Metagenomic Analysis of Dairy Bacteriophages: Extraction Method and Pilot Study on Whey Samples Derived from Using Undefined and Defined Mesophilic Starter Cultures

Running title

Metagenomic Analysis of Dairy Bacteriophages

Key words: abundance; bacteriophages; *Caudovirales*; dairy; diversity; *Lactococcus lactis*; *Leuconostoc*; mesophilic; metavirome; *Siphoviridae*; starter culture; whey

Musemma K. Muhammed[1], Witold Kot[2,3], Horst Neve[4], Jennifer Mahony[5,6], Josué L. Castro-Mejía[1], Lukasz Krych[1], Lars H. Hansen[3], Dennis S. Nielsen[1], Søren J. Sørensen[2], Knut J. Heller[4], Douwe van Sinderen[5,6], Finn K. Vogensen[1]*

[1]Department of Food Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg C, Denmark; [2]Department of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark; [3]Department of Environmental Science, Aarhus University, Frederiksborgvej 399, DK-4000 Roskilde, Denmark; [4]Department of Microbiology and Biotechnology, Max Rubner-Institut, Hermann-Weigmann-Str. 1, D-24103, Kiel, Germany; [5]School of Microbiology, University College Cork, Western Road, Cork, Ireland; [6]APC Microbiome Institute, University College Cork, Western Road, Cork, Ireland.

*Corresponding author: Tel: +45 353 33211; e-mail: fkv@food.ku.dk

Abstract

Despite their huge potential for characterizing the biodiversity of phages, metagenomic studies are currently not available for dairy bacteriophages, partly due to the lack of a standard procedure for phage extraction. We optimized an extraction method that allows to remove the bulk protein from whey and milk samples with losses of less than 50% of spiked phages. The protocol was applied to extract phages from whey in order to test the notion that members of *Lactococcus (Lc.) lactis* 936 (now *Sk1virus*), P335, c2 (now *C2virus*) and *Leuconostoc* phage groups are the most frequently encountered in the dairy environment. The relative abundance and diversity of phages in eight and four whey mixtures from dairies using undefined mesophilic mixed-strain DL-starter cultures and defined cultures, respectively, was assessed. Results obtained from transmission electron microscopy and high-throughput sequence analyses revealed the dominance of *Lc. lactis* 936 phages (order *Caudovirales*, family *Siphoviridae*) in dairies using undefined DL-starter cultures and *Lc. lactis* c2 phages (order *Caudovirales*, family *Siphoviridae*) in dairies using defined cultures. The 936 and *Leuconostoc* phages demonstrated limited diversity. Possible co-induction of temperate P335 prophages and satellite phages in one of the whey mixtures was also observed.

Importance

The method optimized in this study could provide an important basis for understanding the dynamics of the phage community (abundance, development, diversity, evolution, etc.) in dairies of different sizes, locations and production strategies. It may also enable the discovery of previously unknown phages, which is crucial for the development of rapid molecular biology-based methods for phage-burden surveillance systems. The dominance of only a few phage groups in the dairy environment signifies the depth of knowledge gained over the past decades, which served as the basis for designing current phage control strategies. The presence of correlation between phages and the type

of starter cultures being used in dairies might help to improve the selection and/or design of suitable, custom and cost-efficient phage control strategies.

Introduction

Mesophilic cheese production relies on the activity of specific strains of lactic acid bacteria (LAB), employed as so-called starter cultures, to carry out acidification of milk and produce the desired flavor. Starter cultures may contain one or more strains of *Lactococcus (Lc.) lactis* subsp. *lactis*, *Lc. lactis* subsp. *cremoris*, *Lc. lactis* subsp. *lactis* biovar. *diacetylactis*, *Leuconostoc (Le.) mesenteroides* subsp. *cremoris* or *Le. pseudomesenteroides* (1–4). Depending on the starter culture type, the number of strains can be single, multiple (up to six strains) or mixed (usually >50 different undefined strains). Single- and multiple-strain starters are normally defined-strain starters (DSS), unlike mixed-strain starters (MSS), whose strain composition is largely unknown (2, 3, 5–7).

Bacteriophages (phages) attacking starter cultures are of serious economic concern due to their negative impact on fermentations (8). In dairies, phages attacking *Lc. lactis* are the most frequent causes of delayed or arrested fermentations. Currently, *Lc. lactis* phages are divided into ten taxonomic groups, with members of three groups being the most frequently encountered in dairies (9). These are the virulent 936 (now designated as the *Sk1virus* genus) and c2 (now designated as the *C2virus* genus), and the heterogeneous P335 quasi-species constituting both virulent and temperate members. The remaining *Lc. lactis* phage groups (1358, Q54, P087, 949, 1706, P034 and KSY1) have been isolated much less frequently and more from raw milk rather than from failed fermentations (9).

Phages attacking *Leuconostoc* (*Le.*) species can affect the taste and appearance (e.g., eye formation) of the final product, but do not normally interfere with the acidification process. They are grouped

primarily into two major groups, *i.e.*, group I (phages attacking *Le. mesenteroides*) and group II (phages attacking *Le. pseudomesenteroides*) (10, 11).

The development of an early warning system for the detection of phage attacks requires understanding of the development, diversity and evolution of phages in the dairy environment, which in turn necessitates the integration of culture-dependent and culture-independent approaches. Traditionally, plaque assays, morphological characterization, restriction enzyme analysis, DNA-DNA hybridization, PCR-based approaches and sequencing have been primarily employed to characterize emerging phage isolates (6, 9, 12–21). The recent developments of high-throughput sequencing and metagenomics have intensified knowledge on the dynamics and distribution of phages in complex ecosystems (8, 22–26). Although it is possible to generate next-generation sequencing (NGS) libraries from <1 ng DNA (27), dairy phages have so far not been studied by using these approaches. This is partly due to the lack of a defined method for extracting the phage DNA from dairy samples.

An established method for the characterization of individual phage isolates is based on a combination of polyethylene glycol (PEG) precipitation (28, 29) and CsCl-gradient ultracentrifugation (30). Prior enrichment of the phages using susceptible host strains and subsequent purification provides highly concentrated phage stocks. CsCl gradients, particularly designed for phages propagated over susceptible host strains, enable the preparation of ultrapure phage lysates for morphological and genomic analyses. Adaptation of this method for metagenomic analysis of dairy phage communities requires several optimizations, as has previously been shown for studies of phage communities in the human gut (22). Important considerations include the need for adequate sample clarification prior to PEG-precipitation and the maintenance of the diversity of the phages during subsequent purification steps, including CsCl gradient ultracentrifugation. Whey samples from fermentation delays or breaks due to phage attacks of starter strains can, for instance, contain large amount of casein, which is

known to interfere with the phage (DNA) extraction process. Thus, the properties of dairy samples (presence of bacteria, cell debris, various proteins, contaminating nucleic acids, pH variations, etc.) would be the primary challenges for the design of a reliable method.

In the present study, we developed a method for dairy phage metagenomics and investigated the notion that the *Lc. lactis* 936, P335, c2 and *Leuconostoc* phage groups are the most frequently encountered in the dairy environment. We assessed whey samples obtained from whey factories in Denmark (D) and Ireland (I), and from a dairy in Germany (G) at three time points. The whey from Denmark and Germany were produced with undefined mesophilic mixed-strain DL-starter cultures (or simply DL-starter), a subgroup of MSS, whereas those from Ireland were produced with DSS. We analyzed the relative distribution of different *Lc. lactis* and *Leuconostoc* phages and assessed the overall taxonomic composition. Furthermore, we estimated the diversity of 936 and *Leuconostoc* phages based on analysis of the distribution of homologous receptor binding protein (RBP) sequences.

Materials and Methods

Optimization of isolation of phage communities from whey

Whey mixtures used in this study were stored at -60°C until needed to prevent potential inactivation of phages. Test whey samples were thawed (in a water bath, ≤30°C), spiked with ~10^6 mL^{-1} of *Bacillus subtilis* φ29 (31) and *Escherichia coli* T4 (32) phages and centrifuged at 300 x g for 5 min. The supernatant was incubated with 1 M NaCl (w/v) (Sigma Aldrich, USA) for 1 hr at 4°C and the mixture centrifuged at ~28,000 x g for 15 min (*Method 1*), filtered by 100-kDa NMWL Amicon centrifugal filters (Merck Millipore, USA) following the supplier's recommended procedures (*Method 2*) or adjusted to pH ~4.6 using 1 M HCl and/pr 1 M NaOH and centrifuged at ~28,000 x g for 15 min (*Method 3*) (see Fig. 1). The supernatant from *Methods 1* and *3* was incubated with 10%

PEG 6000 (w/v) (Sigma Aldrich) for 1 hr at 4°C and the phages pelleted by centrifugation at 15,000 x g for 15 minutes. The resultant pellet was resuspended by 1 mL SM buffer (18 mM $MgSO_4.7H_2O$, 0.1 M NaCl, 0.05 M Tris-HCl pH 8.0) and incubated overnight at 4°C. The pellet from *Method 3* was resuspended also by 1 mL SM buffer supplemented with 1.3 g mL^{-1} CsCl (Sigma Aldrich) and incubated as above. The phage suspension in SM buffer was purified by CsCl block (106,750 x g for 2.5 hrs) and equilibrium (175,000 x g for 22 hrs) gradient ultracentrifugations (both at 15°C), while that in CsCl-containing SM buffer was purified by a modified two-layer CsCl gradient centrifuged at 106,750 x g for 2.5 hrs at 15°C (Fig. 1). Phages were collected from the portion of the gradient containing visible bands or corresponding to 1.4–1.5 g mL^{-1} density using a needle installed on a syringe (the refractive index was measured using a handheld refractometer (Bellingham + Stanley, UK)). The phage extract was stored at -20°C until needed.

