseudoalterin. (E) The frequency of ARGs
394 and bacteriocin genes in representative MAGs.

395

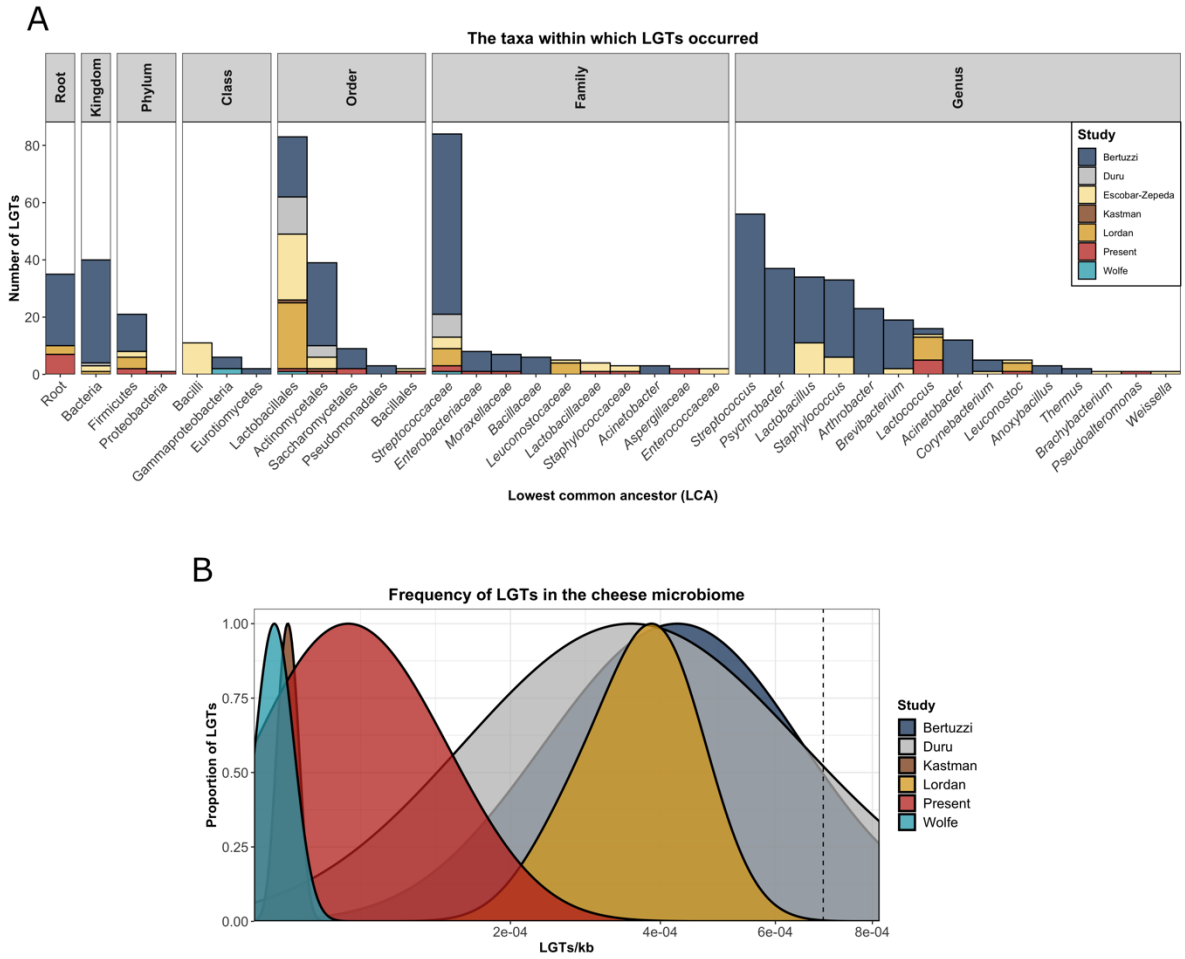
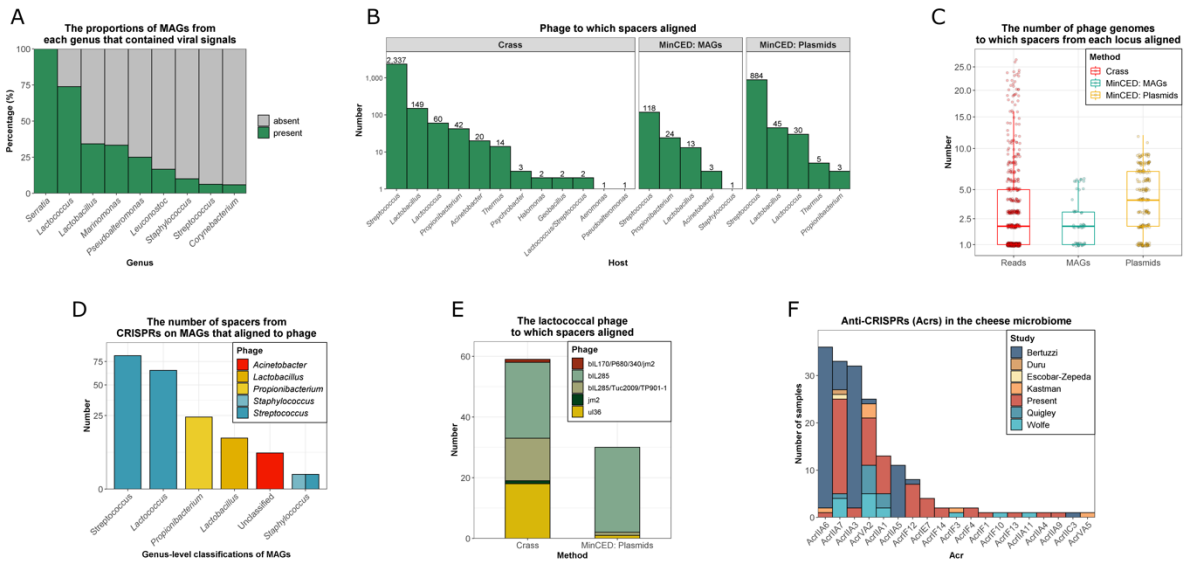


Figure 5. Lateral gene transfer (LGT) in cheese microbiomes. (A) The taxa within in which LGTs were detected. (B) The frequency of LGTs in cheese metagenomes, expressed as LGTs/kb (note: the frequency of LGTs in the Escobar-Zepeda sample is represented by a dashed line).



