Title	Mining of food metagenomes reveals an unexplored diversity of dsDNA bacteriophages
Authors	Walsh, Liam H.;Soni, Viyal;Ancla, Joseph;Somerville, Vincent;Segata, Nicola;Joyce, Susan;Sinderen, Douwe van;Mahony, Jennifer;Shkoporov, Andrey N.;Kenny, John G.;Cotter, Paul D.;O'Sullivan, Orla
Publication date	2026-03-27
Original Citation	Walsh, L H, Soni, V, Ancla, J, Somerville, V, Segata, N, Joyce, S, Sinderen, D V, Mahony, J, Shkoporov, A N, Kenny, J G, Cotter, P D & O'Sullivan, O 2026, 'Mining of food metagenomes reveals an unexplored diversity of dsDNA bacteriophages', npj Biofilms and Microbiomes, vol. 12, no. 1, 104, pp. 1-14. https://doi.org/10.1038/s41522-026-00941-9
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.scopus.com/pages/publications/105040741512 - 10.1038/s41522-026-00941-9
Rights	
Download date	2026-09-12 05:31:51

npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-026-00941-9

Mining of food metagenomes reveals an unexplored diversity of dsDNA bacteriophages

Liam H. Walsh[1,2], Viyal Soni[1,3], Joseph Ancla[1,2], Vincent Somerville[4,5], Nicola Segata[6,7,8], Susan Joyce[3], Douwe van Sinderen[2,9], Jennifer Mahony[2,9], Andrey N. Shkoporov[2,9,10], John G. Kenny[1,9,11], Paul D. Cotter[1,9,11] ✉ & Orla O'Sullivan[1,2,3,9,11] ✉

Bacteriophages are key drivers of microbial ecology, co-existing and co-evolving with bacteria across diverse environments. Limitations in culturing, alongside advances in sequencing and bioinformatics, have driven the use of metagenomics to explore viral diversity. Viral-specific analysis of >3000 food metagenomes from cFMD produced the FVGC, comprising ~3400 metagenome-assembled viruses, most of which belong to novel Caudoviricetes lineages (n = 91), with only ~15% represented in IMG/VR v4. Together, these findings reveal extensive uncharacterized viral diversity in food systems. Beyond serving as a reference, the FVGC facilitates detailed investigation of virus–host interactions. Viral sequences were pervasive across microbial genomes, with several bacterial families exhibiting near-universal associations with viral elements. Bacterial antiviral defence systems were abundant and taxonomically diverse, dominated by restriction–modification systems, while CRISPR–Cas systems showed pronounced lineage-specific distributions; in contrast, viral anti-defence genes were detected at low frequency (<10% of MAVs). Host prediction linked MAVs to clinically relevant taxa, including expanded ESKAPE pathogens such as *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Staphylococcus aureus*, and *Enterobacter* spp., highlighting the ecological connectivity between food-associated viruses and clinically important bacteria. Antimicrobial resistance signals were scarce, suggesting minimal phage-mediated AMR dissemination in food environments. This new publicly available viral database represents a valuable resource for further exploration of viral diversity.

Bacteriophages (phages; bacterial viruses) are ubiquitous across diverse ecosystems. In natural environments, they engage in long-term coexistence with their bacterial hosts, shaping their population dynamics and affecting evolutionary trajectories[1–3]. Many industrial and artisanal food fermentations produce relatively high densities of bacterial biomass, which can, in turn, favour phage proliferation[4]. Phages may impact fermented foods in a variety of ways. Phages are typically viewed as detrimental in fermentations due to their ability to reduce the fermentative capacity of key organisms like lactic acid bacteria (LAB), occasionally leading to fermentation failure[4].

The exact economic impact of phage activity in fermentation industries is difficult to quantify. Although precise estimates remain uncertain, even small disruptions can have large financial consequences given the scale of the global fermented foods market, valued at ~USD 608.6 billion in 2023 and projected to reach nearly USD 900 billion by 2030 (maximizemarketresearch.com)[5]. This highlights the importance of further research to mitigate potential phage-related losses[6,7].

Previously, comprehensive sequence-based analysis of DNA phage communities in microbial habitats has been hindered by several technical and methodological challenges, including the absence of universal

[1]Teagasc Food Research Centre, Moorepark, Fermoy, Ireland. [2]School of Microbiology, University College Cork, Cork, Ireland. [3]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland. [4]Department of Biochemistry, Microbiology and Bioinformatics, Université Laval, Quebec, Canada. [5]Department of Health Sciences and Technology, ETH Zürich, Zürich, Switzerland. [6]Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy. [7]IEO, Istituto Europeo di Oncologia IRCSS, Milan, Italy. [8]Department of Twins Research and Genetic Epidemiology, King's College London, London, UK. [9]APC Microbiome Ireland SFI Research Centre, University College Cork, Cork, Ireland. [10]Department of Medicine, University College Cork, Cork, Ireland. [11]VistaMilk SFI Research Centre, Teagasc, Moorepark, Fermoy, Ireland. ✉e-mail: Paul.Cotter@teagasc.ie; Orla.OSullivan@teagasc.ie

phage markers, difficulties in genome assembly due to high sequence similarity among related phages, and limited annotation resources[8–10]. However, advances in culture-independent high-throughput sequencing, coupled with growing fermented food metagenomic datasets and improved computational tools, now enable systematic characterization of viral communities in these environment[4,10]. Despite these advances, fermented food viruses remain substantially underrepresented in existing reference databases.

The Integrated Microbial Genomes and Viromes (IMG/VR) database[11] provides one of the most comprehensive surveys of environmental viral diversity, including computationally derived host predictions for uncultivated viral genomes (UViGs). However, to our knowledge, only a very small fraction of its catalogue is associated with food or fermentation systems: of the ~5.5 million high-confidence isolates and UViGs in IMG/VR version 4, just 7157 (<1%) correspond to food-related environments[11].This limited coverage constrains our ability to contextualize phage diversity and ecological roles within food microbiomes. Building catalogues containing high-quality phage genome sequences is a crucial step towards characterization of food-based phages, and identification of robust references for next-generation sequencing (NGS) virome studies[12].

Metagenomic sequencing can recover near-complete viral genomes, particularly for double-stranded DNA (dsDNA) viruses, whose genomes are relatively stable, often abundant, and amenable to standard DNA extraction and sequencing workflows[8]. However, viral populations are frequently underrepresented in shotgun metagenomes due to multiple biological and technical factors. Standard 0.22 μm filtration, commonly used for enrichment, can exclude larger viral particles while allowing passage of small microbial cells and extracellular DNA[13]. Additionally, microbial genomes are typically orders of magnitude more abundant than viral genomes in terms of total DNA content, and DNA extraction protocols are often optimized for cellular rather than viral DNA. Single-stranded DNA (ssDNA) and RNA viruses also present additional challenges, as they suffer from inefficient conversion, amplification, or require specialized library preparation methods[14].

In contrast, viromics (virus-enriched metagenomics) physically enriches for viral particles by filtering out cells and treating filtrates with nucleases to digest extracellular host DNA/RNA. This approach enables deeper sequencing of virion-derived genomes and substantially improves detection of viral diversity, genome assembly quality, and identification of rare viral taxa compared to total metagenomes[15]. Nevertheless, total metagenomes offer the advantage of simultaneously capturing microbial and viral components[16], facilitating integrated analyses of host-virus interactions, CRISPR spacer-protospacer linkages, and prophage-host associations within the same dataset. Building on this strength, we introduce the Fermented Food DNA Viral Genome Collection (FVGC), comprising over 3400 putative Metagenome-Assembled Viruses (MAVs)[17], recovered from total metagenomes within the curated FoodMetagenome Database (cFMD). This collection represents the most extensive genome-level investigation of dsDNA phages in food systems to date, encompassing viral diversity, host associations, bacterial defence mechanisms, and viral anti-defence genes.

Results

Genomic catalogue of de novo assembled DNA viruses from the fermented food microbiome

To assess the presence and diversity of dsDNA-encompassing viral populations present in various foods, a total of 3684 food-derived metagenomes were analysed (Fig. 1 and supplemental Fig. S1). Of these, 572 were newly selected research datasets designated for inclusion in future releases of the cFMD, according to criteria as outlined previously[18], while the remainder were sourced from the existing cFMD resource[18] (https://github.com/SegataLab/cFMD). Samples were classified into a range of food categories, with dairy ($n = 2341$) and fermented beverages ($n = 699$) representing the most abundant groups. Smaller numbers of samples were associated with fermented meat ($n = 148$), fermented fruits and vegetables ($n = 53$), fermented grains ($n = 86$), fermented legumes ($n = 64$), fermented seeds ($n = 56$), fermented tubers and roots ($n = 24$), as well as non-fermented foods such as fresh fish ($n = 14$), fruits and vegetables ($n = 60$), meat ($n = 47$), and probiotics ($n = 14$) (Fig. S1).

To identify and characterize viral sequences within the analysed metagenomes, we employed two complementary strategies: a de novo assembly-based approach and a metagenome-assembled genomes (MAGs)-associated viral detection strategy (Fig. 1). The de novo assembly method reconstructed viral genomes directly from metagenomic reads, primarily capturing free viral particles and yielding 147 sequences classified as prophages. To improve the recovery of integrated viruses, we implemented a targeted pipeline to detect viral sequences embedded within MAGs. Using this targeted approach, CheckV classified 2654 of the putative viral sequences as prophages, expanding the catalogue beyond what was recovered through assembly alone. In addition, we analysed unbinned contigs, which were not assigned to any MAGs during the Metabat binning process. These sequences often include short or low-abundance fragments and mobile genetic elements, making them a valuable reservoir of viral diversity. All viral sequences recovered underwent additional screening using CheckV viral and host gene counts alongside VirSorter2 scores (see Methods).

The average number of viral sequences recovered per metagenome varied substantially between food types. Depth-normalized MAV recovery rates ranged from ~0.53 to 2.31 MAVs per million sequencing reads across food categories (Fig. 2A) (see Methods). Fermented fish samples exhibited the highest depth-normalized MAV recovery (2.31 MAVs per million reads; $n = 7$), followed by fermented beverages (1.51; $n = 1365$) and fermented tubers and roots (1.53; $n = 41$). Probiotic products (1.02), dairy-associated samples ($\approx 1.0–1.1$; $n = 2944$), and meat products (0.53) showed intermediate recovery rates, whereas non-fermented fish and alcohol-containing samples displayed lower or comparable depth-normalized MAV recovery ($\approx 1.0–1.5$ MAVs per million reads). Notably, categories with high raw MAV counts did not always retain elevated recovery rates after depth normalization, indicating that sequencing depth partially inflated apparent viral recovery in some food types. In contrast, fermented fish retained the highest depth-normalized MAV recovery despite limited sample size, suggesting enhanced viral assembly recovery in these environments. Nevertheless, the small number of fermented fish and fish product samples limits confident inference regarding viral diversity or generalizability within these categories (Fig. 2A).