The samples were examined visually for clarity and by the double-agar plaque assay (22, 33) for recovery of spiked phages following NaCl-centrifugation, PEG-precipitation and CsCl gradient ultracentrifugation steps.

Isolation of phage communities and electron microscopy

Phages were isolated from 20 mL of whey mixture using *Method 3* combined with the modified two-layer CsCl gradient (see Fig. 1). Whey mixtures D3, D4, D5, G1 and I2 were spiked with ~10^6 mL^{-1} of φ29 and T4 phages prior to isolation, whereas D1, D2, G2, G3, I3, I4 and I5 were isolated without spiking. The phages were analyzed by Transmission Electron Microscopy (TEM) to assess their morphotypes. Adsorption of CsCl-purified phage lysates to an ultra-thin carbon film floated on a freshly coated mica sheet and negative staining with 2% (w/v) uranyl acetate were performed as previously described (34). The film was applied to a 400-mesh copper grid (Agar Scientific, United Kingdom) and images of the phages taken using MegaView G2 CCD camera (Emsis, Germany)

136 installed onto a Tecnai 10 transmission electron microscope (FEI, The Netherlands) operated at an
137 acceleration voltage of 80 kV.

138 **Extraction of phage metaviromes**

139 The phage lysate was dialyzed against dialysis buffer (10 mM NaCl, 50 mM Tris-Cl pH 8.0 and 10
140 mM $MgCl_2$) essentially as described by Sambrook and Russell (35). The pH of the suspension was
141 adjusted to 7.25±0.25 using 1 M HCl and/or 1 M NaOH and 0.1 volume of 10x DNase I reaction
142 buffer (10 mM Tris-HCl, 2.5 mM $MgCl_2$, 0.1 mM $CaCl_2$), 50 units mL^{-1} of DNase I (Sigma Aldrich)
143 and 1 µL mL^{-1} of RNase A solution (R6148, Sigma Aldrich) were added. The mixture was incubated
144 for 30 min at 37°C and DNase I inactivated by the addition of 10 mM EDTA pH 8.0 and 1% SDS
145 (both from Sigma Aldrich). The mixture was treated with 50 µg mL^{-1} Proteinase K and incubated at
146 55°C for 1 hr. Phage DNA was extracted with the GenElute Bacterial Genomic DNA Kit (Sigma
147 Aldrich) following the manufacturer's protocol and eluted with 50 µL elution buffer. The DNA was
148 stored at -20°C until required.

149 **High-throughput sequencing (HTS) and analysis**

150 **Phage sequence databases and phylogenetic tree**. The genomes of *Lc. lactis* (936, P335, c2, 1358,
151 Q54, P087, 949, 1706, P034, KSY1, bIL310, bIL311 and bIL312) and *Leuconostoc* phages available
152 in GenBank (https://www.ncbi.nlm.nih.gov/genome/viruses/) until June 15 2016 were downloaded
153 and, together with genomes obtained from in-house sources, were used to construct phage full-
154 genome sequence database called db-P. Accessible RBP sequences of 936 and *Leuconostoc* phages
155 were used (i) to construct in-house RBP database called db-RBP and (ii) to generate a phylogenetic
156 tree (parameters: neighbor-joining method, Jukes-Cantor distance measure with bootstrap analysis
157 with 100 replicates). According to this phylogenetic tree, the RBPs in the db-RBP were separated
158 into nineteen variants and were later used for estimation of phage diversity (see below).

159 **Optimization of analysis parameters**. Three Illumina NGS sequencing-simulated metaviromes
160 (mock communities) were constructed to optimize parameters for phage relative abundance and
161 diversity estimations. Each mock community comprised of one million reads from selected *Lc. lactis*
162 936, c2 and *Leuconostoc* phages (see Table S1). The mock communities were mapped to db-P (CLC
163 Genomics Workbench 8.5.1, Qiagen, Denmark) and queried against db-RBP variants (Ublast
164 algorithm (36)) through varying the query length and similarity thresholds until the expected phage
165 distribution was achieved. The outputs from db-RBP BLAST (amount of reads aligned to each RBP)
166 were transformed into RPKM (reads per kilobase per million mapped reads) values (37) through
167 normalization of the RBP size variations. The phage diversity in a given mock community was
168 estimated from the number of unique RBP variants formed in that metavirome following combining
169 of RPKM values corresponding to phylogenetically related RBPs (RBPs of the same variant).

170 **Test run**. A HTS sequencing library for D1 was prepared using the Nextera XT DNA kit (Illumina,
171 USA) according to the manufacturer's protocol. The library was deep sequenced as 2 x 250 paired
172 end (PE) reads. Unless indicated, analysis of reads was performed on CLC Genomics Workbench
173 8.5.1. Adapters, low quality reads (quality limit <0.05), reads containing >2 ambiguous nucleotides
174 and short reads (<15 bp) were discarded. The minimum number of reads required for phage relative
175 abundance and diversity estimations was assessed by analysis of an array of randomly selected reads
176 from D1 by (i) mapping to db-P (50% query length and 80% identity thresholds) and (ii) Ublast
177 against db-RBP variants (95% query length and identity thresholds).

178 **Final run**. Library construction, sequencing and trimming of the remaining metaviromes were
179 performed as described above except that I3, I4 and I5 were sequenced as 2 x 300 PE reads. The
180 spiked metaviromes (D3, D4, D5, G1 and I2) were mapped to the φ29 and T4 genomes using 50%
181 query length and 80% identity thresholds. Metaviromes were then assembled using the '*De Novo*
182 Assembly' tool using default parameters, including analysis of coverage (threshold: 50% query

length and 80% identity). Contigs were: (i) assessed with NCBI's BLAST tool (38) using default parameters and (ii) extracted and employed as references for analyzing the metaviromes using the 'Map Reads to References' tool using 50% query length and 80% identity thresholds. The relative abundance and diversity of different phages in each metavirome was estimated by mapping the trimmed reads to db-P or by Ublast against db-RBP, essentially as described above for D1, using one million reads or the entire metavirome as an input, respectively. The output from db-RBP analysis was further treated as described above.

Taxonomic composition. One million randomly selected trimmed reads were uploaded to Metavir (metavir-meb.univ-bpclermont.fr) for BLASTx comparison against the NCBI Refseq complete viral genomes protein sequences database. Taxonomic affiliations were deduced from the best BLAST hit (using threshold of 50 on the BLAST score) and was normalized by the genome length using the GAAS tool (39).

Metavirome Accessibility. The metaviromes can be accessed through the European Nucleotide Archive (40) under accession PRJEB17619.

Results

Sample Clarity and Recovery of Phages

High-speed centrifugation [with and without prior adjustment of pH (*Method 1* and *Method 3*, respectively) and membrane filtration (*Method 2*) were assessed in terms of two performance parameters: (i) the removal of bulk protein (as examined visually) and (ii) the recovery of spiked phages. In terms of removal of bulk proteins, *Method 3* appeared to be the most efficient, followed by *Method 2*, while *Method 1* was the least, indicating that pH adjustment is a simple yet very useful technique to precipitate proteins in milk-based samples. In terms of spiked phages, *Method 1* and *3*

were found to be more efficient than *Method 2*, retaining >50% of both φ29 and T4 compared to retention of ~50% using the latter technique. Thus, we considered that high-speed centrifugation was the most efficient method when coupled with prior pH-adjustment.

The use of 10% PEG 6000 often yielded precipitation of >90% of the phages in the clarified supernatant and hence we did not investigate an alternative phage precipitation technique.

Various CsCl density gradients were assessed in terms of their effectiveness to remove leftover whey proteins as well as to recover adequate amounts of spiked phages. Notably, multiple-layer gradients appeared to be generally less applicable when compared to a modified two-layer CsCl gradient (see the preparation on Fig. 1). The use of multiple-layer gradients failed to provide the intended clarity and recovery due to obstruction of the flow of phages by a band of proteins formed above the 1.3 g mL^{-1}-CsCl layer. The modified gradient, on the other hand, essentially diminished this band by relocating the majority of the proteins to the surface of the gradient. The fraction of spiked phages recovered in the final phage extract was also higher (>20%) when this latter gradient was used, compared to retention of only <20% with multiple-layer gradients. All in all, adequate sample clarity and recovery (37% of φ29 and ~24% of T4) could be achieved by combining *Method 3* with the modified two-layer CsCl gradient.

Phages losses were mainly attributed to pH shock, entrapment to discarded pellet and dispersion within the CsCl gradient. In the most efficient method (Fig. 1), the three factors contributed to losses of up to 6%, 20% and 34% of φ29 and 14%, 35% and 26% of T4, respectively.

