Exploring the diversity of tail tips in Gram-negative Siphophages: defining main puzzle pieces and solving the missing D3-like structure

Abstract

Tailed bacteriophages, which dominate the phage biosphere, play a pivotal role in microbial ecosystems and the evolutionary arms race between viruses and bacteria. Among them, siphophages - characterized by their long, flexible, noncontractile tails - constitute the most abundant group. These elegant tail structures represent complex, multi-protein machines essential for two fundamental purposes: attachment to the bacterial surface and delivery of the viral genome into the host cytoplasm. Siphophage tails are the result of an intricate and finely tuned assembly process that reflects millions of years of structural innovation and optimization.

In Gram-negative hosts, where the cell envelope comprises an outer membrane, a periplasmic space, and a thin peptidoglycan layer, successful genome injection poses unique mechanical and biochemical challenges. Tail structures have evolved to navigate this complexity, with diverse tip morphologies emerging to accommodate variations in host surface architecture and defense strategies. While the assembly and organization of siphophage tails have been studied in increasing detail, the tail tip - the actual point of contact and genome entry - remains less well characterized, particularly across different siphophage lineages.

This study delves into the diversity of tail tip structures in Gram-negative-infecting siphophages, identifying conserved modules and unique adaptations that enable successful host interaction. Special focus is given to solving the D3-like tail tip type, a common but structurally not defined type among siphophages. By integrating comparative genomics, structural modeling, and evolutionary context, we aim to define the main puzzle pieces in a big picture of siphophage tail tips and provide the analysis of the main tail tip types in terms of their structure, organization, and function. Through this analysis, we contribute to a more complete understanding of the modularity, specialization, and evolutionary innovation underlying phage tail morphogenesis and host specificity.

Keywords

Phage-host interactions, tail assembly, comparative genomics, tailed bacteriophages, morphogenesis

1. Introduction

Bacteriophages (phages), the viruses that infect bacteria, are the most abundant biological entities on Earth, with an estimated 10^{31} particles globally. Among them, tailed phages (order Caudoviricetes) dominate both numerically and ecologically, accounting for over 95% of all characterized phages. These viruses are morphologically distinguished by the presence of a tail structure, which serves as a specialized apparatus for host recognition and genome delivery. The three main morphotypes - Sipho-like (siphophages), Myo-like (myophages), and Podo-like (podophages) are differentiated by tail length and contractility. Siphophages, with long, noncontractile, flexible tails, represent the most numerous group across diverse environments, especially within Gram-negative bacterial communities.

The tail of a siphophage is a marvel of molecular engineering. Composed of multiple protein subunits, it is assembled via a highly regulated, sequential process that includes tail shaft and tail tip structures - each adapted to support key steps in the phage infection cycle. Two primary functions define the utility of the tail: anchoring the phage to a specific site on the bacterial surface and acting as a conduit for DNA injection. This dual role demands both structural robustness and precision, especially when infecting Gram-negative bacteria, which present a complex envelope consisting of an outer membrane, a periplasmic space, and a thin peptidoglycan layer. Efficiently breaching this multilayered barrier requires finely tuned molecular machinery capable of both sensing and breaching host defenses.

While much progress has been made in resolving the structure of phage particles through cryo-electron microscopy (cryo-EM) and X-ray crystallography - the tail tip, which represents the interface of host contact and genome delivery, remains comparatively understudied. This structural region is critical not only for host specificity but also for triggering conformational changes necessary to initiate DNA translocation. A growing body of evidence indicates that siphophage tail tips are highly variable, reflecting adaptation to diverse bacterial surface molecules such as lipopolysaccharides, porins, or outer membrane proteins. Despite this functional significance, few tail tip types have been thoroughly characterized at the atomic level, and many common morphotypes, such as the D3-like type, remain undefined in terms of their structural and organizational principles.

The current study focuses on unraveling the diversity and modular design of siphophage tail tips that target Gram-negative bacteria. By integrating comparative genomics, protein domain architecture, and structural prediction tools such as AlphaFold and cryo-EM modeling, we aim to identify conserved modules and lineage-specific innovations that underpin host interaction. In addition to defining main structural tail tip types, special attention is given to the D3-like tail tip - a morphotype frequently encountered in siphophage genome clusters but lacking definitive structural resolution. Our approach combines computational modeling with synteny analusis and functional annotation to propose a structural framework for exploring this enigmatic tail tip machinery.

By mapping the diversity of siphophage tail tips and defining their conserved and divergent elements, this work advances our understanding of how structural modularity enables evolutionary adaptability in phage infection mechanisms. Moreover, it lays the groundwork for the rational design of synthetic phages or tail-like delivery systems, with potential applications in phage therapy, microbiome engineering, and nanobiotechnology.

2. Results

2.1.Defining key tail tip functions using Lambda as a reference

2.1.1. Tail tip organization in phage Lambda

The process of tail assembly has been extensively studied in phage Lambda, a model temperate phage of Gram-negative bacterium *Escherichia coli*. Its host species exhibits a wide range of lifestyles from benign saprophytes that inhabit mammalian gut to highly pathogenic strains that cause severe infections in animal and human populations. Phage Lambda has been widely studied for its genetics, life cycle, and infection mechanisms. As a siphovirus, phage Lambda possesses a long, flexible, non-contractile tail that facilitates host recognition and genome delivery.

Tail formation in phage Lambda is known to undergo three stages: (1) formation of an initiator complex; (2) tail extension; and (3) tail completion. At the first stage, an initiator complex is formed starting from a large protein gpJ that possess host recognition capacity. This protein forms a trimer and requires functions of other proteins including gpI, gpL, and gpK. The results of the previous research indicate that the interactions with other proteins are critical for proper formation of a functional initiator complex that includes three copies of each of the following proteins: gpJ, gpI, and gpL, that interacts with a rod-shaped trimer of gpH shielded with gpG proteins. These interactions are stabilized by a hexameric ring of gpM. Without gpM, the interactions are not stable and the initiator complex comes together and falls apart. Meanwhile, the gpGT is produced by a ribosome slippage resulting in attaching the gpT part to gpG proteins via frameshift (-1) that occurs about 4% of the time. The mix of gpG and gpGT proteins forms likely fragmented spiral-shape shield for gpH that likely maintain it soluble.

The gpGT recruites gpV, main tail tube protein. This interaction initiates polymerization of gpV in the form of hexameric rings stacked around a rod-shaped core of the gpH trimer. The polymerization stops when the required length of the tail tube is reached, and the opposite end of the tail tube (the furthest from the tail tip) is capped by gpU to prevent the further gpV polimerization that would lead to the tail overextension. The bulk of siphophage tails (tail shaft) is made of hexameric tail tube rings that are stacked on a tape measure trimer rode. The C-terminal part of gpH is cleaved in a coordinated fashion with the gpU capping that might pass a signal that extension is finished, and the tail assembly process has entered the final stage. This stage ensures tail completion by making it competent for joining with a phage head. In some siphophages (e.g., phage T5), the C-terminal part of tape measure exhibits muralytic activity. This part can be cleaved and stay in the tail tip. The fact that the C-terminal end of tape measure in phage Lambda is cleaved and does not have this property might provide some insights about the role and position of gpK withins the tail tip structure.

In summary, the tail tip structure is particularly critical for infection, as it mediates attachment to the bacterial receptor and initiates DNA translocation. In phage Lambda, this intricate architecture comprises several key proteins (and the corresponding functional role, see the following sections for more detail about each role): gpM (DTN), gpL (THN), gpK (TNLP), gpI (THI), gpJ (CF), and gpH (TM), each contributing distinct structural and functional roles in the assembly and function of the phage tail tip.

2.1.2. DTN = Distal Tail Gram-Negative (gpM): A Load Strap Belt in Tail Tip

The gpM protein plays a pivotal role as a structural component in the tail tip, aiding in the stabilization and proper positioning of other proteins within the tail tip assembly. While not directly involved in receptor recognition or interactions with tail tube proteins, gpM forms a hexameric ring that serves as a “safety belt” that ensures the correct and secure attachment of other tail tip proteins, particularly gpL, gpI, and gpJ. Studies on related siphophages suggest that homologous structural proteins provide a framework that supports the dynamic functions of receptor-binding proteins during infection.

