Research, Society and Development, v. 11, n. 10, e273111032693, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i10.32693

Expanding the genomes insight of *Streptococcus thermophilus* phages through a multifaceted approach

Expandindo a visão dos genomas de fagos de *Streptococcus thermophilus* através de uma abordagem multifacetada

Ampliación del conocimiento de los genomas de los fagos de *Streptococcus thermophilus* mediante un enfoque multifacético

Received: 07/08/2022 | Reviewed: 07/19/2022 | Accept: 07/24/2022 | Published: 07/30/2022

Laís S. Batalha
ORCID: https://orcid.org/0000-0001-9971-5791
Universidade Federal de Viçosa, Brazil
E-mail: lais.batalha@ufv.br
Pedro Marcus P. Vidigal
ORCID: https://orcid.org/0000-0002-5116-9856
Universidade Federal de Viçosa, Brazil
E-mail: pedro.vidigal@ufv.br
Marco Túllio P. Gontijo
ORCID: https://orcid.org/0000-0001-6605-3402
Universidade Estadual de Campinas, Brazil
E-mail: m264546@dac.unicamp.br
Monique R. Eller
ORCID: https://orcid.org/0000-0002-5412-897X
Universidade Federal de Viçosa, Brazil
E-mail: monique.eller@ufv.br

Abstract
Viruses have complex evolutionary relationships, and several strategies have been used in an attempt to classify the phages that infect *S. thermophilus*. In this study, we used a wide range of complementary methods, including comparative genomics, core genome analysis, and signature genes phylogenetics, to show that the *S. thermophilus* phages are organized into 142 species and five genera (three of them new) and that due to their genetic diversity, the classification at family level varies according to the classification criteria used. No significantly conserved genes were identified among the 183 genomes evaluated. However, the genes encoding holin protein were conserved in more than 95% of genomes. The holins analysis suggests that at least two α-helix are required for protein function within *S. thermophilus* phages. This study expanded the of knowledge about the genetic diversity and evolution of streptococcal phages, both fundamental to promoting control strategies and minimizing failures in milk fermentation processes.
Keywords: Biodiversity; Core genome; Holin; Panvirome; Taxonomy; Signature genes; Starter culture.

Resumo
Os vírus possuem relações evolutivas complexas, e várias estratégias têm sido utilizadas na tentativa de classificar os fagos que infectam *S. thermophilus*. Neste estudo, usamos uma ampla gama de métodos complementares, incluindo genômica comparativa, análise do genoma central e filogenética de genes de assinatura, para mostrar que os fagos de *S. thermophilus* estão organizados em 142 espécies e cinco gêneros (três deles novos) e que devido à sua diversidade genética, a classificação em nível familiar varia de acordo com os critérios de classificação utilizados. Nenhum gene significativamente conservado foi identificado entre os 183 genomas avaliados. No entanto, os genes que codificam a proteína holina foram conservados em mais de 95% dos genomas. A análise das holinas sugere que pelo menos duas α-hélices são necessárias para a função da proteína dentro dos fagos de *S. thermophilus*. Este estudo ampliou o conhecimento sobre a diversidade genética e evolução dos fagos estreptocócicos, ambos fundamentais para promover estratégias de controle e minimizar falhas nos processos de fermentação do leite.
Palavras-chave: Biodiversidade; Genoma core; Holina; Panviroma; Taxonomia; Genes de assinatura; Cultura iniciadora.

Resumen
Los virus tienen relaciones evolutivas complejas y se han utilizado varias estrategias en un intento de clasificar los fagos que infectan a *S. thermophilus*. En este estudio, utilizamos una amplia gama de métodos complementarios, incluida la genómica comparativa, el análisis del genoma central y la filogenética de los genes característicos, para demostrar que los fagos de *S. thermophilus* están organizados en 142 especies y cinco géneros (tres de ellos nuevos) y que debido a su

diversidad genética, la clasificación a nivel de familia varía según el criterio de clasificación utilizado. No se identificaron genes significativamente conservados entre los 183 genomas evaluados. Sin embargo, los genes que codifican la proteína holina se conservaron en más del 95% de los genomas. El análisis de holinas sugiere que se requieren al menos dos hélices α para la función de la proteína dentro de los fagos de *S. thermophilus*. Este estudio amplió el conocimiento sobre la diversidad genética y la evolución de los fagos estreptocócicos, ambos fundamentales para promover estrategias de control y minimizar fallas en los procesos de fermentación de la leche.
Palabras clave: Biodiversidad; Genoma central; Holina; Panviroma; Taxonomía; Genes distintivos; Cultivo iniciador.

1. Introduction

Advances in sequencing technologies and bioinformatics tools applied to metagenomic studies have contributed to the continued expansion of genome databases and understanding of the genetic diversity of bacteriophages (phages) (Dion et al., 2020). This knowledge is particularly important for the phages that infect *Streptococcus thermophilus*, a thermophilic starter culture fundamental in manufacturing dairy products such as cheese and yogurt (de Melo et al., 2018). Infection of starter cultures by phages during food fermentation, particularly in dairy fermentation, can lead to technological, economic, and environmental problems due to delays in fermentation or incomplete fermentation of milk, change in product quality, reduced productivity and disposal of milk, in cases of total process failure (de Melo et al., 2018; Mahony et al., 2020; Pujato et al., 2019).

Streptococcal phages share common features such as the *S. thermophilus* host, genome with modular structure, synteny of gene functions, and genetic mosaicism. However, the constant characterization of isolates, sequencing, and comparative analysis of genomes has shown an impressive genomic diversity among these phages (Lavelle et al., 2018a; McDonnell et al., 2016, 2017; Szymczak et al., 2017, 2019a). The expansion of viral diversity at the genomic level is further highlighted by the discovery of P738 group phages (Philippe et al., 2020), in addition to the other four existing groups: *Moineauvirus* (previously called *cos*-containing phages), *Brussowvirus* (previously called *pac*-containing phages), 5093 and 987 (Le Marrec et al., 1997; McDonnell et al., 2016; Mills et al., 2011; Tremblay and Moineau, 1999). Many of these phages have genes originating from non-dairy environments, further highlighting their complex evolutionary relationships (McDonnell et al., 2016; Mills et al., 2011). From an industrial point of view, this diversification represents a constant threat to milk fermentation processes. Recombination events between phages that infect different bacterial species can result in the extension of the host range and hamper the efficient control of phage infections in starter cultures, which generally consist of mixtures of lactic acid bacteria (LAB) (de Melo et al., 2018; Leroy and De Vuyst, 2004; McDonnell et al., 2016; Quiberoni et al., 2010; Szymczak et al., 2017).

The phage genomes that infect *S. thermophilus* are under a constant process of evolution due to the selective pressure of their hosts, both in response to natural mechanisms of resistance to phage infection, as well as by industrialization and the use of commercial starter cultures. The evolution mechanisms of these phages involve horizontal gene transfer, insertions, deletions, point mutations, and exchange of modules, which promote their genetic diversification and complex evolutionary relationships that do not follow the traditional hierarchical phylogeny (Brussow and Desiere, 2001; Desiere et al., 1998; Lavelle et al., 2018a, 2018b; Lucchini et al., 1999a; Philippe et al., 2020; Szymczak et al., 2017). To encompass the full range of genetic diversity, the taxonomy of streptococcal phages must be updated to provide a stable and standardized classification framework but dynamic, capable of accommodating revisions and reinterpretations of the perceived relationships between these phages as knowledge about new viruses and their genomes advances.

Thus, our classification strategy has included a variety of classification tools that employ very different approaches. Our analyzes ranged from clustering methods based on whole proteome similarity, for example, ViPTree (Nishimura et al., 2017) and VICTOR (Meier-Kolthoff & Göker, 2017), analysis of panvirome concerning gene distribution with Roary (Page et al., 2015), intergenomic similarity using VIRIDIC (Moraru et al., 2020), and detection and phylogeny of signature genes performed with CoreGenes (Zafar et al., 2002) and phylogeny.fr (Dereeper et al., 2008). This multifaceted approach allowed us to gradually descend from clustering at a family level to intra-familial relationships. Despite the diversity of applied methods, their results proved to be complementary and predominantly in agreement. All methods converged in robust results from five genera of *S. thermophilus* phages (three new), distributed in 142 species. With this study, we allowed the expansion of the number of *S. thermophilus* phages genera and species available in the ICTV database, making sense of the genetic diversity existing among them.

2. Methodology

2.1 Genome database

Sequences of phage genomes that infect *S. thermophilus* available in the GenBank/NCBI database (https://www.ncbi.nlm.nih.gov/genbank) (Benson et al., 2012) were retrieved in March 2022 along with other associated information. These data were complemented with bibliographic references linked to the genomes and organized in Supplementary Information (SI). Taxonomic classifications of *S. thermophilus* phages were retrieved from the International Committee on Taxonomy of Viruses (ICTV) (https://talk.ictvonline.org/taxonomy/). The lifestyle of the phages based on the conserved protein domains was predicted using BACPHLIP (BACterioPHage LIfestyle Predictor) (Hockenberry & Wilke, 2021).

2.2 Panvirome analysis and search for core genome

The panvirome of phages that infect *S. thermophilus* was analyzed using Roary 3.13.0 (Page et al., 2015). For this, the phage genomes were reannotated using PROKKA 1.14.5 (Seemann, 2014), selecting the annotation mode for viral genomes (--kingdom Viruses), and the GFF (General Feature Format) files were used as input by the Roary. Protein sequences annotated on the phage genomes were compared and grouped using a minimum percentage identity of 40% (-i 40). After grouping, the binary matrix of presence/absence of genes that are part of the accessory genome was used to group the phages. This matrix corresponds to the distribution of genes that encode proteins that were shared between the analyzed genomes. The accessory genome was considered because no genes were included in the core genome. The dendrogram with the grouping of the binary matrix was visualized using FigTree v1.4.4. (Rambaut, 2012). CoreGenes 5.0 (Zafar et al., 2002) was also used to identify the core genome.

2.3 Genomic organization of phages

Gene synteny and conservation of genomic structure of phages that infect *S. thermophilus* were evaluated using Clinker 0.0.21 (Gilchrist & Chooi, 2020). A reference genome was selected, including phages named: Streptococcus phage DT1 (NC_002072), Streptococcus phage O1205 (NC_004303), Streptococcus phage 5093 (NC_012753), Streptococcus virus 9871 (KU678389), and Streptococcus phage P738 (MK911750). These phages were chosen because they were used as a model for the group in several studies (Deveau et al., 2008; Levesque et al., 2005; McDonnell et al., 2017; Mills et al., 2011; Tremblay and Moineau, 1999) or because they were the first representatives of the group (Le Marrec et al., 1997; Lucchini et al., 1999b; McDonnell et al., 2016; Philippe et al., 2020). The products of the predicted ORFs (Open Reading Frames) for each phage were manually revised based on the genome annotations available in GenBank/NCBI.

2.4 Protein analysis

To evaluate the robustness of the Roary and CoreGenes analysis in detecting orthologous proteins in the panvirome, multiple alignments of holins sequence of the streptococcal phages were performed using MUSCLE 3.8.31 (Edgar, 2004). All holin sequences of the 183 phages were included. In addition, holin families enzymatic catalytic domains (ECDs) of holin families were predicted using HMMER software v3.3.2 (Finn et al., 2011). The structure of representative holins was predicted as described by Gontijo et al., (2022) using RoseTTAFold modeling (Baek et al., 2021). The structures were then compared using FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists) (https://fatcat.godziklab.org/) (Li et al., 2020) and visualized using Mol* Viewer (https://molstar.org/) (Sehnal et al., 2021).