401

402 **Figure 6.** Phage, CRISPRs and anti-CRISPRs in cheese microbiomes (A) The proportions of MAGs from each genus that
403 contained viral signals. (B) Phage to which spacers aligned. (C) The number of phage genomes to which spacers from each
404 locus aligned. (D) The number of spacers from CRISPRs on MAGs that aligned to phage. (E) The lactococcal phage to
405 which spacers aligned. (F) Anti-CRISPRs (Acrs) in the cheese microbiome.

METHODS

Sample collection, preparation and analysis

A total of 55 artisanal cheese samples were obtained across 27 artisanal farm producers and farmer's markets throughout Ireland. The samples included 15 soft cheeses, 16 semi-hard cheeses and 24 hard cheeses, manufactured from unpasteurised or pasteurised cow, goat or sheep milk (Supplementary Table 11). Approaches taken for sample collection, storage, DNA extraction, sequencing and volatile analysis are described in Supplementary Methods.

Bioinformatic analysis

Shotgun metagenomic fastq files were converted to bam files using SAMtools [38], and duplicate reads were subsequently removed using Picard Tools (https://github.com/broadinstitute/picard). Next, low quality reads were removed using the trimBWAstyle.usingBam.pl script (https://github.com/genome/genome/blob/master/lib/perl/Genome/Site/TGI/Hmp/HmpSraProcess/trimBWAstyle.usingBam.pl). Specifically, reads with a quality score less than Q30 were discarded. Resulting fastq files were converted to fasta files using the fq2fa option from IDBA-UD [39]. The number of quality reads for each of the newly sequenced samples is reported in Supplementary Table 11. Compositional analysis was performed using MetaPhlAn2 [40]. Strain-level analysis was performed using PanPhlAn [11]. If a strain was identified by PanPhlAn, the abundance of that strain was inferred as the abundance of the species, as measured by MetaPhlAn2, to which that strain belonged. Functional analysis was performed using SUPER-FOCUS [41].

Metagenome assembly was performed using IDBA-UD. PlasFlow [42] was used to identify plasmid contigs in metagenome assemblies, and Kaiju was used to determine the taxonomy of these contigs [43]. Lateral gene transfer (LGT) events were identified in assembled metagenomes using WAAFLE (http://huttenhower.sph.harvard.edu/waafle), which was run

with contig-level quality control. Genome binning was performed using MetaBAT 2 [44], with default settings. Reads were mapped against assemblies using Bowtie 2 [45]. CheckM [46] was used to assess the quality of metagenome assembled genomes (MAGs). Low quality MAGs (i.e., <80% completeness and/or >10% contamination) were excluded from further analysis. Next, CAT-BAT was used to classify the MAGs, while PhyloPhlAn [47] was used to infer the taxonomy of MAGs. The average nucleotide identity (ANI) of MAGs to references, which were downloaded from RefSeq [48], was calculated using FastANI [49]. dRep [50] was used to cluster genomes into primary clusters based on their relative similarities. Prodigal [51] was used to identify open reading frames on MAGs, which were annotated using eggNOG-mapper [52,53]. The prevalence of carotenoid genes on MAGs was assessed by aligning contigs against the ProCarDB [54] database of bacterial carotenoids. CarveMe [55] was used to build metabolic models from MAGs. The models were initiated under a medium that was designed to replicate cheese agar medium (CAM) [56] (Supplementary Table 14). Flux balance analysis (FBA) was performed using COBRApy [57] (Python 3.6), to simulate the metabolism of the organisms.

RGI was used for the detection of ARGs (antibiotic resistance genes) on contigs (note: only "perfect" matches were considered as ARGs), while IntegronFinder [58] was used to detect ARGs in integrons. The prevalence of bacteriocin genes across MAGs was estimated using BAGEL3 [59]. The prevalence of pseudoalterin [16] was determined by aligning the protein against MAGs. dRep [50] was used to dereplicate MAGs from cheeses to select those that represented the diversity within the microbiome, and the frequencies of genes of interest among representatives from cheeses were compared to those among representatives from other environments (Supplementary Table 15). VirSorter [60] was used to detect prophage in MAGs. CRISPRs were detected from short metagenomic reads using Crass [61], while CRISPRs were identified in MAGs and plasmid contigs using MinCED

(https://github.com/ctSkennerton/minced). BLASTn [62] was used to align CRISPR spacer sequences against the RefSeq viral database. The prevalence of anti-CRISPR proteins was assessed by using tBLASTx to align metagenome assemblies against a database of Acrs [63].

Statistical analysis and data visualisation

The R package vegan [64] was used for alpha diversity analysis, in addition to non-metric multidimensional scaling (nMDS). The Wilcoxon Rank Sum Test was used to measure statistical differences in alpha diversity between groups. PERMANOVA (PERMutational ANalysis Of VAriance) was performed using the adonis function from vegan. The linear discriminant analysis (LDA) effect size (LEfSe) method [65] was used to determine if any taxa, as measured by MetaPhlAn2, or pathways, as measured by SUPER-FOCUS, were differentially abundant between groups. Spearman's test was used to measure the association of strains with metabolites, and p-values were adjusted using the Bonferroni correction. If a cheese was sampled from both regions (i.e., core and rind), its abundance was taken as the average across these regions. Spearman rank correlation coefficients were compared using the Fisher transformation. Nonlinear correlation analysis was performed using the nlcor R package (https://github.com/ProcessMiner/nlcor/). Fisher's exact test was used to determine if genes of interest were enriched in cheese compared to other environments. . Data was visualised using GraPhlAn [66] and the R packages ggplot2 [67] and pheatmap [68].

Data availability

Raw reads have been deposited to the European Nucleotide Archive under the project accession number PRJEB32768, while MAGs are available here https://drive.google.com/file/d/1TCLYBX7kkxNUWn4jr4YGXNL_qV97lc70/view.

Acknowledgements

The authors would like to thank Dr. Jennifer Mahony, Prof. Douwe Van Sinderen and members of the Vision 1 laboratory, particularly Dr. Wiley Barton, for helpful discussions and a critical review of the manuscript, as well as Dr. Fiona Crispie and Laura Finnegan for their contributions to DNA sequencing. This research was conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers SFI/12/RC/2273P1 and SFI/12/RC/2273P2 (APC Microbiome Ireland). Research in the Cotter laboratory is also funded through MASTER, an Innovation Action funded by the European Commission under the Horizon 2020 Programme under grant number 818368, SFI and the Department of Agriculture, Food and Marine under Grant 16/RC/3835 (VistaMilk) and the Enterprise Ireland Technology Centre, Food for Health Ireland**.**

Competing Interests Statement

The authors declare no competing interests.