All viral sequences recovered were quality assessed using CheckV (Figs. 1 and S2), which estimates genome completeness and host contamination based on a curated viral genome database. Viral sequences were classified into five mutually exclusive quality tiers: complete, high-quality, medium-quality, low-quality, and not-determined. Of these, 841 were categorized as complete, 816 as high-quality, and 1788 as medium-quality. These three groups were retained as MAVs[17], collectively forming the Food DNA Viral Genome Collection (FVGC), a curated catalogue of ~3,400 high-confidence viral genomes (Fig. 1).

These quality tiers reflect key differences in genome completeness and contig characteristics. Complete MAVs represent fully assembled genomes with near-100% completeness, typically confirmed by circularization or identification of genome termini. High-quality genomes exceed 90% completeness but may lack small terminal regions, while medium-quality genomes are estimated at 50–90% completeness. Reflecting these designations, complete contigs were the longest (2.3–362 kb, median: 40.0 kb) and encoded the highest number of viral genes (up to 249, median: 30). High-quality and medium-quality contigs showed comparable median lengths (28.1 kb and 25.9 kb, respectively), with maxima of 701 kb and 356 kb, encoding up to 189 and 85 viral genes (medians: 16 and 18) (Fig. S2A, B). Furthermore, host gene contamination was on average minimal, with a median of zero host genes detected across those categories (Fig. S2C). In contrast, the majority of sequences (14,614) were classified as low-quality, and 2743 as not-determined. These contigs were shorter (median lengths: 8.8 kb and 6.1 kb) and encoded fewer viral genes (median: 4 and 0), with the not-determined category typically lacking sufficient similarity to known

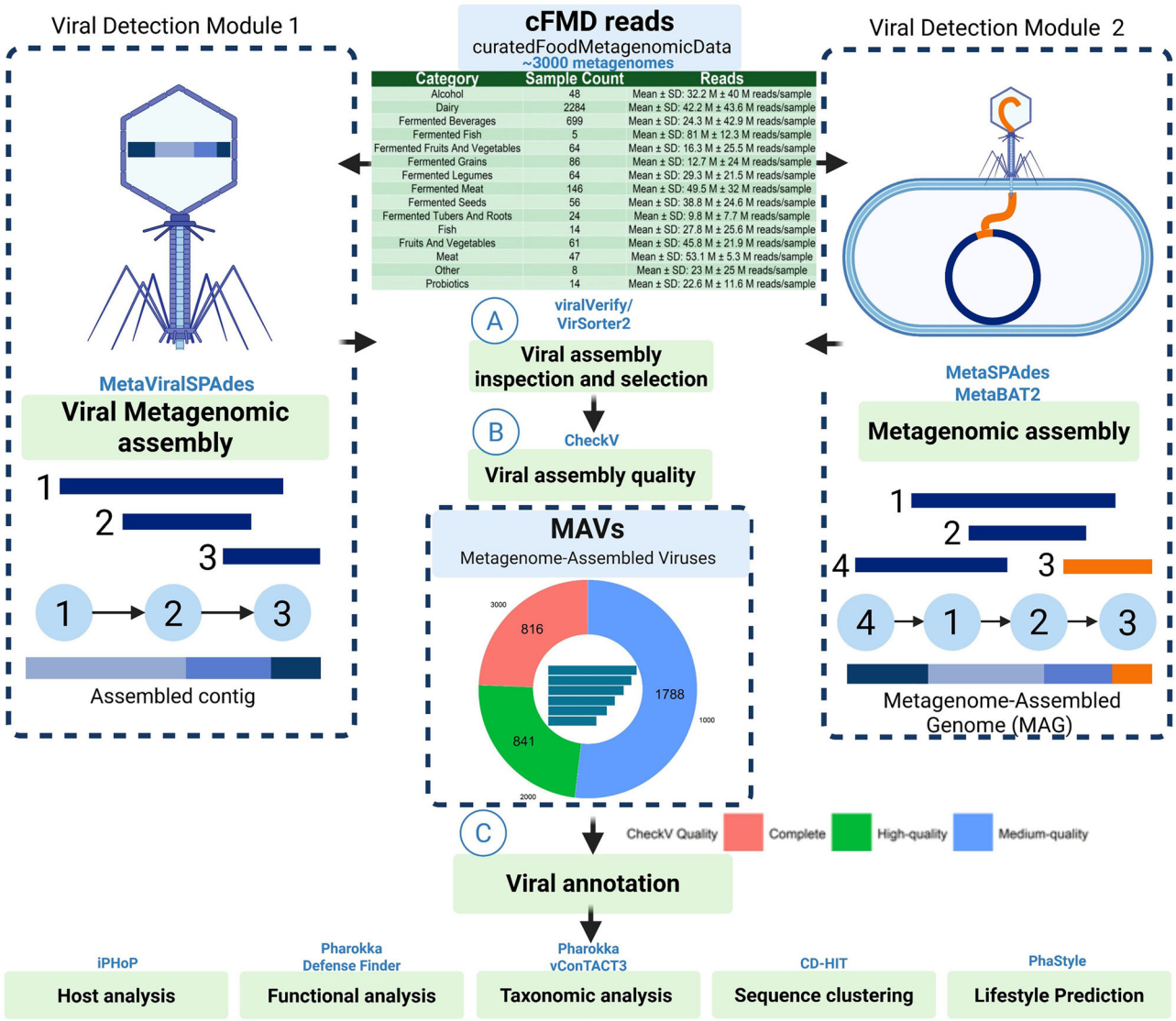


Fig. 1 | Schematic overview of the viral discovery process and creation of the fermented food DNA viral genome collection (FVGC). This figure outlines the integrated workflow for analyzing metagenomic data from fermented food samples. High-quality reads were processed through two complementary viral detection pipelines: a de novo assembly pipeline using SPAdes with metaviral parameters (Viral Detection Module 1), and a MAG-associated viral detection pipeline (Viral Detection Module 2), targeting integrated viral sequences within microbial genomes. **A** The resulting assemblies from both pipelines were evaluated for viral signatures using tools such as ViralVerify and VirSorter2. **B** Quality assessment of the viral assemblies was performed using CheckV, categorizing metagenome-assembled viruses (MAVs) based on completeness and contamination metrics. **C** MAVs meeting or exceeding medium-quality thresholds were annotated using iPHoP, Pharokka, and DefenseFinder, providing insights into their taxonomy, functional potential, and anti-defense systems. MAVs were also clustered with reference genomes from RefSeq phages using vConTACT3.

references or lacking hallmark genes, preventing reliable completeness estimation[19]. Across all quality tiers, a total of 2840 MAVs were identified as prophages by CheckV, with 108 classified as complete, 138 as high-quality, 375 as medium-quality, 2195 as low-quality, and 24 as not-determined (Fig. 2B).

Within the MAV dataset, the vast majority of MAVs annotated using VirSorter2 were predicted to be dsDNA viruses. The average GC content was 42% (SD: 9%), and genomes encoded a mean of 42 predicted CDSs (SD: 39.7). Coding density was high across genomes, averaging 90.2% (SD: 6.8%), consistent with typical viral genome compactness[20] (Fig. 2C).

Viral lifecycle, taxonomy, and host associations vary significantly across fermented food categories

To characterize viral communities across fermented food types, we examined lifestyle strategies, taxonomic composition, and host associations (Fig. 3). We employed PhaStyle to classify MAVs as virulent (lytic) or temperate (lysogenic)[21], identifying 2647 temperate and 798 virulent phages with strong concordance to CheckV provirus predictions (94.0% of proviruses were classified as temperate; 95.1% of virulent phages were non-integrated). Lifestyle distribution varied significantly across food categories (Fisher's exact test, $p < 0.001$; Cramér's $V = 0.269$, indicating medium effect size). While temperate phages predominated overall, their prevalence varied nearly two-fold across categories, ranging from 52.5% in fermented seeds to 92.2% in fermented beverages. Fish products diverged sharply from this pattern with only 27.8% temperate phages (95% CI [9.7%, 53.5%]).

To assess the novelty of MAVs, we performed global alignment using vsearch against 7157 food-associated isolates and UViGs curated from IMG/VR v4. Only 2.1% of MAVs ($n = 72/3445$) returned detectable matches (see Methods). When comparing the FVGC against all high-confidence IMG/VR v4 isolates and UViGs at 95% ANI, ~15% of MAVs showed detectable matches, confirming extensive novel viral diversity. We therefore employed gene-sharing network analysis with vConTACT3 to