2.5 Taxonomic assignment

To determine the family-level classification of the *S. thermophilus* phages, we used (i) the main predicted proteome-based clustering tools: VICTOR (VIrus Classification and Tree building Online Resource) (https://ggdc.dsmz.de/victor.php) (Meier-Kolthoff & Göker, 2017), VirClust (Virus Clusterer) (https://rhea.icbm.uni-oldenburg.de/VIRCLUST/) (Moraru, 2021) and ViPTree (Viral Proteomic Tree) (https://www.genome.jp/viptree/) (Nishimura et al., 2017) and we calculated (ii) shared orthologous genes. On the VICTOR analysis, all pairwise comparisons of the amino acid sequences were conducted using the Genome-BLAST Distance Phylogeny (GBDP) method (Meier-Kolthoff et al., 2013) under settings recommended for prokaryotic viruses (Meier-Kolthoff & Göker, 2017). The resulting intergenomic distances were used to infer a balanced minimum evolution tree with branch support via FastME including SPR postprocessing (Lefort et al., 2015) for the D6 intergenomic distance formula. Branch support was inferred from 100 pseudo-bootstrap replicates. Trees were rooted at the midpoint (Farris, 1972) and visualized with iTOL (Letunic & Bork, 2021). Taxon boundaries at the species, genus, and family level were estimated with the OPTSIL program (Göker et al., 2009), and VICTOR used the following clustering thresholds to suggest taxon boundaries at genus (0.749680), subfamily (0.888940) and family (0.985225) level (Meier-Kolthoff & Göker, 2017). The parameters for VirClust were: (i) protein clustering based on “e-value” after reciprocal BLASTp hits were removed if e-value >0.0001 and bitscore <50; (ii) hierarchical clustering based on protein clusters (PCs) with a bootstrapping of 1000 replicates. The resulting tree was split into viral genome clusters (VGCs) using an 0.98, 0.90, or 0.85 intergenomic distance threshold to determine the family-level classification of the *S. thermophilus* phages (Moraru, 2021). On the ViPTree analysis, a proteomic tree was constructed from the phage genomes based on genome-wide sequence similarities computed by tBLASTx, using dsDNA nucleic acid type and prokaryote host category. The orthologous gene identification was performed with VirClust, based on PCs calculated with the above parameters, CoreGenes and Roary.

In addition to the amino acid-based VICTOR analysis, the intra-familial relationships were analyzed using (i) nucleic acid-based intergenomic similarities calculated with VIRIDIC (Virus Intergenomic Distance Calculator) (http://rhea.icbm.uni-oldenburg.de/VIRIDIC/) (Moraru et al., 2020) and (ii) core protein phylogeny. The species and genus definition thresholds in VIRIDIC were 95% and 70% intergenomic similarity, respectively. The core proteins analysis was conducted as follows: (i) core genes were detected and annotated with Roary, VirClust, and CoreGenes; (ii) multiple alignments of core protein sequences were constructed with MUSCLE, and (iii) the phylogenetic tree was constructed with phylogeny.fr in “one click” mode (Dereeper et al., 2008). Phylogenetic trees of signature genes were rooted using a more distant relative (outgroup) and were accompanied by bootstrap values.

3. Results

3.1 General characteristics of phages that infect *S. thermophilus*

3.1.1 Distribution and current taxonomy of S. thermophilus phages

As of March 2022, 183 complete phage genomes that infect *S. thermophilus* were available in the GenBank database (SI file 1 Table S1). According to the literature, the phages that infect *S. thermophilus* are distributed into five groups: *Moineauvirus* (59.56%, n = 109), *Brussowvirus* (24.59%, n = 45), 5093 (7.10%, n = 13), 987 (7.65%, n = 14) and more recently, the P738 group (1.09%, n = 2)) (Achigar et al., 2017; Ali et al., 2014; Arioli et al., 2018; da Silva Duarte et al., 2018; Desiere et al., 1998; Deveau et al., 2008; Guglielmotti et al., 2009; Hynes et al., 2018; Lavelle et al., 2018a, 2018b; Le Marrec et al., 1997; Levesque et al., 2005; Lucchini et al., 1998; McDonnell et al., 2016, 2017; Mills et al., 2011; Neve et al., 1998; Philippe et al., 2020; Somerville et al., 2019; Stanley et al., 1997; Szymczak et al., 2017, 2019a, 2019b; Tremblay and Moineau, 1999).

At the moment, the ICTV recognizes the existence of 10 species of phages that infect *S. thermophilus*, which are classified into two genera: *Moineauvirus* (*Moineauvirus Abc2*, *Moineauvirus DT1*, *Moineauvirus mv7201*, *Moineauvirus Sfi19*, *Moineauvirus Sfi21*) and *Brussowvirus* (*Brussowvirus ALQ132*, *Brussowvirus bv858*, *Brussowvirus bv2972*, *Brussowvirus bvO1205*, *Brussowvirus Sfi11*), all belonging to the *Caudoviricetes* class (https://talk.ictvonline.org/taxonomy/, 2021 release). This classification is incomplete and does not represent the genetic diversity among streptococcal phage populations, which requires that a new genome-based classification of the group be proposed along with updating the ICTV database.

3.1.2 Lifestyle of S. thermophilus phages

Based on the phage lifestyle assessment (SI file 1 Table S1, Figure 1), a total of 114 (64.05%) and 64 (35.95%) phages were identified as virulent and temperate, respectively, with >50% probability. Five phage genomes from the *Moineauvirus* genus showed undefined classification and 53.84% (n = 56) were classified as temperate. In contrast to the phages of the *Brussowvirus* genus and 5093, 987 and P738 groups, which were classified as virulent in 91.11% (n = 41), 92.30% (n = 12), 78.57% (n = 11) and 100% (n = 2) of the analyzed genomes, respectively, with >50% probability.

3.1.3 Isolation and sampling sources

The main isolation sources of phages that infect *S. thermophilus* were cheese whey (n = 106) and cheese (n = 51), corresponding to almost 86% of the total phages isolated from fermentation processes. Both virulent and temperate phages were mostly isolated from these sources, corresponding to 85.96% (n = 98) and 84.37% (n = 54), respectively. Italy (n = 48), France (n = 39), and Ireland (n = 17) are the countries where the most phages were isolated, characterized, and sequenced (SI file 1 Table S1).

3.1.4 Phage morphology and host recognition

All phages that infect *S. thermophilus* are to the siphovirus morphotype and have an icosahedral capsid connected to a non-contractile tail (Lavelle et al., 2018a; Mahony and van Sinderen, 2014; McDonnell et al., 2017; Philippe et al., 2020). They have a limited host range, as most isolated phages only infect their primary host and a small number infect between two and 14 strains of *S. thermophilus* (Binetti et al., 2005; Lavelle, Martinez, et al., 2018; McDonnell et al., 2017b; Philippe et al., 2020; Zinno et al., 2010). Each group of phages has individual characteristics that reflect the genetic content and morphological characteristics of these agents, including host recognition structures at the tip of the tail. Phages from the *Moineauvirus* and *Brussowvirus* genera are morphologically similar, displaying long tails and often a terminal structure at the tip of the tail like a small plaque accompanied by a fiber (Accolas & Spillmann, 1979; Lavelle, Martinez, et al., 2018; Le Marrec et al., 1997). Group 5093 phages have globular appendages attached to the tips of the long tail, and group 987 phages have tails considerably shorter than those of the other phage groups, with a broad appendage at the tip of the tail that resembles that of the phages that infect *Lactococcus lactis* of group P335 (Accolas and Spillmann, 1979; Lucchini et al., 1998; McDonnell et al., 2016, 2017; Mills et al., 2011; Szymczak et al., 2017). On the other hand, the phages of the newly named group P738 have short tails similar to those of the group 987 phages and twisted tail fibers composed of two or three subfibers (Philippe et al., 2020).

3.1.5 Genome type and organization

Streptococcal phage genomes have a modular structure and synteny of gene functions (Le Marrec et al., 1997; McDonnell et al., 2016; Philippe et al., 2020; Szymczak et al., 2017; Tremblay and Moineau, 1999). They have double-stranded

DNA (dsDNA) in linear form, and the size and GC content of the genomes range from 30 to 48 kb and 36.6 to 40.2%, respectively. The number of predicted ORFs (Open Reading Frames) predicted in the phage genomes ranges from 38 to 68 (SI file 1 Table S1).

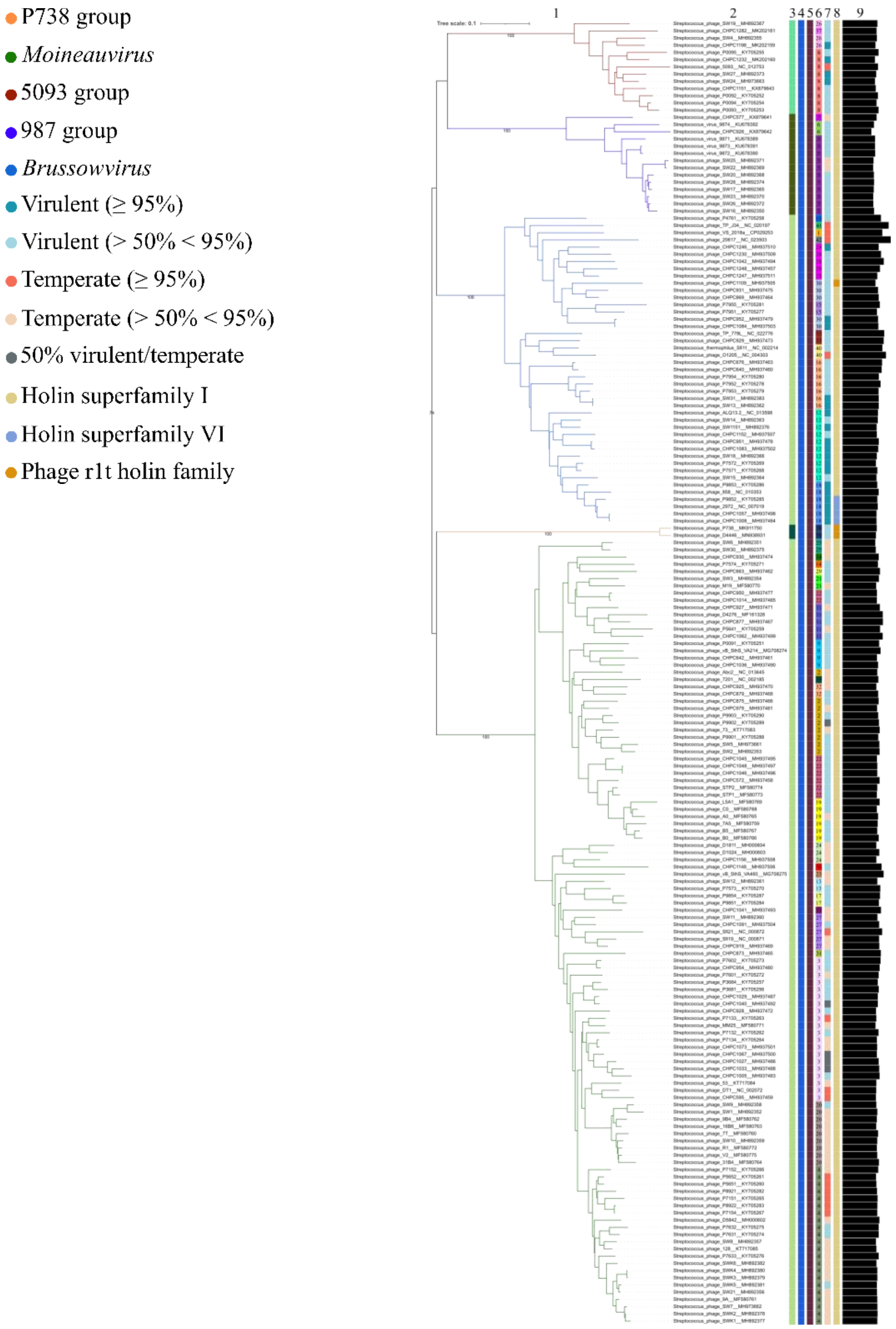
3.2 Genomic diversity of *S. thermophilus* phages

Panvirome analysis based on gene distribution divided the phages into five main clusters (SI file 2), which correspond to the phage groups already described (*Moineauvirus*, *Brussowvirus*, 5093, 987, and P738). In total, 580 genes were identified in the panvirome analysis. Of these, almost 21% (121) are unique genes in the analyzed genomes, which shows the genetic heterogeneity among the phages that infect *S. thermophilus*. The panviromes of the *Moineauvirus* and *Brussowvirus* genera have greater genetic diversity, with a total of 187 and 162 genes, respectively, and are organized into several subclusters, which may be related to the greater number of representatives already characterized. The panviromes of the 5093, 987, and P738 groups have 100, 81, and 50 genes, respectively. Still, the analysis of the genetic diversity of the phages of these groups, in turn, is limited by the smaller number of sequenced genomes. No significantly conserved genes were identified among the 183 genomes, considering a minimum identity of 40%. Likewise, no core genome was identified with CoreGenes. However, using Roary, two genes were present in 176 (96.17%) of the 183 analyzed genomes. Such genes are not present in the genomes of the two phages of the P738 group and five phages of the *Brussowvirus* genus. One of these genes had its putative function identified and corresponded to the genes encoding holin protein.

