Reviewing the main types of siphophage Gram-negative tail tips: the patchwork of Central Fibers and other modular patterns

Tatiana Lenskaia[1,*], Sherwood Casjens[2], and Alan Davidson[1,3]

[1] Molecular Genetics, Temerty Faculty of Medicine, University of Toronto, Toronto, Canada

[2] School of Medicine and School of Biological Sciences, University of Utah, Salt Lake City, USA

[3] Department of Biochemistry, University of Toronto, Toronto, Canada

* Corresponding author: t.lenskaia@utoronto.ca

Abstract

The tail tip of siphophages, particularly those infecting Gram-negative bacteria, exhibits remarkable structural and evolutionary diversity, serving as the critical interface for host recognition and DNA delivery. In this study, we present a comprehensive comparative analysis of siphophage tail tips classified according to distinct structural features and functional organization, with a special focus on the diversity and versatility of Central Fiber proteins and associated structural components. Using Hidden Markov Models curated from validated proteins of model phages, we explored major tail tip types including Lambda-like, D3-like, MP22-like, PY54-like, and T5-like. Each type displays unique patterns of modularity in the organization of the tail tip components and enzymatic domains, reflecting both functional adaptation and evolutionary lineage. Functional annotations, structure prediction, sequence homology, and domain alignments uncover novel fusion and separation patterns. Additionally, host range analysis across siphophage tail tip clusters reveals a spectrum of target taxa highlighting phage-host co-evolution as a key driver of tail tip diversification. Our results illuminate the structural logic and modular evolution of siphophage tail tips and provide a foundational framework for refining phage taxonomy, exploring host specificity, and guiding synthetic phage engineering.

Keyword

Bacteriophage tail, host recognition mechanisms, tail tip complex, phage structural evolution

1. Introduction

Bacteriophages (phages), viruses that infect bacteria, exhibit an extraordinary diversity in their tail organization(Casjens & Molineux, 2012; Davidson et al., 2012; Leiman & Shneider, 2012), particularly at the distal end of the tail, where the interaction with the host cell surface occurs. Among tailed phages, siphophages - characterized by their long, flexible, non-contractile tails - represent a major group with significant ecological and biomedical relevance. The tail tip, composed of specialized structural proteins, is critical for host recognition, binding, and genome delivery, yet remains one of the most structurally and functionally variable and understudied components among phages (Linares et al., 2020).

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 ejection. 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, a thin peptidoglycan layer, and an inner membrane. 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 tips have been thoroughly characterized at the atomic level.

Tail assembly and morphogenesis have been most extensively characterized in bacteriophage Lambda. Structural and genetic studies have determined the molecular organization of Lambda's tail with high precision and accuracy (Casjens & Hendrix, 1974; Tsui & Hendrix, 1983; Wang et al., 2024). Recent cryo-EM structures have extended this foundational knowledge to other siphophages. In 2023, a cryo-EM study resolved the tail tip structure before and after DNA ejection in phage T5 infecting *Escherichia coli* (Linares et al., 2023). Phage Chi, another Gram-negative infecting siphophage, has also yielded high-resolution structural (Sonani et al., 2024). Cryo-EM analysis of Chi-like particles such as Gene Transfer Agent infecting *Rhodobacter capsulatus*, or RcGTA, (Bárdy et al., 2020) demonstrated the structural versatility and diversity among tail machinery for penetrating defenses of Gram-negative bacteria. Most recent addition to the collection is the near-atomic structure of the T-series siphophage T1 cryo-EM (Chen et al., 2025), revealing the architecture of its head, connector complex, flexible tail tube, and cone-shaped tail tip, and uncovering conserved structural features shared with Lambda-like siphophages that shed light on their assembly and function.