2.1.3. THN = Tail Hub Gram-Negative (gpL): A Linker Between Tail Shaft and Tail Tip Proteins

The gpL protein functions as a critical linker protein. Its primary function is to act as a bridge that connects the sixfold-symmetric tail shaft to the threefold-symmetric tail tip. Also, gpL is involved in the recruitment of gpK and facilitates the docking of gpJ, a multi-domain protein that includes the primary receptor-binding domain of phage Lambda. Mutational analyses have demonstrated that the absence of gpL disrupts tail assembly and prevents successful host infection. This indicates its indispensable role in maintaining the integrity of the tail structure.

2.1.4. TNLP = Tail NlpC domain (gpK): An Essential Component in Tail Maturation

The TNLP is another important players that contributes to tail tip maturation. In phage lambda, this role is fulfilled by the gpK protein – one of key components in the tail tip complex assembly – exhibits a two-domain architecture with distinct structural and functional roles. The N-terminal domain adopts a ubiquitin-like fold, which may contribute to molecular stability or facilitate specific protein-protein interactions during tail assembly. Ubiquitin-like domains are often involved in non-covalent binding and can serve as interaction hubs, suggesting a scaffolding or organizational function within the tail tip. The C-terminal domain of gpK, by contrast, contains an NlpC/P60-like peptidoglycan hydrolase domain, consistent with roles in host cell wall degradation. This domain is thought to mediate localized digestion of the bacterial peptidoglycan layer during infection, aiding in the injection of phage DNA into the host cytoplasm. Importantly, the enzymatic activity of this domain is believed to be tightly regulated and spatially confined, to avoid premature degradation. Together, the two-domain structure of gpK reflects a sophisticated functional bifurcation: one domain mediating structural integration within the tail tip and the other facilitating the mechanical breach of host defenses.

2.1.5. THI = Tail Hub Internal (gpI): A Structural Internal Connector

The gpI protein of phage lambda plays a critical structural role as the internal tail hub protein (THI), acting as a connector between the central tail fiber and other tail proteins during tail assembly and DNA delivery. Functionally, gpI is thought to serve as a molecular adapter, ensuring the proper alignment and stabilization of downstream tail tip components. It may also participate in the conformational rearrangements that occur upon host recognition.

2.1.6. CF = Central Fiber (gpJ): Host specificity

Among all tail tip components, gpJ is the most functionally critical for host specificity. This protein mediates the initial interaction between phage λ and its bacterial receptor, LamB, a maltose porin located on the outer membrane of *E. coli*. Structural studies indicate that gpJ consists of an elongated domain with a receptor-binding region at its distal end. Mutations in gpJ can alter host range, emphasizing its role in determining phage tropism. Upon binding to LamB, gpJ likely transduces mechanical and conformational signals that initiate DNA injection.

2.1.7. TM = Tape Measure (gpH): Facilitating DNA Ejection and Tail Stabilization

The determination of tail tube length in bacteriophages of the Siphoviridae and Myoviridae families is primarily governed by the Tape Measure protein. This regulatory role has been elegantly demonstrated through detailed studies on bacteriophage λ underscoring the conserved function of this protein across diverse tailed phage lineages. The gpH protein serves dual functions in tail tip organization and genome delivery. Structurally, gpH contributes to stabilizing the distal end of the tail tube, ensuring proper alignment of the tail tip proteins. Functionally, it plays a role in triggering DNA ejection upon successful receptor binding. Some models propose that gpH undergoes conformational changes upon interaction with LamB, which, in turn, facilitates the opening of the tail tube, allowing DNA passage into the host cytoplasm.

In conclusion, the tail tip of phage λ is a highly specialized structure composed of multiple proteins working in concert to recognize the bacterial receptor and initiate infection: gpM, gpL, gpK, gpI, gpJ, and gpH each contribute to the stability, function, and specificity of this molecular apparatus. While structural studies have provided significant insights, ongoing research using cryo-electron microscopy and mutagenesis continues to unravel the fine details of tail tip dynamics.

2.2.Defining main tail tip types in Gram-negative siphophages

In the study, we define a tail tip type based on the arrangements of key functional components. The tail tip region, which mediates host recognition and initial infection, typically encodes structural components such as the tail spike, tail fiber, baseplate, and tape measure protein, along with associated chaperones and assembly factors. Among Gram-negative-infecting siphophages, distinct gene arrangements have been observed, reflecting different evolutionary strategies for host interaction. For example, in Escherichia coli phage T5, the tail tip genes are organized with the tape measure protein adjacent to the receptor-binding tail fiber, whereas in Pseudomonas siphophages like PA1 and JBD25, the tail spike genes are positioned between structural scaffolds, indicating alternative infection mechanisms. By analyzing conserved gene synteny across siphophages, tail tip types can be delineated, offering insights into host specificity and phage adaptation strategies.

Tail tip types are defined based on a type of Tail Hub protein determined by the corresponding HMM hit.

2.2.1. Lambda-like type

This tail tip type posses six separate proteins to fulfill functions that are represented by the main tail tip functional groups. The Lambda-like tail tip type represents the most prevalent structural class among the siphoviruses analyzed, accounting for approximately 40% of the 467 genomes in the collection. This group is defined by a conserved set of tail tip proteins - gpM, gpL, gpK, and gpI - all of which fall within the similar profile Hidden Markov Model (HMM) families, indicating strong evolutionary conservation and functional coherence. These proteins are well-characterized in bacteriophage Lambda, the prototypical member of this group, where they assemble to form a sophisticated tail tip structure essential for host cell recognition and DNA injection. gpM and gpL form part of the distal tail structure, while gpK and gpI are critical for completing tail initiator complex assembly and mediating attachment to the host receptor, typically the LamB maltoporin in *Escherichia coli*. The architectural conservation of these proteins across the group suggests a shared infection strategy and a relatively constrained evolutionary trajectory compared to other siphoviral lineages. Lambda-like phages are also notable for their lifestyle and modular genome organizations, which facilitate both lysogenic and lytic cycles. Their prevalence, genetic stability, and deep historical study make them a foundational model for phage biology and a benchmark for classifying tail tip diversity across siphoviruses.

2.2.2. D3-like type

The D3-like tail tip type represents a distinct subgroup of siphophages characterized by specific tail tip proteins that deviate from the classical Lambda phage architecture. Within this group, the tail proteins gpM and gpL align closely with those found in the canonical D3 phage, as identified by profile Hidden Markov Models (HMMs), underscoring a conserved core structure. However, the presence of a gpK homolog reveals a more complex evolutionary trajectory. Although a gpK gene is identifiable across this group, it shares homology only with the C-terminal region of the Lambda phage gpK, indicating significant sequence divergence. This divergence is so pronounced that standard lambda gpK HMMs fail to detect all instances, prompting the development of a D3-specific HMM for gpK. Despite this refinement, the D3-specific HMM still does not detect all candidate gpKs, suggesting functional conservation amid substantial sequence variability. Although detection remains challenging, synteny and comparative genomics imply that a gpK-like gene is conserved across members of this group, which accounts for approximately 9% of the analyzed genomes. This pattern of conservation amidst divergence is consistent with findings from studies on phage tail morphogenesis, where tail tip proteins often evolve rapidly to accommodate host-specific interactions. The D3-like architecture thus represents a modular and adaptable phage tail tip lineage, with implications for phage-host specificity and therapeutic applications.

2.2.3. PY54-like type

The PY54-like tail tip type represents a small but distinct phage group comprising approximately 3% of the genome collection. Members of this group are defined by tail proteins gpM and gpL that closely resemble those found in Yersinia phage PY54, as confirmed through HMM-based annotation. These phages also encode a distinct gpK, which shows homology to the C-terminal domain (CTD) of lambda gpK, as identified via HHpred. This structural similarity suggests a conserved role in tail tip organization and potentially DNA ejection processes, despite divergent primary sequences. Similar to other lambdoid phages, PY54 encodes proteins that suggest a modular tail tip organization, but with unique evolutionary paths.These differences highlight the evolutionary plasticity of tail modules and the potential of the PY54-like group to reveal novel host-interaction strategies in siphoviruses.

2.2.4. T5-like type

The T5-like tail tip type represents a distinct and more prevalent type among siphoviruses with large genomes (>100 kb) representing about 18% of the analyzed genomes. Although these phages share some homology with the PY54-like group - most notably, a gpM that clusters within the PY54 HMM family - they differ significantly in tail architecture. Specifically, their gpL (THN) is afused to the central fiber (CF), forming a chimeric structural element that sets this group apart from all other tail tip types. This fusion likely represents an evolutionary adaptation to accommodate larger genome packaging requirements or more complex infection mechanisms. Notably, there is no distinct gpK protein (TNLP) in this group, suggesting a streamlined or functionally divergent tail tip module compared to other siphophage groups. Phage T5, the prototypical member of this group, is well-known for its complex tail machinery and two-step DNA injection system, where pre-early genes are transcribed before the full genome enters the host. The presence of this fused gpL in large-genome siphoviruses points to a broader structural strategy among complex phages for optimizing host recognition and genome delivery. These features justify the classification of T5-like phages as a separate structural type.