Comparative analyzes of the genetic organization and genome content of the reference phages showed that the genomes of these phages share common features such as the modular structure, synteny of gene functions, and high genetic variation between nucleotide sequences (Figure 2). However, no gene product was conserved in the five genomes. The lysis module was the only partially conserved genomic region among the phages from *Moineauvirus* and *Brussowvirus* genera, and 5093 and 987 groups phages. Notably, the reference phage of the P738 group was revealed to be genetically distinct from the others evaluated, sharing some gene products in a limited way with the reference phage of the 987 group. This result confirms the previous analysis that the lysis module is the only partially conserved genomic region among the phages from *Moineauvirus* and *Brussowvirus* genera, and 5093 and 987 groups phages.

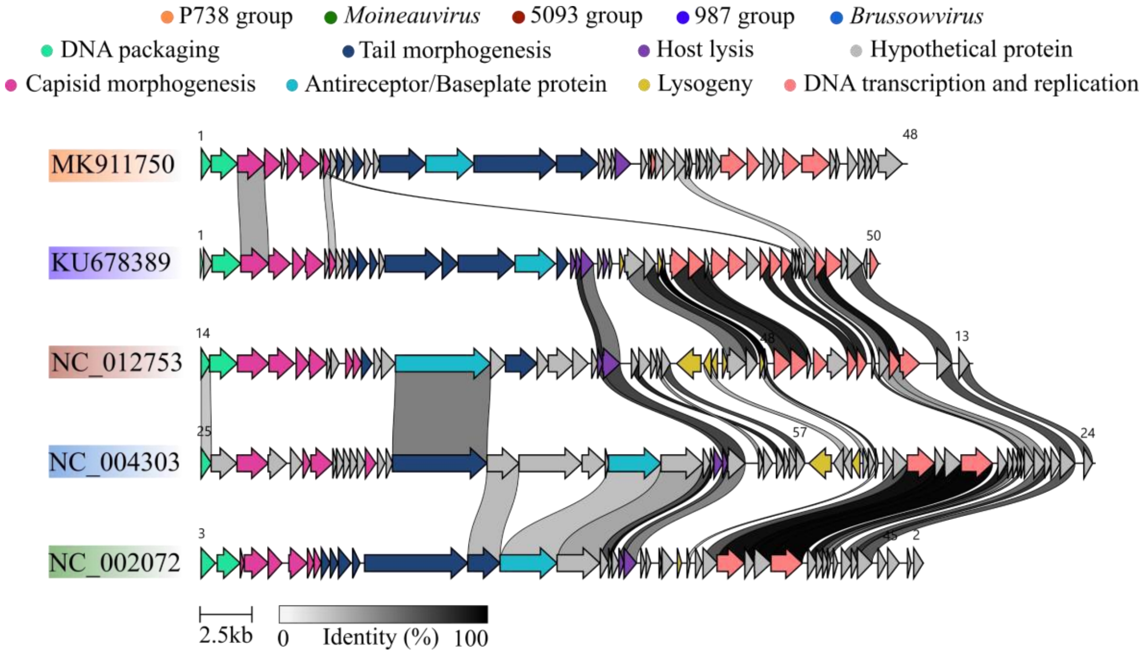
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i10.32693

Figure 1. Phylogenomic tree generated by VICTOR using the amino-acid sequences of the *Streptococcus thermophilus* phages.



Proteome-based tree (1), phage name (2), VICTOR genus cluster (3), VICTOR subfamily cluster (4), VICTOR family cluster (5), VIRIDIC genus cluster (6), BACPHLIP lifestyle prediction (7), phage holins classification (8), amino-acids sequence length - Min. (8,861), Max. (14,848) (9). See SI file 3 for the high-resolution version. During step 'check_and_map_input', the VICTOR analysis skipped duplicate genomes files 'Streptococcus_phage_SW32__MH892384' and 'Streptococcus_phage_SW33__MH892385' (duplicates of 'Streptococcus_phage_SW13__MH892362'). Although these phages are identical, they were isolated from different samples in three countries (SI file 1 Table S1). Source: Authors.

Figure 2. Comparative analysis of the genetic organization and content of phage genomes that infect *Streptococcus thermophilus*.



Predicted ORFs (indicated by arrows) and gene products (putative function indicated by color coding) are aligned with adjacent genomes according to percent amino acid identity (indicated by shaded boxes). Source: Authors.

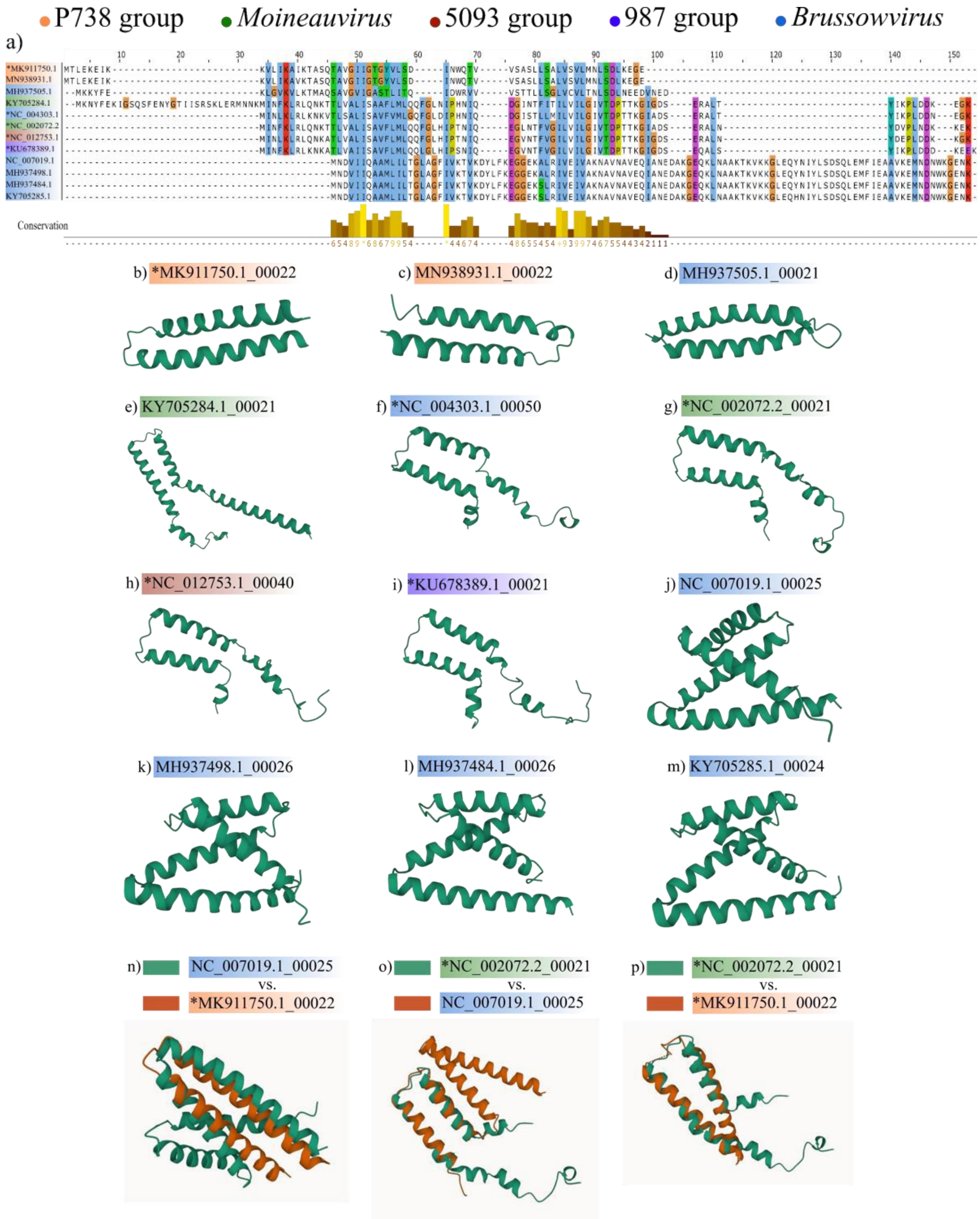
3.3 Characterization of *S. thermophilus* phages-encoded holins

Multiple alignments of holin sequences encoded by the 183 *S. thermophilus* phages revealed that this protein is conserved among the *Moineauvirus* genus and 5093 and 987 groups (SI file 4). Holin sequences are also conserved in phages of the *Brussowvirus* genus, except for five (MH937505, KY705285, NC_007019, MH937498, and MH937484). The holins encoded by the two phages of the P738 group are conserved with each other, presenting, notably, sequence similarity to the holin encoded by the phage MH937505 (Streptococcus phage CHPC1109, *Brussowvirus* genus). The P738 group phages (MK911750.1_00022 and MN938931.1_00022) encode holins belonging to phage r1t holin family (PF16945), as well as one phage (MH937505.1_00021) from the *Brussowvirus* genus (Figure 1). Still, in the *Brussowvirus* genus, four other phages (KY705285.1_00024, NC_007019.1_00025, MH937498.1_00026, and MH937484.1_00026) encode holins from holin superfamily VI (PF09682). The other bacteriophages (n = 176) encode holins from holin superfamily I (PF04531).

The two sequences of the holins encoded by the P738 group phages and five sequences of the *Brussowvirus* genus were selected to verify the structural prediction and determine possible structural conservation despite low sequence similarity (Figure 3a). Furthermore, these sequences were compared with the sequences of the holins encoded by the reference phages and representative holins from each one of the predicted protein families (PF16945, PF09682, and PF04531). Holins belonging to the phage r1t holin family (PF16945) have two α-helix (Figure 3b-d), while superfamily VI holins (PF09682) have four (Figure 3j-m). The structure of holins from holin superfamily I (PF04531) is conserved, with representatives in *Moineauvirus* and *Brussowvirus* genera and 5093 and 987 groups phages (Figure 3e-i). These holins possess two α-helix, and the superfamily I holin encoded by some *Moineauvirus* genus also have an additional helix at N-terminus (Figure 3e). The angstrom error estimates for the predictions are shown in SI file 5 Figure S1.

The pairwise alignment of the predicted structures revealed a significant pairwise alignment (*p*-value 1.67e-05) of two α-helix in holins from superfamily VI and r1t holin family (Figure 3n). Similar results are found for holins from superfamilies I and VI (*p*-value 8.13e-04; Figure 3o) and holins from superfamily I and r1t holin family (*p*-value 1.42e-03; Figure 3p). The graphic representation of the alignments and the differential distance matrix decomposition are shown in SI file 5 Figure S2. The conserved structural region in all the alignments is within the functional domains of each holin protein family, suggesting that at least two α-helix are required for protein function within *S. thermophilus* phages.

Figure 3. Structural prediction of representative holins encoded by *Streptococcus thermophilus phages*.



Multiple alignments of representative holin sequences (a). Phage r1t holin family (PF16945) from P738 group (b, c), and Brussowvirus genus (d). Holin superfamily I (PF04531) from Moineauvirus genus (e, g), Brussowvirus genus (f), 5093 group (h), and 987 group (i). Superfamily VI holins (PF09682) from the Brussowvirus genus (j, k, l, m). Pairwise alignment of two α-helix in holins from superfamily VI and r1t holin family (n). Pairwise alignment of two α-helix in holins from superfamilies I and VI (o). Pairwise alignment of two α-helix in holins from superfamily I and r1t holin family (p). Source: Authors.