Despite these recent advances, our understanding of siphophage tail tip architecture remains fragmented, particularly in terms of the modular diversity and evolutionary logic underlying these structures. While high-resolution studies of Lambda, T5, Chi, and T1 have illuminated individual examples, there is currently no unifying framework that categorizes the structural motifs - such as central fiber domains and other mosaic tail tip components - across siphophages infecting Gram-negative bacteria. The functional implications of this modularity, especially in host recognition, penetration strategies, and DNA delivery, remain largely unexplored. Moreover, the extent to which these tail tip components are conserved, recombined, or specialized among different phage lineages is poorly understood. These gaps highlight the need for a comparative structural perspective that can synthesize emerging data into coherent patterns. Our study addresses this need by systematically surveying the known structural types and proposing a framework for understanding tail tip architecture as a modular and evolutionarily dynamic system.

The current study focuses on exploring mosaic design of siphophage Gram-negative tail tips by integrating insights from comparative genomics, protein domain architecture, structural prediction tools such as AlphaFold/ESMFold, and the cryo-EM modeling results. We aim to identify conserved modules and lineage-specific innovations that underpin host interaction. Our approach combines computational modeling with synteny analysis and functional annotation to propose a structural framework for exploring this complex 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.

The current study aims to systematically explore and classify tail tip architectures across a curated dataset of 467 Gram-negative siphophages. The core of the dataset includes genomes annotated using tail morphology as a criterion in the ICTV Master Species List 2020. Although subsequent taxonomy revisions have removed this classification, the legacy dataset offers a unique snapshot of structural diversity that can be reanalyzed with modern bioinformatic tools. By leveraging hidden Markov models (HMMs) built from canonical tail tip proteins of model phages, we define and categorize major types of tail tips: Lambda-like, T5-like, MP22-like, PY54-like, and D3-like. Each of these types represents a distinct “puzzle piece” in the evolutionary and structural landscape of siphophage infection mechanisms.

This work establishes a foundational framework for interpreting the structural and functional diversity and versatility of phage tail tips and underscores the importance of modularity and versatility in phage evolution. It also opens new directions for phage classification, synthetic biology, and therapeutic design by revealing the architectural logic behind tail tip specialization.

2. Results

2.1.Major Tail Tip Types Among Gram-Negative Siphophages

Our comparative analysis reveals that Gram-negative siphophages possess several major tail tip types, corresponding to the following structural paradigms: Lambda-like (Sipho-1), D3-like (Sipho-2), PY54-like (Sipho-3), MP22-like (Sipho-4), and T5-like(Sipho-5). These types represent recurring frameworks across diverse phages infecting Gram-negative hosts.

- **Lambda-like (Sipho-1):** Prototypical examples in the lamboid supercluster include phages *Lambda*, *ES18*, and *HK97*, which define subtypes 1a, 1b, and 1c, respectively.
- **D3-like (Sipho-2):** This type includes the lamboid phages such as *CobraSix* and *KPP5665-2*, which define subtypes 2a and 2b.
- **PY54-like (Sipho-3):** Represented in lamboid phages solely by *FSL_SP-016*.

- **MP22-like (Sipho-4):** Represented by Chi and R4C. Solved structures of this type are as follows JBD30, Gene Transfer Agent, and phage Chi.
- **T5-like (Sipho-5):** Prototypical example is phage T5. Prevalent among phages with large genomes (> 100 Kbp).

In a paradigm phage Lambda, tail tip types encode proteins to fulfill five key functional roles - DTN (gpM), THN (gpL), TNLP (gpK), THI (gpI), and CF(gpJ)**.** In other tail tip types, these functional roles can occur in distinct structural forms and domain organizations. These differences help define each type and subtype and reflect their unique structural and functional adaptations.

2.2. Defining key tail tip functions using phage Lambda as a reference

2.2.1. 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 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 unstable, and the initiator complex comes together and falls apart. Meanwhile, the gpGT is produced by a ribosome slippage resulting in attaching the gpT sequence 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, or tape measure protein, to 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 end of the tail tube that is the furthest from the tail tip becomes capped by a hexameric ring of gpU to prevent the further gpV polimerization that would lead to the tail overextension. The C-terminal part of gpH is cleaved in a coordinated fashion with the gpU capping that might pass a signal that the extension stage is finished, and the tail assembly process has entered the final stage of tail completion. This stage ensures the tail is complete and 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.