Phage Morphologies

TEM analysis was conducted in order to examine the morphologies of the most predominant phages. Phage particles were observed only in six samples. Phages could not be observed in the remaining

samples due to sub-threshold phage titers (the limit of detection for TEM is ~10^5 – 10^6 phages mL^{-1}) (41, 42). Representative TEM micrographs are presented in Fig. 2. Notably, except for one sample (I4), all the other samples contained phages with isometric capsids and long non-contractile tails. The majority of the isometric-headed phages exhibited tail lengths of 141 nm (D3, phage on the right side) to 151 nm (D4, phage on the left side), but shorter tail lengths were also measured ranging from 120 nm (D5, phage on the right side) to 139 nm (D2, phage on the right side). In sample I4, a prolate-headed phage was observed with a tail length of 100 nm. The phages did not demonstrate distinct baseplates but rather exhibited slightly enlarged tail terminal ends. Six of the phages demonstrated neck passage structures (NPS) at the head-tail junction. The overall dimensions and morphologies of the isometric-headed phages are generally similar to those documented for previously isolated phages of the 936 group (15, 43–46).

Optimum Parameters

Mock communities were mapped to db-P or db-RBP databases in order to optimize parameters for analysis of sample metaviromes. Reads were mapped to db-P as expected when the length and similarity thresholds were adjusted to 50% and 80%, respectively (Fig. S1). Some reads from the 936 phages were mapped to the genomes of P335 and 949 phages due to relatively short regions of DNA sequence homology between the phages (Fig. S2). Such regions correspond to, for instance, *orf11* and *orf47* (*nps*) of phage TP901-1 (a P335 species) and *gp047*, *gp062*, *gt004*, *gp089-gt005* intergenic space, *gt005* and *gp128* of phage 949 (a 949 species) (Fig. S2).

Ublast analysis of the mock communities against db-RBP yielded the expected profile when a stringent mapping threshold was used, *i.e.* 95% minimum query length and similarity (Fig. S1B and S1C). Importantly, this analysis also revealed RBP sequences that were not represented in the mock communities, such as the 1727 and Q49 RBP variants (Fig. S1B). Further analysis of these RBP

variants indicated that they show considerable sequence similarity to the fd13 (e.g., fd13 and CaseusJM1) (75-80%) and the HD6 (e.g., HD6 and jm3) (69-82%) RBP clusters, despite being phylogenetically unrelated. Additionally, most of the RBP sequences of the 936 phages exhibit high sequence similarity towards the 5' half, which could have increased the possibility of detection of sequences that were not included in the mock communities.

Analysis of Spiked Phages

Reference mapping. Metavirome reads were mapped to the spiked phage genomes as summarized in Table 1. Nearly the entire φ29 genome was mapped in all cases (99.9%), with significant fluctuation in depth of coverage along the reference genome. At the terminal ends of φ29, regions with no coverage were observed encompassing 11, 18, 6, 8 and 17 bases (D3, D4, D5, G1 and I2, respectively) (Fig. S3A). The fraction of reads mapped to φ29 ranged from 0.2% (D4) to 7.9% (D5), which was generally higher compared to the fraction of reads mapped to T4 [0.2 (D4) to 4.3% (D5)]. This was also consistent with the average coverage (Table 1). With regard to T4, although the mapped reads fraction and the average coverage was relatively low, the mapping was nearly complete in D5 (99.5%) and G1 (99.4%). The unmapped portion of the genome corresponded to the terminal 837 (D5) and 957 (G1) bases (Fig. S3B). A very high mapping coverage was also seen in D3 and D4 (94.6% and 93.2%, respectively), although multiple gaps were formed along the alignment. I2 provided the least T4 coverage (58%), which is reasonable given a very small metavirome size (Fig. S3B).

***De novo* assembly**. The method's power to yield adequate sequences for full genome characterization was inferred from *de novo* assembly of the spiked phages. Nearly the entire φ29 genome was assembled, except the proximate terminal regions, where 56, 1, 6, 8 and 17 bases were absent (D3, D4, D5, G1 and I2, respectively). The assembly was further confirmed by BLAST query

to have significant similarity to the φ29 genome (Table S3). Unlike φ29, T4 was assembled into several short contigs [except in D5 (168,032 bases) and G1 (104,664 bases)], which was apparently due to inadequate amount of reads to cover the very large T4 genome (Table S3). Importantly, many of the BLAST hits corresponding to the T4 contigs were described as *Enterobacteria* phage RB55 or RB59 (see Table S5). These phages belong to the *T4virus* phages group with 99.96% identity to each other and 99.80% similarity to the T4 phage (47).

Analysis of HTS Metagenome Sequences

Input sequence threshold. To test if variations in the amount of input sequences can affect estimation of the relative abundance of phage species and RBP diversity, a test sequencing run and preliminary analysis was carried out with D1 prior to the other metaviromes. With deep sequencing, a total of ~9.5 million PE reads were generated, ~5% of which were removed by trimming. In order to assess the impact of varying the amount of input sequences on phage relative abundance and diversity estimations, an array of randomly selected sequences were tested (Table S2). Eventually, the minimum requirement for accurate estimation of relative abundance and diversity of phages was found to be ~50,000 and ~1,000,000 reads, respectively, indicating that the latter demands large number of input sequences. Accordingly, we devised to generate at least 1 million reads per each of the study samples. Sequencing of eleven samples generated a total of >12.6 million PE reads (Fig. S4), of which 0.04% was discarded by trimming.

Comparison of phage content between pairs of metaviromes. In order to assess the overall similarity in phage content between the metaviromes, they were mapped to each other in a pairwise manner as outlined in Table 2, resulting in a matrix of 121 mapping scores. The matrix clearly indicated that DSS and DL starter-derived whey samples vary considerably in terms of overall phage composition. The similarity score from comparison of samples derived from the same starter culture

296 type (DSS or DL) was generally higher, compared to when DSS-starter derived samples were
297 compared to DL starter-derived samples. This suggests the presence of significant overlap in phage
298 composition among samples corresponding to the same starter culture type. Nonetheless, the overall
299 composition of I2 appeared to be relatively similar to DL-starter derived samples, which was
300 unexpected for a sample that was derived from DSS starters.

301 **BLAST comparison of selected contigs**. To test whether certain phage strains were responsible for
302 the observed sequence similarity among samples derived by the same type of starter, a set of contigs
303 selected of being the highest in read coverage in the respective metavirome assemblies were
304 analyzed by BLAST query against the NCBI nr database. Notably, the contigs selected from ten out
305 of the eleven metaviromes displayed the greatest similarity to the lactococcal phage sequences in the
306 database (Table S4). Accordingly, the dominant sequence entities in the Danish and German
307 metaviromes appeared to be those of *Lc. lactis* 936 phages, whereas in the Irish metaviromes
308 sequences of *Lc. lactis* 936 phages (I2), bacteria (particularly *Streptococcus pyogenes*) (I3) or *Lc.*
309 *lactis* c2 phages (I4 and I5) were dominant. Contigs with no sequence similarity to available
310 sequences were also represented in G3 and I3 (Table S4).

311 As majority of the selected contigs were rather short in length, we tested the universality of the
312 observed finding by analyzing a set of contigs selected of being the longest. To this end, sequences
313 of lactococcal phages (in D2-D5 and I2), bacteria (in G2, G3 and I3-I5) or a mixture of both phage
314 and bacteria (in G1) appeared to represent the dominant sequence entities in the metaviromes (Table
315 S5). Contigs with no sequence similarity to available sequences were also represented in G1 and G3,
316 possibly indicating sequences of novel phages or viruses, or of contaminating host cells.

317 **Mapping reads to reference genomes**. To estimate the fraction of phages associated to mesophilic
318 cheese production, namely *Lc. lactis* and *Leuconostoc* phages, in the metaviromes, reads were

mapped to db-P (in-house database of reference genomes). A total of one million randomly selected trimmed reads were used, except for I2, which was mapped entirely due to its small size (~0.33 million reads). Of the eleven metaviromes, nine presented very high proportion of mapped reads, with the average being ~74% (Fig. S4). The two exceptional metaviromes presenting relatively low proportion of reads mapping to the database, namely I3 and I5 (~9.6% and ~51.5%, respectively), were both from the Irish whey factories (Fig. S4).

Of the eleven metaviromes, eight [encompassing all the Danish and German samples and the Irish sample I2] demonstrated the dominance of sequences of the 936 phage group (Table 3). The fraction of mapped reads ranged from 90.0% to 91.5% (in the Danish samples), 85.0 to 93.0% (in the German samples) and 0.6% to 95.1% (in the Irish samples). In the three Irish metaviromes, where 936 phages were a minority, namely I3, I4 and I5, sequences of the c2 phage group appeared to be the most frequent. The fraction of mapped reads ranged from 0.2% (I2) to 97.8% (I4) and averaged ~20.8%. The mapping fraction for I3, I4 and I5 alone was significantly higher than the rest of the samples, averaging ~80.6%. Therefore, phages of the 936 group were the most frequent in DL-starter-derived samples, while phages of the c2 group represented the vast majority in DSS-starter-derived samples.