3.4 Taxonomic assignment

3.4.1 Support for the description of 142 species of streptococcal phages

The intergenomic similarity between pairs of the 183 phage genomes that infect *S. thermophilus* and the information on the length ratio and aligned fraction of the genomes calculated by VIRIDIC are presented in SI file 6. The evaluated phages were divided into five clusters, according to the phage groups already described in the literature. Additional divisions into subclusters (internal triangles) in the *Moineauvirus* and *Brussowvirus* genera again highlight the genetic diversity among their members. VIRIDIC integrates the ICTV virus classification criteria based on nucleotide identity thresholds of 95% for species. The similarity data of the genome pairs quantified by VIRIDIC suggest that the 183 phages that infect *S. thermophilus* are organized into 142 species. The two phages of the P738 group were considered two different species (SI file 5 Table S1, Figure S3). The 109 phages of the *Moineauvirus* genus were organized into 86 species (SI file 5 Tables S2 and S3, Figure S4), while the 45 phages of the *Brussowvirus* genus were classified into 35 species (SI file 5 Tables S4 and S5, Figure S5). The 14 phages of the 987 group were distributed in seven species (SI file 5 Tables S6 and S7, Figure S6), and of the 13 phages of the 5093 group, 12 formed different species (SI file 5 Tables S8 and S9, Figure S7). With this study, we allowed the expansion of the number of *S. thermophilus* phage species available in the ICTV database from 10 to 142, including their taxonomic classification.

3.4.2 Support for the description of five genera of streptococcal phages

The ICTV has established 70% nucleotide identity of the entire genome length as the cut-off for genera and the presence of homologous conserved 'signature or core genes' and evaluated using phylogenetics. Based on the proteome-based VICTOR analysis, the division into the five groups is also found in the VICTOR phylogeny (Figure 1). Even though VICTOR has recommended four genera, the light-green-colored one is not monophyletic and was split into two distinct genera. That is, the subtree comprising phages "Streptococcus_phage_SW6__MH892351" to "Streptococcus_phage_SWK1__MH892377" formed the fifth genus.

Each genus identified in the VICTOR analysis was investigated for the presence of core genes. The Roary, VirClust, and CoreGenes analysis results were similar (Table 1, SI file 1 Tables S2-S16). The P738 group phages shared at least 43 homologs proteins (91.49%). These homologs proteins included: small/large terminase subunits, portal protein, minor/major capsid proteins, head scaffolding protein, head-tail adaptor, major tail protein, tape measure protein, host-specificity tail protein, holin, endolysin, DNA helicase, and primase. The phages from this genus have genomes of, on average, 33.85 kb (37.05 %GC), encoding 47 proteins. At the DNA level, these phages share 90.71% DNA similarity (SI file 1 Tables S2-S4, SI file 5 Table S1, SI file 6). The *Moineauvirus* phages shared at least 15 homologs proteins (33.33%). These homologs proteins included: small/large terminase subunits, portal protein, major capsid protein, head scaffolding protein, head-tail connector protein, major tail protein, antireceptor, and holin. The phages from this genus have genomes of, on average, 35.90 kb (38.74 %GC) encoding 45 proteins. At the DNA level, these phages share at least 44.35% DNA similarity (SI file 1 Tables S5-S7, SI file 5 Table S2, SI file 6). The *Brussowvirus* phages shared at least 15 homologs proteins (31.25%). These homologs proteins included: portal protein, major/minor capsid proteins, head scaffolding protein, major tail protein, and tape measure protein. The phages from this genus have genomes of, on average, 37.32 kb (39.34 %GC) encoding 48 proteins. At the DNA level, these phages share at least 37.0% DNA similarity (SI file 1 Tables S8-S10, SI file 5 Table S4, SI file 6). The 987 group phages shared 23 homologs proteins (50%). These homologs proteins included: large terminase subunit, portal protein, minor/major capsid proteins, head-tail connector protein, head scaffolding protein, antireceptor, major tail protein, tape measure protein, holin, endolysin, and DNA binding protein. The phages from this genus have genomes of, on average, 32.26 kb (36.94 %GC) encoding 46 proteins. At the DNA level, these phages share at least 56.14% DNA similarity (SI file 1 Tables S11-S13, SI file 5 Table S6, SI file 6). The 5093 group phages shared at least 22 homologs proteins (46.80%). These homologs proteins included: small/large terminase subunits, portal protein, minor/major capsid proteins, head scaffolding protein, antireceptor, major tail protein, holin, endolysin, and DNA binding protein. The phages from this genus have genomes of, on average, 34.25 kb (38.26 %GC) encoding 47 proteins. At the DNA level, these phages share at least 64.44% DNA similarity (SI file 1 Tables S14-S16, SI file 5 Table S8, SI file 6).

Table 1. Core genes with putative functions identified for each group/genus using the Roary, VirClust, and CoreGenes tools.

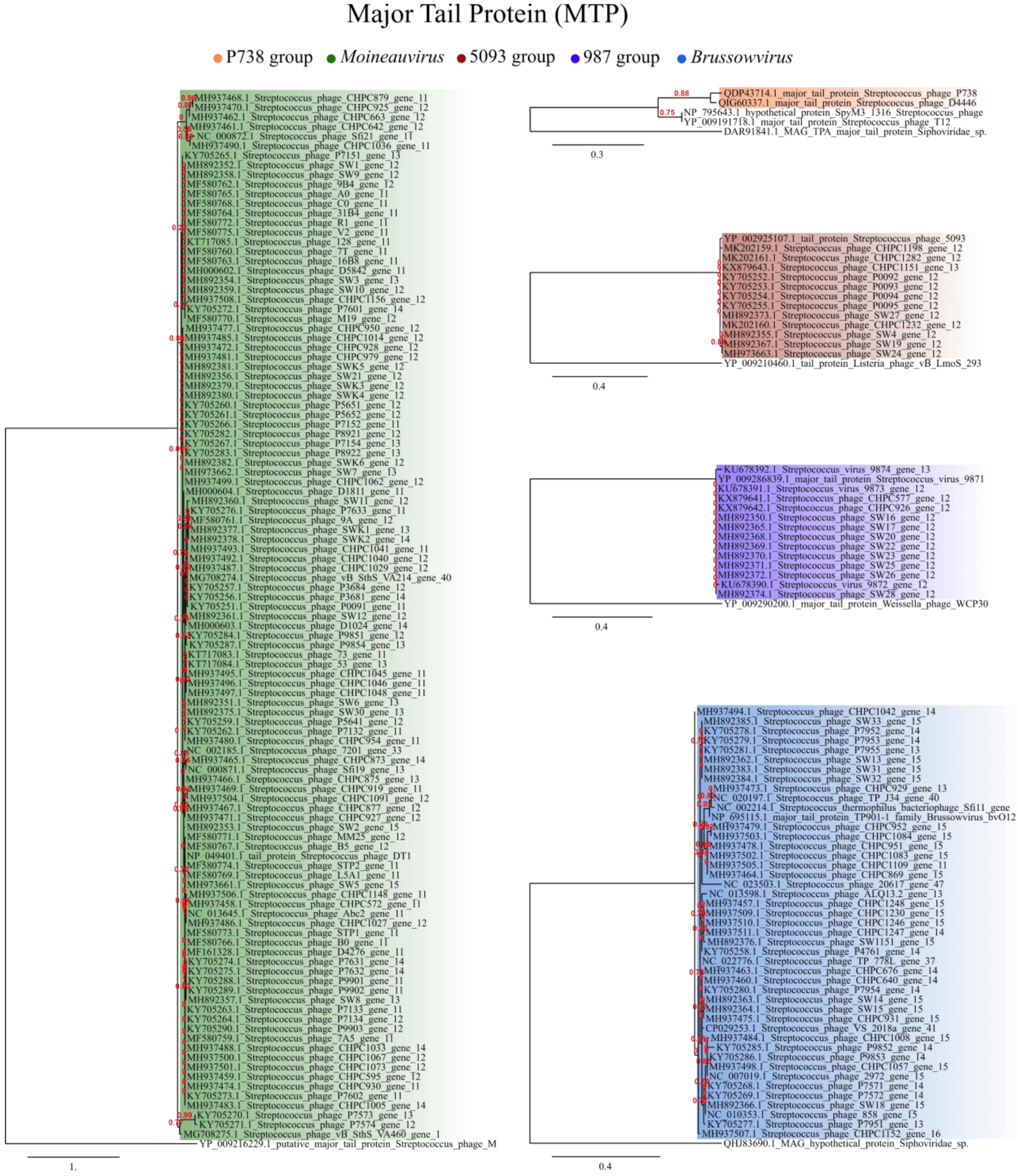
Genus (reference genome)	Roary infomation			VirClust information			CoreGenes information	
	Gene_ID	Protein length	BLAST_annot	Gene_ID	Protein length	PHROGS_annot	ID	annot
	gene_00001	159	terminase small subunit	gene_1	160	terminase small subunit	QDP43702.1	putative terminase small subunit
	gene_00002	425	terminase large subunit	gene_2	426	terminase large subunit	QDP43703.1	putative terminase large subunit
	gene_00003	430	portal protein	gene_3	431	portal protein	QDP43704.1	putative portal protein
	gene_00004	271	head protein	gene_4	272	minor head protein	QDP43705.1	putative head morphogenesis protein
	-	-	-	gene_5	55	no_hit	QDP43706.1	hypothetical protein
	gene_00006	195	scaffold protein	gene_6	196	head scaffolding protein	QDP43707.1	putative scaffold protein
	gene_00007	302	major head protein	gene_7	303	major head protein	QDP43708.1	major head protein
	gene_00008	50	hypothetical protein	gene_8	51	Arc-like repressor	QDP43709.1	hypothetical protein
	gene_00009	108	head-tail connector	gene_9	109	head-tail adaptor	QDP43710.1	putative head-tail connector
	gene_00010	103	hypothetical protein	gene_10	104	no_hit	QDP43711.1	hypothetical protein
	gene_00011	116	tail protein	gene_11	117	neck protein Ne1	QDP43712.1	putative tail protein
	gene_00012	136	hypothetical protein	gene_12	137	tail completion Tc1	QDP43713.1	hypothetical protein
	gene_00013	165	major tail protein	gene_13	166	major tail protein	QDP43714.1	major tail protein
	gene_00014	104	hypothetical protein	gene_14	105	no_hit	QDP43715.1	hypothetical protein
	gene_00015	97	hypothetical protein	gene_15	98	no_hit	QDP43716.1	hypothetical protein
	gene_00016	738	tape-measure protein	gene_16	739	tail length tape measure protein	QDP43717.1	putative tape-measure protein
	gene_00017	773	host-specificity tail protein	gene_17	774	minor tail protein	QDP43718.1	putative host-specificity tail protein
	gene_00018	1323	tail fiber protein	gene_18	1324	tail fiber protein and host specificity	QDP43719.1	putative tail fiber protein
	gene_00019	666	tail protein	gene_19	667	tail protein	QDP43720.1	putative tail protein
	gene_00020	90	hypothetical protein	gene_20	91	no_hit	QDP43721.1	hypothetical protein
	gene_00021	116	hypothetical protein	gene_21	117	no_hit	QDP43722.1	hypothetical protein
P738 (MK911750)	gene_00022	63	holin	gene_22	64	holin	QDP43723.1	hypothetical protein
	gene_00023	248	endolysin	gene_23	249	endolysin	QDP43724.1	putative endolysin
	gene_00024	88	hypothetical protein	gene_24	89	no_hit	QDP43725.1	hypothetical protein
	gene_00025	41	hypothetical protein	gene_25	42	no_hit	QDP43726.1	hypothetical protein
	gene_00026	75	transcriptional regulator	gene_26	76	transcriptional repressor	QDP43727.1	putative transcriptional regulator
	gene_00027	120	hypothetical protein	gene_27	123	nucleotide kinase	QDP43728.1	hypothetical protein
	gene_00031	48	hypothetical protein	gene_31	49	no_hit	QDP43732.1	hypothetical protein
	gene_00033	113	hypothetical protein	gene_33	114	no_hit	QDP43734.1	hypothetical protein
	gene_00034	75	hypothetical protein	gene_34	76	no_hit	QDP43735.1	hypothetical protein
	gene_00035	154	hypothetical protein	gene_35	155	no_hit	QDP43736.1	hypothetical protein
	gene_00036	395	helicase	gene_36	396	DNA helicase	QDP43737.1	putative helicase
	gene_00037	215	recombinase	gene_37	207	Erf-like ssDNA annealing protein	QDP43738.1	putative DNA recombination protein
	gene_00038	140	hypothetical protein	gene_38	141	no_hit	QDP43739.1	hypothetical protein
	gene_00039	112	hypothetical protein	gene_39	113	no_hit	QDP43740.1	hypothetical protein
	gene_00040	274	primase	gene_40	275	DNA polymerase/primase	QDP43741.1	putative replication protein
	gene_00041	449	virulence-associated protein E	gene_41	450	DNA helicase	QDP43742.1	putative virulence-associated protein E
	gene_00042	96	hypothetical protein	gene_42	97	no_hit	QDP43743.1	hypothetical protein
	gene_00043	57	hypothetical protein	gene_43	58	no_hit	QDP43744.1	hypothetical protein
	gene_00044	138	hypothetical protein	gene_44	139	no_hit	QDP43745.1	hypothetical protein
	gene_00045	106	hypothetical protein	gene_45	107	no_hit	QDP43746.1	hypothetical protein
	gene_00046	92	hypothetical protein	gene_46	93	no_hit	QDP43747.1	hypothetical protein
	gene_00047	112	hypothetical protein	gene_47	113	no_hit	QDP43748.1	hypothetical protein
	gene_00048	388	hypothetical protein	gene_48	390	DNA repair exonuclease	QDP43749.1	hypothetical protein
	gene_00002	153	terminase small subunit	gene_2	154	terminase small subunit	NP_049390.1	terminase small subunit P27 family
	gene_00003	229	terminase large subunit	gene_3	230	terminase large subunit	NP_049392.1	putative terminase large subunit
	-	-	-	gene_5	60	no_hit		
	gene_00005	386	portal protein	gene_6	387	portal protein	NP_049394.1	portal protein
	gene_00006	222	scaffolding protein	gene_7	223	head maturation protease		
	gene_00007	397	major capsid protein	gene_8	398	major head protein	NP_049396.1	major head protein
	gene_00008	104	head-tail connector protein	gene_9	105	head-tail adaptor Ad1		
	gene_00009	116	head-tail connector protein	gene_10	117	no_hit	NP_049398.1	head closure protein
	gene_00010	140	tail protein	gene_11	141	no_hit	NP_049399.1	tail protein
	gene_00011	123	tail protein	gene_12	124	tail protein		
Moineauvirus (NC_002072)	gene_00012	203	major tail protein	gene_13	204	major tail protein	NP_049401.1	tail protein
	gene_00013	117	tail chaperone protein	gene_14	118	tail protein	NP_049402.1	tail protein
	gene_00014	1656	tape measure protein	gene_15	1657	minor tail protein	NP_049403.2	putative tail component protein
	-	-	-	gene_16	519	minor tail protein	NP_049405.1	tail family protein
	gene_00016	914	antireceptor	gene_17	915	tail fiber protein and host specificity	NP_049406.1	tail-host specificity protein
	-	-	-	gene_18	686	minor tail protein		
	-	-	-	gene_19	132	no_hit	NP_049409.1	DUF1366 domain-containing protein
	gene_00021	80	holin	gene_22	81	holin	NP_049412.1	holin
	-	-	-	gene_43	133	transcriptional regulator	NP_049433.1	DUF1492 domain-containing protein
	gene_00039	165	DNA binding protein	-	-	-		
	gene_00041	172	HNH endonuclease	-	-	-	NP_049434.1	HNH endonuclease
	gene_00043	235	hypothetical protein	-	-	-		
	gene_00008	82	hypothetical protein	gene_8	83	no_hit		
	-	-	-	gene_9	52	no_hit		
	gene_00012	170	DNA binding protein	gene_12	171	DNA binding protein	YP_002925093.1	DNA binding protein
	gene_00014	97	hypothetical protein	gene_14	98	no_hit		
	gene_00017	235	domain-containing protein	gene_17	236	no_hit	YP_002925095.1	DUF1340 domain-containing protein
	gene_00019	148	terminase small subunit	gene_19	149	terminase small subunit	YP_002925097.1	terminase small subunit
	gene_00020	434	terminase large subunit	gene_20	435	terminase large subunit	YP_002925098.1	PBSX family terminase large subunit
	gene_00021	502	portal protein	gene_21	503	portal protein	YP_002925099.1	portal protein
	gene_00022	407	minor capsid protein	gene_22	408	minor head protein	YP_002925100.1	minor capsid protein
	gene_00023	205	scaffolding protein	gene_23	206	head scaffolding protein	YP_002925101.1	scaffolding protein
	gene_00024	281	major capsid protein	gene_24	282	major head protein	YP_002925102.1	N4-gp56 family major capsid protein
	gene_00025	57	hypothetical protein	gene_25	62	no_hit	YP_002925103.1	hypothetical protein
5093 (NC_012753)	gene_00026	129	head-tail connector protein	gene_26	130	head-tail adaptor	YP_002925104.1	hypothetical protein
	gene_00027	111	minor capsid protein	gene_27	112	head-tail adaptor		
	gene_00028	119	minor capsid protein	gene_28	120	minor head protein	YP_002925105.1	minor capsid protein
	gene_00029	134	minor capsid protein	gene_29	135	minor head protein	YP_002925106.1	minor capsid protein
	gene_00030	167	major tail protein	gene_30	168	major tail protein	YP_002925107.1	tail protein
	gene_00031	120	tail assembly chaperone protein	gene_31	121	no_hit	YP_002925108.1	hypothetical protein
	gene_00032	219	hypothetical protein	gene_32	220	no_hit	YP_002925109.1	Gp15 family bacterioprotein
	gene_00033	1528	tape-measure protein	gene_33	1529	tail protein	YP_002925110.1	putative antireceptor protein
	gene_00034	239	distal tail protein	gene_34	240	no_hit	YP_002925111.1	hypothetical protein
	gene_00035	508	tail-associated lysin	gene_35	509	tail protein	YP_002925112.1	endolysin
	-	-	-	gene_37	395	no_hit	YP_002925114.1	antireceptor
	gene_00039	91	holin	gene_39	92	holin	YP_002925116.1	hypothetical protein
	gene_00040	80	holin	gene_40	81	holin	YP_002925117.1	holin
	-	-	-	gene_41	282	endolysin	YP_002925118.1	peptidoglycan hydrolase
987 (KU678389)	gene_00002	462	terminase large subunit	gene_2	463	terminase large subunit	AMQ65697.1	TerL
	gene_00003	446	portal protein	gene_3	447	portal protein	AMQ65698.1	portal protein