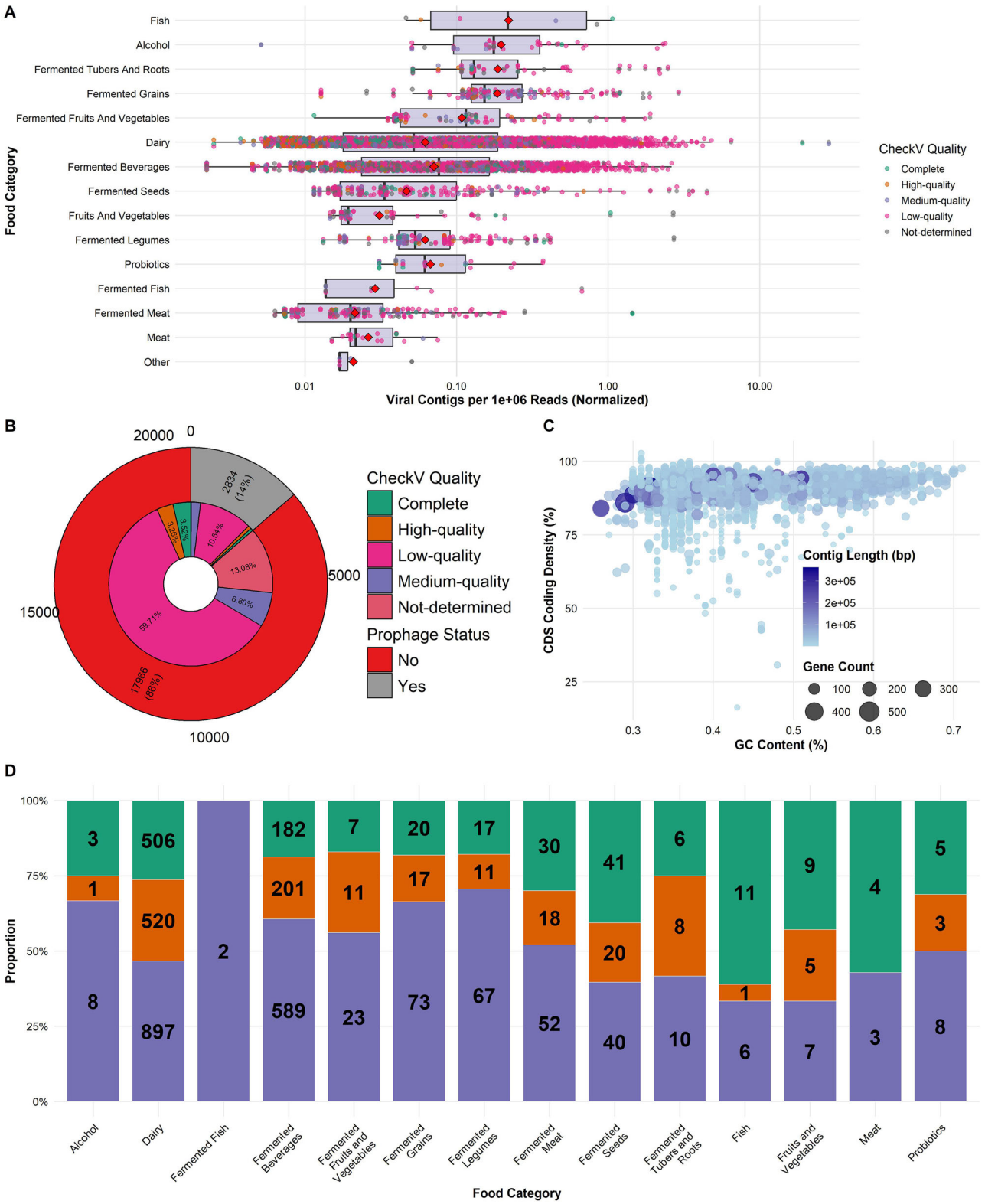


Fig. 2 | Distribution and characteristics of viral sequences within the cFMD. This figure presents an overview of viral sequences identified within the curated Fermented Microbiome Database (cFMD). The panels are described as follows: **A** Boxplots show the distribution of depth-normalized metagenome-assembled viral (MAV) recovery rates across food metagenomes, stratified by food category and CheckV genome quality class (complete, high-quality, medium-quality, low-quality, and not determined). MAV recovery was normalized to sequencing depth and is expressed as MAVs recovered per million sequencing reads. Jittered points represent individual metagenomes, colored by CheckV quality, and red diamonds indicate the mean depth-normalized MAV recovery for each food category. **B** depicts a donut chart showing the distribution of CheckV genome quality categories based on prophage presence (Yes/No). **C** Relationship between GC content (*X* axis) and CDS coding density (*Y* axis) across MAVs. Each point represents a MAV, with point size indicating gene count and colour representing contig length (bp). **D** presents a stacked bar chart showing the proportional distribution of MAVs across various fermented food categories, with colours denoting CheckV quality levels and absolute MAV counts displayed within each bar segment.

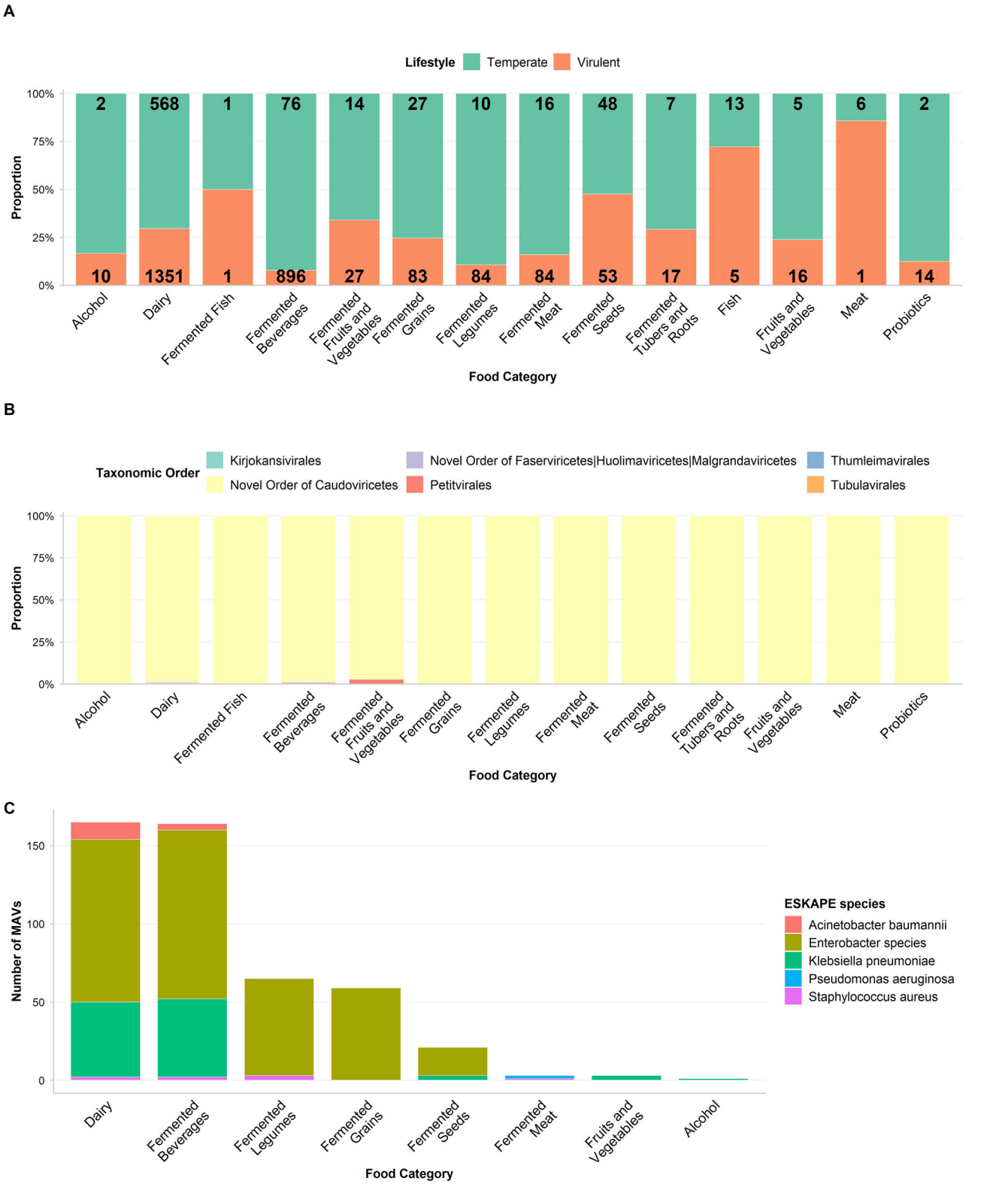


Fig. 3 | Lifestyle profiles, taxonomic composition, and ESKAPE-associated MAV distributions across food categories. A Proportion of temperate and virulent (lytic) MAVs were detected across food categories. Bars represent the relative contribution of each lifestyle class within a category, with counts displayed above or below each segment. **B** Viral order-level composition of MAVs recovered from each food category. The distribution was highly similar across food categories, indicating broadly conserved taxonomic profiles. **C** Total number of MAVs predicted to be associated with ESKAPE pathogens across food categories. Each bar is coloured by the predicted host species.

taxonomically classify these novel viruses. Using vConTACT3-based predictions, we classified 3042 high-quality MAVs into 99 distinct viral orders across 10 food categories with sufficient sample sizes ($n \geq 10$)(Fig. S3). Viral order distribution differed significantly across food categories (Fisher's exact test: $p < 0.001$), though Cramér's V could not be calculated due to data sparsity, reflecting the extensive diversity of novel viral lineages. The dominant viral orders varied substantially by food type, with novel orders of Caudoviricetes being most abundant. Specifically, novel_order_32 predominated across categories but varied from 18.3% in dairy to 28.9% in fermented grains (Fig. 3B).

To characterize prokaryotic host associations, we applied iPHoP v1.3.3, a two-step host prediction framework integrating multiple homology- and composition-based approaches. We constructed a custom host database by supplementing the default iPHoP Aug_2023_pub_rw database with dereplicated MAGs recovered from the same food metagenomes. MAGs were integrated into the iPHoP phylogenetic framework by clustering with existing reference genomes at 95% average nucleotide identity (ANI), with representative genomes selected for each species cluster. This approach enabled host prediction for MAVs infecting fermented food-associated bacteria underrepresented in standard reference databases.

Of the 3445 MAVs analyzed, 2530 (73%) were confidently assigned to bacterial hosts and 25 (0.7%) to archaeal hosts, while 890 (26%) remained unclassified at the domain level. Host predictions were dominated by Bacillota (1783 MAVs; 52%) and Pseudomonadota (633 MAVs; 18%), with Lactobacillales being the most frequent order (1621 MAVs), consistent with the prevalence of LAB in fermented foods (Fig. S4). At the family level, Lactobacillaceae (978 MAVs) and Streptococcaceae (608 MAVs) were most common, with predictions spanning 64 distinct host families. Notably, iPHoP identified novel viral-host associations for 1545 MAGs recovered in this study, all representing putative novel bacterial species. Additional taxonomic details of host associations at all ranks (phylum through species, including novel ranks) are provided in Supplementary Result S2.

Host family distribution differed significantly across food categories (Fisher's exact test: $p < 0.001$, Cramér's $V = 0.330$, indicating medium effect size). The most abundant predicted host families varied markedly by food type: dairy viruses predominantly targeted Streptococcaceae (52.8%), Lactobacillaceae (15.4%), and Pseudomonadaceae (9.3%), while fermented beverages showed distinct patterns with Lactobacillaceae (45.0%), Acetobacteraceae (15.4%), and Enterobacteriaceae (13.4%) as dominant hosts. Alcoholic fermentations exhibited the highest relative abundance of Acetobacteraceae-infecting viruses (37.8%), reflecting the importance of acetic acid bacteria in ethanol oxidation, while fermented fruits and vegetables showed strong dominance of Lactobacillaceae-targeting viruses (86.1%), consistent with lactic acid fermentation pathways.

Finally, we identified MAVs with predicted associations to clinically relevant ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species), which represent priority targets for antimicrobial resistance interventions (Fig. 3C). Of the six ESKAPE pathogen groups, four were detected as predicted hosts: *S. aureus* (8 MAVs), *K. pneumoniae* (105 MAVs), *Ac. baumannii* (15 MAVs), and *P. aeruginosa* (2 MAVs). Notably, the majority of ESKAPE-associated MAVs (472 total, representing 13.8% of the catalog) were linked to Enterobacter species, an increasingly recognized opportunistic pathogen group within the expanded ESKAPE framework. Among 17 Enterobacter species detected, *E. hormaechei* (89 MAVs) and *E. cloacae* (78 MAVs) were most prevalent, both recognized as emerging multidrug-resistant nosocomial pathogens. *E. roggenkampii* (43 MAVs), associated with carbapenem resistance, also showed substantial phage representation. Other notable species included *E. asburiae* (17 MAVs), *E. bugandensis* (18 MAVs), *E. cancerogenus* (13 MAVs), and *E. kobei* (16 MAVs), reflecting taxonomic diversity within this genus.