2.2.5. MP22-like type

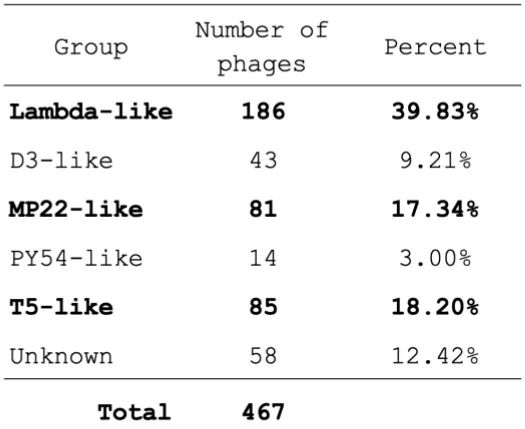
The MP22-like tail tip type (correlates with Sipho-4 in the lamboid phages) represents a major and structurally distinct lineage among siphophages, comprising approximately 17% of the genome collection. Unlike the canonical lambdoid phages, this group is entirely absent from the classical lambda-related lineages, and is instead typified by phages such as Chi, JBD30, and gene transfer agents (GTAs). These phages possess unique tail tip organization defined by distinct gpM and gpL homologs, as well as a divergent gpI-like protein. Structural studies of members in this group - including cryo-EM resolved tail tips from Chi-like phages and GTAs - have revealed novel folds and modular arrangements that differ markedly from those observed in lambda or D3.

2.3. Exploring main tail tip types among Gram-negative siphophages

2.3.1. Summary statistics

Our analysis of 467 siphophages infecting Gram-negative hosts revealed five major categories of tail tip structures (**Table 1**.), reflecting both conserved modules and lineage-specific innovations. The most prevalent type was the Lambda-like tail tip, found in 186 phages (39.83%), underscoring its widespread use among Enterobacteria phages. The T5-like group was the second most common, comprising 85 phages (18.20%), followed closely by the MP22-like group with 81 phages (17.34%), both representing distinct structural and functional solutions for host interaction. The D3-like architecture, which has previously been under-characterized, was identified in 43 phages (9.21%), expanding our understanding of this unique structural variant. Less common was the PY54-like group, present in 14 phages (3.00%), likely reflecting a more specialized evolutionary niche. This tail tip type is somewhat similar to T5-like type (but it has distinct characteristics), and it is found among phages with genome length below 100 Kbp. Notably, 58 phages (12.42%) did not fit into any of the defined categories and were classified as Unknown, highlighting the potential for novel tail tip types yet to be characterized and the ongoing evolutionary diversification of these structures.

Table 1. Description of main tail tip types.

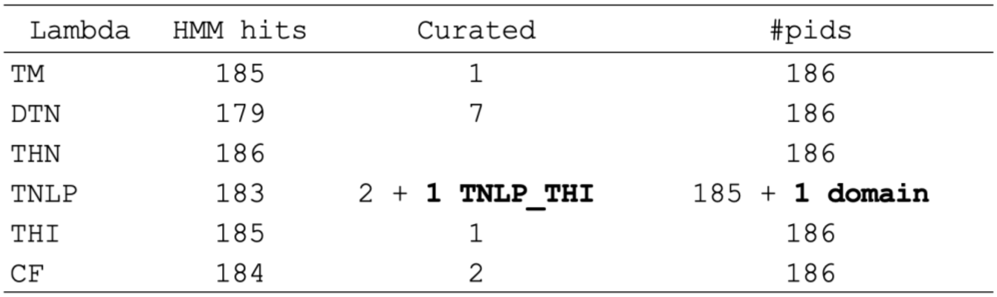


Group	Number of phages	Percent
Lambda-like	**186**	**39.83%**
D3-like	43	9.21%
MP22-like	**81**	**17.34%**
PY54-like	14	3.00%
T5-like	**85**	**18.20%**
Unknown	58	12.42%
Total	**467**	

2.3.2. Lambda-like phages

This tail tip type account for almost 40% of all Gram-negative siphophages in the dataset. There are 186 phages in total that are Lambda-like. The annotation summary are shown in **Table 2**.

Table 2. Annotation summary for the Lambda-like phages.



Lambda	HMM hits	Curated	#pids
TM	185	1	186
DTN	179	7	186
THN	186		186
TNLP	183	2 + **1 TNLP_THI**	185 + **1 domain**
THI	185	1	186
CF	184	2	186

For DNT among Lambda-like phages, there were 186 proteins identified as DTN based on HMM hits and synteny (179 HMM+synteny; 7 synteny+curation). There proteins were separated into four clusters by MMseq2 (0 0). Cluser1 contains 118 protein sequences including DTN of phage T1, Cluster 2 contains 63 protein sequences including DTN of phage Lambda, and Cluster 3 and 4 are singletons containing more divergent DTN protein sequences of Caulobacter phage Seuss (gp023, YP_009785533.1) and Colwellia phage 9A (YP_006489243.1) respectively. The alignment of four representative protein sequences (Jalview, coroled by percent identity top row to bottom row: T1, Lambda, Seuss, and 9A) and their structural representations are shown on **Fig. 1.**