	gene_00004	346	minor capsid protein	gene_4	347	minor head protein	AMQ65699.1	minor capsid protein
	gene_00005	199	domain-containing protein	gene_5	200	head scaffolding protein	AMQ65700.1	scaffolding protein
	gene_00006	287	major capsid protein	gene_6	288	major head protein	AMQ65701.1	major capsid protein
	gene_00007	63	Ig domain containing protein	gene_7	64	structural protein with Ig domain	AMQ65702.1	hypothetical protein
	gene_00008	110	head-tail connector protein	gene_8	111	head-tail adaptor	AMQ65703.1	head-tail connector protein
	gene_00009	103	hypothetical protein	gene_9	104	minor head protein	AMQ65704.1	hypothetical protein
	gene_00010	108	capsid and scaffold protein	gene_10	109	neck protein Ne1	AMQ65705.1	hypothetical protein
	gene_00011	129	capsid protein	gene_11	130	tail completion Tc1	AMQ65706.1	tail terminator protein
	gene_00012	165	major tail protein	gene_12	166	major tail protein	AMQ65707.1	major tail protein
	gene_00013	116	tail chaperone protein	gene_13	117	tail protein	AMQ65708.1	tail chaperone protein
	gene_00014	89	hypothetical protein	gene_14	90	tail assembly chaperone	AMQ65709.1	hypothetical protein
	gene_00015	916	tape measure protein	gene_15	917	no_hit	AMQ65710.1	tape measure protein
	gene_00016	253	tail family protein	gene_16	254	minor tail protein	AMQ65711.1	distal tail protein
	gene_00017	910	tail-associated lysin	gene_17	911	minor head protein	AMQ65712.1	tail-associated lysin
	gene_00018	647	antireceptor	gene_18	648	virion structural protein	AMQ65713.1	antireceptor
	gene_00021	81	holin	gene_21	82	holin	AMQ65716.1	holin
	gene_00022	200	lysin	gene_22	201	endolysin	AMQ65717.1	lysin
	gene_00040	51	hypothetical protein	gene_40	52	no_hit	AMQ65737.1	hypothetical protein
	gene_00042	170	DNA-binding protein	gene_42	171	DNA binding protein	AMQ65739.1	DNA-binding protein
	gene_00044	79	hypothetical protein	gene_44	97	no_hit	AMQ65741.1	hypothetical protein
	gene_00045	235	hypothetical protein	gene_45	236	no_hit	AMQ65742.1	hypothetical protein
	-	-	-	gene_20	182	DNA binding protein	NP_695098.1	hypothetical protein
	gene_00022	236	hypothetical protein	gene_22	237	no_hit	NP_695101.1	hypothetical protein
	gene_00025	107	terminase small subunit	gene_25	108	no_hit		
	-	-	-	gene_26	412	terminase large subunit	NP_695104.1	PBSX family terminase large subunit
	gene_00027	501	portal protein	gene_27	503	portal protein	NP_695105.1	portal protein
	gene_00028	297	minor capsid protein	gene_28	298	minor head protein	NP_695106.1	minor capsid protein
	gene_00029	193	capsid and scaffold protein	gene_29	194	head scaffolding protein	NP_695107.1	DUF4355 domain-containing protein
	gene_00030	119	major capsid protein	gene_30	120	major head protein	NP_695108.1	putative structural protein
Brussowvirus	gene_00035	114	hypothetical protein	gene_35	115	neck protein Ne1	NP_695113.1	HK97 gp10 family protein
(NC_004303)	gene_00036	128	hypothetical protein	gene_36	129	tail completion Tc1	NP_695114.1	DUF3168 domain-containing protein
	gene_00037	168	major tail protein	gene_37	169	major tail protein	NP_695115.1	major tail protein
	gene_00038	117	tail chaperone protein	gene_38	118	tail protein	NP_695116.1	tail assembly chaperone
	gene_00039	105	hypothetical protein	gene_39	106	tail assembly chaperone	NP_695117.1	hypothetical protein
	gene_00040	1517	tape measure protein	gene_40	1518	tail protein	NP_695118.1	putative tail protein
	gene_00041	512	tail protein	gene_41	513	minor tail protein	NP_695119.1	tail family protein
	-	-	-	gene_45	844	host range and adsorption protein	NP_695120.1	putative tail protein
	gene_00046	654	tail protein	gene_46	670	tail protein	NP_695124.1	hypothetical protein
	gene_00047	117	hypothetical protein	gene_47	118	no_hit	NP_695125.1	hypothetical protein

Source: Authors.

The genes encoding the major tail protein (MTP) and portal protein were detected and annotated in the five genera by all tools (Roary, VirClust, and CoreGenes) and, therefore, were used to visualize the clades of genera in the phylogenetic analysis. All genes encoding the MTP and Portal proteins produced a phylogenetic tree in which the genus was represented by a well-supported clade (Figure 4, SI file 7). Based on the results of the phage genomes clustering analyses, detection, and phylogeny of the core genes, the *S. thermophilus* phages were classified at the genus level according to the criteria described below. The criteria for demarcating a new genus for the phages of the P738 group were: 70% DNA sequence identity and monophyly in the terminase large subunit (TerL), portal, and holin proteins phylogenetic trees (SI file 5 Figures S3 and S8). The criteria for demarcating the *Moineauvirus* genus were: at least 44.35% DNA sequence identity and monophyly in the terminase small subunit (TerS), major tail protein (MTP), and portal proteins phylogenetic trees (SI file 5 Figures S4 and S9). The criteria for demarcating the *Brussowvirus* genus were: at least 37.0% DNA sequence identity and monophyly in the major capsid protein (MCP), portal protein, and major tail protein (MTP) proteins phylogenetic trees (SI file 5 Figures S5 and S10). The criteria for demarcating a new genus for the phages of the 987 group were: at least 56.3% DNA sequence identity and monophyly in the major capsid protein (MCP), portal protein, and major tail protein (MTP) proteins phylogenetic trees (SI file 5 Figures S6 and S11). The criteria for demarcating a new genus for the phages of the 5093 group were: at least 64.44% DNA sequence identity and monophyly in the terminase large subunit (TerL), major capsid protein (MCP), and major tail protein (MTP) proteins phylogenetic trees (SI file 5 Figures S7 and S12).