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

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

The DTN 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, DTN forms a hexameric ring that serves as a “strap belt” that ensures the correct and secure attachment of other tail tip proteins downstream. Previous research indicates that DTM proteins can serve as a utility belt that provides support to the dynamic functions of side fibers and receptor-binding proteins during infection.

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

The THN 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, THN is involved in the recruitment of TNLP and facilitates the docking of CF, a multi-domain protein that often includes the primary receptor-binding domain, e.g. in phage Lambda. The previous mutational analysis has demonstrated that the absence of THN (gpL) in Lambda disrupts tail assembly and prevents successful host infection. This indicates THN's indispensable role in maintaining the integrity of the tail structure.

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

The TNLP is another important player 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. It exhibits a two-domain organization with distinct structural and functional roles. The N-terminal domain adopts the Mov34-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 penetration. This domain is thought to mediate localized digestion of the bacterial peptidoglycan layer during infection, aiding in the ejection 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.

2.2.5. 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, THI 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.

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

The determination of tail length in bacteriophages of the Sipho-like and also Myo-like families is primarily governed by the TM protein. This regulatory role has been elegantly demonstrated through detailed studies on bacteriophage Lambda and later several other phages 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.

2.2.7. CF = Central Fiber (gpJ): Framework for modular assembly and host interaction

Among all tail tip components, CF is the most functionally critical for tail tip organization. This protein in phage Lambda (gpJ) mediates the initial interaction between phage Lambda and its bacterial receptor, LamB, a maltose porin located on the outer membrane of *E. coli*. Structural studies indicate that gpJ consists of multiple domains 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 ejection.

2.3. Anatomy of Central Fiber

The central fiber (CF) is a critical structural component of siphophage tail tips, where it plays an essential role in host interaction and infection initiation. This long structure extends from the distal end of the tail shaft and is typically composed of a repeating arrangement of tail-associated proteins organized into a helical or pseudo-helical architecture. Central fibers function as a mechanical and molecular bridge between the phage particle and the bacterial cell surface, frequently terminating in specialized receptor-binding domains that mediate the initial attachment to host receptors. The modular nature of central fiber proteins, combined with the structural diversity of their terminal regions, contributes to the host specificity of the phage and represents a key target for engineering efforts aimed at redirecting phage tropism. Despite their functional importance, the genetic and structural determinants governing central fiber assembly and receptor binding remain poorly characterized, underscoring the need for integrated computational and experimental approaches to systematically investigate their organization and evolutionary diversity across phage lineages.

The CF of phage lambda, encoded by gene J (protein gpJ), is a trimeric tail tip protein essential for host recognition and DNA delivery. Structurally, gpJ comprises a conserved core and variable peripheral domains. The core of gpJ is responsible for trimerization and structural integrity, facilitating the proper assembly of the tail tip complex. This core ensures the CF stability during infection. Also, the core domains form the structural and functional backbone of the tail tip, enabling host recognition and initiating infection. These core domains are conserved across many siphophages, providing a scaffold that supports diverse peripheral or variable modules (**Fig.1**). The description of core and variables domains is shown in **Table 1**.