References

1 Yeluri Jonnala, B. R., McSweeney, P. L. H., Sheehan, J. J. & Cotter, P. D. Sequencing of the Cheese Microbiome and Its Relevance to Industry. *Frontiers in Microbiology* **9**, doi:10.3389/fmicb.2018.01020 (2018).

2 De Filippis, F., Genovese, A., Ferranti, P., Gilbert, J. A. & Ercolini, D. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. *Scientific reports* **6** (2016).

3 Bertuzzi, A. S. *et al.* Omics-Based Insights into Flavor Development and Microbial Succession within Surface-Ripened Cheese. *mSystems* **3**, doi:10.1128/mSystems.00211-17 (2018).

4 Dugat-Bony, E. *et al.* Overview of a surface-ripened cheese community functioning by meta-omics analyses. *PLoS One* **10**, e0124360 (2015).

5 Wolfe, B. E., Button, J. E., Santarelli, M. & Dutton, R. J. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. *Cell* **158**, 422-433, doi:10.1016/j.cell.2014.05.041 (2014).

6 Morin, M., Pierce, E. C. & Dutton, R. J. Changes in the genetic requirements for microbial interactions with increasing community complexity. *eLife* **7**, e37072, doi:10.7554/eLife.37072 (2018).

7 Bonham, K. S., Wolfe, B. E. & Dutton, R. J. Extensive horizontal gene transfer in cheese-associated bacteria. *eLife* **6**, e22144, doi:10.7554/eLife.22144 (2017).

8 De Filippis, F., Parente, E. & Ercolini, D. Metagenomics insights into food fermentations. *Microbial biotechnology* **10**, 91-102 (2017).

9 Franzosa, E. A. *et al.* Sequencing and beyond: integrating molecular'omics' for microbial community profiling. *Nature Reviews Microbiology* **13**, 360-372 (2015).

10 de Melo, A. G., Levesque, S. & Moineau, S. Phages as friends and enemies in food processing. *Current Opinion in Biotechnology* **49**, 185-190, doi:https://doi.org/10.1016/j.copbio.2017.09.004 (2018).

11 Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature methods* (2016).

12 Thierry, A. *et al.* Strain-to-strain differences within lactic and propionic acid bacteria species strongly impact the properties of cheese–A review. *Dairy Science & Technology* **95**, 895-918 (2015).

13 Quigley, L. *et al.* Thermus and the Pink Discoloration Defect in Cheese. *mSystems* **1**, e00023-00016 (2016).

14 Kamelamela, N., Zalesne, M., Morimoto, J., Robbat, A. & Wolfe, B. E. Indigo- and indirubin-producing strains of Proteus and Psychrobacter are associated with purple rind defect in a surface-ripened cheese. *Food Microbiol* **76**, 543-552, doi:10.1016/j.fm.2018.07.011 (2018).

15 Cotter, P. D., Hill, C. & Ross, R. P. Bacteriocins: developing innate immunity for food. *Nat Rev Microbiol* **3**, 777-788, doi:10.1038/nrmicro1273 (2005).

16 Tang, B. L. *et al.* A predator-prey interaction between a marine Pseudoalteromonas sp. and Gram-positive bacteria. *Nat Commun* **11**, 285, doi:10.1038/s41467-019-14133-x (2020).

17 Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology* **35**, 833 (2017).

18 Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications* **9**, 870, doi:10.1038/s41467-018-03317-6 (2018).

19 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e620, doi:https://doi.org/10.1016/j.cell.2019.01.001 (2019).

20 Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203, doi:10.1038/sdata.2017.203 (2018).

21 Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* **9**, 870, doi:10.1038/s41467-018-03317-6 (2018).

22 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662 e620, doi:10.1016/j.cell.2019.01.001 (2019).

23 Quigley, L. *et al.* High-throughput sequencing for detection of subpopulations of bacteria not previously associated with artisanal cheeses. *Appl Environ Microbiol* **78**, 5717-5723, doi:10.1128/AEM.00918-12 (2012).

24 Walsh, A. M. *et al.* Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. *mSystems* **1**, doi:10.1128/mSystems.00052-16 (2016).

25 Niccum, B. A., Kastman, E. K., Kfoury, N., Robbat, A. & Wolfe, B. E. Strain-level diversity impacts cheese rind microbiome assembly and function. *bioRxiv* **652768**, doi:10.1101/652768 (2019).

26 Dugat-Bony, E. *et al.* Viral metagenomic analysis of the cheese surface: a comparative study of rapid procedures for extracting virus-like particles. *bioRxiv*, 503599, doi:10.1101/503599 (2018).

27 Gobbetti, M. *et al.* Drivers that establish and assembly the lactic acid bacteria biota in cheeses. *Trends in Food Science & Technology* **78**, 244-254, doi:https://doi.org/10.1016/j.tifs.2018.06.010 (2018).

28 Mahony, J. & van Sinderen, D. Novel strategies to prevent or exploit phages in fermentations, insights from phage–host interactions. *Current opinion in biotechnology* **32**, 8-13 (2015).

29 Millen, A. M., Horvath, P., Boyaval, P. & Romero, D. A. Mobile CRISPR/Cas-Mediated Bacteriophage Resistance in Lactococcus lactis. *PLOS ONE* **7**, e51663, doi:10.1371/journal.pone.0051663 (2012).