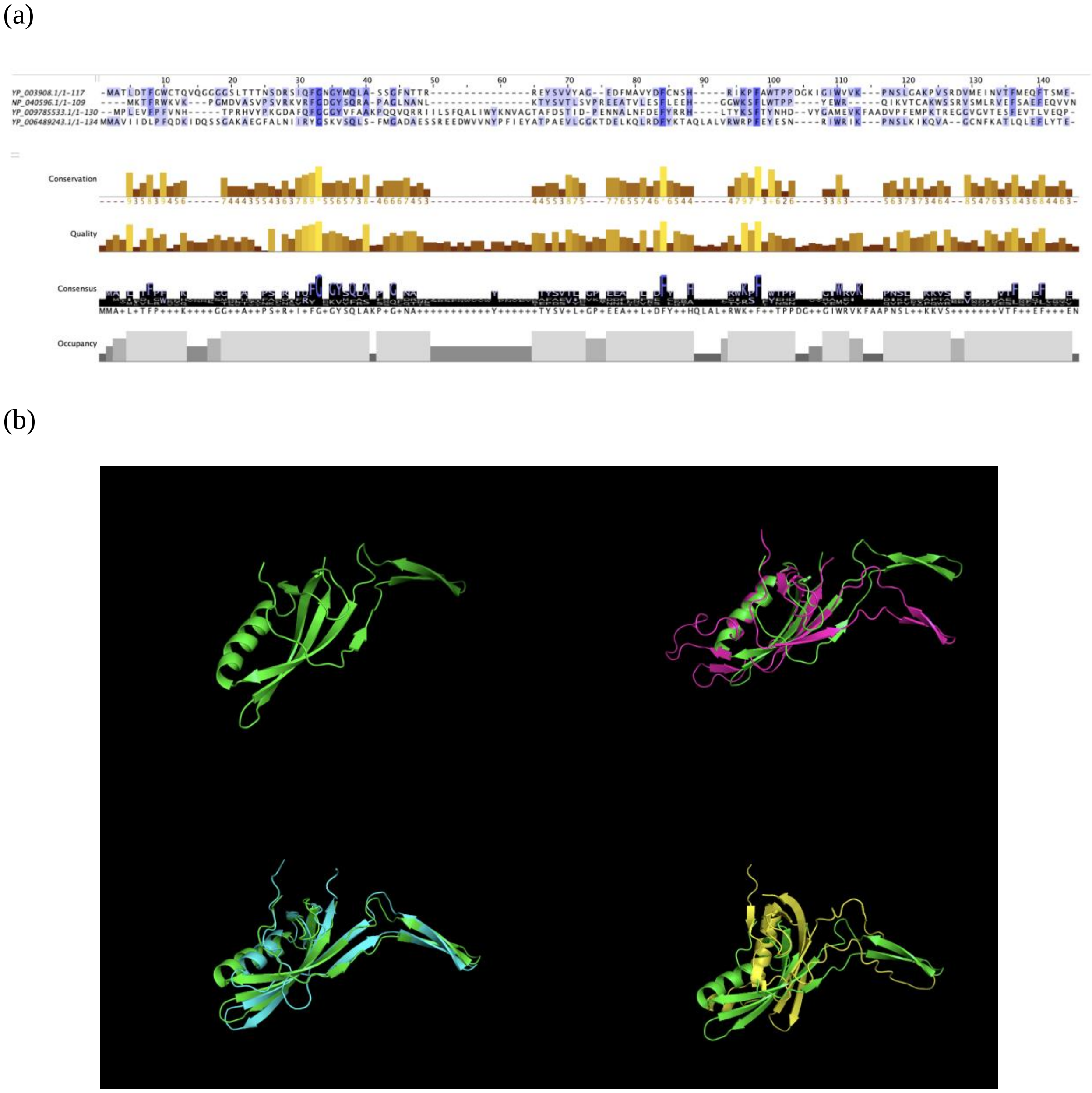


Fig.1. (a) MSA for the five DTN representative sequences: T1, Lambda, Seuss, and 9A (Jalview, colored by percent identity); (b) DTN representative sequence structures: DTN Lambda (green) – top left, DTN Lambda (green) aligned with DTN T1 (cyan) – bottom left, DTN Lambda (green) aligned with DTN Seuss (magenta) – top right, and DTN Lambda (green) aligned with DTN 9A (yellow).

There were three phages that were missing DTN ORF in their GenBank files (**Table 3**.). In these cases, TM was followed by THN without a separate ORF detected between them. It was possible that DTN role had been played by another protein or a domain fused to another protein. However, TM proteins of these three phages without apparent DTN had within length within the expected range in comparison with TMs in other Lambda-like phages. Also, THN proteins were structurally similar to THN of Lambda even though all three of “no apparent DTN” phages had a small extra helix sticking out (**Fig.2(a)**). THN of phage Lambda was missing that small helix. It was unlikely that this small structural change could compensate for the lack of DTN.

Table 3. Description of three siphophages without apparent DTN.

Genome_ID	Genome_name	Genome_length	Genome_GC
NC_049830.1	Escherichia phage PGN590	49043	0.4381
NC_049844.1	Klebsiella phage 13	43094	0.5063
NC_019934.1	Cronobacter phage ENT39118	39012	0.5306

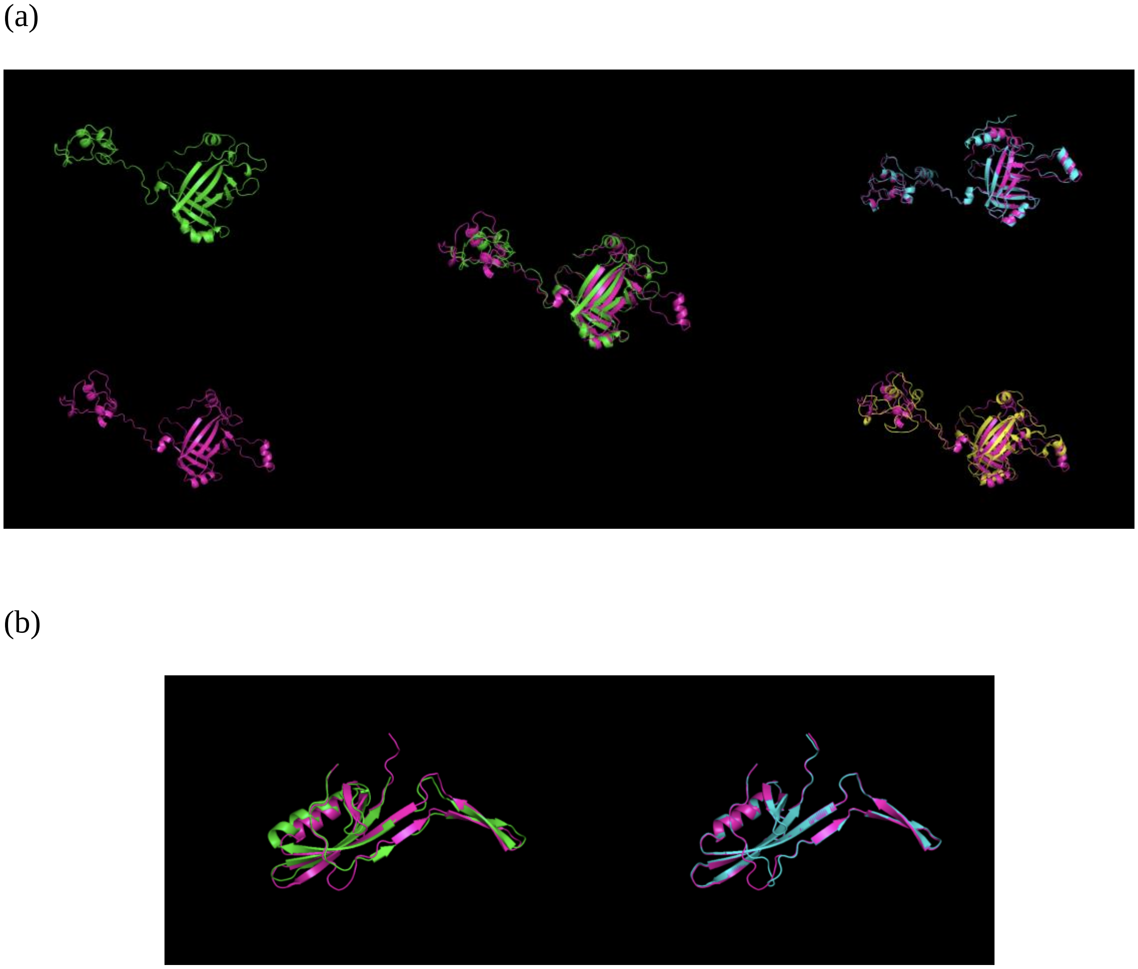


Fig.2. (a) Comparison between THN of Klebsiella phage 13 (magenta, bottom left) and Lambda THN (green, top left) - center and THNs of the remaining two phages with “no DTN”: Escherichia phage PGN590 (blue, top right) and Cronobacter phage ENT39118 (yellow, bottom right); **(b)** comparison between DTN of Klebsiella phage 13 (magenta) and Lambda DTN (green) – on the left and two Klebsiella phages: phage 13 and phage Shelby (blue) – on the right.

We performed in-depth analysis of the three THN and found that one of these THN proteins matched exactly to THN of several other Lambda-like phages that did have DTNs including Klebsiella phage Shelby (NC_049846.1). Moreover, all three "no apparent DTN" phages had a gap of about 400 nucleotides between the annotated ORFs of TM and THN that could potentially accommodate the missing DTN. In one of these three "no apparent DTN" phages, Klebsiella phage 13 (NC_049844.1) there was a region of length 345 nucleotides (or 114 aa) denoted "unsure" and located between TM (YP_009903215.1) and THN (YP_009903214.1). The start codon in this coding region was GTG (V) not ATG (M), and it might be a possible reason for misannotation. The length of the region was right for housing DTN and the predicted structure of this region translated into protein aligned well with DTN of Lambda and Klebsiella phage Shelby (**Fig.2(b)**). Similar analysis revealed the missing DTN in the remaining two phages.

For THN, the vast majority of THN proteins among the Lambda-like phages were clustered together with Lambda THN (NP_040597.1) by MMseq2 (0 0) with only two exceptions that formed singleton clusters: THN from Colwellia phage 9A (YP_006489242.1) and Caulobacter phage Seuss (YP_009785534.1). These two proteins were on the lower end of the length range for THN among the Lamba-like phages. The alignment of the three representative THNs and the predicted structures are shown on **Fig. 3**.