We evaluated how the distribution of ESKAPE pathogen-associated MAVs varied across fermented food categories. Five categories contained sufficient sample sizes for statistical testing ($n \geq 10$): dairy ($n = 165$), fermented beverages ($n = 164$), fermented legumes ($n = 65$), fermented grains ($n = 59$), and fermented seeds ($n = 21$) (Fig. 3C). A global Fisher's exact test detected significant differences in host-species composition across all categories ($p < 0.001$). The result remained significant when restricting analysis to well-sampled categories ($p < 0.001$). Cramér's V could not be computed due to sparse counts, so interpretation is based on the Fisher test.

Clear food-category-specific structuring was observed. Dairy and fermented beverages showed similar profiles and contained the highest diversity and abundance of ESKAPE-associated MAVs. In dairy, the most represented hosts were *Klebsiella pneumoniae* (48 MAVs, 29.1%), *Enterobacter hormaechei* (32 MAVs, 19.4%), and *Enterobacter roggenkampii* (28 MAVs, 17.0%). Fermented beverages showed a comparable pattern, with *K. pneumoniae* (50 MAVs, 30.5%), *E. hormaechei* (42 MAVs, 25.6%), and *E. cloacae* (23 MAVs, 14.0%) dominating. Fermented grains were strongly enriched for *E. cloacae* (17 MAVs, 28.8%) and contained moderate levels of *E. hormaechei* (8 MAVs, 13.6%) and multiple species present at lower proportions, including *E. cancerogenus*, *E. kobei*, *E. ludwigii*, and *E. roggenkampii* (5 MAVs each). Fermented legumes exhibited a broad distribution in which eleven Enterobacter species each contributed four MAVs (6.2%), resulting in a more even community structure compared to other food matrices.

Defense system prevalence reflects viral pressure across fermented food microbiota

To expand the viral association network within food systems beyond the FVGC, we applied a complementary approach using high-quality and medium-quality MAGs from bacterial families with at least 10 representatives. Through VirSorter2 we screened these MAGs for viral sequences to assess family-specific viral prevalence patterns. This analysis revealed substantial variation across bacterial lineages. Several families exhibited near-ubiquitous viral associations, including Thermaceae (94%, 67/71 MAGs), Bacillaceae (96%, 24/25), Acetobacteraceae (88%, 85/97), Staphylococcaceae (88%, 164/187), and Enterococcaceae (88%, 35/40). Other families also showed high levels of viral associations, such as Alteromonadaceae (88%, 14/16), Bifidobacteriaceae (84%, 16/19), Moraxellaceae (82%, 79/96), and Sphingomonadaceae (83%, 10/12). Moderate prevalence was observed in families like Lactobacillaceae (75%, 668/888), Streptococcaceae (68%, 342/506), and Micrococcaceae (68%, 112/166), while relatively lower detection rates were found in Halomonadaceae (49%, 18/37), Dermabacteraceae (47%, 16/34), and Microbacteriaceae (58%, 22/38). The lowest viral associations were observed in Vibrionaceae, with only 5% (4/76) of MAGs containing viral signatures (Fig. 4A).

To assess the distribution of antiviral defense systems across bacterial lineages in the cFMD dataset, we profiled CRISPR-Cas loci and other defense systems in high- and medium-quality MAGs from bacterial families with at least 10 representatives. In line with the hypothesis that viral predation drives the retention of defense systems[22], most families showed high prevalence. More than half of the families had over 80% of their MAGs encoding at least one defense system. Although genome fragmentation in MAGs likely leads to some underreporting, many families still exhibited high detection rates. For example, Alteromonadaceae, Dermatophilaceae, and Sphingomonadaceae had defense systems in 100% of their MAGs, while Enterococcaceae, Bifidobacteriaceae, and Brevibacteriaceae had detection rates exceeding 90%. Broadly distributed families such as Lactobacillaceae and Staphylococcaceae also showed high prevalence, with over 80% of MAGs carrying at least one system (Fig. 4B). We also observed variation in the number of defense systems per MAG across families. Alteromonadaceae and Propionibacteriaceae had the highest averages, with 8 and 6 systems per MAG, respectively. Notably, both Alteromonadaceae and Propionibacteriaceae also showed high rates of phage association (Fig. 4A). Sphingomonadaceae, Vibrionaceae, Thermaceae, and Brevibacteriaceae followed, averaging between 4 and 5 systems per MAG. Notably, multiple systems were also found in the most numerous families. For example, Staphylococcaceae, Streptococcaceae, and Lactobacillaceae averaged 4, 4, and 3 systems per MAG, respectively.

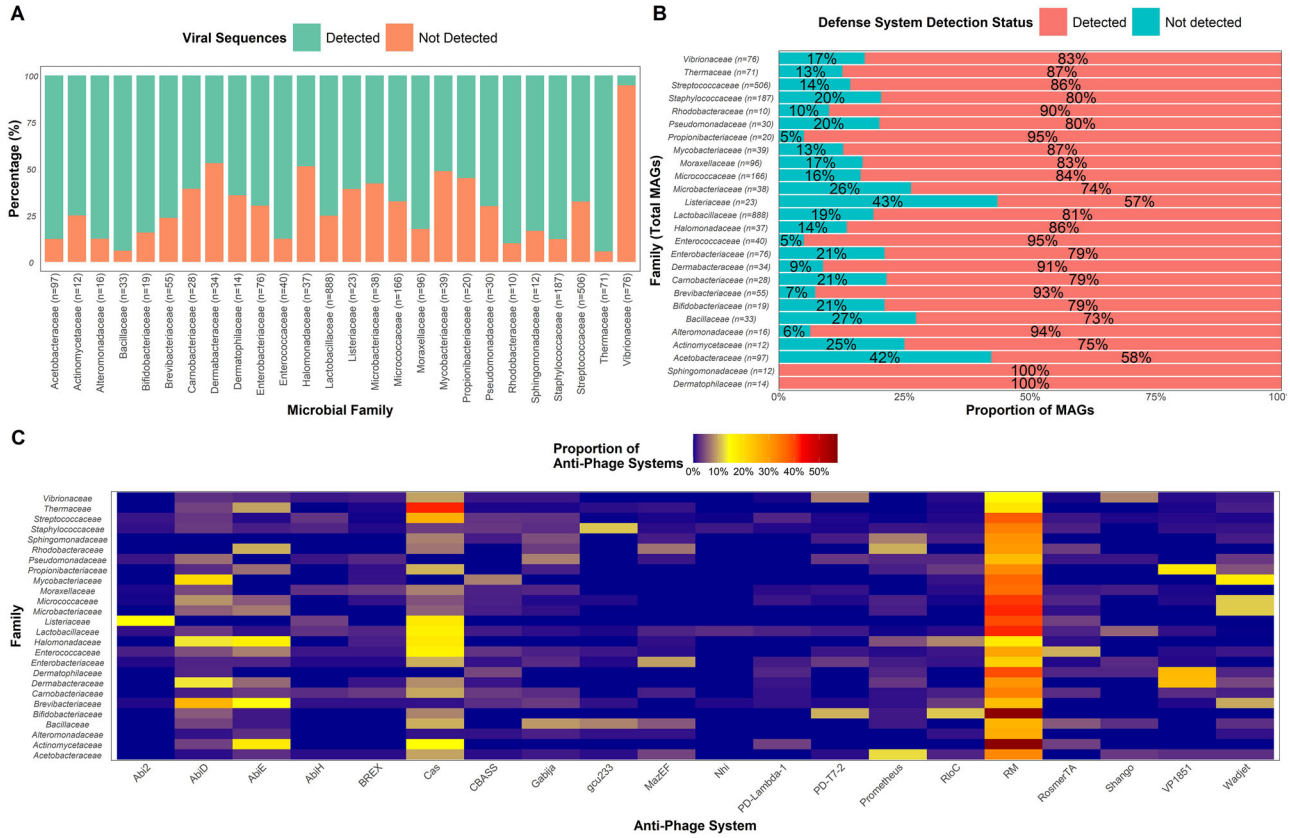


Fig. 4 | Viral associations and antiviral defence system distributions across microbial families in the cFMD. The panels are described as follows: **A** Bar chart displaying the proportion of medium-quality or greater MAGs containing viral sequences across each microbial family (*X* axis). Viral detection status is indicated by colour: Green bars represent MAGs with detected viral sequences, and orange bars indicate MAGs without detected viral sequences. **B** Stacked bar plot showing the proportion of MAGs (*X* axis) within each microbial family (*Y* axis), limited to families with ≥10 representative MAGs. Bar segments indicate antiviral defense system status: red represents MAGs with at least one detected antiviral defense system, while blue represents MAGs with no detected antiviral defense systems. **C** Proportion of antiviral defense systems detected across microbial families (*X* axis) with ≥10 representative MAGs. Only the most numerous defense systems, representing the top 80th %, are shown.

The most frequently detected antiviral system was the Restriction Modification (RM) system, which appeared in 5,097 instances across the dataset. RM systems were detected in all families and were especially common in Lactobacillaceae (1202 of 3002 MAGs; 40.0%), Streptococcaceae (704 of 1,888; 37.3%), Staphylococcaceae (238 of 736; 32.3%), and Micrococcaceae (197 of 510; 38.6%). Some families had particularly high proportions of RM-positive MAGs, such as Bifidobacteriaceae (57%), Actinomycetaceae (52%), and Mycobacteriaceae (37%). CRISPR-Cas systems were the second most commonly detected, with 1999 instances. These were unevenly distributed. Higher frequencies were observed in Streptococcaceae (488 of 1888; 26%), Lactobacillaceae (454 of 3002; 15%), Thermaceae (136 of 333; 41%), and Enterococcaceae (29 of 181; 16.0%). In contrast, some families, such as Dermatophilaceae, Mycobacteriaceae, and Pseudomonadaceae showed no detectable Cas systems. Other systems, abortive infection systems such as AbiD ($n = 573$), AbiE ($n = 296$), and AbiH ($n = 254$) were commonly detected (Fig. 4C). Additional systems with more limited distributions included Shango ($n = 271$), CBASS ($n = 256$), Wadjet ($n = 256$), and BREX ($n = 145$), reflecting the diversity of antiviral strategies in microbial populations found in fermented foods (Fig. 4C).

Functional capacity of the fermented food virome reveals high proportion of unknown genes, minimal virulence and AMR, and diverse anti-defense strategies

Although the functional potential of bacteria and eukaryotes in fermented foods has been well studied[18], much less is known about their associated viruses. To explore this, we used Pharokka to assess the functional potential of MAVs within the FVGC. Functional annotation analysis revealed substantial variation in gene content across MAVs. To account for differences in genome size (range: 1–903 CDS per MAV; median: 37), we normalized annotation counts by expressing them as percentages of total CDS per genome.