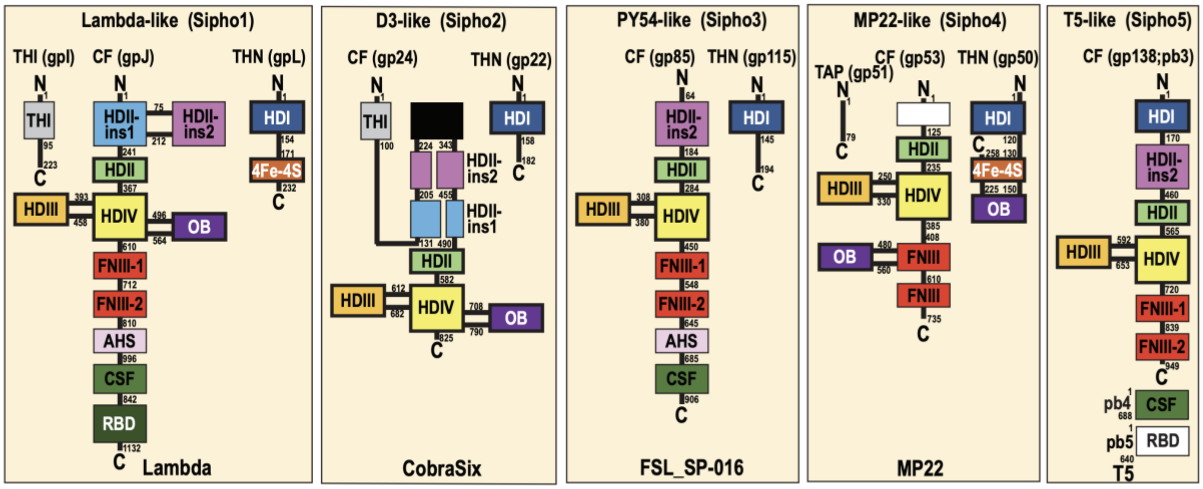
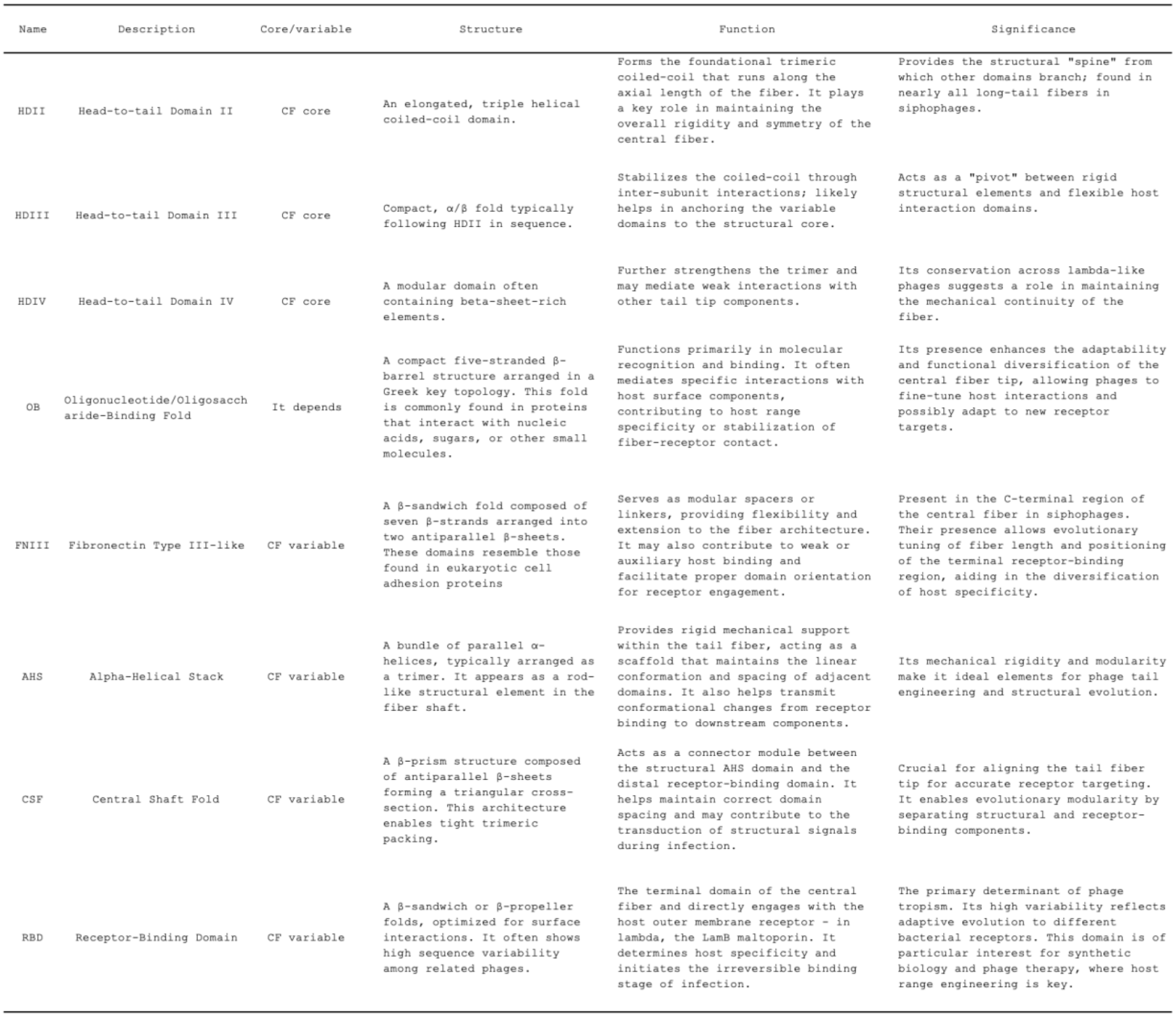