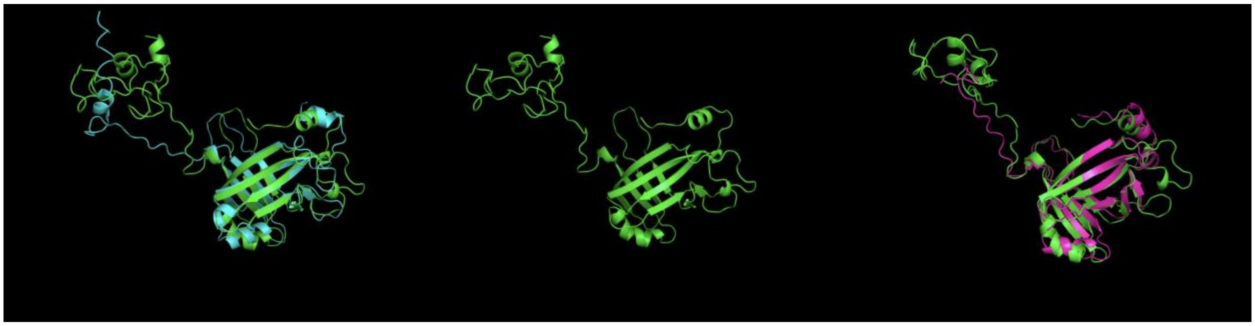


Fig.3. Lambda THN (green) – center; Lambda THN compared with THN of Colwellia phage 9A(cyan) – left; and Lambda THN (green) compared with THN of Caulobacter phage Seuss (magenta) – right.

For TNLP, the vast majority of TNLP proteins among the Lambda-like phages were clustered together with Lambda TNLP (NP_040598.1) by MMseq2 (0 0) with one exception that formed a singleton cluster: Escherichia phage PGN590 (YP_009902212.1). Interestingly, in Citrobacter phage HCF1 (NC_054897.1), TNLP was fused to THI (**Fig. 4.**) forming a protein of 338 aa long (with usual TNLP in the Lambda-like phages having length of approximately 200 aa).

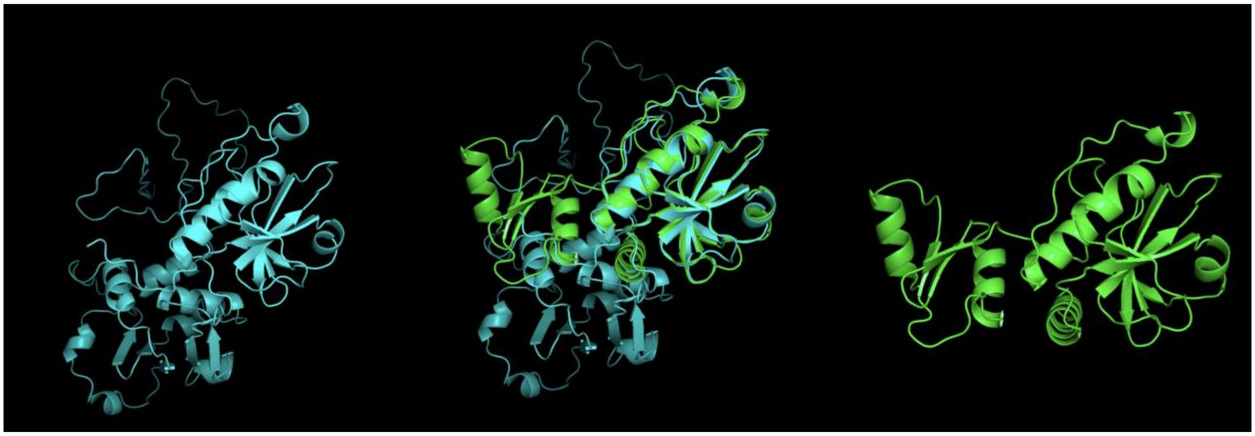


Fig.4. TNLP fused to the N-terminal end of THI of Citrobacter phage HCF1 (cyan, left); Lambda TNLP (green, right); and the comparison between Lambda TNLP and Citrobacter phage HCF1 TNLP fused to THI (center).

For THI, the vast majority of THI proteins among the Lambda-like phages were clustered together with Lambda THI (NP_040599.1) by MMseq2 (0 0) with one exception of Colwellia phage 9A (P_006489240.1) that formed a singleton cluster.

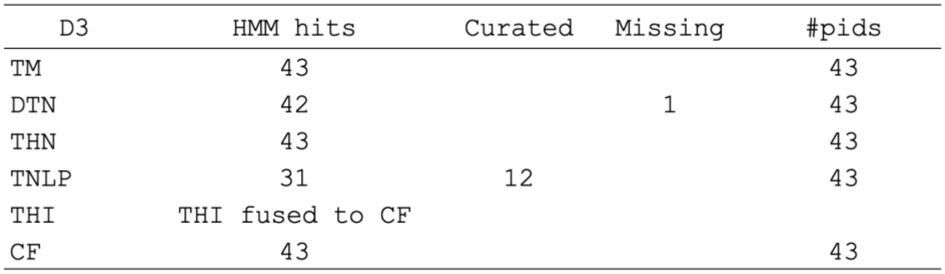
For CF, the vast majority of CF proteins among the Lambda-like phages were clustered together with Lambda CF (NP_040600.1) by MMseq2 (0 0) with two exception of Klebsiella phage 1513 (YP_009197878.1; 659 aa) and Escherichia phage PGN590 (YP_009902204.1; 748 aa) that formed singleton clusters. The outlier clusters have CF that are shorter than it is expected (~ 1200 aa for the Lambda-like phages). Also, these CF were not captured by HMMs and were curated. It is possible that these two phages have CF that consists of multiple fragments. For example, there were four genomes with CF split into two separate proteins (NC_029071, NC_048150, NC_049838, and NC_049942).

In conclusion, the Lambda-like phages form a rather homogeneous group with a few exceptions that are quite rare. All of these phages have proteins or domains that fulfill the five tail tip roles (DTN, THN, TNLP, THI, and CF) described in phage Lambda.

2.3.3. D3-like phages

This tail tip type account for almost 10% of all Gram-negative siphophages in the dataset. There are 43 phages in total that are D3-like. The annotation summary are shown in **Table 4**.

Table 4. Annotation summary for the D3-like phages.



D3	HMM hits	Curated	Missing	#pids
TM	43			43
DTN	42		1	43
THN	43			43
TNLP	31	12		43
THI	THI fused to CF			
CF	43			43

For DTN, one phage, Acinetobacter phage YMC11/11/R3177 (NC_041866.1) was missing DTN. This might be due to sequencing and/or assembly artifacts since a DTN ORF is usually located right after the TM ORF and before the THN ORF. The GenBank file of Acinetobacter phage mentioned start with THN ORF, ends with TM ORF, and DTN ORF is likely "lost in translation".

For THN, most of THNs (32 protein sequences) are clustered with THN of Pseudomonas phage D3 (NC_002484.2) by MMseq2 (0 0). Also, there are four other smaller clusters, two of which are singletons: Rhodobacter phage RcapMu (NC_016165.1) and Pseudomonas phage nickie (NC_042091.1). The structural alignments of 5 representative sequences (one for each cluster) are shown on **Fig. 5.** Notably, THN of Pseudomonas phage nickie is better aligned with THN Lambda than with D3 even though it was hit by HMM THN-D3 with high E-value and did not have any hit by HMM THN-Lam. However, phage nickie does not have a separate THI as other Lambda-like phages do, and its THI is fused to CF like in D3-like phages (see the THI-related paragraph below). These finding suggest that Pseudomonas phage nickie represents a transitional stage between Lambda-like and D3-like phages.