30 McDonnell, B. *et al.* Identification and Analysis of a Novel Group of Bacteriophages Infecting the Lactic Acid Bacterium Streptococcus thermophilus. *Applied and environmental microbiology* **82**, 5153-5165, doi:10.1128/AEM.00835-16 (2016).
31 Pawluk, A., Davidson, A. R. & Maxwell, K. L. Anti-CRISPR: discovery, mechanism and function. *Nature Reviews Microbiology* **16**, 12, doi:10.1038/nrmicro.2017.120 (2017).
32 Hynes, A. P. *et al.* Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. *Nat Commun* **9**, 2919, doi:10.1038/s41467-018-05092-w (2018).
33 Rauch, B. J. *et al.* Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell* **168**, 150-158.e110, doi:10.1016/j.cell.2016.12.009 (2017).
34 Marino, N. D. *et al.* Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science* **362**, 240-242, doi:10.1126/science.aau5174 (2018).
35 Verraes, C. *et al.* Antimicrobial resistance in the food chain: a review. *International journal of environmental research and public health* **10**, 2643-2669, doi:10.3390/ijerph10072643 (2013).
36 García-Bayona, L. & Comstock, L. E. Bacterial antagonism in host-associated microbial communities. *Science* **361**, eaat2456, doi:10.1126/science.aat2456 (2018).
37 Favaro, L., Barretto Penna, A. L. & Todorov, S. D. Bacteriocinogenic LAB from cheeses – Application in biopreservation? *Trends in Food Science & Technology* **41**, 37-48, doi:https://doi.org/10.1016/j.tifs.2014.09.001 (2015).
38 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
39 Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428 (2012).
40 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* **12**, 902-903 (2015).
41 Silva, G. G. Z., Green, K. T., Dutilh, B. E. & Edwards, R. A. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* **32**, 354-361 (2016).
42 Lipinski, L., Dziembowski, A. & Krawczyk, P. S. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research* **46**, e35-e35, doi:10.1093/nar/gkx1321 (2018).
43 Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**, 11257, doi:10.1038/ncomms11257 (2016).
44 Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
45 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
46 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).
47 Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**, 2304, doi:10.1038/ncomms3304 (2013).
48 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61-D65 (2006).
49 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).
50 Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864-2868, doi:10.1038/ismej.2017.126 (2017).
51 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).
52 Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309-D314, doi:10.1093/nar/gky1085 (2019).
53 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115-2122, doi:10.1093/molbev/msx148 (2017).
54 Nupur, L. N. *et al.* ProCarDB: a database of bacterial carotenoids. *BMC Microbiol* **16**, 96, doi:10.1186/s12866-016-0715-6 (2016).
55 Machado, D., Tramontano, M., Andrejev, S. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research* **46**, 7542-7553, doi:10.1093/nar/gky537 (2018).
56 Neviani, E., Juliano De Dea, L., Bernini, V. & Gatti, M. Recovery and differentiation of long ripened cheese microflora through a new cheese-based cultural medium. *Food Microbiology* **26**, 240-245, doi:https://doi.org/10.1016/j.fm.2009.01.004 (2009).
57 Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology* **7**, 74, doi:10.1186/1752-0509-7-74 (2013).

58 Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic acids research* **44**, 4539-4550, doi:10.1093/nar/gkw319 (2016).
59 van Heel, A. J., de Jong, A., Montalban-Lopez, M., Kok, J. & Kuipers, O. P. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res* **41**, W448-453, doi:10.1093/nar/gkt391 (2013).
60 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985, doi:10.7717/peerj.985 (2015).
61 Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Research* **41**, e105-e105, doi:10.1093/nar/gkt183 (2013).
62 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410, doi:https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
63 Yin, Y., Yang, B. & Entwistle, S. Bioinformatics Identification of Anti-CRISPR Loci by Using Homology, Guilt-by-Association, and CRISPR Self-Targeting Spacer Approaches. *mSystems* **4**, doi:10.1128/mSystems.00455-19 (2019).
64 Oksanen, J. *et al.* The vegan package. *Community ecology package* **10**, 631-637 (2007).
65 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome biology* **12**, 1 (2011).
66 Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029, doi:10.7717/peerj.1029 (2015).
67 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).
68 Kolde, R. Pheatmap: pretty heatmaps. *R package version* **61**, 926 (2012).

Supplemental Information

Supplementary Results

Sample collection and DNA sequencing

Samples were collected in sterile bags, transported and maintained at 4°C until the analyses were performed. For rind analysis, a total of 22 sub-samples from across the 55 cheeses were collected by scraping the rind surface with a sterile razor blade and transferring rind samples in sterile bags. Up to three wheels for each cheese were tested to account for production variation. A subsample of 100g from each cheese was stored at −20°C prior to flavour analysis. To facilitate the culture independent analysis of the bacterial composition of the cheeses, their associated rinds, naturally developed or smear-ripened cheese rinds, 2g of cheese or 2g of cheese rind was combined with 8ml 2% tri-sodium citrate and homogenised before DNA was extracted using the PowerFood Total Microbial DNA Isolation kit as described in the manufacturer's protocol (MoBio Laboratories Inc., USA). Total DNA was quantified with the Qubit high-sensitivity DNA assay (Bio-Sciences, Dublin, Ireland). Whole-metagenome shotgun libraries were prepared in accordance with the Nextera XT DNA library preparation guide from Illumina. Libraries were sequenced on the Illumina NextSeq 500 with a v2 NextSeq 500/550 high-output reagent kit (300 cycles). All sequencing was done in the Teagasc sequencing facility (Moorepark, Cork, Ireland) in accordance with standard Illumina sequencing protocols.

HS-SPME GCMS Analysis

Sample preparation

4 g of sample was weighed into to a 20 ml amber screw capped SPME vial (Apex Scientific Ltd, Maynooth, Ireland). The vial was sealed with a magnetic screw capped silicone/PTFE liner septum and equilibrated to 40°C for 10 mins with pulsed agitation of 5 sec at 500 rpm using a Shimadzu AOC 5000 autosampler (Mason Technology Ltd, Dublin, Ireland). The samples were analysed in duplicate. A single 50/30 µm Carboxen^{TM}/divinylbenzene/polydimethylsiloxane (DVB/CAR/PDMS) fibre (Agilent Technologies Ireland Ltd, Cork, Ireland) was used. The fibre was exposed to the headspace above the samples for 20 min at depth of 1 cm at 40°C.

GCMS Method

The fibre was retracted and injected into the GC inlet and desorbed for 2 min at 250°C using a Shimadzu 2010 Plus GC (Mason Technologies Ltd, Ireland) with a DB-5MS (60m x 0.25mm x 0.25µm) column (Agilent Technologies Ireland Ltd) using a split/splitless injector in splitless mode with a merlin microseal (Merck, Arklow, Ireland). The temperature of the column oven was set at 35°C, held for 0.5 min, increased at 6.5°C/min to 230°C then increased at 15°C/min to 320°C, yielding at total GC run time of 41.5 min. The carrier gas was helium held at a constant pressure of 23 psi. The detector was a Shimadzu TQ8030 mass spectrometer detector (Mason Technologies Ltd), ran in single quad mode. The ion source temperature was 220°C and the interface temperature was set at 280°C. The MS mode was electronic ionization (70 V) with the mass range scanned between 35 and 250 amu. Compounds were identified using mass spectra comparisons to the NIST 2014 mass spectral library and an in-house library created in Shimadzu GCMS Solutions software (Mason Technologies Ltd) with target and qualifier ions and linear retention indices [17] for each compound. Spectral deconvolution was also performed to confirm identification of compounds using R (www.r-project.org/) and AMDIS (www.amdis.net/). An auto-tune of the GCMS was carried out prior to the analysis to ensure optimal GCMS performance. A set of

external standards was ran at the start and end of the sample set and abundances were compared to known amounts to ensure that both the SPME extraction and MS detection was performing within specification.