Compared to the 936 and c2 phages, sequences of the P335 phage group demonstrated a more consistent frequency. The fraction of mapped reads was in the range of 0.9% (I4) to 6.8% (G1), except in I3 (36.1), and averaged ~6.7%. Contrary to the finding observed in the mock communities (Fig. S2), the vast majority of the P335 genomes were covered by reads, although numerous gaps were also present (see Fig. S5). This is suggestive of mapping of primarily P335 sequences, although it is impossible to rule out non-specific mapping of sequences from the 936 phages. Furthermore, reads that mapped to the 949 genomes averaged ~1%: however, many of these reads were concentrated around regions corresponding to *gp047*, gp062, *gt004*, *gp089-gt005* intergenic space, *gt005*, gt006 and *gp128* of phage 949 (Fig. S5). This might indicate non-specific mapping of reads

343 potentially originating from 936 phages. The remaining phages were found to be insignificant,
344 except *Leuconostoc* and satellite phages that represented 4.1% and 11.6% in G1 and I3, respectively.
345 These results generally indicate that the 936 species is highly dominant in dairies using DL-starters,
346 as is the c2 species in most dairies using DSS.

347 The presence of a large fraction of unclassified reads in some of the metaviromes, such as D3 and I5,
348 prompted us to perform BLAST analysis of selected *de novo*-assembled contigs. The results
349 indicated lactococcal phage sequences to be the most frequent hits for a set of contigs selected of
350 being the highest in read coverage (Table S6). Of these, sequences of 936 phages appeared to be the
351 majority, except that sequence of c2 phages occasionally encountered in I4 and I5. A set of contigs
352 selected of being the longest presented the greatest similarity to phages or viral sequences. Many of
353 these corresponded to sequences of spiked phages (Table S7), while the fraction of lactococcal
354 phages appeared to be rather low (~5%). BLAST also revealed bacterial sequences (likely from
355 contamination) in the metavirome extracts, mainly in the German and Irish samples (Table S7).

356 **Taxonomic composition**. To estimate the taxonomic composition of the metaviromes, we executed
357 a BLAST comparison with the RefSeq complete viral genomes proteins. Taxonomic affiliations were
358 deduced from the best BLAST hit (threshold of 50 on the BLAST score). The ratio of taxonomic
359 affiliation was very high in the majority of the samples, although it ranged from 12.5% (I3) to 91.8%
360 (D2), with an average affiliation of 63.8%. Notably, no obvious correlation between the degree of
361 taxonomic affiliation and the origin of the samples was observed.

362 To estimate the proportion of each virotype in the initial sample in terms of the number of viral
363 particles, taxonomic composition normalized by the genome length of the virotypes was performed,
364 computed via the GAAS tool (39). Notably, certain virotypes appeared to dominate in the
365 metaviromes (Table 4). dsDNA viruses belonging to the order *Caudovirales* and the family

366 *Siphoviridae* accounted for majority of the viruses identified. At lower taxonomic ranks, lactococcal
367 phages were found to be highly abundant in all the samples. In all the Danish and German plus the
368 Irish sample I2, *Lc. lactis* 936 phages P008, jm2, P680, jm3, biL170, 340, phage 7, jj50, biBB29,
369 712 and sk1 appeared to be evenly distributed, whereas *Lc. lactis* c2 phages bIL67 and c2 dominated
370 in the remaining samples, which were all from Ireland (Table 4). P335 phages generally exhibited
371 low abundance, except in I3 (~7%). Sequences corresponding to phages TP901-1, BK5-T, Tuc2009,
372 ul36, bIL286 and φLC3 were relatively abundant in I3 (Table 4). In this sample, satellite phages
373 represented ~17% of the viruses (bIL310 ~12%, bIL311 ~1% and bIL312 ~4%). These phages
374 however appeared insignificant in the other metaviromes (Table 4).

375 **Diversity of 936 and *Leuconostoc* phages.** To estimate the diversity of phages, we analyzed reads
376 by Ublast against db-RBP (in-house database of RBP sequences). The result is summarized in Table
377 S8. Of nineteen RBP variants representing 936 phages, a maximum of four appeared to be
378 predominant per sample. Two variants, namely SCH and phage 7, were found to be widely
379 distributed, whereas others including 645, ASCC406, 1727 and ΦL.6 seemed to be associated to
380 many of the samples. Notably, SCH was almost the sole variant detected in G1, while phage 7 was
381 the major variant in D5, G3 and I2. In the Irish samples I3, I4 and I5, generally very small quantity
382 of RBPs were detected, which is consistent to the low prevalence of 936 phages in these samples
383 (Table S8).

384 Sequences corresponding to the RBP sequences of *Leuconostoc* phages were detected in just three
385 samples (D5, G1 and G3) (Table S9), with the LN6B variant being more frequent than the others.
386 Very small amount of sequences corresponding to the P793 variant was also detected in G1 and G3
387 (Table S9). The LN6B variant, together with the P793 and φLN23 variants, represents *Le.*
388 *pseudomesenteroides* phages, possibly suggesting the dominance of these phages over *Le.*
389 *mesenteroides* phages in the underlying dairies.

Discussion

One of the primary challenges with the extraction of phages from different dairy metagenomic samples is the difficulty to separate various residues (proteins, bacterial cells and cell debris) before concentrating the phages. The pH variation among different dairy samples (bulk starter, milk, whey, etc.) partly explains the processing challenges. The extraction method described in this study included a step, where samples were adjusted to pH ~4.6, which has been shown to precipitate ~80% of the total nitrogen in dairy samples (48). Incorporation of this step facilitated removal of a significant fraction of residues by low-speed centrifugation. However, there was unexpected loss of phages during this initial sample processing stage, which was mainly attributed to pH shock (loss of infectivity) and entrapment of phages in the pellet. The former loss seemed reasonable as the spiked phages were not adapted to the dairy environment and hence could have been relatively sensitive to low pH. Conversely, dairy phages tolerate low pH and hence minimal (if any) losses may be expected.

Up on PEG precipitation of phages, smaller proteins likely pose the greatest challenge as they precipitate concurrently with the phages. These proteins are impossible to separate by standard CsCl gradient ultracentrifugation (35), as they tend to aggregate at the interface between the phage suspension and the CsCl layer, preventing the flow of phages to their buoyant density (isopycnic point). Thus, a modified two-layer CsCl gradient was optimized that could lower the blocking effect of proteins on the flow of the phages. It consists of an upper layer of phage suspension (prepared by resuspedning the PEG-pellet with 1.3 g mL^{-1} CsCl in SM buffer) and a lower layer of 1.7 g mL^{-1} CsCl. The usefulness of this gradient lies on its potential to alter the direction of flow of proteins (density <1.3 g mL^{-1}) without affecting the direction of flow of (most) phages (>1.3 g mL^{-1}), which was proved by the aggregation of proteins on the top of the gradient. The higher density of the phage suspension increases the physical distance between the phages and smaller proteins within the

414 gradient, which ultimately increases the sample clarity. The absence of intermediate CsCl layers
415 (1.45 and/or 1.5 g mL^{-1}) increases the phage diversity being extracted as it promotes the gathering of
416 the phages just above the 1.7 g mL^{-1} CsCl layer.

417 Overall, the extraction method enables progressive removal of bulk proteins while retaining adequate
418 fraction of phages for metagenomic studies. The recovery of adequate φ29 and T4 phages for
419 assembly of almost the entire genomes, as demonstrated by reference mapping (both phages) and *de*
420 *novo* assembly (only φ29), demonstrates this. Of note, the observed drop in sequence coverage
421 towards the terminal ends of the genomes is attributed to the Nextera XT transposome technology
422 (49). The protocol's ability to recover T4, which has a very large genome (>8 times the size of φ29)
423 (31, 32), proves its suitability for metagenomic characterization of a wide variety of phages,
424 including dairy phages. Yet, the low success rate with *de novo* assembly of the T4 genome suggests
425 its strength towards phages of smaller genome sizes. Besides, the relatively low efficiency of
426 recovery would not allow the protocol to be combined with, for instance, TEM examination due to a
427 detection threshold of 10^5-10^6 phages mL^{-1} (41, 42). Thus, further optimizations are needed to
428 improve the efficiency of recovery, thereby to extend the methods' applicability.

429 The influence of starter culture type on the composition of phages is evident from the predominance
430 of 936 phages in whey from DL-starters and of primarily c2 phages in whey from DSS. This,
431 combined with the relatively conserved nature of the genomes of 936 and c2 phages, could have
432 contributed to the observed overall similarity among the different metaviromes. The situation might
433 have been different, if the heterogeneous P335 phages were dominant, since these phages show much
434 less overall genome conservation (50, 51). Nevertheless, the dominance of 936 phages in Danish and
435 German whey appears to be due to the employment of undefined cultures, whereas the dominance of
436 c2 phages in most Irish whey is likely due to application of strains sensitive to specific phages rather
437 than the starter type itself.

438 The dominance of 936 phages in the present metavirome study is consistent with most previous
439 lactococcal phage isolation studies (6, 18, 46, 52–59). Detection of mostly SCH and phage 7 and
440 occasionally 645, ASCC406, 1727 and ΦL.6 RBP variants may suggest the presence of limited
441 diversity of these phages in the dairies. Phages within these RBP variants mostly infect strains of *Lc.*
442 *lactis* subsp. *cremoris*, while the 645 RBP variant infects strains of both *Lc. lactis* subsp. *lactis* and
443 *Lc. lactis* subsp. *cremoris* (45). Furthermore, the presence of NPS in the majority of phages is
444 consistent with the recent increase in the frequency of phages displaying this structure (18, 46, 58,
445 60). This highlights the view that NPS might enhance host range and adaptation (43), though the
446 NPS of the temperate TP901-1 phage does not appear to be important for infection, assembly and
447 stability (61). Thus, the function of NPS for 936 phages is yet to be elucidated.