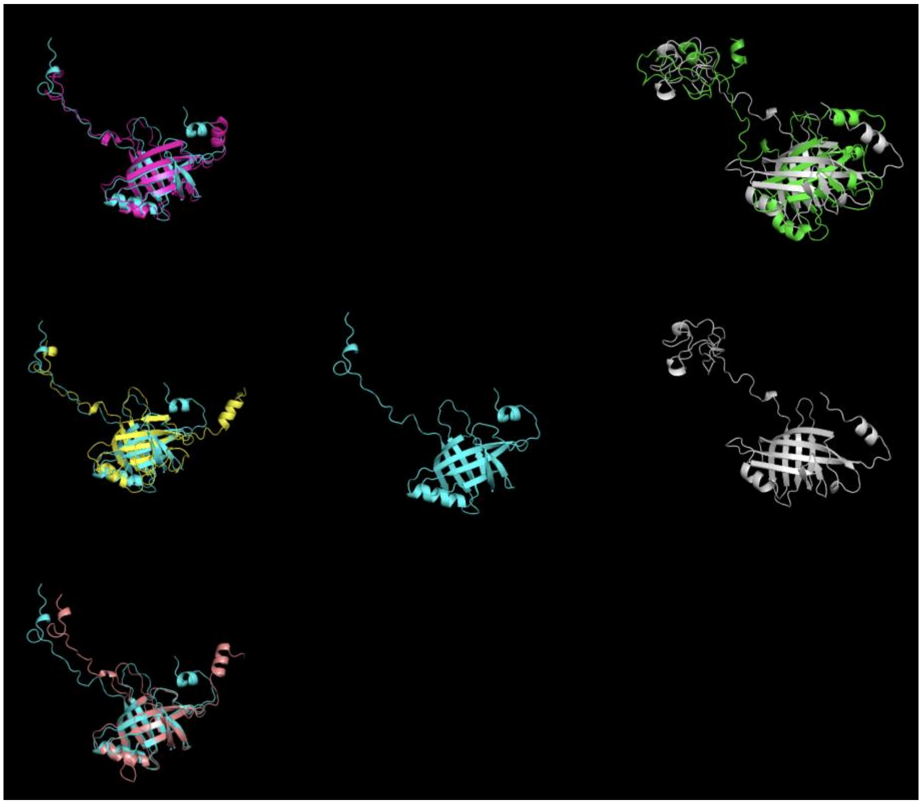


Fig. 5. THN of Pseudomonas phage D3 (cyan) – center; comparison of D3 THN (cyan) with Xp10 (magenta, top left), RcapMu (yellow, center left), and JWX (salmon, bottom left); THN of Pseudomonas phage nickie (grey, center right); nickie THN (grey) is compared with Lambda THN (green) – top right (visualized and aligned in PyMOL 3.1.1.).

For TNLP, in D3-like phages, there are 5 clusters identified by MMseq2 (0 0), three relatively big clusters and two small ones. It is worth noting that D3 TNLP is clustered with nickie TNLP. Importantly, there is a significant difference between Lambda TNLP and D3 TNLP. The TNLP protein in phage Lambda is a two-domain protein, the N-terminal domain is similar to the Mov34 family of peptidases and the C-terminal domain has peptidoglycan degrading capacity. However, in D3-like phages, TNLP has one domain that is similar to the C-terminal domain of Lambda TNLP (**Fig.6.**).

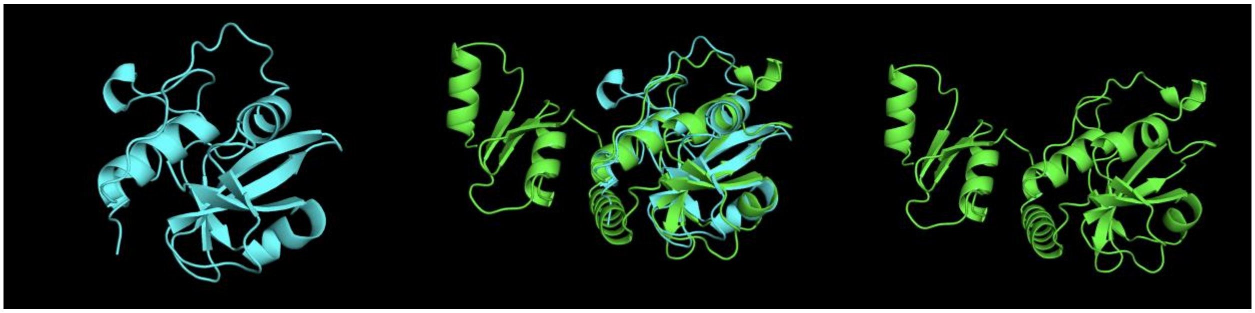


Fig.6. The comparison (center) of D3-like TNLP from Pseudomonas phage nickie (cyan, left) and Lambda TNLP (green, right) (visualized and aligned in PyMOL 3.1.1.).

For THI, unlike the Lambda-like phages that have a separate protein that fulfills the THI role (e.g., gpI in phage Lambda), in D3-like phages this role is delegated to a specific domain that is located at the N-terminal end of CF (this domain is usually from 100 to 200 residues long). This THI-related domain aligns with the N-terminal part of Lambda THI (**Fig. 7**.).

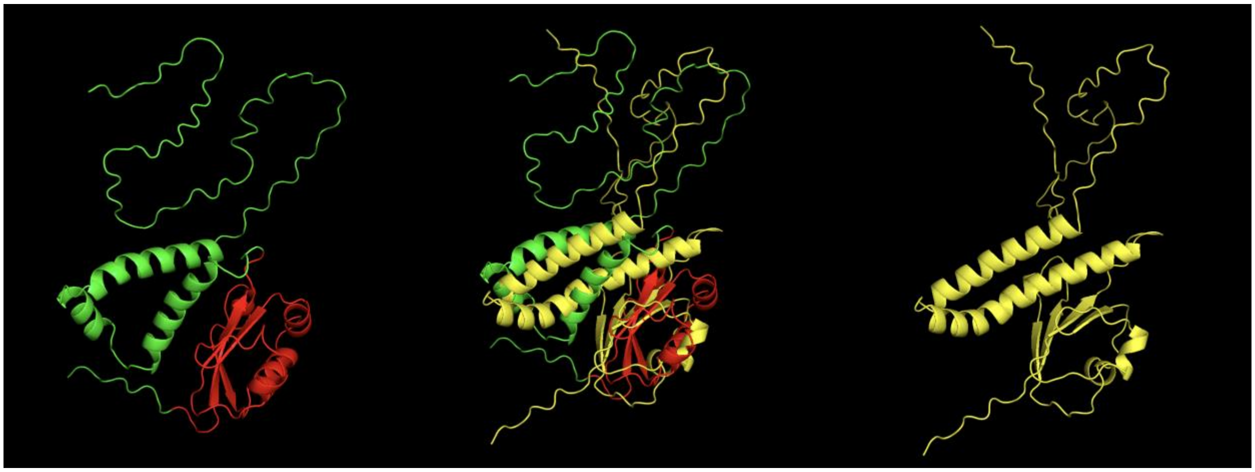


Fig.7. The comparison (center) between Lambda THI (green-red, left) and the N-terminal domain of CF from D3-like phage Pseudomonas phage nickie (yellow, right). The red color indicates the N-terminal part of THI that is similar to the THI domain of CF for D3-like phages.

In summary, the D3-like tail tip type, present in approximately 10% of the Gram-negative siphophages in our dataset (43 phages), represents a structurally distinct and evolutionarily informative group. Most THN proteins clustered with that of Pseudomonas phage D3, but several outliers, including Pseudomonas phage nickie, exhibited features that bridge the Lambda-like and D3-like structural paradigms – suggesting nickie may represent a transitional form. TNLP proteins in D3-like phages also diverge from the Lambda-type by lacking the N-terminal ubiquitin-like domain, retaining only the peptidoglycan-degrading domain. Similarly, while Lambda-like phages encode THI as a standalone protein, D3-like phages embed this functionality within the N-terminal domain of the CF protein. Together, these findings highlight not only the structural cohesiveness within the D3-like group but also its nuanced divergence from the Lambda model, offering insight into modular innovation and evolutionary transitions in tail tip architecture.

2.3.4. PY54-like phages

The PY54-like tail tip type represents one of the least abundant but structurally intriguing architectures identified in our analysis, accounting for only 3% (14 out of 467) of the Gram-negative siphophages. Despite its rarity, this group exhibits a highly conserved gene module reminiscent of the Yersinia phage PY54, characterized by a compact tail tip operon with distinct domain arrangements not found in more common types such as Lambda-like or T5-like phages.

Previous studies on PY54 have shown that its tail fiber assembly and infection mechanisms diverge from canonical enterobacterial siphophages, suggesting adaptation to specific host niches. In our dataset, PY54-like phages are often associated with environmental and non-model hosts, further supporting the notion that this tail tip type represents a specialized evolutionary lineage. This section explores the defining features of PY54-like tail modules, their gene synteny patterns, and possible functional implications, placing them in context with both well-characterized and transitional tail tip architectures.

For DTN, the vast majority of DTN proteins among the PY54-like phages were clustered together with PY54 DTN (NP_892062.1) by MMseq2 (0 0) with three exceptions that formed singleton clusters: DTN from Rhizobium phage 16-3 (YP_002117577.1), Paracoccus phage Shpa (YP_009593493.1), and Escherichia phage DTL (YP_009795730.1). These three proteins were on either on the lower or higher ends of the length range for DTNs among the PY54-like phages (~200 aa).