Figure 4. Phylogenetic trees constructed for each genus identified in the VICTOR analysis of the *Streptococcus thermophilus* phages.



The trees were constructed using the Major Tail Protein (MTP) amino acid sequences. Source: Authors.

The similarity data of the genome pairs quantified by VIRIDIC suggested that the 183 phages that infect *S. thermophilus* are organized into 42 possible genera (Figure 1, SI file 6). Phages from the *Moineauvirus* and *Brussowvirus* were distributed in 23 and 12 genera, respectively, while the phages from groups 5093 and 987 were reorganized into three genera each. However, the VIRIDIC analysis was based only on nucleotide identity thresholds of 70% for genus level demarcation. The presence of a set of conserved genes and monophyly in the signature genes, added to the morphological, genomic, and ecological characteristics shared by all the phages in the same group, allow us to classify the five groups described in the literature as being in fact, five genera of streptococcal phages. With these results, we allowed the expansion of the number of genera of *S.*

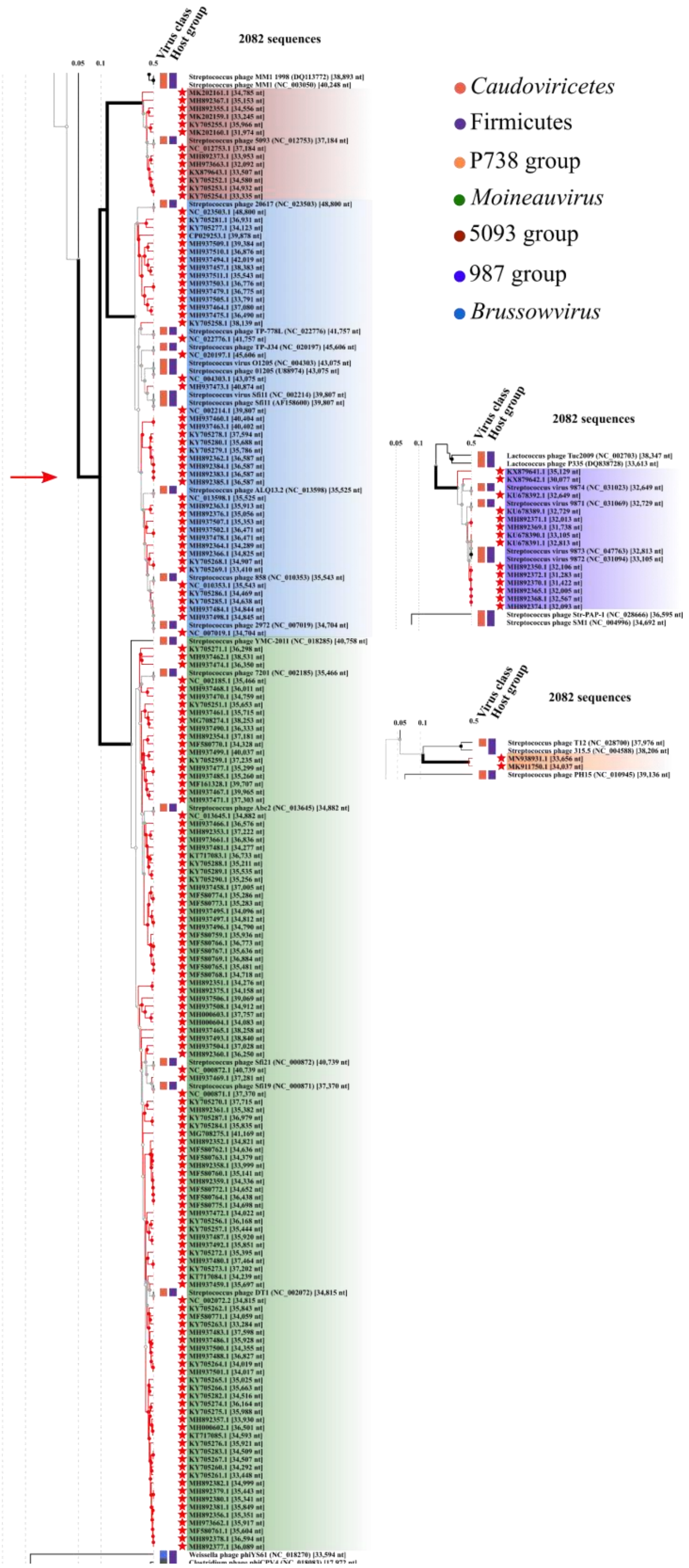
thermophilus phages taxonomically classified by ICTV from two to five, adding three new genera within the *Caudoviricetes* class.

3.4.3 Family level clustering of streptococcal phages

ICTV is currently defining criteria for the distinction of new families and orders. However, it has been proposed that the family-level represents a cohesive and monophyletic group in the main predicted proteome-based clustering tools, whose members share a significant number of orthologous genes. The amino acid-based VICTOR analysis showed that *S. thermophilus* phages are organized into one subfamily and one family (Figure 1). However, VICTOR analysis does not identify the orthologous proteins that contributed to the clustering of phages at subfamily and family levels.

It was described that a 0.90 intergenomic distance threshold applied to the PC tree in VirClust delineated most families within the *Duplodnaviria* realm (Moraru, 2021). Here, different distance thresholds (0.98, 0.90, and 0.85) were applied to the hierarchical trees produced by VirClust, to determine if one of them is suitable for defining family level clusters. When a 0.98 intergenomic distance threshold was applied to the PC tree in VirClust, the 183 phage genomes clustered into a single VGC (SI file 5 Figure S13). However, the core proteins shared by the phages that defined the VGC were not identified, that is, no PC was formed, generating uncertainty regarding the clustering for demarcation of the family level. Potentially, 0.98 could be used for order level delineation. When protein clustering was performed with the default parameters of VirClust (0.90 intergenomic distance threshold), the resulting tree was split into three VGCs based on PCs (SI file 5 Figure S14). The core proteins defining each VGC were identified and annotated (SI file 1 Tables S17-S19). Phage genomes from P738 and *Moineauvirus* genera formed one family each, sharing 44 and 19 PCs, respectively. While phage genomes from *Brussowvirus*, 987 and 5093 genera formed a third family, sharing two PCs. One of these shared PCs had its putative function identified and corresponded to the gene encoding DNA binding protein. However, as only a small proportion of proteins were shared with other phages in the dataset, it increases clustering uncertainty and can indicate incorrect clustering (Moraru, 2021). Identical results were obtained with CoreGenes (SI file 1 Table S20). Roary detected a single core gene among the phages of these three genera, encoding a hypothetical protein, also detected by the other tools (SI file 1 Table S21). A total of 107 shell genes (15% <= strains < 95%) were also found among the phages of these genera. A threshold of 0.85 intergenomic distance applied on the hierarchical tree split the phage genomes into four VGCs (SI file 5 Figure S15). Phage genomes from P738, *Moineauvirus*, and 5093 genera formed one family each, sharing 44, 19, and 25 PCs, respectively (SI file Tables S22-S24). Phage from *Brussowvirus* and 987 genera formed a fourth family, sharing seven PCs (SI file 1 Table S25). Shared PCs by the phages of the *Brussowvirus* and 987 genera include the DNA binding protein, portal protein, minor head protein, connector, major tail protein, tail protein, and other hypothetical proteins. Similar results were obtained with CoreGenes (SI file 1 Table S26), which identified eight orthologous proteins shared by these phage genera. Roary detected two core genes among the phages of these genera, encoding a hypothetical protein and major tail protein, also detected by the other tools (SI file 1 Table S27). A total of 93 shell genes (15% <= strains < 95%) were also found among the phages of these genera.

Figure 5. Parts of the proteomic tree constructed by ViPTree.



The tree includes 2082 related sequences with the 183 *Streptococcus thermophilus phages*. The phages of interest in this study are indicated with red stars. The red arrow indicates the possible family-level clustering for phages of Moineauvirus, Brussowvirus, and 5093 genera. The class of the related phages and the phylum of bacterial hosts are also annotated. The colored boxes delineate the genera. The complete phylogenetic tree created by ViPTree can be seen in SI file 8. Source: Authors.

Similar clades were obtained using the main predicted proteome-based clustering tools, confirming that *S. thermophilus* phages are grouped separately into five genera in the hierarchical clustering. However, these clades were assigned to a different number of families by the VICTOR (a single family) and VirClust analysis (one, three, or four families, according to the intergenomic distance threshold). Phylogenetic analysis of phage proteomes using ViPTree revealed that three of the five genera of phages are more closely related to each other than to any other phages (Figure 5). The phages from *Moineauvirus*, *Brussowvirus*, and 5093 genera formed a cohesive and monophyletic group and appeared to form a family separately from the P738 and 987 genera phages. Orthologous proteins shared by the phages of these three genera were identified using Roary. A single core gene (gene encoding hypothetical protein) was included (SI file 1 Table S28). A total of 79 shell genes (15% <= strains < 95%) and a single soft core gene (95% <= strains < 99%) were also found among the phages of these genera. On the other hand, the next closest relatives of P738 and 987 genera were phages that infect the pathogenic bacteria *Streptococcus pyogenes* (Streptococcus phage T12, NC_028700 and Streptococcus prophage 315.5, NC_004588) and the lactic acid bacterium *Lactococcus lactis* (Lactococcus phage Tuc2009, NC_002703 and Lactococcus phage P335, DQ838728), respectively. *Moineauvirus* and *Brussowvirus* genera shared five orthologous proteins using CoreGenes (SI file 1 Tables S29) and two core genes detected with Roary, encoding hypothetical proteins (SI file 1 Tables S30). Phages from these genera also shared 70 shell genes (15% <= strains < 95%) and two soft core genes (95% <= strains < 99%).

4. Discussion

Phages have complex evolutionary relationships (Dion et al., 2020; Koonin et al., 2020), and several strategies have been used in an attempt to classify the phages that infect *S. thermophilus*. Historically, the classification of streptococcal phages was based on morphology and the combination of the DNA packaging mode and the number of major structural proteins (Desiere et al., 1999; Le Marrec et al., 1997; Lucchini et al., 1999a, 1999b). This classification limited the division of phages that infect *S. thermophilus* into two groups: *cos*-containing phages, whose phages had cohesive genomic ends when treated with restriction enzymes (*Moineauvirus* genus) and two main structural proteins; and *pac*-containing phages, whose phages showed absence of cohesive ends (*Brussowvirus* genus) and three main structural proteins (Le Marrec et al., 1997). The discovery of phages with new properties resulted in the need to update this classification, which led to the proposition of two new groups, 5093 (Mills et al., 2011) and 987 (McDonnell et al., 2016). Although group 5093 phages have a genomic architecture similar to members of the *Moineauvirus* genus, with two main structural proteins, their genome sequences are more similar to those of phages from the *Brussowvirus* genus and a streptococcal prophage of non-dairy origin (Mills et al., 2011). The genomes of group 987 phages, on the other hand, show genetic exchange events, exhibiting a DNA sequence relationship with the morphogenesis modules of certain phages that infect *L. lactis* of the P335 group and with the replication modules of phages that infect *S. thermophilus* (McDonnell et al., 2016). A fifth group, P738, formed by two genetically distinct phages from the other phages that infect *S. thermophilus* was described. These phages are closely related to each other and share similarities with non-dairy streptococcal phages (Philippe et al., 2020).

Changes in virus classification, taxonomy, and nomenclature occur annually (Walker et al., 2021) as a result of an increasing increase in sequenced genomes and the rise of metagenomic studies (Al-Shayeb et al., 2020; Dion et al., 2020; Simmonds et al., 2017; Turner et al., 2021). To accommodate the full spectrum of virus genetic divergence, ICTV changed the hierarchical taxonomic classification structure to 15 positions, including eight major classifications (realm, kingdom, phylum, class, order, family, genus, and species) and seven classifications derived from the main (ICTV, 2020). At the realm level, viruses were grouped into *Adnaviria*, *Duplodnaviria*, *Monodnaviria*, *Riboviria*, *Ribozyviria,* and *Varidnaviria* (Walker et al., 2021). For tailed phages, it was proposed to eliminate the order *Caudovirales* and the families *Myoviridae*, *Siphoviridae,* and *Podoviridae* (Turner et al., 2021). Without these taxonomic levels, the *Caudoviricetes* class automatically integrated all tailed phages, and ICTV is currently defining the criteria for the distinction of new families and orders. However, it has been proposed that they should be based on analyzes of entire viral proteomes and consider shared orthologous proteins (Koonin et al., 2020; Turner et al., 2021).