Unknown function genes constituted the largest proportion (median: 62.5%, IQR: 51.3%–75.0%),

consistent with the novelty of fermented food viral lineages. Among characterized functions,

Head & Packaging showed the highest relative abundance (median: 9.5%, IQR: 5.2%–14.7%), followed by Nucleotide Metabolism (median: 6.6%) and Tail Proteins (median: 6.1%) (Fig. 5A). Virulence and antimicrobial resistance genes were very uncommon (Fig. 5A). Six contigs, including seq67, seq2722, seq811, seq832, seq941, and seq434, carried high-confidence virulence gene hits. Five of these encoded GtrB, a bactoprenol glucosyl transferase involved in LPS modification, with ~87% identity to *Shigella flexneri* (VFDB ID: VFG000670). One contig, seq434, carried a Type VI secretion protein TssM (80% identity, VFDB ID: VFG048748). Only one contig, seq2176, carried an antimicrobial resistance gene: FosD, a fosfomycin resistance determinant originally described in *Staphylococcus aureus* (90% identity, CARD ARO:3004674).

Given the widespread distribution of defense systems across microbial families, we investigated whether viruses encode potential counter-defense mechanisms. Specifically, we screened the MAV dataset for anti-defense systems using AntiDefenseFinder. A total of 314 MAVs carried at least one predicted anti-defense element (Fig. 5B), while the remaining 91% of the FVGC dataset showed no detectable hits. Across these 314 MAVs, we identified anti-defense elements spanning 11 distinct system types, with

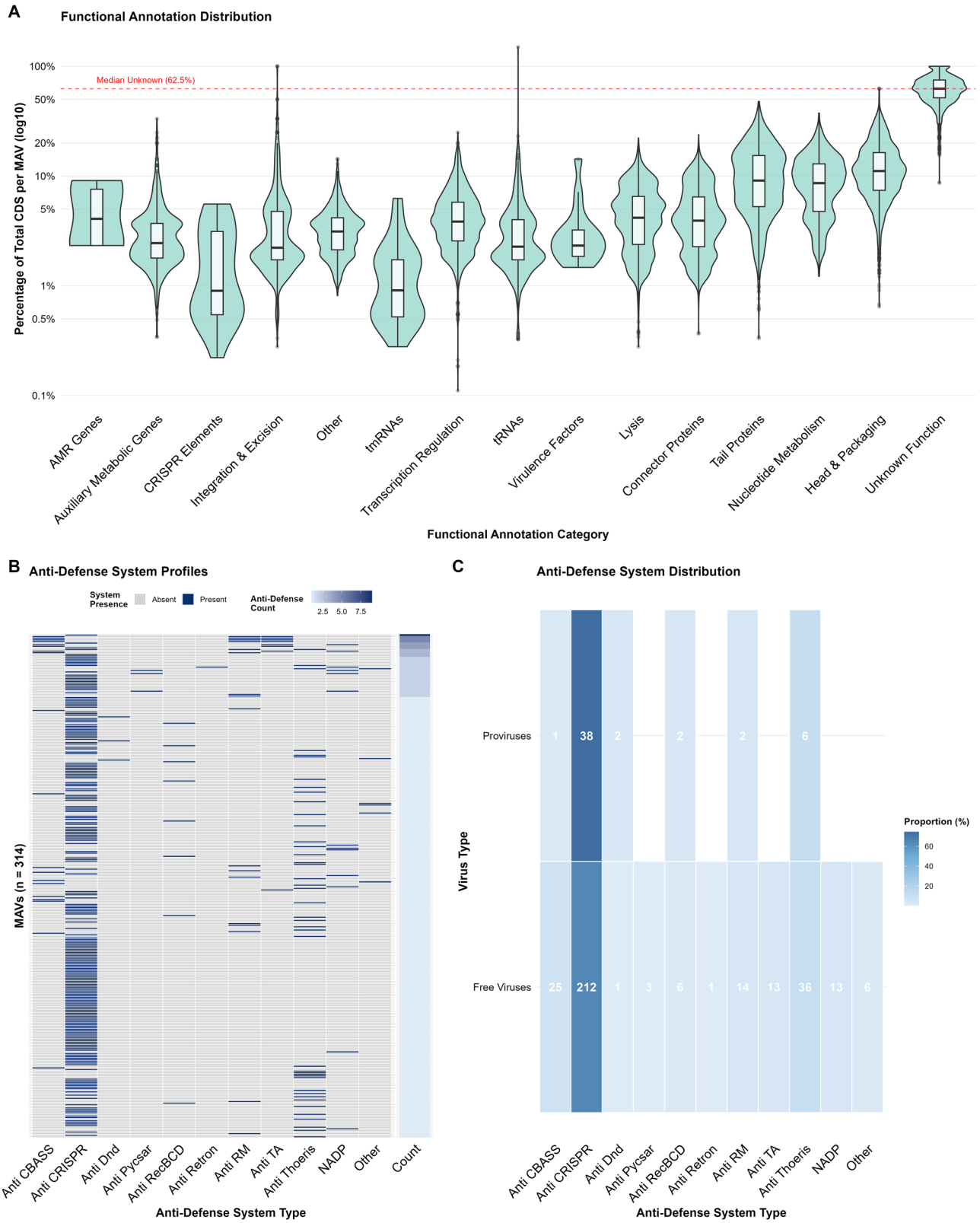


Fig. 5 | Functional capacity and anti-defence system composition in dsDNA food-associated phage. The panels are described as follows: **A** Violin plots show the distribution of CDS counts (*Y* axis) across functional annotation (*X* axis). Each violin is overlaid with a boxplot showing the median and interquartile range; counts are shown on a $\log_{10}$ scale to accommodate variation across categories. Transparent red horizontal lines represent CDS counts for each MAV, providing a reference for coding density. Functional categories are ordered by median annotation count to aid interpretation. **B** Heatmap illustrating the presence (blue) or absence (light grey) of anti-defence systems (*X* axis) across individual MAVs (Y-axis). An "Other" column consolidates anti-defence systems not included in the ten primary types. **C** Heatmap illustrating the proportion (of anti-defence systems (*X* axis) across viral types (*Y* axis). On the right, a gradient sidebar shows the total number of unique anti-defence systems per MAV, with darker shades indicating a higher count. An "Other" column consolidates anti-defence systems not included in the ten primary types.

individual MAVs encoding between 1 and 9 features (Median value of 1) (Fig. 5B). The most prevalent were Anti-CRISPR proteins ($n = 251$), other notable systems included Anti-Thoeris ($n = 42$), Anti-RM ($n = 19$), and Anti-RecBCD ($n = 8$) (Fig. 5B). Less frequent systems included Anti-CBASS ($n = 26$), Anti-Dnd ($n = 3$), Anti-Pycsar ($n = 3$), Anti-TA ($n = 13$), Anti-Retron ($n = 1$), and NADP-associated systems ($n = 13$) (Fig. 5B).

Importantly, the prevalence of anti-defense systems differed significantly between virus types (Fisher's exact test: $p = 0.020$). Free viruses showed higher anti-defense system prevalence (9.6%) compared to proviruses (6.6%), representing a 1.5-fold enrichment in free viruses (OR = 0.67, 95% CI [0.47–0.95]). However, among MAVs encoding anti-defense systems, the number of systems per genome was similar between free viruses and proviruses (median: 1 system for both groups; Wilcoxon test: $p = 0.123$). The distribution of specific anti-defense system types showed no significant enrichment in either group after multiple testing correction (Bonferroni-adjusted $p > 0.05$ for all systems) (Fig. 5C). Anti-defense systems were distributed across MAVs infecting a broad range of bacterial families. The majority were associated with Lactobacillaceae ($n = 104$) and Streptococcaceae ($n = 127$), followed by Enterobacteriaceae ($n = 18$), Clostridiaceae ($n = 5$), Burkholderiaceae_B ($n = 5$), Moraxellaceae ($n = 6$), and several other families including Acetobacteraceae, Halomonadaceae, and Pseudomonadaceae (each with ≤3 associations).

Discussion

In this study, we investigated the genomic diversity of dsDNA viruses of food microbial environments within the cFMD using two complementary viral detection pipelines (Fig. 1). Although the metagenomes within the cFMD were not specifically enriched for viral particles, we recovered a substantial number of viral assemblies classified as complete, high-quality, or medium-quality MAVs ($n = 3445$) spanning a broad spectrum of food categories (Fig. 2A). Comparative analysis against the IMG/VR v4 at 95% ANI revealed that ~15% of MAVs showed detectable matches. Furthermore, taxonomic classification revealed that the vast majority of MAVs belonged to the class Caudoviricetes, dominated by putative novel viral orders (Fig. 3B and S3). These results suggest that food environments may harbour extensive and largely undocumented phage diversity, particularly within the class Caudoviricetes.

The assembled MAVs have been integrated into the latest FVGC release, enhancing our ability to assess viral diversity across metagenomic datasets. The curated FVGC now serves as a resource for virome research and studies of microbial ecology (Fig. 6). To illustrate its utility, we demonstrated how the FVGC supported analyses of dsDNA-based virome features across food substrates (Fig. 3) and virus–host interactions within food-associated microbial ecosystems (Fig. 4).

Predicted bacterial hosts were primarily members of the phylum Bacillota, especially the class Bacilli and the order Lactobacillales (Fig. S4), consistent with the dominance of LAB in fermented food systems[23]. In contrast, archaeal-infecting MAVs were extremely rare (0.7%, $n = 25$) (Fig. S4), a pattern that likely reflects both genuine ecological scarcity and limitations in current host-prediction databases. Archaea are typically minor constituents of food microbiomes, usually <0.1% of community composition based on our estimates, which are in line with previous reported estimates based on 16S rRNA surveys[24].

Database underrepresentation further constrains archaeal host prediction. The iPHoP_db_Aug23 database (GTDB Release 202) includes 254,090 bacterial genomes but only 4,316 archaeal genomes[25]. Notably, 24 of the 25 archaeal-host MAVs identified were only detectable because of the supplementation of the default iPHoP Aug_2023_pub_rw database with dereplicated MAGs recovered from the same food metagenomes. More broadly, among all MAVs with confident host predictions, 1545 were linked to novel host MAGs from this study, representing putative novel bacterial and archaeal species absent from existing references. These newly represented hosts spanned 68 genera, covering 2 archaeal phyla (24 MAVs) and 4 bacterial phyla (1521 MAVs). Within Archaea, 23 MAVs were predicted to infect unclassified Thermoproteota, and one targeted Halarchaeum (Halobacteriota). These findings highlight that even modest improvements in genome representation from understudied environments can substantially enhance viral host-prediction accuracy. A notable finding in our analysis was the presence of MAVs predicted to infect several clinically important ESKAPE pathogens. Four of the six ESKAPE groups appeared in our catalogue, with predicted hosts including *S. aureus* (8 MAVs), *K. pneumoniae* (105 MAVs), *Ac. baumannii* (15 MAVs), and *P. aeruginosa* (2 MAVs). These pathogens are well-recognized reservoirs and vectors of AMR, and previous studies have shown that foodborne ESKAPE group strains, particularly *S. aureus*, frequently harbor multiple AMR genes[26]. Biofilm formation is also widespread across these bacteria, and the resulting structured communities are markedly more tolerant than planktonic cells to antibiotics, disinfectants, and sanitizing agents[27]. This resilience has driven interest in alternative and more sustainable control strategies. Among these, host-specific bacteriophages have emerged as particularly promising[28] because they are inherently safe for humans[29,30] and can be applied without altering the sensory qualities of food products[31,32].