448 Probably, the most striking finding of the analysis of the c2 phages is their dominance in the majority
449 of samples derived from defined cultures, which is in stark contrast to the notion that the frequency
450 of isolation of members of c2 species has dropped lately due to the adoption of anti-phage strategies
451 (14). As far as we are aware, no previous study has published similar findings in Ireland. Elsewhere,
452 c2 phages have occasionally been reported to be more frequent than other lactococcal phages (62,
453 63). The finding may indicate that the c2 phages in Irish dairies are relatively tolerant to cheese milk
454 thermal treatments, which is in accordance with a recent study that reported that many c2 phages
455 survive traditional cheese milk heat treatments (64). Phages of the c2 group require the host receptor
456 phage infection protein (PIP) in order to attack sensitive strains (65, 66). Strains that carry mutations
457 in *pip* acquire complete resistance against c2 phages (67). It appears that starter culture
458 manufacturers in Ireland do not generally select for *pip* mutants or that other genes such as *yjaE* (68)
459 could substitute *pip* during infection.

460 The relatively low abundance of P335 phages appears to be in accordance with the reported decrease
461 in the frequency of isolation of P335 phages due to the adoption of anti-phage strategies (14). It is

462 possible that the increased frequency of P335 phages in I3 was primarily due to induction of
463 temperate P335 prophages during cheese production. The increased frequency of satellite phages in
464 the same whey mixture might thus be due to co-induction of P335 and satellite phages. Chopin *et al.*
465 suggested that satellite phages of *Lc. lactis* possibly rely on phages from the P335 group for
466 multiplication (50). It is, therefore, possible that the satellite phages in I3 have acquired certain
467 modules from P335 phages prior to induction, which may indicate the presence of co-evolution of
468 these phages in the corresponding dairy.

469 *Lc. lactis* phages other than 936, P335 and c2 were generally detected much less frequently.
470 Although the 949 phage group was detected in all whey mixtures, further evaluation of individual
471 mappings revealed mostly localized mapping of reads at the regions corresponding to *gp047*, *gp062*,
472 between *gp089* and *gt005* and *gp128* in 949, likely suggesting non-specific mapping of reads
473 originating potentially from 936 phages such as CB13 (69). This is consistent with the taxonomic
474 composition analysis, which revealed just a small fraction of 949 phages. Furthermore, *Leuconostoc*
475 phages were mostly represented by the LN6B RBP variant, indicating that *Le. pseudomesenteroides*
476 phages were relatively more common. The exceptional abundance of *Leuconostoc* phages in G1 was
477 reflected also by detection of large number of *Leuconostoc* phages RBP in this metavirome.

478 In conclusion, the method described in this study allows for metagenomic studies of dairy phages.
479 Particularly, the CsCl gradient enables isolation of phages with a wide range of genome sizes. The
480 findings from this study support the previous notion that *Lc. lactis* phages of the 936, P335 and c2
481 species are the most frequently encountered in the dairy environment. The composition of the phage
482 population was somehow linked to the starter culture propagation regime, as shown from the
483 dominance of 936 phages in whey from DL-starters and that of c2 phages in (most) whey from DSS.
484 Concurrent increases in the frequency of P335 and satellite phages may indicate co-induction and
485 evolution of these phages. Future work would possibly describe the mechanisms on how starter

culture types influence the composition of the phage population.

Acknowledgements

This work was funded by FTP (Project No.: 0602-022170B) and was carried out as part of the MetaPhageLAB project. We would like to acknowledge the whey factories for their cooperation, and Angela Back (Max-Rubner Institut [MRI], Germany) and Bashir Aideh (University of Copenhagen, Denmark) for technical assistance. Musemma K. Muhammed has been the recipient of a PhD stipend from the MetaPhageLAB project. Jennifer Mahony is the recipient of a Starting Investigator Research Grant (SIRG) (Ref. No. 15/SIRG/3430) funded by Science Foundation Ireland (SFI). Josué L. Castro-Mejía was supported by CALM. Douwe van Sinderen is supported by a Principal Investigator award (Ref. No. 13/IA/1953) through SFI.

References

1. **Atamer Z**, **Ali Y**, **Neve H**, **Heller KJ**, **Hinrichs J**. 2011. Thermal resistance of bacteriophages attacking flavour-producing dairy *Leuconostoc* starter cultures. Int Dairy J **21**:327–334.

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

3. **Smid EJ**, **Erkus O**, **Spus M**, **Wolkers-Rooijackers JC**, **Alexeeva S**, **Kleerebezem M**. 2014. Functional implications of the microbial community structure of undefined mesophilic starter cultures. Microb Cell Fact **13 Suppl 1**:S2.

4. **Tariq MA**, **Louise F**, **Everest C**, **Cowley L**, **Soyza A De**, **Holt GS**, **Bridge H**, **Perry A**,

Perry JD, **Bourke S**, **Cummings S**, **Lanyon CV**, **Barr JJ**, **Smith DL**. 2015. A Metagenomic approach to characterize temperate bacteriophage populations from cystic fibrosis and non-cystic fibrosis bronchiectasis patients. Front Microbiol **6**:1–12.

5. **Stadhouders J**, **de Vos WM**. 1991. Starter cultures for cheese production, p. 77–112. *In* Currell, BC, Dam-Mieras, RCE (eds.), Biotechnological Innovations in Food Processing, 1st ed. Butterworth-Heinemann Ltd, Oxford.

6. **Kleppen HP**, **Bang T**, **Nes IF**, **Holo H**. 2011. Bacteriophages in milk fermentations: Diversity fluctuations of normal and failed fermentations. Int Dairy J **21**:592–600.

7. **Nielsen EW**. 1999. Long term use of a cheddar starter and development of phages with homology to its bacteria. Int Dairy J **8**:1003–1009.

8. **Park EJ**, **Kim KH**, **Abell GCJ**, **Kim MS**, **Roh SW**, **Bae JW**. 2011. Metagenomic analysis of the viral communities in fermented foods. Appl Environ Microbiol **77**:1284–1291.

9. **Deveau H**, **Labrie SJ**, **Chopin MC**, **Moineau S**. 2006. Biodiversity and classification of lactococcal phages. Appl Environ Microbiol **72**:4338–4346.

10. **Ali Y**, **Kot W**, **Atamer Z**, **Hinrichs J**, **Vogensen FK**, **Heller KJ**, **Neve H**. 2013. Classification of lytic bacteriophages attacking dairy *Leuconostoc* starter strains. Appl Environ Microbiol **79**:3628–3636.

11. **Kot W**, **Hansen LH**, **Neve H**, **Hammer K**, **Jacobsen S**, **Pedersen PD**, **Sørensen SJ**, **Heller KJ**, **Vogensen FK**. 2014. Sequence and comparative analysis of *Leuconostoc* dairy bacteriophages. Int J Food Microbiol **176**:29–37.

12. **Azaïez SRC**, **Fliss I**, **Simard RE**, **Moinau S**. 1998. Monoclonal antibodies raised against

native major capsid proteins of lactococcal c2-like bacteriophages. Appl Environemntal Microbiol **64**:4251–4259.

13. **Emond E**, **Holler BJ**, **Boucher I**, **Vandenbergh PA**, **Vedamuthu ER**, **Kondo JK**, **Moineau S**. 1997. Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from *Lactococcus lactis*. Appl Environ Microbiol **63**:1274–1283.

14. **Garneau JE**, **Tremblay DM**, **Moineau S**. 2008. Characterization of 1706, a virulent phage from *Lactococcus lactis* with similarities to prophages from other *Firmicutes*. Virology **373**:298–309.

15. **Jarvis AW**. 1984. Differentiation of lactic streptococcal phages into phage species by DNA-DNA homology. Appl Environ Microbiol **47**:343–349.

16. **Jarvis AW**. 1978. Serological studies of a host range mutant of a lactic streptococcal bacteriophage. Appl Environ Microbiol **36**:785–789.

17. **Moineau S**, **Pandian S**, **Klaenhammer TR**. 1994. Evolution of a lytic bacteriophage via DNA acquisition from the *Lactococcus lactis* chromosome. Appl Environ Microbiol **60**:1832–1841.

18. **Rousseau GM**, **Moineau S**. 2009. Evolution of *Lactococcus lactis* phages within a cheese factory. Appl Environ Microbiol **75**:5336–5344.

19. **Verreault D**, **Gendron L**, **Rousseau GM**, **Veillette M**, **Massé D**, **Lindsley WG**, **Moineau S**, **Duchaine C**. 2011. Detection of airborne lactococcal bacteriophages in cheese manufacturing plants. Appl Environ Microbiol **77**:491–497.

20. **Mahony J**, **van Sinderen D**. 2014. Current taxonomy of phages infecting lactic acid bacteria.

549 Front Microbiol **5**:1–7.

550 21. **Kot W**, **Neve H**, **Heller KJ**, **Vogensen FK**. 2014. Bacteriophages of *Leuconostoc*,
551 *Oenococcus*, and *Weissella*. Front Microbiol **5**:1–9.