In this study, we used a wide range of complementary methods, including comparative genomics, core genome analysis, and signature genes phylogenetics, to show that the *S. thermophilus* phages are organized into 142 species and five genera and that due to their genetic diversity, the classification at family level varies according to the classification criteria used. In the process of the taxon evaluation, we explored the feasibility of different demarcation criteria and critically evaluated the usefulness of our methods for phage classification. The convergence of results, drawing a consistent and comprehensive picture of five well-supported clades, regardless of method, demonstrates that the tools applied here are particularly useful in *S. thermophilus* phages taxonomy at species and genus levels. At the family level classification, the *S. thermophilus* phages from *Brussowvirus*, 987 and 5093 genera formed a group sharing at least one ortholog gene, or yet, *Brussowvirus* and 987 genera formed a group sharing at least two orthologous genes. These groupings corroborate a recent study that suggested the evolution/emergence of 5093 and 987 phages through recombination with temperate *Brussowviruses*, based on phylogenetic analyzes of streptococcal phage replication modules (Hanemaaijer et al., 2021). On the other hand, the phages from *Moineauvirus*, *Brussowvirus*, and 5093 genera formed a monophyletic group in the viral proteomic tree sharing at least one ortholog gene. The relationship among the phages of these three genera has been described previously (Mills et al., 2011). It was reported that the first phage of the genus 5093 (Streptococcus phage 5093, NC_012753) evolved from gene exchanges with *pac*-containing phages (*Brussowvirus* genus), representing a hybrid phage and that although has a genomic architecture similar to members of the *Moineauvirus* genus, their genome sequences are more similar to those of phages from the *Brussowvirus* genus.

Five well-supported clades were obtained using the main classification tools, and the convergence of their results for the analyzed taxa, confirms that *S. thermophilus* phages are grouped separately into five genera in the hierarchical clustering, whose members do not share a sufficiently conserved gene. However, the genes encoding holin protein are conserved in more than 95% of genomes. Holin superfamily I, mainly, is widespread among LAB (Fujimoto et al., 2020). (Mills et al., 2011b) also observed that endolysin and holin genes are highly conserved in *S. thermophilus* phages, mainly holin superfamily I. Our study highlights two other holin families in *S. thermophilus* phages, the rIt family and holin superfamily VI. (Labrie et al., 2004) were the authors that first described the r1t holin family in *Lactococcus lactis* phages. Finally, the authors observed that the lysis cassette of the phage r1t has the same general lysis module features as other *Lactococcus lactis* phages.

Comparative genomic analyzes have shown that the genomes of these phages exhibit a high degree of conservation within the genera defined here, particularly in the structural modules (Lavelle et al., 2018a; McDonnell et al., 2017; Philippe et al., 2020; Szymczak et al., 2019a), but also exhibit nucleotide divergence between genera and between phage within the same genus. This particularity in genomes of sharing regions of high sequence similarity with abrupt transitions into adjacent regions without detectable similarity is described as genetic mosaicism (Dion et al., 2020; Hendrix et al., 1999). These regions (genes and gene blocks) have distinct evolutionary histories due to multiple genetic exchange events that suffer in response to the selective pressure of their hosts, which drives their diversity (Dion et al., 2020). Streptococcal phage genomes are undergoing a process of evolution through, as studies indicate, horizontal gene transfer, insertions, deletions, point mutations, and exchange of genetic modules (Brussow and Desiere, 2001; Desiere et al., 1998; Lavelle et al., 2018a, 2018b; Lucchini et al., 1999a; Philippe et al., 2020; Szymczak et al., 2017). From an industrial point of view, this diversification represents a constant threat to milk fermentation processes, as it can result in the extension of the host range and make it difficult to efficiently control phage infections in starter cultures, which generally consist of mixtures of lactic acid bacteria (LAB) (de Melo et al., 2018; Leroy and De Vuyst, 2004; McDonnell et al., 2016; Quiberoni et al., 2010; Szymczak et al., 2017). As the 183 phages evaluated were mainly isolated from industrial fermentation environments, industrialization and the use of commercial starter cultures have likely limited the diversity of bacterial strains in dairy environments, resulting in a high gene flow capable of generating phage recombinants with potentially expanded lytic activity.

5. Conclusion

The phages that infect *S. thermophilus* have a remarkable genetic diversity that can result in the extension of their host range, which poses a threat to milk fermentation processes. Considering all the advantages and limitations of the classification tools used here and the convergence of their results for the analyzed taxa, our study contributes to expanding knowledge about the genetic diversity and evolution of streptococcal phages. Our work also highlights the importance of monitoring, isolating, and sequencing streptococcal phages in industrial environments to promote control strategies and minimize failures in milk fermentation processes. The sets of core genes of the five genera described in this study are promissory targets for developing future strategies for biocontrol of *S. thermophilus* phages.

Supplementary Information

All supplementary files are temporally available here: https://figshare.com/s/f3e0595904c2fdd0cfdb; and will be publish under the following: https://doi.org/10.6084/m9.figshare.20359347

Acknowledgments

This research was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant ID 409972/2018-5). Laís S. Batalha received a PhD Sandwich Scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). We are grateful to the Núcleo de Análise de Biomoléculas (NuBioMol) of the Universidade Federal de Viçosa (UFV) for providing the facilities for the conduction of the experiments and data analysis. NuBioMol is financially supported by the following Brazilian agencies: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos (Finep) and Sistema Nacional de Laboratórios em Nanotecnologias (SisNANO)/Ministério da Ciência, Tecnologia e Informação (MCTI).

References

Accolas, J.-P., & Spillmann, H. (1979). The Morphology of Six Bacteriophages of *Streptococcus thermophilus*. *Journal of Applied Bacteriology*, *47*(1), 135–144. https://doi.org/10.1111/j.1365-2672.1979.tb01177.x

Achigar, R., Magadán, A. H., Tremblay, D. M., Julia Pianzzola, M., & Moineau, S. (2017). Phage-host interactions in Streptococcus thermophilus: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. *Scientific Reports*, *7*(1), 43438. https://doi.org/10.1038/srep43438

Ali, Y., Koberg, S., HeÃŸner, S., Sun, X., Rabe, B., Back, A., Neve, H., & Heller, K. J. (2014). Temperate Streptococcus thermophilus phages expressing superinfection exclusion proteins of the Ltp type. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00098

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., … Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems. *Nature*, *578*(7795), 425–431. https://doi.org/10.1038/s41586-020-2007-4

Arioli, S., Eraclio, G., Della Scala, G., Neri, E., Colombo, S., Scaloni, A., Fortina, M. G., & Mora, D. (2018). Role of Temperate Bacteriophage ϕ20617 on Streptococcus thermophilus DSM 20617T Autolysis and Biology. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.02719

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

Binetti, A. G., Del Río, B., Martín, M. C., & Álvarez, M. A. (2005). Detection and Characterization of Streptococcus thermophilus Bacteriophages by Use of the Antireceptor Gene Sequence. *Applied and Environmental Microbiology*, *71*(10), 6096–6103. https://doi.org/10.1128/AEM.71.10.6096-6103.2005

Brussow, H., & Desiere, F. (2001). Comparative phage genomics and the evolution of Siphoviridae: Insights from dairy phages. *Molecular Microbiology*, *39*(2), 213–223. https://doi.org/10.1046/j.1365-2958.2001.02228.x

da Silva Duarte, V., Giaretta, S., Treu, L., Campanaro, S., Pereira Vidigal, P. M., Tarrah, A., Giacomini, A., & Corich, V. (2018). Draft Genome Sequences of Three Virulent *Streptococcus thermophilus* Bacteriophages Isolated from the Dairy Environment in the Veneto Region of Italy. *Genome Announcements*, *6*(10), e00045-18, /ga/6/10/e00045-18.atom. https://doi.org/10.1128/genomeA.00045-18

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

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., & Gascuel, O. (2008). Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, *36*(Web Server issue), W465-469. https://doi.org/10.1093/nar/gkn180

Desiere, F., Lucchini, S., & Brüssow, H. (1998). Evolution ofStreptococcus thermophilusBacteriophage Genomes by Modular Exchanges Followed by Point Mutations and Small Deletions and Insertions. *Virology*, *241*(2), 345–356. https://doi.org/10.1006/viro.1997.8959

Desiere, F., Lucchini, S., & Brüssow, H. (1999). Comparative Sequence Analysis of the DNA Packaging, Head, and Tail Morphogenesis Modules in the Temperate cos-Site Streptococcus thermophilus Bacteriophage Sfi21. *Virology*, *260*(2), 244–253. https://doi.org/10.1006/viro.1999.9830

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., & Moineau, S. (2008). Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. *Journal of Bacteriology*, *190*(4), 1390–1400. https://doi.org/10.1128/JB.01412-07

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology*, *18*(3), 125–138. https://doi.org/10.1038/s41579-019-0311-5

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Farris, J. S. (1972). Estimating Phylogenetic Trees from Distance Matrices. *The American Naturalist*, *106*(951), 645–668. https://www.jstor.org/stable/2459725

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl_2), W29–W37. https://doi.org/10.1093/nar/gkr367

Fujimoto, K., Kimura, Y., Shimohigoshi, M., Satoh, T., Sato, S., Tremmel, G., Uematsu, M., Kawaguchi, Y., Usui, Y., Nakano, Y., Hayashi, T., Kashima, K., Yuki, Y., Yamaguchi, K., Furukawa, Y., Kakuta, M., Akiyama, Y., Yamaguchi, R., Crowe, S. E., … Uematsu, S. (2020). Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts. *Cell Host & Microbe*, *28*(3), 380-389.e9. https://doi.org/10.1016/j.chom.2020.06.005

Gilchrist, C. L. M., & Chooi, Y.-H. (2020). clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *BioRxiv*, 2020.11.08.370650. https://doi.org/10.1101/2020.11.08.370650

Göker, M., García-Blázquez, G., Voglmayr, H., Tellería, M. T., & Martín, M. P. (2009). Molecular Taxonomy of Phytopathogenic Fungi: A Case Study in Peronospora. *PLOS ONE*, *4*(7), e6319. https://doi.org/10.1371/journal.pone.0006319

Gontijo, M. T. P., Teles, M. P., Vidigal, P. M. P., & Brocchi, M. (2022). Expanding the Database of Signal-Anchor-Release Domain Endolysins Through Metagenomics. *Probiotics and Antimicrobial Proteins*. https://doi.org/10.1007/s12602-022-09948-y

Guglielmotti, D. M., Deveau, H., Binetti, A. G., Reinheimer, J. A., Moineau, S., & Quiberoni, A. (2009). Genome analysis of two virulent Streptococcus thermophilus phages isolated in Argentina. *International Journal of Food Microbiology*, *136*(1), 101–109. https://doi.org/10.1016/j.ijfoodmicro.2009.09.005

Hanemaaijer, L., Kelleher, P., Neve, H., Franz, C. M. A. P., de Waal, P. P., van Peij, N. N. M. E., van Sinderen, D., & Mahony, J. (2021). Biodiversity of Phages Infecting the Dairy Bacterium Streptococcus thermophilus. *Microorganisms*, *9*(9), 1822. https://doi.org/10.3390/microorganisms9091822

Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E., & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(5), 2192–2197. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26759/

Hockenberry, A. J., & Wilke, C. O. (2021). BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains. *PeerJ*, *9*, e11396. https://doi.org/10.7717/peerj.11396

Hynes, A. P., Rousseau, G. M., Agudelo, D., Goulet, A., Amigues, B., Loehr, J., Romero, D. A., Fremaux, C., Horvath, P., Doyon, Y., Cambillau, C., & Moineau, S. (2018). Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. *Nature Communications*, *9*(1), 2919. https://doi.org/10.1038/s41467-018-05092-w