Fig.1. Comparison between the five major tail tip types with respect to the CF, THN, and THI organization.

Table 1. Description of CF core and variable domains.



Name	Description	Core/variable	Structure	Function	Significance
HDII	Head-to-tail Domain II	CF core	An elongated, triple helical coiled-coil domain.	Forms the foundational trimeric coiled-coil that runs along the axial length of the fiber. It plays a key role in maintaining the overall rigidity and symmetry of the central fiber.	Provides the structural "spine" from which other domains branch; found in nearly all long-tail fibers in siphophages.
HDIII	Head-to-tail Domain III	CF core	Compact, α/β fold typically following HDII in sequence.	Stabilizes the coiled-coil through inter-subunit interactions; likely helps in anchoring the variable domains to the structural core.	Acts as a "pivot" between rigid structural elements and flexible host interaction domains.
HDIV	Head-to-tail Domain IV	CF core	A modular domain often containing beta-sheet-rich elements.	Further strengthens the trimer and may mediate weak interactions with other tail tip components.	Its conservation across lambda-like phages suggests a role in maintaining the mechanical continuity of the fiber.
OB	Oligonucleotide/Oligosaccharide-Binding Fold	It depends	A compact five-stranded β-barrel structure arranged in a Greek key topology. This fold is commonly found in proteins that interact with nucleic acids, sugars, or other small molecules.	Functions primarily in molecular recognition and binding. It often mediates specific interactions with host surface components, contributing to host range specificity or stabilization of fiber-receptor contact.	Its presence enhances the adaptability and functional diversification of the central fiber tip, allowing phages to fine-tune host interactions and possibly adapt to new receptor targets.
FNIII	Fibronectin Type III-like	CF variable	A β-sandwich fold composed of seven β-strands arranged into two antiparallel β-sheets. These domains resemble those found in eukaryotic cell adhesion proteins	Serves as modular spacers or linkers, providing flexibility and extension to the fiber architecture. It may also contribute to weak or auxiliary host binding and facilitate proper domain orientation for receptor engagement.	Present in the C-terminal region of the central fiber in siphophages. Their presence allows evolutionary tuning of fiber length and positioning of the terminal receptor-binding region, aiding in the diversification of host specificity.
AHS	Alpha-Helical Stack	CF variable	A bundle of parallel α-helices, typically arranged as a trimer. It appears as a rod-like structural element in the fiber shaft.	Provides rigid mechanical support within the tail fiber, acting as a scaffold that maintains the linear conformation and spacing of adjacent domains. It also helps transmit conformational changes from receptor binding to downstream components.	Its mechanical rigidity and modularity make it ideal elements for phage tail engineering and structural evolution.
CSF	Central Shaft Fold	CF variable	A β-prism structure composed of antiparallel β-sheets forming a triangular cross-section. This architecture enables tight trimeric packing.	Acts as a connector module between the structural AHS domain and the distal receptor-binding domain. It helps maintain correct domain spacing and may contribute to the transduction of structural signals during infection.	Crucial for aligning the tail fiber tip for accurate receptor targeting. It enables evolutionary modularity by separating structural and receptor-binding components.
RBD	Receptor-Binding Domain	CF variable	A β-sandwich or β-propeller folds, optimized for surface interactions. It often shows high sequence variability among related phages.	The terminal domain of the central fiber and directly engages with the host outer membrane receptor - in lambda, the LamB maltoporin. It determines host specificity and initiates the irreversible binding stage of infection.	The primary determinant of phage tropism. Its high variability reflects adaptive evolution to different bacterial receptors. This domain is of particular interest for synthetic biology and phage therapy, where host range engineering is key.