552 22. **Castro-Mejía JL**, **Muhammed MK**, **Kot W**, **Neve H**, **Franz CMAP**, **Hansen LH**,
553 **Vogensen FK**, **Nielsen DS**. 2015. Optimizing protocols for extraction of bacteriophages prior
554 to metagenomic analyses of phage communities in the human gut. Microbiome **3**:1–14.

555 23. **Dutilh BE**, **Cassman N**, **McNair K**, **Sanchez SE**, **Silva GGZ**, **Boling L**, **Barr JJ**, **Speth**
556 **DR**, **Seguritan V**, **Aziz RK**, **Felts B**, **Dinsdale E a.**, **Mokili JL**, **Edwards RA**. 2014. A
557 highly abundant bacteriophage discovered in the unknown sequences of human faecal
558 metagenomes. Nat Commun **5**:1–11.

559 24. **Willner D**, **Thurber RV**, **Rohwer F**. 2009. Metagenomic signatures of 86 microbial and viral
560 metagenomes. Environ Microbiol **11**:1752–1766.

561 25. **Cottrell MT**, **Kirchman DL**. 2012. Virus genes in Arctic marine bacteria identified by
562 metagenomic analysis. Aquat Microb Ecol **66**:107–116.

563 26. **Sharon I**, **Battchikova N**, **Aro E-M**, **Giglione C**, **Meinnel T**, **Glaser F**, **Pinter RY**,
564 **Breitbart M**, **Rohwer F**, **Béjà O**. 2011. Comparative metagenomics of microbial traits within
565 oceanic viral communities. ISME J **5**:1178–1190.

566 27. **Illumina**. 2014. Nextera library validation and cluster density optimization. Illumina, Inc.

567 28. **Albertsson P**, **Frick G**. 1960. Partition of virus particles in a liquid two-phase system.
568 Biochim Biophys Acta **15**:230–7.

569 29. **Philipson L**, **Albertsson PÅ**, **Frick G**. 1960. The purification and concentration of viruses by
570 aqueous polymer phase systems. Virology **11**:553–571.

571 30. **Bachrach U**, **Friedmann A**. 1971. Practical procedures for the purification of bacterial
572 viruses. Appl Microbiol **22**:706–715.

573 31. **Meijer WJJ**, **Horcajadas JA**, **Salas M**. 2001. phi29 family of phages. Microbiol Mol Biol
574 Rev **65**:261–287.

575 32. **Miller ES**, **Kutter E**, **Mosig G**, **Kunisawa T**, **Rüger W**, **Arisaka F**, **Ru W**. 2003.
576 Bacteriophage T4 genome. Microbiol Mol Biol Rev **67**:86–156.

577 33. **Lillehaug D**. 1997. An improved plaque assay for poor plaque-producing temperate
578 lactococcal bacteriophages. J Appl Microbiol **83**:85–90.

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

582 35. **Sambrook J**, **Russell DW**. 2001. Molecular cloning: A laboratory manual, 3rd ed. Cold
583 Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

584 36. **Edgar RC**, **Bateman A**. 2010. Search and clustering orders of magnitude faster than BLAST.
585 Bioinforma Appl NOTE **26**:2460–246110.

586 37. **Mortazavi A**, **Williams BA**, **McCue K**, **Schaeffer L**, **Wold B**. 2008. Mapping and
587 quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5**:621–628.

588 38. **Altschup SF**, **Gish W**, **Miller W**, **Myers EW**, **Lipman DJ**. 1990. Basic local alignment

589 search tool. J Mol Biol **215**:403–410.

590 39. **Angly FE**, **Willner D**, **Prieto-Davo A**, **Edwards RA**, **Schmieder R**, **Vega-Thurber R**,
591 **Antonopoulos DA**, **Barott K**, **Cottrell MT**, **Desnues C**, **Dinsdale EA**, **Furlan M**, **Haynes**
592 **M**, **Henn MR**, **Hu Y**, **Kirchman DL**, **McDole T**, **McPherson JD**, **Meyer F**, **Miller RM**,
593 **Mundt E**, **Naviaux RK**, **Rodriguez-Mueller B**, **Stevens R**, **Wegley L**, **Zhang L**, **Zhu B**,
594 **Rohwer F**. 2009. The GAAS metagenomic tool and its estimations of viral and microbial
595 average genome size in four major biomes. PLoS Comput Biol **5**:1–13.

596 40. **European Nucleotide Archive**. http://www.ebi.ac.uk/ena.

597 41. **Laue M**, **Bannert N**. 2010. Detection limit of negative staining electron microscopy for the
598 diagnosis of bioterrorism-related micro-organisms. J Appl Microbiol **109**:1159–1168.

599 42. **Ackermann H-W**. 2009. Basic phage electron microscopy, p. 113–126. *In* Martha R. J.
600 Clokie, Andrew M. Kropinski (eds.), Bacteriophages: Methods and Protocols, Vol. 1:
601 Isolation, Characterization and Interaction. Humana Press, New York.

602 43. **Crutz-Le Coq A-M**, **Cantele F**, **Lanzavecchia S**, **Marco S**. 2006. Insights into structural
603 proteins of 936-type virulent lactococcal bacteriophages. Arch Virol **151**:1039–1053.

604 44. **Dupont K**, **Vogensen FK**, **Neve H**, **Bresciani J**, **Josephsen J**. 2004. Identification of the
605 receptor-binding protein in 936-species lactococcal bacteriophages. Appl Environ Microbiol
606 **70**:5818–5824.

607 45. **Mahony J**, **Kot W**, **Murphy J**, **Ainsworth S**, **Neve H**, **Hansen LH**, **Heller KJ**, **Sørensen**
608 **SJ**, **Hammer K**, **Cambillau C**, **Vogensen FK**, **Van Sinderen D**. 2013. Investigation of the
609 relationship between lactococcal host cell wall polysaccharide genotype and 936 phage

receptor binding protein phylogeny. Appl Environ Microbiol **79**:4385–4392.

46. **Murphy J**, **Bottacini F**, **Mahony J**, **Kelleher P**, **Neve H**, **Zomer A**, **Nauta A**, **van Sinderen D**. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. Sci Rep **6**:1–13.

47. **Yaung SJ**, **Esvelt KM**, **Church GM**. 2015. Complete genome sequences of T4-Like Bacteriophages RB3, RB5, RB6, RB7, RB9, RB10, RB27, RB33, RB55, RB59, and RB68. Genome Announc **3**:1–2.

48. **Mulvihill DM**, **Ennis MP**. 2003. Functional milk proteins: Production and Utilization, p. 1175–1228. *In* Fox, PF, McSweeney, PLH (eds.), Advanced Dairy Chemistry Volume 1: Proteins, 3rd ed. Springer US, Boston, MA.

49. **Kot W**, **Vogensen FK**, **Sørensen SJ**, **Hansen LH**. 2014. DPS - A rapid method for genome sequencing of DNA-containing bacteriophages directly from a single plaque. J Virol Methods **196**:152–156.

50. **Chopin A**, **Bolotin A**, **Sorokin A**, **Ehrlich SD**, **Chopin M**. 2001. Analysis of six prophages in *Lactococcus lactis* IL1403: Different genetic structure of temperate and virulent phage populations. Nucleic Acids Res **29**:644–651.

51. **Mahony J**, **Martel B**, **Tremblay DM**, **Neve H**, **Heller KJ**, **Moineau S**, **Van Sinderen D**. 2013. Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. Appl Environ Microbiol **79**:4401–4409.

52. **Atamer Z**, **Dietrich J**, **Müller-Merbach M**, **Neve H**, **Heller KJ**, **Hinrichs J**. 2009. Screening for and characterization of *Lactococcus lactis* bacteriophages with high thermal

631 resistance. Int Dairy J **19**:228–235.

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

635 54. **Casey CN**, **Morgan E**, **Daly C**, **Fitzgerald GF**. 1993. Characterization and classification of
636 virulent lactococcal bacteriophages isolated from a Cheddar cheese plant. J Appl Bacteriol
637 **74**:268–275.

638 55. **Mahony J**, **Murphy J**, **van Sinderen D**. 2012. Lactococcal 936-type phages and dairy
639 fermentation problems: From detection to evolution and prevention. Front Microbiol **3**:335.

640 56. **Miklic A**, **Rogelj I**. 2003. Characterization of lactococcal bacteriophages isolated from
641 slovenian dairies. Int J Food Sci Technol **38**:305–311.

642 57. **Moineau S**, **Borkaev M**, **Holler BJ**, **Walker S a**, **Kondo JK**, **Vedamuthu ER**,
643 **Vandenbergh PA**. 1996. Isolation and characterization of lactococcal bacteriophages from
644 cultured buttermilk plants in the United States. J Dairy Sci **79**:2104–2111.

645 58. **Murphy J**, **Royer B**, **Mahony J**, **Hoyles L**, **Heller K**, **Neve H**, **Bonestroo M**, **Nauta A**, **van**
646 **Sinderen D**. 2013. Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type
647 cheese-producing plants. J Dairy Sci **96**:4945–57.

648 59. **Szczepańska AK**, **Hejnowicz MS**, **Kołakowski P**, **Bardowski J**. 2007. Biodiversity of
649 *Lactococcus lactis* bacteriophages in Polish dairy environment. Acta Biochim Pol **54**:151–
650 158.