ICTV. (2020). The new scope of virus taxonomy: Partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, *5*(5), 668–674. https://doi.org/10.1038/s41564-020-0709-x

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2), e00061-19, /mmbr/84/2/MMBR.00061-19.atom. https://doi.org/10.1128/MMBR.00061-19

Labrie, S., Vukov, N., Loessner, M. J., & Moineau, S. (2004). Distribution and composition of the lysis cassette of Lactococcus lactis phages and functional analysis of bacteriophage ul36 holin. *FEMS Microbiology Letters*, *233*(1), 37–43. https://doi.org/10.1016/j.femsle.2004.01.038

Lavelle, K., Martinez, I., Neve, H., Lugli, G., Franz, C., Ventura, M., Bello, F., Sinderen, D., & Mahony, J. (2018). Biodiversity of Streptococcus thermophilus Phages in Global Dairy Fermentations. *Viruses*, *10*(10), 577. https://doi.org/10.3390/v10100577

Lavelle, K., Murphy, J., Fitzgerald, B., Lugli, G. A., Zomer, A., Neve, H., Ventura, M., Franz, C. M., Cambillau, C., van Sinderen, D., & Mahony, J. (2018). A Decade of *Streptococcus thermophilus* Phage Evolution in an Irish Dairy Plant. *Applied and Environmental Microbiology*, *84*(10), e02855-17, /aem/84/10/e02855-17.atom. https://doi.org/10.1128/AEM.02855-17

Le Marrec, C., van Sinderen, D., Walsh, L., Stanley, E., Vlegels, E., Moineau, S., Heinze, P., Fitzgerald, G., & Fayard, B. (1997). Two groups of bacteriophages infecting Streptococcus thermophilus can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. *Applied and Environmental Microbiology*, *63*(8), 3246–3253. https://doi.org/10.1128/AEM.63.8.3246-3253.1997

Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, *32*(10), 2798–2800. https://doi.org/10.1093/molbev/msv150

Leroy, F., & De Vuyst, L. (2004). Lactic acid bacteria as functional starter cultures for the food fermentation industry. *Trends in Food Science & Technology*, *15*(2), 67–78. https://doi.org/10.1016/j.tifs.2003.09.004

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Levesque, C., Duplessis, M., Labonte, J., Labrie, S., Fremaux, C., Tremblay, D., & Moineau, S. (2005). Genomic Organization and Molecular Analysis of Virulent Bacteriophage 2972 Infecting an Exopolysaccharide-Producing Streptococcus thermophilus Strain. *APPL. ENVIRON. MICROBIOL.*, *71*, 12.

Li, Z., Jaroszewski, L., Iyer, M., Sedova, M., & Godzik, A. (2020). FATCAT 2.0: Towards a better understanding of the structural diversity of proteins. *Nucleic Acids Research*, *48*(W1), W60–W64. https://doi.org/10.1093/nar/gkaa443

Lucchini, S., Desiere, F., & Brüssow, H. (1998). The Structural Gene Module inStreptococcus thermophilusBacteriophage φSfi11 Shows a Hierarchy of Relatedness to Siphoviridae from a Wide Range of Bacterial Hosts. *Virology*, *246*(1), 63–73. https://doi.org/10.1006/viro.1998.9190

Lucchini, S., Desiere, F., & Brüssow, H. (1999a). The Genetic Relationship between Virulent and Temperate Streptococcus thermophilus Bacteriophages: Whole Genome Comparison of cos-Site Phages Sfi19 and Sfi21. *Virology*, *260*(2), 232–243. https://doi.org/10.1006/viro.1999.9814

Lucchini, S., Desiere, F., & Brüssow, H. (1999b). Comparative Genomics of Streptococcus thermophilus Phage Species Supports a Modular Evolution Theory. *Journal of Virology*, *73*(10), 8647–8656. https://doi.org/10.1128/JVI.73.10.8647-8656.1999

Mahony, J., Casey, E., & van Sinderen, D. (2020). The Impact and Applications of Phages in the Food Industry and Agriculture. *Viruses*, *12*(2), 210. https://doi.org/10.3390/v12020210

Mahony, J., & van Sinderen, D. (2014). Current taxonomy of phages infecting lactic acid bacteria. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00007

McDonnell, B., Mahony, J., Hanemaaijer, L., Neve, H., Noben, J.-P., Lugli, G. A., Ventura, M., Kouwen, T. R., & van Sinderen, D. (2017a). Global Survey and Genome Exploration of Bacteriophages Infecting the Lactic Acid Bacterium Streptococcus thermophilus. *Frontiers in Microbiology*, *8*, 1754. https://doi.org/10.3389/fmicb.2017.01754

McDonnell, B., Mahony, J., Hanemaaijer, L., Neve, H., Noben, J.-P., Lugli, G. A., Ventura, M., Kouwen, T. R., & van Sinderen, D. (2017b). Global Survey and Genome Exploration of Bacteriophages Infecting the Lactic Acid Bacterium Streptococcus thermophilus. *Frontiers in Microbiology*, *8*, 1754. https://doi.org/10.3389/fmicb.2017.01754

McDonnell, B., Mahony, J., Neve, H., Hanemaaijer, L., Noben, J.-P., Kouwen, T., & Sinderen, D. van. (2016). Identification and Analysis of a Novel Group of Bacteriophages Infecting the Lactic Acid Bacterium Streptococcus thermophilus. *Applied and Environmental Microbiology*, *82*(17), 5153–5165. https://doi.org/10.1128/AEM.00835-16

McDonnell, B., Mahony, J., Neve, H., Hanemaaijer, L., Noben, J.-P., Kouwen, T., & van Sinderen, D. (2016). Identification and Analysis of a Novel Group of Bacteriophages Infecting the Lactic Acid Bacterium Streptococcus thermophilus. *Applied and Environmental Microbiology*, *82*(17), 5153–5165. https://doi.org/10.1128/AEM.00835-16

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, *14*(1), 60. https://doi.org/10.1186/1471-2105-14-60

Meier-Kolthoff, J. P., & Göker, M. (2017). VICTOR: Genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*, *33*(21), 3396–3404. https://doi.org/10.1093/bioinformatics/btx440

Mills, S., Griffin, C., O'Sullivan, O., Coffey, A., McAuliffe, O. E., Meijer, W. C., Serrano, L. M., & Ross, R. P. (2011a). A new phage on the 'Mozzarella' block: Bacteriophage 5093 shares a low level of homology with other Streptococcus thermophilus phages. *International Dairy Journal*, *21*(12), 963–969. https://doi.org/10.1016/j.idairyj.2011.06.003

Mills, S., Griffin, C., O'Sullivan, O., Coffey, A., McAuliffe, O. E., Meijer, W. C., Serrano, L. M., & Ross, R. P. (2011b). A new phage on the 'Mozzarella' block: Bacteriophage 5093 shares a low level of homology with other Streptococcus thermophilus phages. *International Dairy Journal*, *21*(12), 963–969. https://doi.org/10.1016/j.idairyj.2011.06.003

Moraru, C. (2021). VirClust – a tool for hierarchical clustering, core gene detection and annotation of (prokaryotic) viruses. *BioRxiv*, 2021.06.14.448304. https://doi.org/10.1101/2021.06.14.448304

Moraru, C., Varsani, A., & Kropinski, A. M. (2020). VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses*, *12*(11), 1268. https://doi.org/10.3390/v12111268

Neve, H., Zenz, K. I., Desiere, F., Koch, A., Heller, K. J., & Brüssow, H. (1998). Comparison of the Lysogeny Modules from the TemperateStreptococcus thermophilusBacteriophages TP-J34 and Sfi21: Implications for the Modular Theory of Phage Evolution. *Virology*, *241*(1), 61–72. https://doi.org/10.1006/viro.1997.8960

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., & Goto, S. (2017). ViPTree: The viral proteomic tree server. *Bioinformatics*, *33*(15), 2379–2380. https://doi.org/10.1093/bioinformatics/btx157

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Philippe, C., Levesque, S., Dion, M. B., Tremblay, D. M., Horvath, P., Lüth, N., Cambillau, C., Franz, C., Neve, H., Fremaux, C., Heller, K. J., & Moineau, S. (2020). Novel Genus of Phages Infecting *Streptococcus thermophilus*: Genomic and Morphological Characterization. *Applied and Environmental Microbiology*, *86*(13), e00227-20, /aem/86/13/AEM.00227-20.atom. https://doi.org/10.1128/AEM.00227-20

Pujato, S. A., Quiberoni, A., & Mercanti, D. J. (2019). Bacteriophages on dairy foods. *Journal of Applied Microbiology*, *126*(1), 14–30. https://doi.org/10.1111/jam.14062

Quiberoni, A., Moineau, S., Rousseau, G. M., Reinheimer, J., & Ackermann, H.-W. (2010). Streptococcus thermophilus bacteriophages. *International Dairy Journal*, *20*(10), 657–664. https://doi.org/10.1016/j.idairyj.2010.03.012

Rambaut, A. (2012). *FigTree*. http://tree.bio.ed.ac.uk/software/figtree/

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koča, J., & Rose, A. S. (2021). Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, *49*(W1), W431–W437. https://doi.org/10.1093/nar/gkab314

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., … Zerbini, F. M. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168. https://doi.org/10.1038/nrmicro.2016.177

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., Frey, J. E., & Ahrens, C. H. (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiology*, *19*(1), 143. https://doi.org/10.1186/s12866-019-1500-0

Stanley, E., Fitzgerald, G., Le Marrec, C., Fayard, B., & Van Sinderen, D. (1997). Sequence analysis and characterization of ∅O1205, a temperate bacteriophage infecting Streptococcus thermophilus CNRZ1205. *Microbiology (Reading, England)*, *143 (Pt 11)*, 3417–3429. https://doi.org/10.1099/00221287-143-11-3417

Szymczak, P., Janzen, T., Neves, A. R., Kot, W., Hansen, L. H., Lametsch, R., Neve, H., Franz, C. M. A. P., & Vogensen, F. K. (2017). Novel Variants of Streptococcus thermophilus Bacteriophages Are Indicative of Genetic Recombination among Phages from Different Bacterial Species. *Applied and Environmental Microbiology*, *83*(5), e02748-16, e02748-16. https://doi.org/10.1128/AEM.02748-16

Szymczak, P., Rau, M. H., Monteiro, J. M., Pinho, M. G., Filipe, S. R., Vogensen, F. K., Zeidan, A. A., & Janzen, T. (2019). A comparative genomics approach for identifying host-range determinants in Streptococcus thermophilus bacteriophages. *Scientific Reports*, *9*(1), 7991. https://doi.org/10.1038/s41598-019-44481-z

Szymczak, P., Vogensen, F. K., & Janzen, T. (2019). Novel isolates of Streptococcus thermophilus bacteriophages from group 5093 identified with an improved multiplex PCR typing method. *International Dairy Journal*, *91*, 18–24. https://doi.org/10.1016/j.idairyj.2018.12.001

Tremblay, D. M., & Moineau, S. (1999). Complete Genomic Sequence of the Lytic Bacteriophage DT1 ofStreptococcus thermophilus. *Virology*, *255*(1), 63–76. https://doi.org/10.1006/viro.1998.9525

Turner, D., Kropinski, A. M., & Adriaenssens, E. M. (2021). A Roadmap for Genome-Based Phage Taxonomy. *Viruses*, *13*(3), 506. https://doi.org/10.3390/v13030506

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Davison, A. J., Dempsey, D. M., Dutilh, B. E., García, M. L., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., … Zerbini, F. M. (2021). Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Archives of Virology*. https://doi.org/10.1007/s00705-021-05156-1

Zafar, N., Mazumder, R., & Seto, D. (2002). CoreGenes: A computational tool for identifying and cataloging "core" genes in a set of small genomes. *BMC Bioinformatics*, *3*(1), 12. https://doi.org/10.1186/1471-2105-3-12

Zinno, P., Janzen, T., Bennedsen, M., Ercolini, D., & Mauriello, G. (2010). Characterization of Streptococcus thermophilus lytic bacteriophages from mozzarella cheese plants. *International Journal of Food Microbiology*, *138*(1–2), 137–144. https://doi.org/10.1016/j.ijfoodmicro.2009.12.008