The OB domain (oligonucleotide/oligosaccharide-binding fold) is also a part of the conserved CF core in phage Lambda. In phages, it has been co-opted for diverse protein-protein or protein-host interactions. However, Chi-like phages (MP22-like, Sipho-4) include an OB-fold domain at the C-terminus of THN, suggesting an accessory role in host recognition or structural stabilization. In Lambda, the canonical CF ends in fibronectin-like and receptor-binding domains, but in some phage variants, OB domains appear as additional modules, often tailored to specific host interactions. It is not present in all phages, nor is it essential for the CF backbone. When present, it adds functional specificity, likely involved in fine-tuned host binding or adapting to new host receptors.

The core domains work together to form a trimeric, elongated rod-like structure that supports the C-terminal variable domains (such as AHS, CSF, and RBD in phage Lambda), which are responsible for host receptor binding. The HDII–HDIV domains are highly conserved in Lambda-like Sipho-1 phages and are key to tail fiber assembly, strength, and alignment with the rest of the tail machinery. Extending from the core are variable domains that mediate host interactions. The C-terminal region of gpJ includes fibronectin type III (FNIII) domains, an alpha-helical stack (AHS), a central shaft fold (CSF), and a receptor-binding domain (RBD). The FNIII domains provide structural support, while the AHS stabilizes the trimeric structure. The CSF, a mixed β-sheet prism, connects the AHS to the RBD, which directly interacts with the LamB receptor of *Escherichia coli*.

Upon binding to LamB, gpJ undergoes conformational changes that facilitate DNA ejection into the host cell. These structural rearrangements are crucial for the transition from reversible to irreversible binding, ensuring successful infection. In summary, the domain organization of CF integrates structural versatility with functional specificity, enabling bacteriophages to effectively recognize and infect their host.

2.4. Summary statistics of main tail tip types

Our analysis of 467 siphophages infecting Gram-negative hosts revealed five major categories of tail tip structures, reflecting both conserved modules and lineage-specific innovations. The most prevalent type was the Lambda-like tail tip, found in 187 phages (40%), underscoring its widespread use among siphophages. The T5-like group was the second most common, comprising 84 phages (18%), followed closely by the MP22-like group with 81 phages (17%), 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%), expanding our understanding of this unique structural organization. Less common was the PY54-like group, present in 13 phages (3.00%), likely reflecting a more specialized evolutionary niche. This tail tip type is sharing some features with T5-like type (but it has distinct characteristics), and it is found among phages with genome length not exceeding 100 Kbp. Notably, 59 phages (~12.6%) did not fall into the main tail tip types. Among them, there were 16 phages of Proteobacteria and 43 phages with hosts beyond Proteobacteria (see **Section 4.** for detail).

2.4.1. Lambda-like (Sipho-1) Tail Tips

The Lambda-like tail tip type, the most prevalent structural class among the siphophage genomes analyzed, exhibits a high degree of conservation in functional organization. Defined by a consistent set of distinct proteins corresponding to the principal tail tip-associated functional groups, this type likely exemplifies a mainstream in terms of evolutionary coherence and strategic efficiency in host infection.

At the core of this conserved architecture are genes typically encoded in a contiguous arrangement for DTN, THN, TNLP, and THI. The CF component is especially noteworthy, displaying a distinct domain organization. In the prototypical member, bacteriophage Lambda, the N-terminal region of the CF protein includes insertions such as HDII-ins-1 and HDII-ins-2, followed by four core domains: HDII, HDII, HDIV, and OB, while the C-terminal end carries two fibronectin type III domains, FNIII-1 and FNIII-2, along with other variable domains consisting of AHS, CSF, and RBD. These structural elements collectively shape a tail tip complex tailored for precise host recognition and genome delivery.