In terms of safety, known virulence factors and antimicrobial resistance genes were exceptionally rare but not entirely absent. We identified six genomes with high-confidence virulence factors (predominantly gtrB homologues involved in LPS modification) and a single MAV encoding fosD. This low frequency aligns with studies showing that phage-genome-encoded ARGs are generally uncommon and that apparent ARG signals in viromes can reflect non-genomic associations (for example, packaged bacterial DNA or outer-membrane vesicles), rather than stable phage carriage[33,34]. Nevertheless, environmental viromes have been shown to contain small numbers of novel, functional ARGs, so the large fraction of uncharacterized genes across MAVs (Fig. 5A) means we cannot exclude the presence of as-yet-unreported resistance determinants[34]. Similar proportions of uncharacterised genes have been reported in marine, soil, human gut, and engineered ecosystems, indicating that incomplete functional annotation is a pervasive feature of viral metagenomes rather than a peculiarity of food-associated systems[35,36]. Together, these lines of evidence suggest that phage-mediated dissemination of currently known AMR genes in these food environments is likely limited, though continued surveillance and functional validation remain warranted.

To investigate viral prevalence across bacterial lineages, high- and medium-quality MAGs were screened for viral associations. While viral signatures were consistently detected, their prevalence varied across families. Bacillaceae, Thermaceae, and Staphylococcaceae exhibited near-universal viral associations, whereas families like Vibrionaceae and Dermabacteraceae showed much lower detection rates (Fig. 4A). The widespread presence of antiviral defence systems (Fig. 4B, C). supports the idea of ongoing viral pressure in fermented food microbiomes[37]. In more than half of the bacterial families examined, over 80% of MAGs encoded at least one defence mechanism, suggesting that viral predation plays an important evolutionary role. While incomplete MAGs may limit detection, the high prevalence and variety of defence systems point to their biological relevance. The number and types of systems per genome varied across lineages (Fig. 4C), indicating different strategies for resisting phage infection and supporting the pan-immunity model.

For example, Alteromonadaceae and Propionibacteriaceae exhibited particularly high numbers of defence systems, consistent with more complex or layered defence repertoires. Notably, both families also showed relatively high levels of phage association as detected using VirSorter2 (Fig. 4A). The co-occurrence of frequent phage associations and elevated defence system counts is compatible with increased phage exposure in these lineages. However, defence system abundance alone is insufficient to infer viral pressure, as it may also reflect differences in genome size, lineage-specific evolutionary history, or horizontal gene transfer. Discriminating among these possibilities will require additional data, such as quantitative measurements of viral abundance, CRISPR spacer–protospacer matching, or experimental infection assays.

Restriction-Modification (RM) systems were the most common and widely distributed, found in over half of MAGs from families such as

1. Summary of MAV Features

1.1. Genome: seq1021 | Sample: TG | Dataset: ID052_MG_T1_S290_Kefir4All

1.1.1. MAV Properties

Contig Length (bp)	GC Content (%)	CDS Coding Density (%)	Total Genes	Provirus?	Viral Genes	Host Genes	CheckV Quality	Food Category	Food type	Food subtype
129,472	0.32	89.7	192	No	64	1	Complete	dairy	kefir	milk_kefir_grain

1.1.2. Viral Taxonomy (vConTACT v.3.0)

Predicted Phylum	Predicted Class	Predicted Order	Predicted Family	Predicted Subfamily	Predicted Genus
Uroviricota	Caudoviricetes	novel_order_98_of_Caudoviricetes	novel_family_3_of_novel_order_98_of_Caudoviricetes	novel_subfamily_2_of_novel_family_3_of_novel_order_98_of_Caudoviricetes	novel_genus_0_of_novel_subfamily_2_of_novel_family_3_of_novel_order_98_of_Caudoviricetes

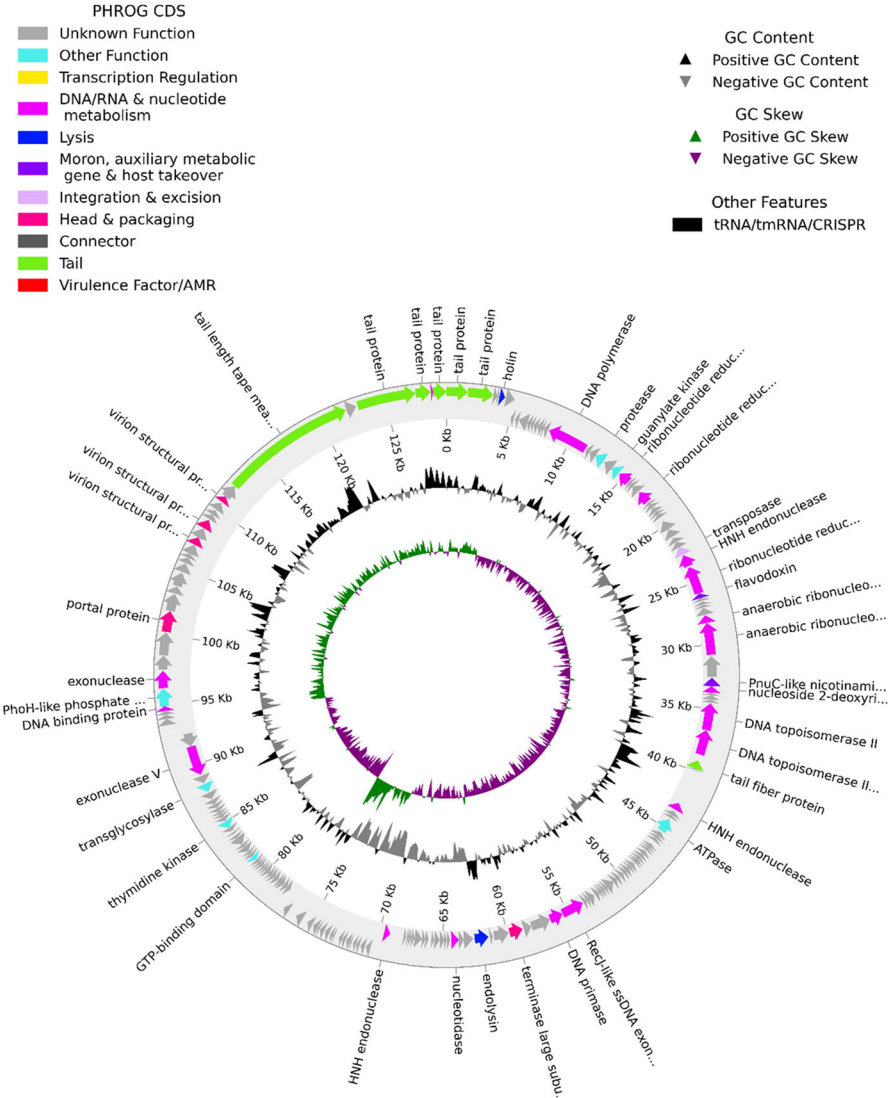
1.1.3. Viral Taxonomy (Pharokka)

Description	Jumbophage	Host	Lowest Taxa	Genus	Subfamily	Family	Order	Class	Phylum	Kingdom	Realm	Baltimore Group
Lactococcus phage phiL47	FALSE	Lactococcus	Audreyjarvisvirus	Audreyjarvisvirus	Unclassified	Unclassified	Unclassified	Caudoviricetes	Uroviricota	Heunggongvirae	Duplodnaviria	Group I

1.1.4. Predicted Host Taxonomy

Host Genome	Host Domain	Host Phylum	Host Class	Host Order	Host Family	Host Genus	Host Species
RS_GCF_002441655.1	Bacteria	Bacillota	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus	Lactococcus fujiensis
RS_GCF_002441885.1	Bacteria	Bacillota	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus_A	Lactococcus_A chungangensis
RS_GCF_001591705.1	Bacteria	Bacillota	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus	Lactococcus cremoris

1.1.5. Genome Overview Plot: genome features and functional annotation



1.1.6. GBK File Link

Download GBK File: seq1021.gbk

Fig. 6 | Representative viral genome annotation and characteristics. This figure shows a representative metagenome-assembled virus (MAV) from the FVGC, including a *Pharokka*-generated circular genome plot with PHROG-annotated coding sequences and an accompanying table summarising MAV properties (length, GC content, CDS density, gene counts, CheckV quality, provirus status), predicted host taxonomy (via iPHoP), and viral taxonomy (via vConTACT3 and Pharokka), with full MAV figures available on OSF:https://osf.io/c7uwn/overview.

Bifidobacteriaceae and Actinomycetaceae. CRISPR-Cas systems were also frequently identified, particularly in Streptococcaceae, Thermaceae, and Enterococcaceae. Other systems, including Abi, CBASS, BREX and Shango, further highlight the diversity of antiviral defences present in these microbial communities. These patterns are consistent with large-scale genomic surveys showing that RM and CRISPR–Cas constitute two of the most widespread antiviral systems in bacteria, that many prokaryotic genomes encode multiple, mechanistically distinct defence systems, and that defence repertoires vary considerably across lineages and environments[38].