For THN, 10 of 14 protein sequences were clustered with PY54 THN. The remaining 4 sequences were split into three clusters with two of these clusters being singletons. Only one of the divergent four sequences was structurally dissimilar to THN, the protein sequence from Escherichia phage DTL (YP_009795731.1). The in-depth analysis revealed that this phage was incorrectly classified as PY54-like phage (based of additional THN hit with high E-value) when it was more similar to phage Lambda. The protein was in an expected spot for THN among the Lambda-like phages (i.e., between DTN and TNLP), but was not hit by THN-Lam HMMs. Notably, the predicted structure of this protein by ESMFold did not reassemble THN in phage Lambda. This outlier protein annotated in GenBank file as “HNH endonuclease” was next to TNLP that has much longer length (252 aa) than expected (~130-150 aa). It seemed possible that the THI of Escherichia phage DTL was fused to TNLP. Later, this hypothesis was confirmed by structure prediction and comparative analysis. For TNLP, the remaining thirteen proteins were within the expected length range not exceeding 150 aa. They all clustered nicely together with PY54 TNLP.

For THI, in PY54 phage, THI is fused to the C-terminus of CF based on the HMM analysis (~first 50 residues of the fused protein). Also, CF from other phages (except for the misclassified DL) are clustered beautifully with the PY54 CF (including the potential fused region) indicating that a similar fusion event is relevant to other PY54-like phages.

2.3.5. T5-like

The T5-like tail tip type is one of the most frequently encountered architectures in our dataset, found in 85 of the 467 Gram-negative siphophages (18.2%). This tail tip type is notably associated with large-genome phages, typically exceeding 100 Kbp, such as Escherichia phage T5. These phages are distinguished by a sophisticated DNA injection mechanism, which involves the sequential delivery of a “pre-early” genome segment that modulates host defenses before full genome entry. Structurally, the T5-like tail tip features an extended tail fiber and an elaborate core ejection system, reflecting a high degree of functional specialization. Consistent gene synteny and domain organization across T5-like phages suggest a conserved infection strategy that has been evolutionarily successful across diverse hosts – including Escherichia, Salmonella, Shigella, and Klebsiella. The prevalence of this architecture in large-genome phages points to a correlation between genome size and structural complexity, highlighting the T5-like module as a hallmark of expansive, multi-functional phage genomes. This section explores the defining characteristics, structural conservation, and host associations of T5-like tail tips in the context of genome architecture and evolutionary strategy.

For DTN, 83 of 85 phages have DTN protein sequences that are clustered neatly with T5 DTN and their lengths are very similar as well (~200 aa). The remaining two phages required further analysis due to no apparent DTN candidates detected based on both synteny and HMM analysis. The results of the further analysis indicate that one of these phages, Aeromonas phage AhSzw-1 (NC_047950.1), was misclassified based on PY-54 DTN HMM hit with relatively high E-value and another one, Salmonella phage 1-19 (NC_048819.1), was missing a DTN ORF due to potential sequencing and/or assembly artefacts.

For THN, in phage T5, THN is fused to the C-terminus of CF. The fusion is evident by joint HMM hits to the resulting fused protein by both CF HMMs and THN HMMs with low E-values. The other T5-like phages (except for misclassified Aeromonas phage AhSzw-1) have a similar fused protein, and these fused proteins are aligned well with the canonical THN-CF in T5. Also, T5-like phages do not have a separate TNLP protein. However, the C-terminus of TM in T5 has a muralytic activity. It is highly likely that other T5-like phages share this feature.

2.3.6. MP22-like phages

The MP22-like tail tip type represents a structurally distinct and moderately abundant group in our analysis, present in 81 of the 467 Gram-negative siphophages (17.34%). This tail tip architecture, named after Pseudomonas phage MP22, is characterized by a set of tail-associated genes that differ significantly from the classical Lambda-like or T5-like modules. MP22-like tail tips are frequently associated with medium-sized phages, and are enriched among Pseudomonas, Burkholderia, and Acinetobacter phages, suggesting a strong ecological linkage to non-enterobacterial hosts. MP22-like phages often display gene arrangements that include fused or multi-domain tail components, indicative of structural streamlining or functional innovation.

Preliminary structural predictions and HMM analyses suggest that MP22-like tips may deploy alternative host recognition or DNA injection strategies, possibly adapted to the diverse outer membrane organization of their host taxa. This section explores the conserved features and structural organization of MP22-like tail tips highlighting their role in broadening our understanding of tail module diversity beyond traditional enteric models.

For DTN, DTN proteins among the MP22-like phages were split into three clusters by MMseq2 (0 0). Also, the DTN length range varied from 200 to 800 aa. It indicates that the MP22-like phages represent a heterogeneous group that can be further divided into sub-groups based on the features of tail tip organization.

2.4. Fitting the puzzle pieces together: D3-like type tail tip structure solved

The D3-like tail tip type, long recognized for its distinct structural and genomic features, has now been elucidated with unprecedented clarity. Building upon our bioinformatic classification of tail tip types across Gram-negative siphophages, the D3-like architecture stood out as a unique yet understudied module. By integrating homology-based predictions, gene synteny, and structural alignment data, we have assembled a coherent and testable model of the D3-like tail tip complex – effectively placing a critical piece into the larger mosaic of phage structural biology.

Our analysis revealed that the D3-like configuration diverges significantly from canonical Lambda-like systems. Most notably, the tail hub internal (THI) domain is not encoded as a separate polypeptide, as in Lambda, but is instead fused to the N-terminal region of the central fiber (CF) protein. This fusion likely represents a functional adaptation that streamlines tail assembly or DNA delivery. Similarly, the tail NlpC protein (TNLP) in D3-like phages possesses a simplified single-domain structure focused on peptidoglycan degradation, in contrast to the bipartite domain arrangement seen in Lambda. These rearrangements, though subtle at the sequence level, have substantial implications for phage infectivity, host range, and evolutionary plasticity.

The structural organization of the D3-like type was corroborated by clustering and alignment of key components across 43 phages in our dataset. The majority of tail hub proteins (THNs) in this group aligned closely with Pseudomonas phage D3, reinforcing the conservation of this architecture. Intriguingly, Pseudomonas phage nickie was identified as a transitional case – it exhibits a THN more similar to Lambda in sequence yet retains the fused THI-CF structure characteristic of D3-like phages. Such intermediates shed light on possible evolutionary pathways between distinct tail tip types, suggesting that modular recombination and domain fusion have driven tail innovation.

Solving the D3-like structure is not just a technical accomplishment – it expands our conceptual framework for phage tail organization. This model can now guide experimental validation, structural determination, and functional assays. It also offers a reference point for detecting related but divergent tail modules in future genome mining efforts. As we continue to explore the boundaries of structural diversity in phage biology, the D3-like tail tip serves as a reminder that even well-trodden phage systems harbor surprises waiting to be decoded.

3. Materials and methods

3.1. Gram-negative Siphophage Tail Tip HMM collection

To characterize the structural diversity of tail tip complexes in siphophages infecting Gram-negative hosts, we compiled a curated set of 25 Hidden Markov Models (HMMs) representing distinct functional categories based on canonical components from *Escherichia coli* phage lambda (**Table 5.**). These categories include: Distal Tail (DTN; 5 HMMs), corresponding to gpM; Tail Hub (THN; 6 HMMs), corresponding to gpL; Tail NlpC domain proteins (TNLP; 3 HMMs), modeled after gpK; Tail Hub Internal (THI; 3 HMMs), related to gpI; and Central Fibre (CF; 8 HMMs), corresponding to gpJ. Each HMM captures conserved sequence features of its respective structural module, enabling sensitive detection of homologous proteins across diverse phage genomes. This HMM collection formed the basis for systematic screening and classification of tail tip architectures in our dataset.

Table 5. Description of Tail Tip Sipho Gram-negative HMM collection (TTC SiphoN HMMs).

Category	Description	Number of HMMs
DTN	Distal Tail siphophage Gram-Negative; E. coli phage lambda gpM	5
THN	Tail Hub siphophage Gram-Negative; E. coli phage lambda gpL	6
TNLP	Tail NlpC domain siphophage Gram-negative; E. coli phage lambda gpK	3
THI	Tail Hub Internal siphophage Gram-negative; E. coli phage lambda gpI	3
CF	Central Fibre siphophage Gram-negative; E. coli phage lambda gpJ	8
	Total	25

3.2. Gram-negative siphophage dataset

We assembled a dataset of 467 siphophages infecting Gram-negative hosts (**Fig.8.**). The core of the dataset (416 phages) is derived from the ICTV Master Species List 2020, the last release in which tail morphology served as a formal classification criterion. This unique historical snapshot captures a morphologically informed view of phage taxonomy before the ICTV moved toward exclusively genome-based classifications. Also, the dataset includes 51 phage sequences from our in-house PAT database. The dataset reveals considerable diversity in both phage lineage and host range.