Supplementary Results

The present study included 55 newly sequenced artisanal Irish cheeses, comprising 55 core cheese samples and 22 rind samples of these cheeses were also sequenced. Thus, the dataset included 77 samples in total, and the microbial profiles of subgroupings of these samples were compared. Firstly, the core microbiota was compared to the rind microbiota and, despite there being no significant difference in alpha diversity measures between the sites (Figure S1A), PERMANOVA detected a significant dissimilarity between the regions ($p=0.002$, $R^2=0.09$) (Figure S1B).

LEfSe determined that 58 taxa, including 16 species, were differentially abundant between the regions (Figure S1C; Table S10). *L. lactis* (LDA=5.21) was highest in the core, whereas *B. linens* (LDA=4.68) was highest on the rind. Among phage, the *Lactococcus* phage ul36 (LDA=4.37) was highest in the core. Several halophiles were enriched on the rind, including species from the genera *Halomonas* (LDA=4.32), *Psychrobacter* (LDA=3.50), and *Tetragenococcus* (LDA=3.51). PERMANOVA did not detect a significant dissimilarity between cheeses produced with pasteurised versus unpasteurised milks (Figure S2A), either in the core ($p=0.194$, $R^2=0.031$) or in the rind ($p=0.219$, $R^2=0.085$), and pasteurisation of the milk did not have a significant impact on the alpha diversity of the resultant cheeses (Figure S2B). However, LEfSe revealed that 7 species (Figure S2C) were differentially abundant between these groups. In the core, *Lactococcus raffinolactis* (LDA=3.10) was highest in pasteurised cheeses, whereas *Lactobacillus casei paracasei* (LDA=3.24), *Lactobacillus otakiensis* (LDA=2.52) and *Pediococcus pentosaceus* (LDA=2.52) were highest in unpasteurised cheeses. In the rind, *Lactococcus* phage P680 (LDA=5.03) and *Staphylococcus saprophyticus* (LDA=4.48) were highest in pasteurised cheeses, whereas *Arthrobacter arilaitensis* (LDA=4.52) was highest in unpasteurised cheeses. Next, the relationship between cheese maturity/hardness and the microbiota was assessed. PERMANOVA revealed that the core microbiota of soft cheese was significantly dissimilar to that of more mature semi-hard ($p=0.034$, $R^2=0.074$) and hard ($p=0.026$, $R^2=0.056$) cheeses (Figure S3A). Similarly, the rind microbiota of soft cheeses was also significantly dissimilar to that of semi-hard ($p=0.003$, $R^2=0.154$) and hard ($p=0.007$, $R^2=0.130$) cheeses. However, maturity did not have a significant effect on the alpha diversity of the cheeses (Figure S3B). LEfSe determined that 29 taxa, including 9 species, were discriminative between cheeses of different maturity/moisture content (Figure S3C). Notably, in the core, several lactobacilli were enriched in soft cheeses (Table S11), while on the rind, *B. linens* (LDA=4.98) was highest in

hard cheeses, *S. thermophilus* (LDA=4.53) was highest in semi-hard cheeses, and eukaryotes were highest in soft cheeses (LDA=4.03). The microbiota of cheeses produced with milk from different animals was also compared. However, the low number of samples from cheeses produced with either buffalo or sheep milk meant that pairwise comparisons between cheeses made with milk from these animals were not meaningful. PERMANOVA indicated that, overall, there were significant dissimilarities between the core microbiota of different animal milk cheeses ($p=0.042$, $R^2=0.091$).

Functional analysis was performed using SUPER-FOCUS. PERMANOVA detected a significant dissimilarity between the core versus the rind microbiota with respect to the abundances of pathways as predicted by SUPER-FOCUS ($p=0.001$, $R^2=197$) (Figure S4A), and LEfSe determined that 22 level 1 subsystems (Figure S4B), in addition to 93 level 2 subsystems, were differentially abundant between these regions (Table S12). Pathways associated with fermentation (LDA=3.22) were highest in the core, whereas those associated with amino acids and derivatives (LDA=4.10), fatty acids (LDA=3.47) and sulphur metabolism (LDA=3.28) were highest in the rind. Several interesting differences in the abundances of pathways associated with niche-specific adaptation were also observed. Specifically, genes associated with bacteriocins (LDA=2.59) were highest in the core, whereas those associated with iron acquisition and metabolism (LDA=3.32) were highest in the rind. Additionally, genes associated with osmotic stress (LDA=3.51) and oxidative stress (LDA=3.39) were highest in the rind. PERMANOVA did not detect significant overall functional dissimilarities between cheeses of different levels of maturity (Figure S5A), or cheeses produced using pasteurised versus unpasteurised milks (Figure S5B). However, LEfSe determined that 6 level 1 subsystems, in addition to 26 level 2 subsystems, were differentially abundant between cheeses of different maturity (Table S13), and determined that 1 level 1 subsystems, in addition to 10 level 2 subsystems, were differentially abundant between cheeses produced with pasteurised versus unpasteurised milks (Table S14). Again, LEfSe identified differentially abundant pathways that highlighted adaptations to cheese. Notably, genes associated with desiccation stress (LDA=2.46) were highest on the rind of hard cheeses, while those associated with cold shock (LDA=2.03) were highest on the rind of unpasteurised cheeses. No significant dissimilarity was observed between cheeses produced from the milks of different animals ($p=0.607$, $R^2=0.030$).