In contrast, anti-defence systems were much less common. Only 314 MAVs (MACs) were predicted to encode any such function, suggesting limited but detectable viral adaptation to host immunity. Most hits were Anti-CRISPR proteins ($n = 251$), which is consistent with the widespread presence of CRISPR-Cas systems among bacterial hosts in fermented food environments[39]. Other predicted systems, such as Anti-Thoeris, Anti-RM and Anti-RecBCD, point to a broader range of virus-host interactions (Fig. 5B, C). Despite extensive annotation efforts, the majority of viral genomes remained functionally uncharacterised, underlining the current limitations in reference databases and viral gene annotation (Fig. 5A). The low prevalence and skewed distribution of anti-defence systems in food-associated viromes represent a deviation from patterns observed in environmental phage populations. Tesson et al.[38] reported that ~20% of diverse phages encode anti-defence systems, with Anti-RM being the most abundant category given the widespread prevalence of RM systems in bacteria. In contrast, we observed <10% (314) of food-associated MAVs encoding anti-defence functions, with a paradoxical dominance of Anti-CRISPR (80%) over Anti-RM systems despite RM being the primary bacterial defence in our dataset (Fig. 4C). These deviations are likely ecologically meaningful rather than methodological artifacts. Food fermentation environments differ fundamentally from the diverse ecosystems (soil, marine, clinical, gut) analyzed by Tesson et al.[40] in several key aspects. First, food fermentations harbor simplified, low-diversity microbial communities dominated by specific bacterial lineages adapted to controlled conditions. Second, the reduced environmental variability and selective pressure in fermentation may favor stable phage-host relationships over antagonistic arms races[41]. Third, many food-associated phages are temperate (Fig. 3A), requiring long-term compatibility with host defences rather than active inhibition. The scarcity of anti-RM systems, despite high RM prevalence, warrants future investigation. Additional research could include exploring viral–plasmid–host interactions at a systems level. Plasmids frequently encode defence and counter-defence functions, including anti-CRISPR, anti-restriction, and other anti-defence proteins, which are often enriched in plasmid leading regions and facilitate conjugative transfer by evading host immune systems[42]. Through horizontal gene transfer, recombination within shared hosts, and the existence of hybrid phage–plasmid elements, these plasmid-associated defence and counter-defence genes can also contribute to the broader mobile gene pool accessible to bacteriophages[43].

Overall, this study reveals largely uncharacterized diversity of dsDNA viruses in fermented foods, including numerous novel lineages and a wealth of poorly annotated genomic elements. The FVGC serves as a foundational resource for advancing the study of viral ecology in food-associated microbial communities. With high prevalence of unknown gene functions, rare but notable occurrences of virulence and AMR genes, and diverse but limited distribution of anti-defense systems targeting bacterial immunity, our findings provide insights into the complex evolutionary dynamics between phages and their microbial hosts. By linking viral sequences to host lineages and defense features, the FVGC enables future investigations into virus–host interactions, horizontal gene transfer, and the functional roles of dsDNA viruses in food fermentation ecosystems.

Methods

Metagenomic sample collection

The cFMD is a publicly accessible database, hosted on GitHub (https://github.com/SegataLab/cFMD), providing curated metadata along with taxonomic and functional profiles for thousands of fermented food-derived metagenomes[18]. For this analysis, we utilized a total of 3684 metagenomes across 88 datasets, including 78 publicly available datasets curated in the cFMD v1.2.0 and an additional 10 from ongoing research, representing 777 newly sequenced metagenomes.

Metagenomics data processing, taxonomic profiling, and functional analysis

All metagenomic data processing was performed on the Teagasc high-performance computing cluster. Initially, paired-end reads were retrieved from ENA (European Nucleotide Archive). All metagenomic datasets underwent quality control using a validated pipeline (preprocessing). Adapter trimming and per-base quality trimming were performed using Trim Galore v0.6.6 (Trim Galore), which applies Cutadapt to remove sequencing adapters and trims low-quality bases according to Phred scores. Reads with an average quality score <20, length <75 bp, or containing more than two ambiguous bases ($N > 2$) were subsequently discarded. Human (hg19) and phiX174 contaminants were removed by mapping reads against their reference genomes using BowTie2 v2.2.9 with the --sensitive-local parameter[44,45]. Reads passing QC were then sorted into forward, reverse, and unpaired files. Taxonomic profiling was conducted using MetaPhlAn 4 v4.0.6[46], with the MetaPhlAn 4 database v0.6. For functional profiling. The bioinformatics tools used in this analysis were implemented via custom scripts available at the GitHub repository: https://github.com/SegataLab/MASTER-WP5-pipelines, which were adapted from the repository's collection. Full methodology details are described immediately below.

MAG recovery, taxonomy, and antiviral defense profiling

Metagenomic assembly was conducted using metaSPAdes v3.15.5[47], while genome binning was performed with MetaBAT2 v2.15[48], both using default settings. To assess the quality of the MAGs, we applied CheckM2 v0.1.3[49]. MAGs with less than 80% completeness and/or more than 5% contamination were excluded from further analysis.

GTDB-Tk v2.4.0[50] was utilized to assign taxonomy to the recovered MAGs and to identify potentially novel species. Defence systems were predicted from MAGs using DefenseFinder v2.0.0[38].

Extraction and quality control of metagenome-assembled viruses

To comprehensively characterise DNA-derived viral populations in food, we employed two complementary pipelines to recover metagenome-assembled viruses (MAVs): (i) a de novo assembly-based approach, and (ii) a MAG-associated viral detection pipeline (Fig. 1). De novo viral assembly was conducted using SPAdes v3.15.3[47], applying the metaviral parameter, which is specifically optimized for viral genomes. Unlike microbial genomes, viral assemblies are typically not binned due to several unique computational and biological challenges[51]. Viral genomes are often short and assembled into scaffolds that typically represent most or all of the genome, reducing the need for binning[51,52]. The resulting assemblies were classified using viralVerify v1.1[10] employing the Pfam37.0 database as the reference for viral protein domain identification. Assembles classified as "Viral" were retained from further analysis.

To complement the de novo assembly pipeline and improve detection of integrated elements, we implemented a MAG-associated viral detection pipeline. Specifically, we applied VirSorter2 to identify potential viral sequences in MAGs and unbinned contigs. A cutoff score of 0.5 (default) was selected to maximize sensitivity, and potential viral sequences shorter than 1500 bp were excluded from further analysis. Both sets of viral sequences- those from de novo assembly and MAG-associated pipelines underwent further quality control. To reduce false positives from both pipelines, we applied additional screening criteria (https://doi.org/10.17504/protocols.io.bwm5pc86) using CheckV viral and host gene counts alongside VirSorter2 scores. Specifically, contigs were retained if they contained at least one viral gene identified by CheckV. If no viral genes were detected, contigs were only kept if they had zero host genes or met conservative thresholds (VirSorter2 score ≥0.95 or >2 hallmark genes).

To ensure sequence integrity, CheckV v1.0.3[19] was used to evaluate viral genome completeness, and prophage status, leveraging the checkv-db-v1.5 database. Based on this analysis, the resulting MAVs were classified into quality tiers: complete (defined by the presence of direct terminal repeats [DTR] and 100% completeness), high quality (≥ 90% complete), medium quality (50–80% complete), and low quality (<50% complete)[19]. MAVs falling into the low-quality category were excluded from the FVGC collection. All tools were applied with default parameters unless otherwise specified.

Taxonomic, lifestyle and functional annotation of MAVs

MAVs were subsequently annotated using a multi-tool approach. First, Pharokka v1.7.3[53], a viral genome annotation tool specifically designed for phage and prophage genomes, was employed to predict genes, functional elements, and other key genomic features within the MAVs. Furthermore, taxonomic classification was considered from the INPHARED search tool embedded in Pharokka, vConTACT3 v3.0.3[54] with database version 230, which is derived from NCBI Virus RefSeq release 230 and incorporates Virus-Host DB data from GenBank release 257 (Zenodo DOI: 10.5281/zenodo.10035619). This database provides a comprehensive taxonomic context through established reference viral genomes against which our MAVs were compared for novelty assessment. Host taxonomy was determined using iPHoP v1.3.3[25]. Phage lifestyles (temperate, virulent or chronic) were inferred using ProkBERT-based PhaStyle v0.1, employing the PhaStyle-mini model (neuralbioinfo/PhaStyle-mini) for sequence-level classification[21]. Phage anti-defence systems were identified using DefenseFinder v2.0.0, run using the -A parameter[38].

Clustering-based approach for viral classification and grouping

To cluster MAVs at species-level resolution, we implemented a rapid ANI-based clustering workflow, following the pairwise ANI clustering strategy outlined in the CheckV[19] toolkit (https://bitbucket.org/berkeleylab/checkv/src/master/). Briefly, MAV nucleotide sequences were first compiled and formatted into a BLAST+ database using makeblastdb (BLAST v2.14.1). An all-vs-all similarity search was then conducted using blastn with the megablast algorithm, allowing up to 10,000 target hits per query. Pairwise ANI was subsequently calculated by combining local alignments between each sequence pair, using the anicalc.py script from the same workflow. Finally, genomes were clustered into species-level groups using aniclust.py, which applies CD-HIT-like single-linkage clustering based on both ANI and alignment coverage. Clusters were defined using thresholds recommended by MIUViG standards: 95% ANI and at least 85% alignment fraction (AF) across the target genome[55].

A total of 3445 phage MAVs were compared against the IMG/VR database (version 4). MAVs were aligned to all high-confidence isolates and Uncultivated Virus Genomes (UViGs). To identify viral genomes originating from food-associated environments, IMG/VR ecological metadata were manually queried using the GOLD-based "Ecological classification" fields[56]. Searches were filtered using keywords such as food, ferment, dairy, cheese, and related terms, resulting in the identification of 7157 high-confidence isolates and UViGs associated with food environments.

Sequence similarity searches were conducted using vsearch (usearch_global)[57] with global pairwise alignment. Matches were retained when exhibiting ≥95% sequence identity, and for each query contig, the top-scoring hit was extracted along with its associated IMG/VR ecological metadata. The "top environment" category was derived from the hierarchical IMG/VR–GOLD ecosystem classifications (e.g., Engineered → Food production → Dairy products). Counts of contig-to-environment assignments were summarized, and the distribution of matches across food-associated categories. All tools outlined above were applied using default parameters unless otherwise specified.

Quantification and statistical analysis

All statistical analysis was performed in R v4.4.2[58] implemented through R studio. To compare MAV recovery across diverse food metagenomes while controlling for differences in sequencing depth, we quantified the number of MAVs per sample and normalized these counts by total reads. Depth-normalized MAV recovery was calculated as (1),

$$\text{Depth} - \text{normalized MAV recovery} = \frac{\text{Number of MAV spersample}}{\text{Total sequencing reads per sample}} \times 10^6$$

These values represent the rate of MAV recovery per unit of sequencing effort and are not equivalent to absolute viral abundance. Fractional values occur when MAV counts are low relative to sequencing depth.

To evaluate whether a given feature (e.g., viral lifestyle, taxonomic group, or other categorical trait) varied systematically across food categories, we applied Fisher's exact test[59] to contingency tables of feature levels by food category. For tables with larger sample sizes, p values were estimated via Monte Carlo simulation with 100,000 replicates to reduce computational burden.

Food categories with fewer than 10 observations were excluded from statistical testing to ensure adequate power, though all categories were retained in descriptive visualizations. Effect size was quantified using Cramér's V: (2),

$$V = \sqrt{\frac{\chi^2}{n \times (\min(r, c) - 1)}}$$

where χ^2 is the chi-squared statistic, n is the total sample size, and r and c are the numbers of rows and columns in the contingency table. Cramér's V values were interpreted as negligible (<0.1), small (0.1–0.3), medium (0.3–0.5), or large (>0.5).