651 60. **Castro-Nallar E**, **Chen H**, **Gladman S**, **Moore SC**, **Seemann T**, **Powell IB**, **Hillier A**,

652 **Crandall K a.**, **Chandr PS**. 2012. Population genomics and phylogeography of an Australian
653 dairy factory derived lytic bacteriophage. Genome Biol Evol **4**:382–393.

654 61. **Vegge CS**, **Neve H**, **Brøndsted L**, **Heller KJ**, **Vogensen FK**. 2006. Analysis of the collar-
655 whisker structure of temperate lactococcal bacteriophage TP901-1. Appl Environ Microbiol
656 **72**:6815–6818.

657 62. **Moineau S**, **Fortier J**, **Ackermann HW**, **Pandian S**. 1992. Characterization of lactococcal
658 bacteriophages from Quebec cheese plants. Can J Microbiol **38**:875–882.

659 63. **Raiski A**, **Belyasova N**. 2009. Biodiversity of *Lactococcus lactis* bacteriophages in the
660 Republic of Belarus. Int J Food Microbiol **130**:1–5.

661 64. **Marvig CL**, **Aideh B**, **Neve H**, **Heller KJ**, **Knøchel S**, **Vogensen FK**. 2011. Heat tolerance
662 of dairy lactococcal c2 phages. Int Dairy J **21**:556–560.

663 65. **Geller BL**, **Ivey RG**, **Trempy JE**, **Hettinger-Smith B**. 1993. Cloning of a chromosomal
664 gene required for phage infection of *Lactococcus lactis* subsp. *lactis* C2. J Bacteriol
665 **175**:5510–5519.

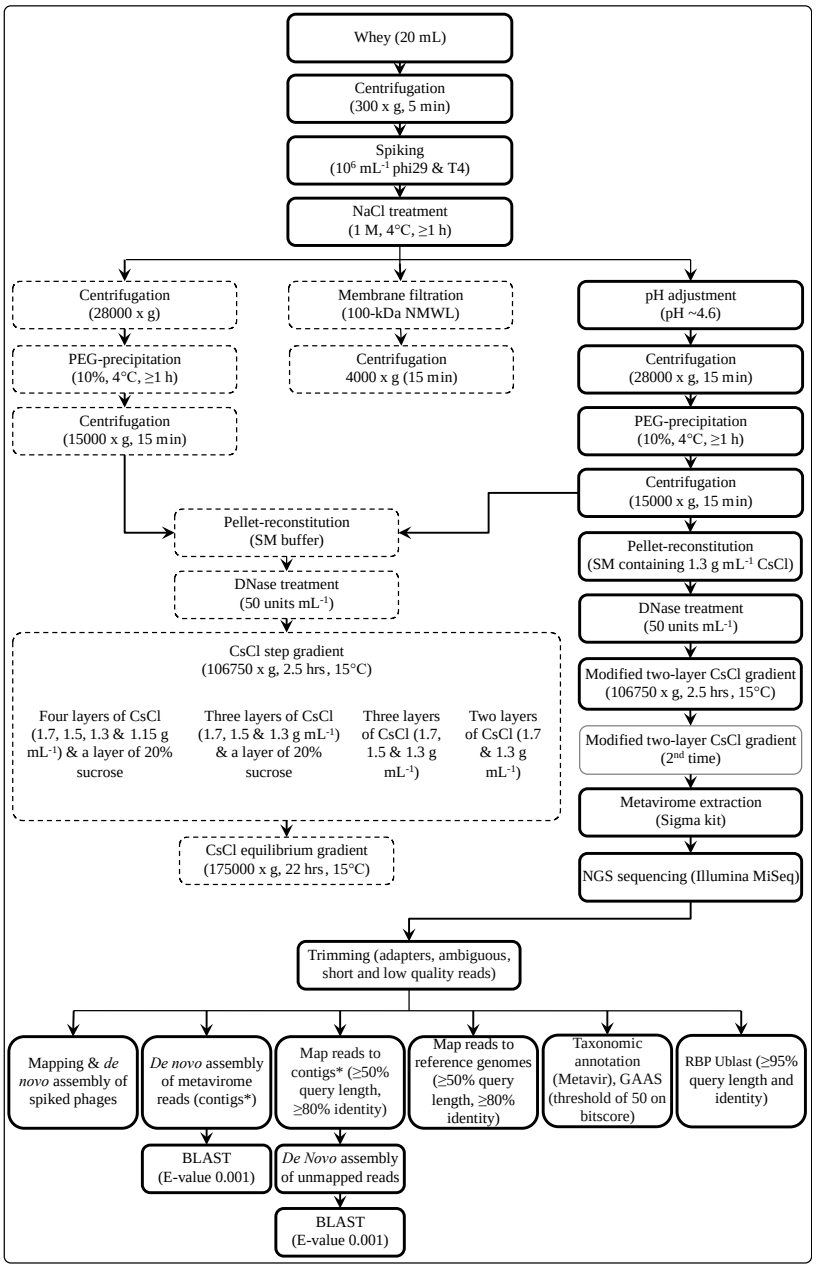
666 66. **Valyasevi R**, **Sandine WE**, **Geller BL**. 1991. A membrane protein is required for
667 bacteriophage c2 infection of *Lactococcus lactis* subsp. *lactis* C2. J Bacteriol **173**:6095–6100.

668 67. **Babu KS**, **Spence WS**, **Monteville MR**, **Geller BL**. 1995. Characterization of a cloned gene
669 (*pip*) from *Lactococcus lactis* required for phage infection. Dev Biol Stand **85**:569–75.

670 68. **Stuer-Lauridsen B**, **Janzen T**. 2006. Bacteriophage resistant lactic acid bacteria. WO
671 2006/072631 A1. World Intellectual Property organization.

69. **Samson JE**, **Moineau S**. 2010. Characterization of *Lactococcus lactis* phage 949 and comparison with other lactococcal phages. Appl Environ Microbiol **76**:6843–52.

689 **Figures**



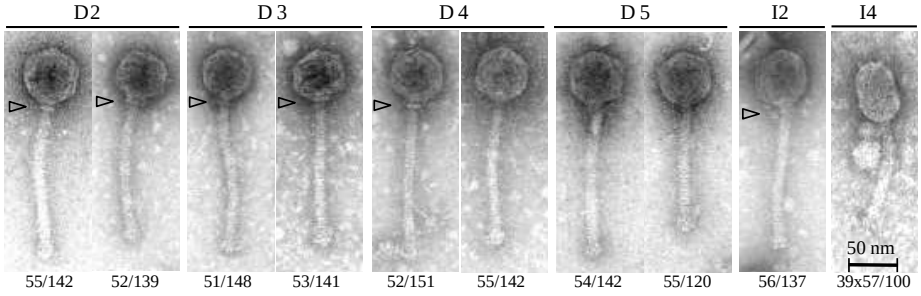
690

691 **Fig. 1**. Dairy phages metagenomic analysis workflow: overview of method optimization and

692 analysis. Different techniques were assessed in terms of ability to separate bulk proteins and to

693 recover spiked *Bacillus subtilis* phi29 and *Escherichia coli* T4 phages from whey samples. Broken-
694 line boxes represent pathways with inadequate protein removal or heavy loss of phi29 and T4.
695 Continuous bold-line boxes represent pathways with removal of reasonable amount of proteins
696 and/or recovery of majority of the spiked phages. A continuous faded-line box represents an optional
697 step that depends on the outcome of the previous step. The modified two-layer CsCl gradient was
698 made of a layer of 1.7 g mL^{-1} CsCl in SM buffer overlaid by a layer of phage suspension (PEG-pellet
699 in 1.3 g mL^{-1} CsCl solution). The method's overall recovery efficiency was assessed following CsCl
700 equilibrium gradient or the last modified CsCl gradient ultracentrifugation. Local databases
701 containing the genomes of dairy *Lc. Lactis* and *Leuconostoc* phages and phylogenetically organized
702 receptor-binding protein (RBP) groups were used to perform the respective mapping and Ublast
703 analyses. NMWL: nominal molecular weight limit, PEG: polyethylene glycol, SM: sodium-
704 magnesium (buffer), NGS: next-generation sequencing, and GAAS: genome relative abundance and
705 average size.

706



707 **Fig. 2**. Representative transmission electron micrographs of phages extracted from whey mixtures.
708 All the micrographs except I4 show 936-like phages (I4 is a c2-like phage). Arrowheads indicate
709 neck passage structures (NPS). Numbers below the micrographs are the head dimensions and tail
710 lengths of the phage particles, respectively. Micrographs are shown at identical magnifications
711 (bar=50 nm).

712 **Tables**

713 **Table 1.** Summary of the sequencing results for spiked whey samples

Sample	Total reads	Reads after trimming	Average reads length		Reads mapped to reference genomes (%)		Average coverage	
			Before trimming	After trimming	φ29	T4	φ29	T4
D3	2,135,308	2,135,098	251.0	152.0	0.4	0.4	78.4	9.08
D4	2,731,990	2,731,728	251.0	175.9	0.2	0.2	74.0	7.53
D5	2,699,702	2,699,300	251.0	209.8	7.9	4.3	2353.14	142.26
G1	3,673,882	3,673,338	251.0	201.2	6.8	2.1	2594.51	78.29
I2	331,230	331,166	251.0	204.3	3.5	0.4	132.54	1.78

714 D: Denmark, G: Germany, I: Ireland

715

716

717

718

719

720 **Table 2.** Pairwise comparison between metaviromes as carried out by reads-to-contigs mapping. The overall similarity in phage population
721 between whey samples was assessed by pairwise mapping of the metavirome sequences, *i.e.*, by mapping trimmed reads (rows) to contigs
722 from *de novo* assembly (columns). The mapping threshold was set to a minimum of 50% query length and 80% identity. Entries are shown
723 in bold to indicate that reads and contigs are cognate, *i.e.*, contigs used as references were assembled from the same metavirome, which
724 yielded an expected high average score of ~97.8±1.9%. All the other entries represent the similarity between unrelated (non-cognate)
725 metaviromes. The key next to the table highlights the grouping of the metaviromes according to overall similarity.