Supplementary Discussion

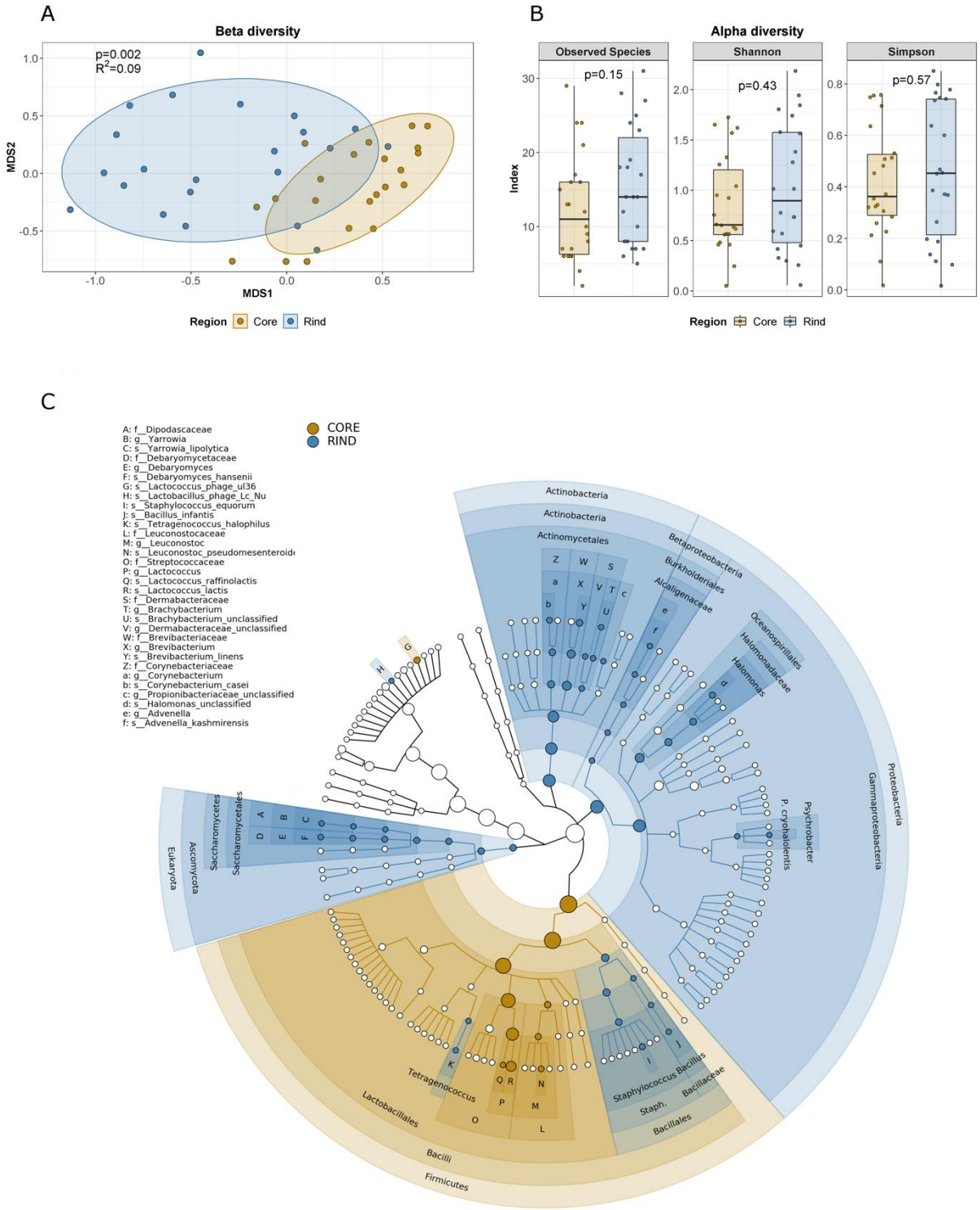
It has been demonstrated that the rind represents a model in which to study the formation of microbiota [1]. In the present study, a subset of cheeses were sampled from the core in addition to the rind, and the microbiome of the two regions was compared. Our results indicated that abiotic factors predictably shaped the microbiome of these microenvironments. For example, it was observed that halophiles, which are associated with saltwater, were enriched on the rind, and it is likely that these halophiles were present in the brine used to wash the rinds, as reported elsewhere [1,2]. Genes associated with osmostic stress response were highest on the rind, which provides further evidence that microorganisms on the rind are adapted to salinity. Additionally, genes associated with the oxidative stress response were also highest on the rind, which is to be expected since microorganisms on the surface are in contact with oxygen. We also found that genes associated with the acquisition of iron were highest on the rind, which reflects the fact that iron is limiting on the surface [3]. Other differences were observed between different groupings of cheese, which provided further examples of adaptations to these foods. Notably, comparison of cheeses based on their moisture content revealed that genes associated with desiccation stress responses were highest in cheeses with low moisture contents, while comparison of samples based on pasteurisation revealed that genes associated with cold shock responses were highest in cheeses that were made with unpasteurised milks. Overall, these results are predictable, since they confirm what is already known about the environment, but this predictability is remarkable in that it reemphasises the suitability of cheese as a model. Specifically, we demonstrated that the cheese microbiome was adapted to the cheese environment, which suggests that the microbiome can be manipulated by manipulating the environment itself.

Supplementary References

1. Wolfe, B. E., Button, J. E., Santarelli, M. & Dutton, R. J. Cheese Rind Communities Provide TracTable Systems for In Situ and In Vitro Studies of Microbial Diversity. *Cell* **158**, 422–433 (2014).

2. Quigley, L. *et al.* High-Throughput Sequencing for Detection of Subpopulations of Bacteria Not Previously Associated with Artisanal Cheeses. *Appl. Environ. Microbiol.* **78**, 5717 LP – 5723 (2012).

3. Monnet, C., Back, A. & Irlinger, F. Growth of aerobic ripening bacteria at the cheese surface is limited by the availability of iron. *Appl. Environ. Microbiol.* **78**, 3185–3192 (2012).



152

153 Figure S1. The (A) beta and (B) alpha diversity of the core and rind of newly sequenced samples. (C) Taxa that were
154 differentially abundant between the core and rind, as determined by LEfSe.