This machinery is finely orchestrated: in Lambda, gpM and gpL form part of the distal tail structure, and gpK and gpI complete the tail initiator complex. This arrangement, echoed across the Lambda-like phages, underlies a conserved infection strategy and reflects a constrained evolutionary path compared to other siphoviral lineages.

For DNT among Lambda-like phages, there were 187 proteins identified as DTN based on HMM hits and synteny (179 HMM and synteny; 8 synteny and curation). There proteins were separated into four clusters by MMseq2 (0 0). Two large clusters: one included DTN of phage T1 and the other included DTN of phage Lambda, and two 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 and their structural representations are shown on **Fig. 2.**

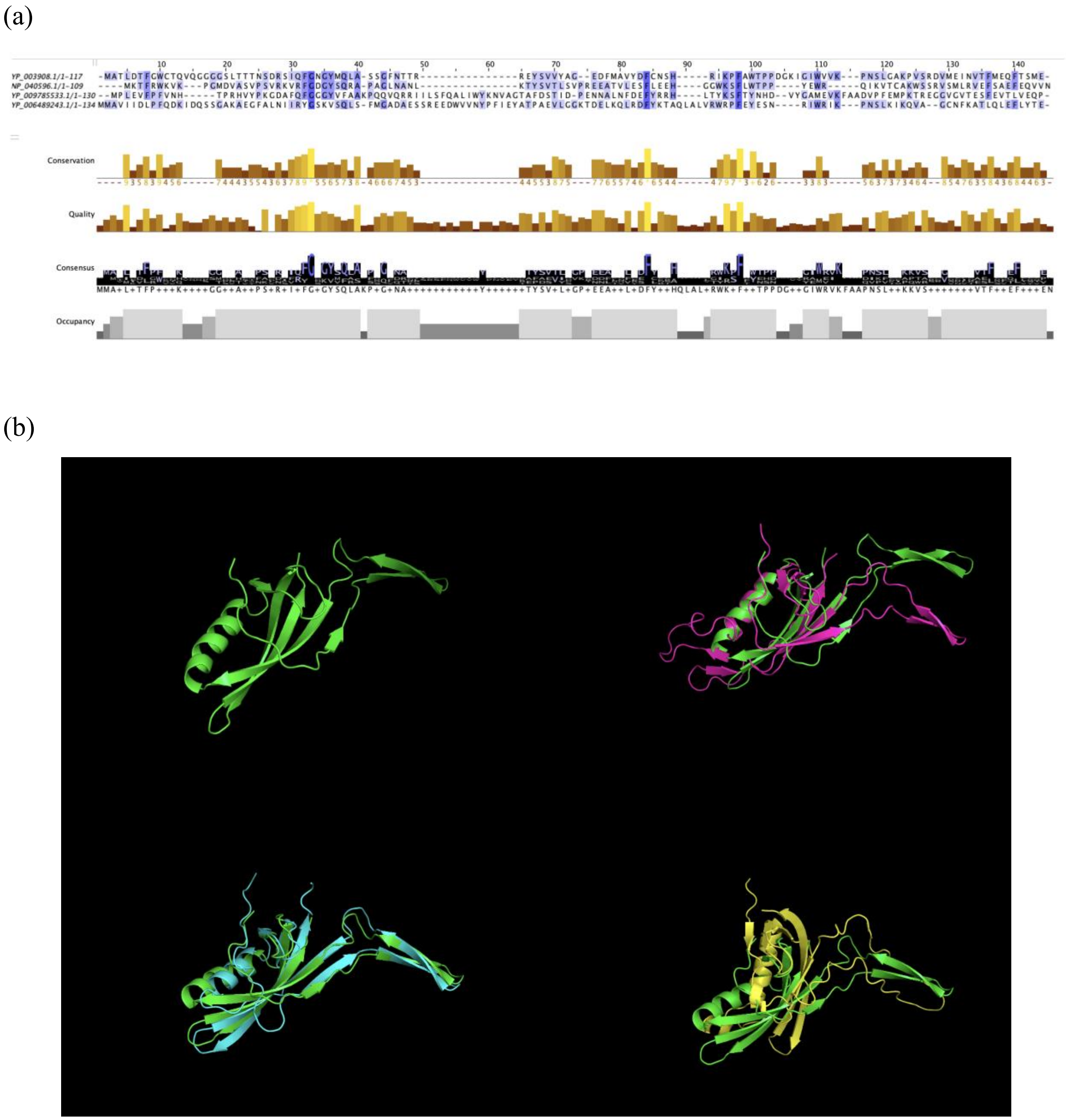
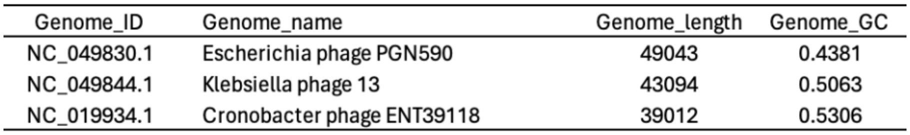