Reads from metavirome sequencing \ Contigs from *de novo* assembly of reads	D2	D3	D4	D5	G1	G2	G3	I2	I3	I4	I5
D2	**98.9**	90.9	90.6	97.0	90.2	91.9	91.8	86.0	77.0	65.1	64.0
D3	97.7	**99.1**	98.9	96.6	90.7	92.7	93.6	86.5	77.9	62.4	67.6
D4	98.5	98.9	**99.5**	97.1	90.6	93.5	94.4	87.0	77.5	65.1	67.2
D5	90.7	90.9	90.6	**97.6**	89.7	86.6	80.0	84.5	66.4	55.3	56.8
G1	85.0	83.1	83.3	83.2	**98.7**	97.2	92.6	78.6	69.8	59.5	60.5
G2	85.0	83.1	83.3	83.2	92.1	**97.2**	92.6	78.6	69.8	59.5	60.5
G3	91.1	88.3	88.1	91.3	94.4	96.1	**99.4**	83.4	75.5	61.9	65.2
I2	96.3	92.0	92.1	94.6	93.1	94.0	92.7	**99.1**	73.2	64.5	63.8
I3	5.9	15.1	8.1	24.3	14.0	14.8	17.7	2.5	**96.5**	85.5	91.1
I4	83.6	1.9	1.1	80.0	10.5	21.0	60.3	19.6	98.6	**98.5**	97.7
I5	50.5	4.5	2.5	53.0	15.1	21.5	41.2	11.1	89.2	82.1	**92.5**

Key

>80%
65-80%
50-65%
<50%

726

727 D: Denmark, G: Germany, I: Ireland

Applied and Environmental Microbiology

728 **Table 3.** Relative abundance of *Lc. lactis* and *Leuconostoc* phages in whey mixture metavirome sequences. Entries (scores) represent the
729 percentage of reads mapped to reference genomes and were obtained from statistics generated from mapping of reads to the genome
730 sequences of *Lc. lactis* and *Leuconostoc* phages. The mapping threshold was set to a minimum of 50% query length and 80% identity.

Reference genome	Mapped reads (%)										
	D2	D3	D4	D5	G1	G2	G3	I2	I3	I4	I5
936	90.4	91.4	91.5	90.0	85.0	92.2	93.0	95.1	2.1	0.7	0.6
P335	5.9	4.1	4.5	5.1	6.8	3.9	3.7	2.3	36.1	0.9	1.9
c2	2.2	0.9	0.8	1.2	1.0	0.8	0.4	0.2	47.6	97.8	96.3
1358	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P034	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
1706	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.3	0.0	0.1
Q54	0.3	1.4	0.9	0.7	0.6	0.7	0.3	0.9	0.2	0.0	0.0
KSY1	0.0	0.0	0.0	0.0	0.1	0.6	0.0	0.0	0.6	0.1	0.1
P087	0.1	0.6	0.6	0.1	0.0	0.0	0.2	0.0	0.2	0.0	0.1
949	1.0	1.5	1.4	1.7	1.7	1.6	1.2	1.4	0.6	0.1	0.1
Leuconostoc	0.1	0.1	0.1	1.0	4.1	0.1	1.3	0.0	0.5	0.1	0.1
Satellites	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.6	0.3	0.7

731

732 D: Denmark, G: Germany, I: Ireland

733

734

735 **Table 4.** Relative abundance of sequences of different phage strains in whey mixture metavirome sequences. Relative abundance was
736 considered as the fraction of reads affiliated to NCBI's RefSeq database (Threshold of 50 on the BLAST score) accessed via the Metavir
737 analysis pipeline. The first four hit categories correspond to *Lc. lactis* 936, P335, c2 and satellite phages, respectively. All hits with an
738 abundance level of <0.25% in all metaviromes were excluded for brevity.

BLAST hit	Sequence similarity (%)										
	D2	D3	D4	D5	G1	G2	G3	I2	I3	I4	I5
Lactococcus phage P008	15.4	24.2	23.0	15.4	11.3	21.8	16.2	8.4	0.0	0.0	0.0
Lactococcus phage jm2	15.6	11.0	11.3	14.0	9.8	14.3	18.3	46.6	0.1	0.0	0.0
Lactococcus phage P680	10.7	8.5	9.2	7.3	8.5	7.2	8.1	3.3	0.0	0.0	0.0
Lactococcus phage jm3	11.5	9.5	9.3	8.9	18.1	9.1	7.5	4.0	0.0	0.0	0.0
Lactococcus phage bIL170	8.4	8.3	9.6	7.0	7.6	9.2	9.2	6.9	0.0	0.0	0.0
Lactococcus phage 340	7.7	9.0	8.3	5.8	4.6	8.2	5.8	2.4	0.0	0.0	0.0
Lactococcus phage φ7	8.3	7.6	8.4	7.3	7.4	9.9	10.1	2.3	0.1	0.2	0.1
Lactococcus phage jj50	5.9	7.6	6.4	5.9	4.6	7.1	8.6	4.5	0.0	0.0	0.0
Lactococcus phage bIBB29	4.9	5.7	6.3	3.4	2.1	3.3	5.2	0.8	0.0	0.0	0.0
Lactococcus phage 712	3.6	3.8	3.6	2.6	2.4	3.8	3.2	1.8	0.1	0.0	0.0
Lactococcus phage sk1	2.4	2.3	2.2	2.4	1.5	3.5	2.2	0.0	0.1	0.0	0.0
Lactococcus phage Q54	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
Lactococcus phage SL4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.6	0.0	0.0	0.0
Lactococcus Phage ASCC191	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9	0.0	0.0	0.0
Lactococcus phage sk1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.0
Lactococcus phage TP901-1	0.3	0.2	0.3	0.2	1.2	0.4	0.1	0.3	2.0	0.0	0.1
Lactococcus phage BK5-T	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	1.7	0.0	0.2
Lactococcus phage Tuc2009	0.3	0.1	0.1	0.3	0.1	0.4	0.0	0.1	1.6	0.0	0.0
Lactococcus phage φLC3	0.4	0.2	0.4	0.4	1.1	0.3	0.1	0.1	0.8	0.0	0.0
Lactococcus phage bIL286	0.3	0.1	0.1	0.3	0.3	0.0	0.1	0.1	1.3	0.2	0.2
Lactococcus phage r1t	0.5	0.1	0.1	0.4	0.0	0.1	1.2	0.1	0.1	0.0	0.0
Lactococcus phage ul36	0.1	0.1	0.1	0.3	0.2	0.4	1.0	0.0	1.5	0.0	0.1
Lactococcus phage BM13	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.3	0.0	0.0
Lactococcus phage P335	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.1
Lactococcus phage bIL285	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.6	0.0	0.0
Lactococcus phage bIL309	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0
Lactococcus phage c2	0.6	0.1	0.1	0.3	0.0	0.0	0.1	0.0	27.6	66.9	52.0

Lactococcus phage bIL67	1.6	0.0	0.0	0.5	0.5	0.0	0.0	0.0	21.2	31.6	44.9
Lactococcus phage bIL310	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.7	0.3	0.6
Lactococcus phage bIL311	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.2
Lactococcus phage bIL312	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	0.1	0.2
Leuconostoc phage φLN25	0.0	0.0	0.0	0.1	0.0	0.0	0.5	0.0	0.0	0.0	0.0
Leuconostoc phage P793	0.0	0.0	0.0	0.2	1.2	0.0	0.4	0.0	0.0	0.0	0.0
Leuconostoc phage φLN04	0.0	0.0	0.0	0.2	1.4	0.0	0.5	0.0	0.0	0.0	0.0
Leuconostoc phage Lmd1	0.0	0.0	0.0	0.4	0.8	0.0	0.3	0.0	0.0	0.0	0.0
Leuconostoc phage φLN03	0.0	0.0	0.0	0.1	0.7	0.0	0.2	0.0	0.0	0.0	0.0
Leuconostoc phage φLN6B	0.0	0.0	0.0	0.2	0.3	0.0	0.1	0.0	0.0	0.0	0.0
Leuconostoc phage φLN12	0.0	0.0	0.0	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0
Lactobacillus phage Ld25A	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Lactobacillus phage Lc-Nu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.1
Lactobacillus phage φAQ113	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
Bacillus phage φ29	0.0	0.7	0.5	14.0	12.7	0.0	0.0	5.8	0.0	0.0	0.0
Shigella phage SfIV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.9	0.0	0.2
Enterobacteria phage fiAA91-ss	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	0.0	0.0
Others	0.3	0.1	0.1	1.2	0.5	0.4	0.2	0.7	6.1	0.1	0.3

739 D: Denmark, G: Germany, I: Ireland