For binary features, we conducted pairwise Fisher's exact tests between all category pairs, with p-values adjusted using the Bonferroni correction to control the family-wise error rate at $\alpha = 0.05$. For each category, we also calculated the proportion of samples with the target feature, with 95% confidence intervals estimated using the Wilson score method to provide more accurate coverage for proportions near 0 or 1[60].

Data visualization and analysis

Data visualization was performed in R v4.4.2[58]. Visualization techniques included ordination plots, correlation scatter plots, bar charts, Sankey, and Venn diagrams, which were generated using ggplot2 v3.3.6[61]. Heatmaps were also created using ggplot2, while phylogenetic trees were visualized with ggtree v3.2.168[62]. The experimental design figure was produced using BioRender[63] (Fig. 1).

Data availability

Details on accessing the metagenomic data utilized in this study can be found on the cFMD GitHub page: https://github.com/SegataLab/cFMD. All medium-quality and higher MAV assemblies produced here have been deposited in the Open Science Framework (OSF): https://osf.io/c7uwn/overviews. Bioinformatic open-source tools and parameters used in this present study are defined in the "Methods" section, and scripts can be found in the following GitHub repository: https://github.com/LiamHWalsh/Food-DNA-Viral-Genome-Collection-FVGC. More details regarding the code to reproduce the analyses are available upon request.

Received: 22 July 2025; Accepted: 11 February 2026;
Published online: 27 March 2026

References

1. Brown, T. L., Charity, O. J. & Adriaenssens, E. M. Ecological and functional roles of bacteriophages in contrasting environments: marine, terrestrial and human gut. *Curr. Opin. Microbiol.* **70**, 102229 (2022).
2. Castledine, M. & Buckling, A. Critically evaluating the relative importance of phage in shaping microbial community composition. *Trends Microbiol.* **32**, 957–969 (2024).

3. Shkoporov, A. N., Turkington, C. J. & Hill, C. Mutualistic interplay between bacteriophages and bacteria in the human gut. *Nat. Rev. Microbiol.* **20**, 737–749 (2022).
4. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617 (2012).
5. Ranveer, S. A. et al. Positive and negative aspects of bacteriophages and their immense role in the food chain. *npj Sci. Food* **8**, 1 (2024).
6. Xu, X. & Gu, P. Overview of phage defense systems in bacteria and their applications. *Int. J. Mol. Sci.* **25**, 13316 (2024).
7. Manrique, P. et al. Healthy human gut phageome. *Proc. Natl. Acad. Sci. USA* **113**, 10400–10405 (2016).
8. Roux, S., Matthijnssens, J. & Dutilh, B. E. Metagenomics in Virology. *Encyclopedia of Virology* 133–140 (2021).
9. Walsh, L. H., Coakley, M., Walsh, A. M., O'Toole, P. W. & Cotter, P. D. Bioinformatic approaches for studying the microbiome of fermented food. *Crit. Rev. Microbiol.* **49**, 693–725 (2023).
10. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).
11. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–d743 (2023).
12. Zolfo, M. et al. Discovering and exploring the hidden diversity of human gut viruses using highly enriched virome samples. *bioRxiv*, https://doi.org/10.1101/2024.02.19.580813 (2024).
13. Palermo, C. N., Fulthorpe, R. R., Saati, R. & Short, S. M. Metagenomic analysis of virus diversity and relative abundance in a eutrophic freshwater harbour. *Viruses* **11**, 792 (2019).
14. Garmaeva, S. et al. Studying the gut virome in the metagenomic era: challenges and perspectives. *BMC Biol.* **17**, 84 (2019).
15. Santos-Medellin, C. et al. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.* **15**, 1956–1970 (2021).
16. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-age characterization of soil viruses: a user's guide to virus isolation, detection within metagenomes, and viromics. *Soil Syst.* **4**, 23 (2020).
17. Antipov, D., Rayko, M., Kolmogorov, M. & Pevzner, P. A. viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data. *Genome Biol.* **23**, 57 (2022).
18. Carlino, N. et al. Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. *Cell* **187**, 5775–5795.e15 (2024).
19. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
20. Minakhin, L. et al. Genome comparison and proteomic characterization of Thermus thermophilus bacteriophages P23-45 and P74-26: siphoviruses with triplex-forming sequences and the longest known tails. *J. Mol. Biol.* **378**, 468–480 (2008).
21. Juhász, J. et al. ProkBERT PhaStyle: accurate phage lifestyle prediction with pretrained genomic language models. *Bioinformatics Adv.* **5**, vbaf188 (2025).
22. Stern, A. & Sorek, R. The phage-host arms race: shaping the evolution of microbes. *Bioessays* **33**, 43–51 (2011).
23. Murata, R., Shim, Y. Y., Kim, Y. J., Reaney, M. J. T. & Tse, T. J. Exploring lactic acid bacteria in food, human health, and agriculture. *Crit. Rev. Food Sci. Nutr.* 1–27 (2023).
24. Mohammadpour, H., Cardin, M., Carraro, L., Fasolato, L. & Cardazzo, B. Characterization of the archaeal community in foods: the neglected part of the food microbiota. *Int. J. Food Microbiol.* **401**, 110275 (2023).
25. Roux, S. et al. iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
26. Quijada, N. M. et al. The food-associated resistome is shaped by processing and production environments. *Nat. Microbiol.* **10**, 1854–1867 (2025).
27. Abee, T., Kovács, A. T., Kuipers, O. P. & van der Veen, S. Biofilm formation and dispersal in Gram-positive bacteria. *Curr. Opin. Biotechnol.* **22**, 172–179 (2011).
28. Jayamanne, M. N. & Foddai, A. C. G. Use of bacteriophages for biocontrol of pathogens in food and food-contact surfaces: a systematic review of the literature. *Sustain. Microbiol.* **2**, qvaf005 (2025).
29. Meaden, S. & Koskella, B. Exploring the risks of phage application in the environment. *Front. Microbiol.* **4**, 358 (2013).
30. Kasman, L. M. & Porter, L. D. Bacteriophages (2018).
31. Aprea, G. et al. The applications of bacteriophages and their lysins as biocontrol agents against the foodborne pathogens Listeria monocytogenes and Campylobacter: an updated look. *Vet. Ital.* **54**, 293–303 (2018).
32. Moye, Z. D., Woolston, J. & Sulakvelidze, A. Bacteriophage applications for food production and processing. *Viruses* **10**, 205 (2018).
33. Enault, F. et al. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).
34. Moon, K. et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).
35. Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
36. Hatfull, G. F. Dark matter of the biosphere: the amazing world of bacteriophage diversity. *J. Virol.* **89**, 8107–8110 (2015).
37. Somerville, V. et al. Extensive diversity and rapid turnover of phage defense repertoires in cheese-associated bacterial communities. *Microbiome* **10**, 137 (2022).
38. Tesson, F. et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
39. Cui, Y. & Qu, X. CRISPR-Cas systems of lactic acid bacteria and applications in food science. *Biotechnol. Adv.* **71**, 108323 (2024).
40. Wolfe, B. E. & Dutton, R. J. Fermented foods as experimentally tractable microbial ecosystems. *Cell* **161**, 49–55 (2015).
41. Somerville, V. et al. Functional strain redundancy and persistent phage infection in Swiss hard cheese starter cultures. *Isme J.* **16**, 388–399 (2022).
42. Samuel, B., Mittelman, K., Croitoru, S. Y., Ben Haim, M. & Burstein, D. Diverse anti-defence systems are encoded in the leading region of plasmids. *Nature* **635**, 186–192 (2024).
43. Pfeifer, E. & Rocha, E. P. C. Phage-plasmids promote recombination and emergence of phages and plasmids. *Nat. Commun.* **15**, 1545 (2024).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
46. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
47. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
48. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
49. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

50. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).
51. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Res.* **50**, e83 (2022).
52. Mallawaarachchi, V. et al. Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* **39**, btad586 (2023).
53. Bouras, G. et al. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, btac776 (2022).
54. Bolduc, B. et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243 (2017).
55. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
56. Mukherjee, S. et al. Genomes OnLine Database (GOLD) v.10: new features and updates. *Nucleic Acids Res.* **53**, D989–D997 (2024).
57. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
58. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
59. Agresti, A. *Categorical Data Analysis,* 2nd edn (John Wiley & Sons, 2002).
60. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).
61. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
62. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinform.* **69**, e96 (2020).
63. BioRender. *BioRender: scientific image and illustration software* (2024).

Acknowledgements

The MASTER EU Consortium was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 818368. The work was also partially supported by the European Union's Horizon Europe programme (project DOMINO-101060218), the Italian Ministry of Foreign Affairs and International Cooperation (project FOODMICROHERITAGE-VN21GR09), the European Research Council (ERC-STG project MetaPG-716575 and ERC-CoG project microTOUCH-101045015) to N.S., the European Union's Horizon 2020 programme (projects ONCOBIOME-825410 and IHMCSA-964590) to N.S., the MUR PNRR INEST-Interconnected Nord-Est Innovation Ecosystem by Next-GenerationEU (project ECS00000043) to N.S., the National Cancer Institute of the National Institutes of Health (projects 1U01CA230551 and U01CA261961) to N.S., the Premio Internazionale Lombardia e Ricerca 2019 to N.S, the European Union's Horizon 2020 programme (Marie Skłodowska-Curie project 101034371) to N.M.Q., the Spanish Ministry of Science and Innovation (Juan de la Cierva postdoctoral contract FJC2019-042125-I). Research in the Vision 1 laboratory is also funded by Taighde Éireann-Research Ireland, formerly Science Foundation Ireland (SFI) under grant number SFI/12/RC/2273_P2 (APC Microbiome Ireland), by Taighde Éireann with the Irish Department of Agriculture, Food and the Marine under grant number SFI/16/RC/3835 (VistaMilk) and by Enterprise Ireland and industry in the Food for Health Ireland (FHI)-3 project, under grant number TC/2018/0025.

Author contributions

Conceptualization: L.H.W., V.S., N.S., D.v.S., J.M., V.Som., P.D.C. Methodology: O.O'S., L.H.W., V.S., A.N.S. Software: O.O'S., L.H.W. Validation, Supervision, Funding Acquisition: O.O'S., P.D.C., N.S.Formal Analysis, Writing, Original Draft Preparation: L.H.W., V.S.Data Curation, Visualization: L.H.W., V.S.Writing, Review and Editing: L.H.W., P.D.C., O.O'S., N.S., D.v.S., J.M., A.N.S., J.G.K., V.S., V.Som., J.A. Project Administration: O.O'S., P.D.C., L.H.W.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-026-00941-9.

Correspondence and requests for materials should be addressed to Paul D. Cotter or Orla O'Sullivan.