155

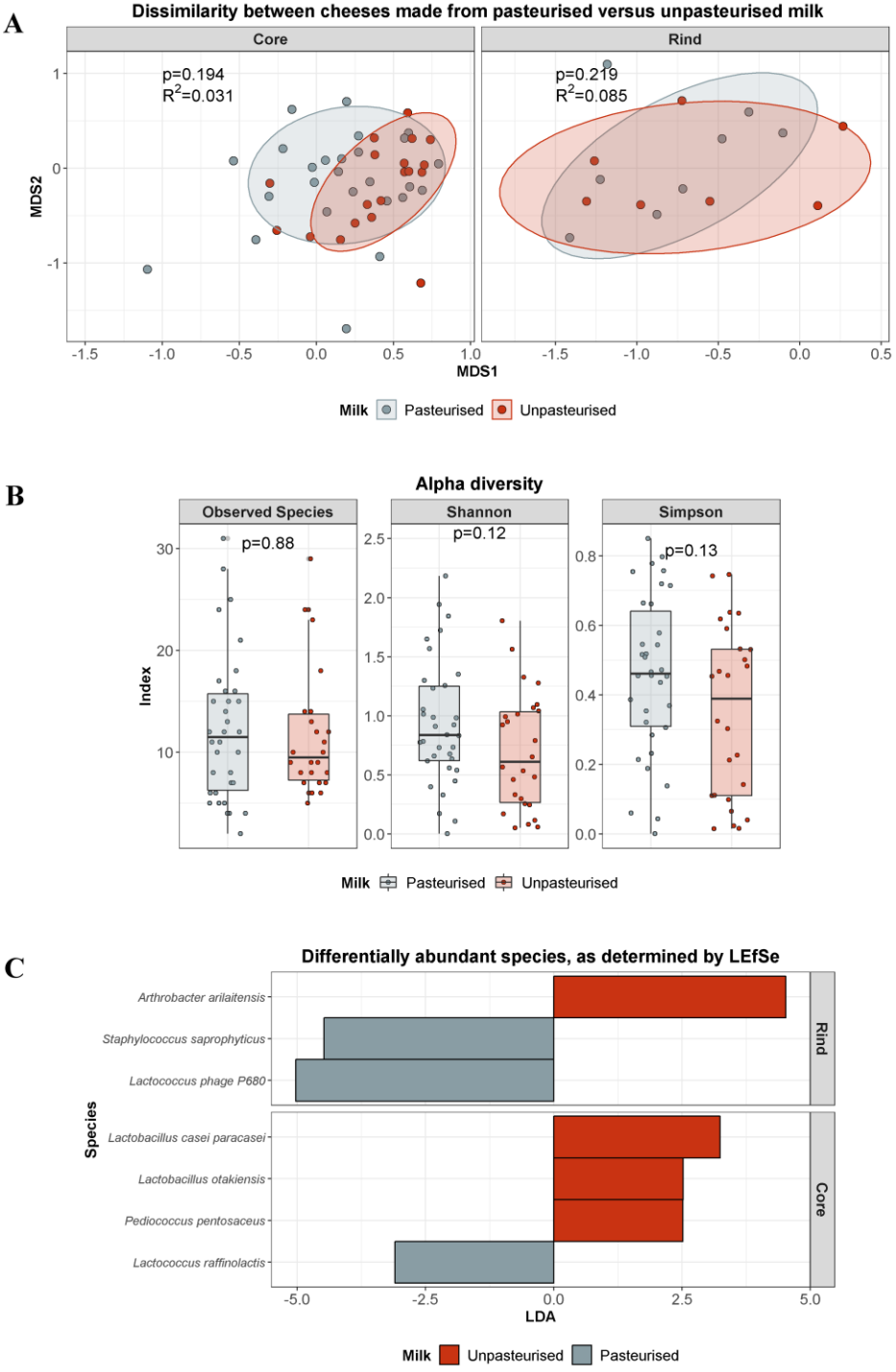


Figure S2. Compositional analysis of newly sequenced cheeses produced with pasteurized versus unpasteurized cheeses. (A) Beta and (B) alpha diversity of pasteurized and unpasteurized cheeses. (C) Species that were differentially abundant between pasteurized versus unpasteurized cheeses.

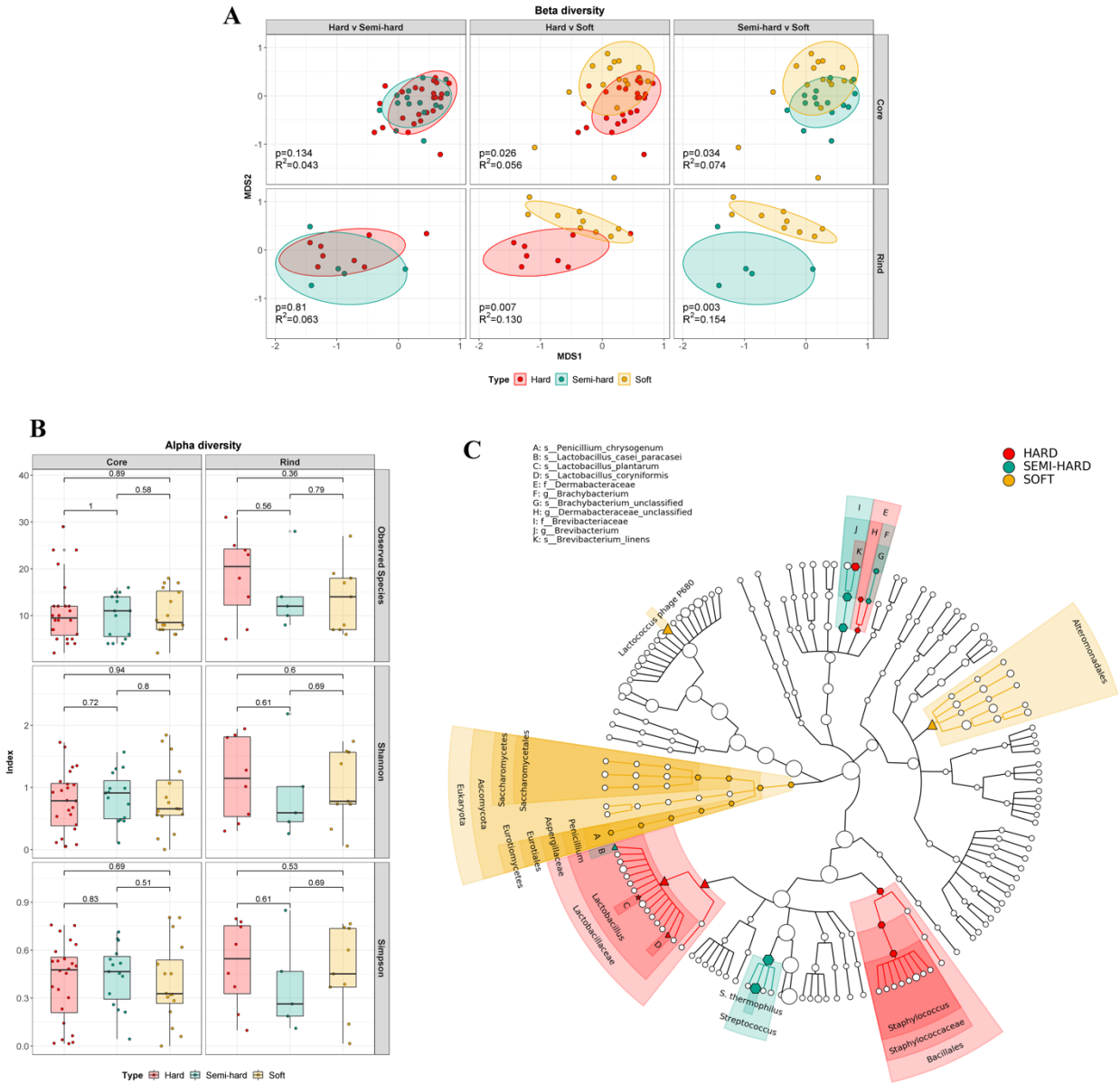


Figure S3. The (A) beta and (B) alpha diversity of cheeses of different maturity. (C) Taxa that were differentially abundant between cheeses of different maturity.