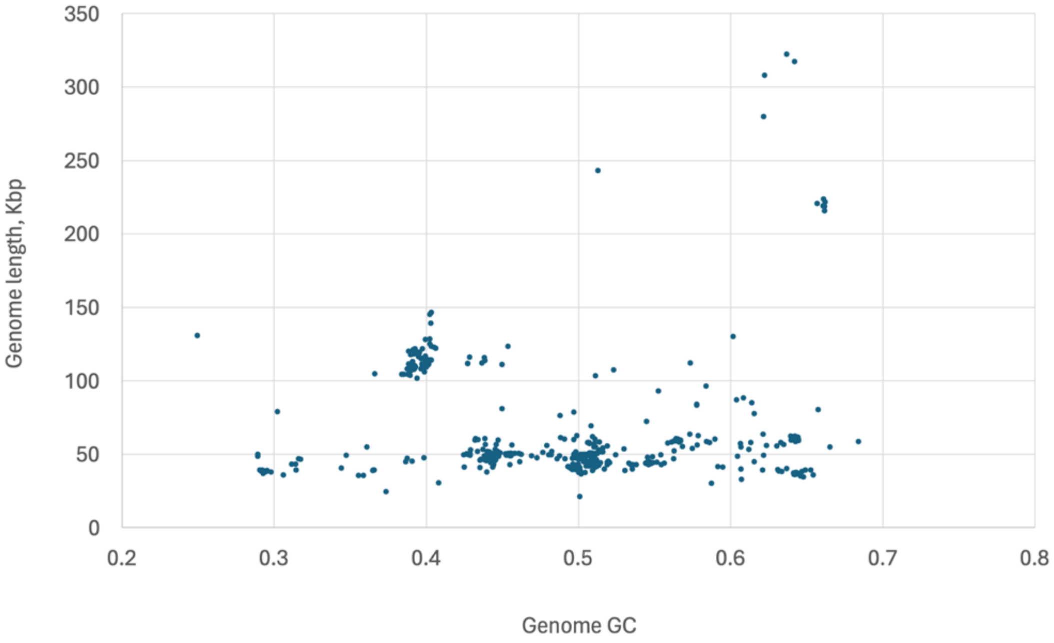
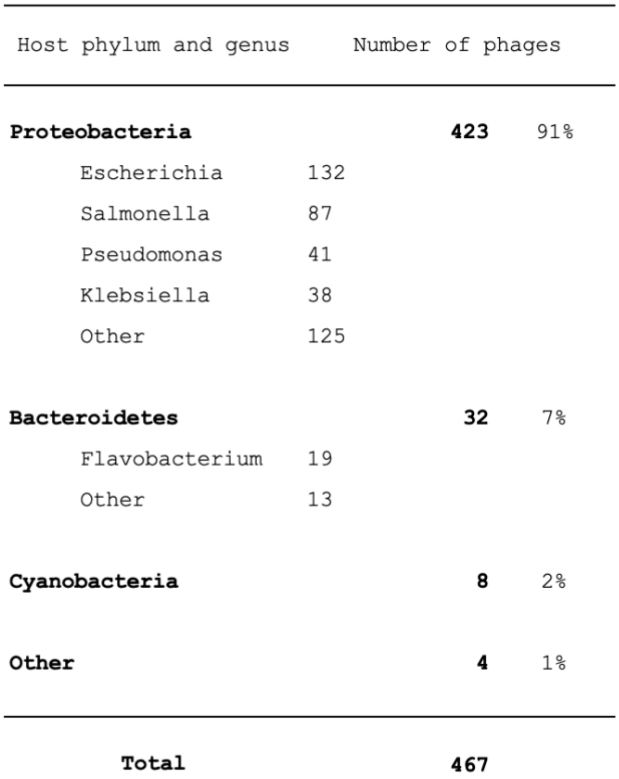


Fig. 8. Distribution of 467 phage sequences.

Host taxonomy analysis shows a strong focus on medically and environmentally significant genera. Escherichia coli accounted for the largest portion of phages (~28%), followed by Salmonella (~19%), Pseudomonas (~9%), and Klebsiella (~8%), reflecting both clinical importance and research interest. This host spectrum underscores the ecological and evolutionary versatility of siphophages infecting Gram-negative bacteria.

The inclusion of diverse bacterial hosts (**Table 6**.) enhances the value of this dataset for comparative genomics and for identifying lineage-specific tail tip innovations shaped by distinct host interactions. The host range spans a broad spectrum of bacterial phyla, with a strong representation of diverse Proteobacteria (including Alpha-, Beta-, and Gammaproteobacteria), but also includes phages infecting members of Bacteroidetes, Firmicutes, Cyanobacteria, and even Thermophilic lineages, reflecting the ecological breadth and adaptive versatility of Gram-negative siphophages.

Table 6. Description of bacterial hosts for 467 Gram-negative siphophages.



Host phylum and genus	Number of phages		
Proteobacteria		**423**	91%
Escherichia	132		
Salmonella	87		
Pseudomonas	41		
Klebsiella	38		
Other	125		
Bacteroidetes		**32**	7%
Flavobacterium	19		
Other	13		
Cyanobacteria		**8**	2%
Other		**4**	1%
Total		**467**	

4. Discussion

Our characterization of five major tail tip types in siphophages infecting Gram-negative hosts provides a new framework for understanding tail tip diversity and its functional implications in host recognition and infection. This work complements and extends prior studies that have primarily focused on morphological classification or relied on sparse experimental data, by leveraging a comprehensive comparative genomics approach. The use of profile HMM hits enabled the identification of conserved functional domains associated with tail tip components such as receptor-binding proteins, tail spikes, and baseplate-like structures, while gene synteny analysis provided contextual evidence for structural and functional organization.

A key strength of our bioinformatic pipeline is its scalability and ability to detect divergent homologs with low sequence identity, which is particularly valuable for phage structural genes. However, limitations include the reliance on available HMM profiles, which may miss highly novel or lineage-specific proteins, and the inherent uncertainty in inferring function solely from homology without experimental validation. Alternative approaches such as cryo-electron microscopy, proteomics of purified virions, or transposon mutagenesis could offer direct structural or functional insights, albeit with higher resource requirements. Building on our findings, future work should aim to integrate structural prediction tools (e.g., AlphaFold) with experimental validation to refine functional annotations, and to explore the ecological and evolutionary drivers of tail tip diversity across environmental and clinical phage populations.

While our classification of five major tail tip types in siphophages infecting Gram-negative hosts captures a broad and functionally coherent landscape, it is by no means exhaustive. The remarkable plasticity of phage genomes and their capacity for modular evolution suggest that tail tip architectures can undergo significant diversification, potentially giving rise to novel types that deviate substantially from the mainstream and might evade the current framework. Evolutionary innovations - such as gene fusion events, acquisition of host-derived domains, or horizontal gene transfers from unrelated phages - can generate tail tip modules that escape detection by existing HMM profiles and synteny patterns. Rather than undermining our classification, the discovery of such outliers would enrich the framework by extending its boundaries and challenging its assumptions. Our work provides a reference map - against which new variants can be evaluated, classified, or identified as new lineages. These future discoveries will not only refine the phylogenetic and functional resolution of phage tail tips but also deepen our understanding of the evolutionary mechanisms driving host adaptation and structural innovation in phage biology.

5. Conclusion

Our study offers a foundational framework for understanding the modular diversity and evolutionary logic of siphophage tail tip organization. By integrating HMM-based domain detection with synteny analysis across a broad genomic dataset, we identified five major tail tip types, each representing distinct functional strategies for host interaction. Our resolution of the previously enigmatic D3-like structure not only fills a critical gap in the structural landscape but also demonstrates the power of comparative genomics in decoding phage biology. While the system we propose is necessarily incomplete - acknowledging the likelihood of further undiscovered variants - the classification scheme and evolutionary insights laid out here serve as a conceptual map for future discovery. Our work thus makes an important contribution to the field by translating tail tip diversity into structural and functional understanding. The results of this study are laying the groundwork for a comprehensive taxonomy of phage infection machinery in Gram-negative hosts. Ultimately, by charting the known and illuminating the unknown, we provide a lasting foundation for a living atlas of phage evolution.