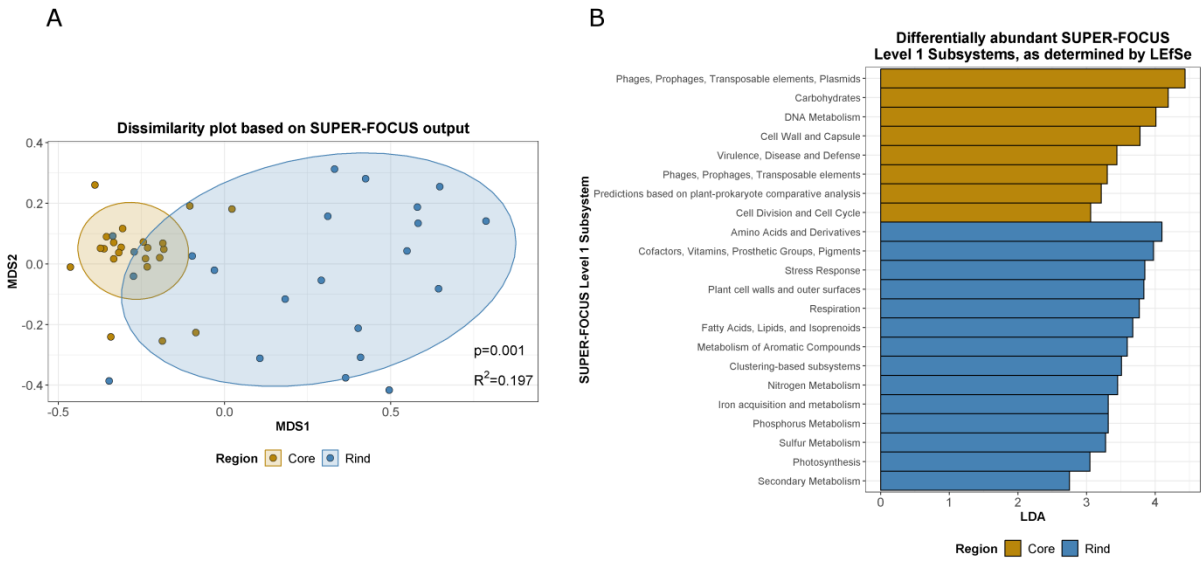
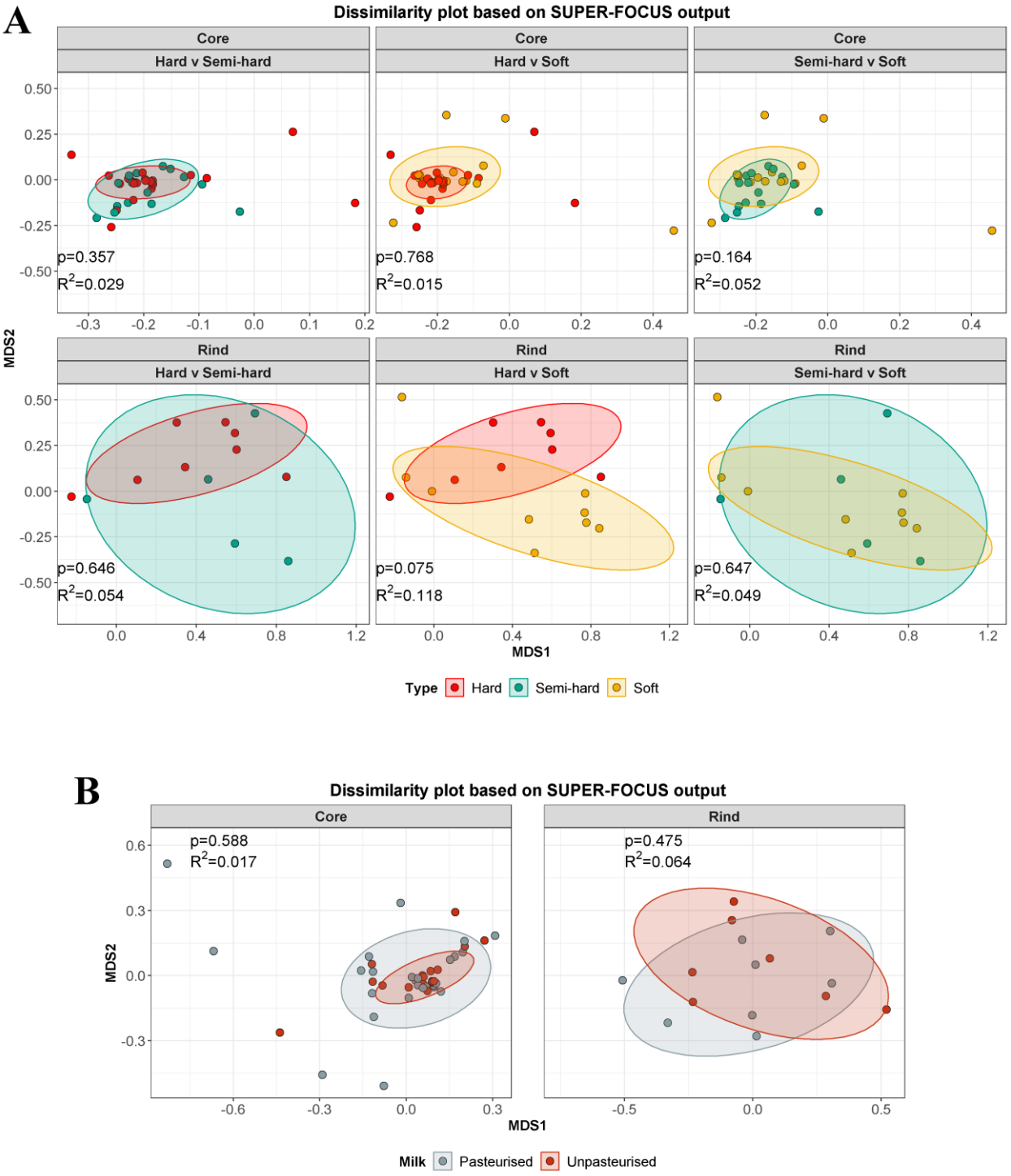


Figure S4. (A) The functional dissimilarity between the core and rind. (B) SUPER-FOCUS level 1 subsystems that were differentially abundant between the core and rind, as determined by LEfSe.



169

170 Figure S5. The functional dissimilarity between (A) cheeses of different maturity and (B) cheeses of pasteurized versus
171 unpasteurized cheeses.