Fig.2. (a) MSA for the four DTN representative sequences in the Lambda-like phages: T1, Lambda, Seuss, and 9A (Jalview, colored by percent identity); (b) DTN representative sequence structures: DTN Lambda (green) – top left, DTN Lambda (green) aligned with DTN T1 (cyan) – bottom left,   DTN Lambda (green) aligned with DTN Seuss (magenta) – top right, and DTN Lambda (green) aligned with DTN 9A (yellow).

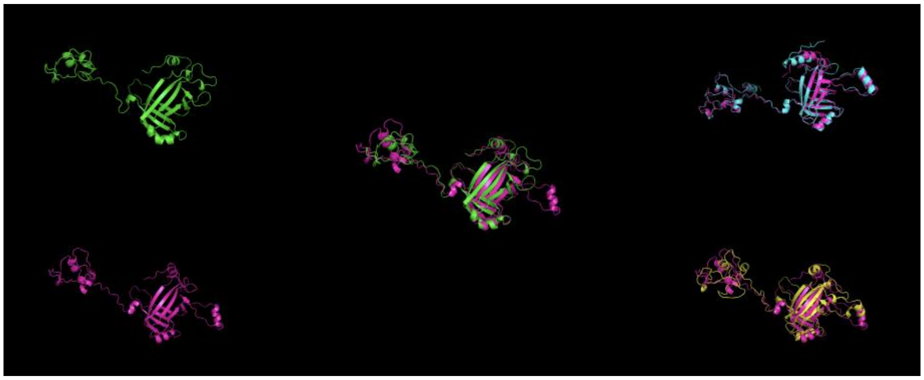
There were three phages that were missing apparent DTN ORF in their GenBank files (**Table 2**.). In these cases, TM ORF was followed by THN ORF without a separate ORF detected between them. It was possible that DTN role had been played by another protein or a domain fused to another protein. However, TM proteins of these three phages without apparent DTN had 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.3(a)**), and THN of phage Lambda was missing that small helix. It was very unlikely that this small structural change could compensate for the lack of DTN.

Table 2. 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

(a)



(b)

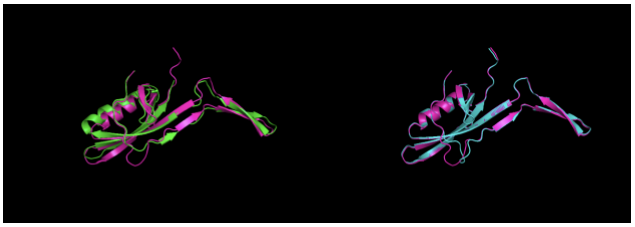


Fig.3. (a) Comparison between THN of Klebsiella phage 13 (magenta, bottom left) and Lambda THN (green, top left), and their aligned structures are shown in the middle; and the pairwise alignments of THN of Klebsiella phage 13 (magenta) and THNs of the remaining two phages with "no apparent DTN": Escherichia 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 (magenta) 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 its misannotation. The length of the region was consistent with housing DTN, and the predicted structure of this region translated into protein aligned well with DTN of Lambda and Klebsiella phage Shelby (**Fig.3(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 pairwise alignments of the predicted structures for the three representative THNs are shown on **Fig. 4**.

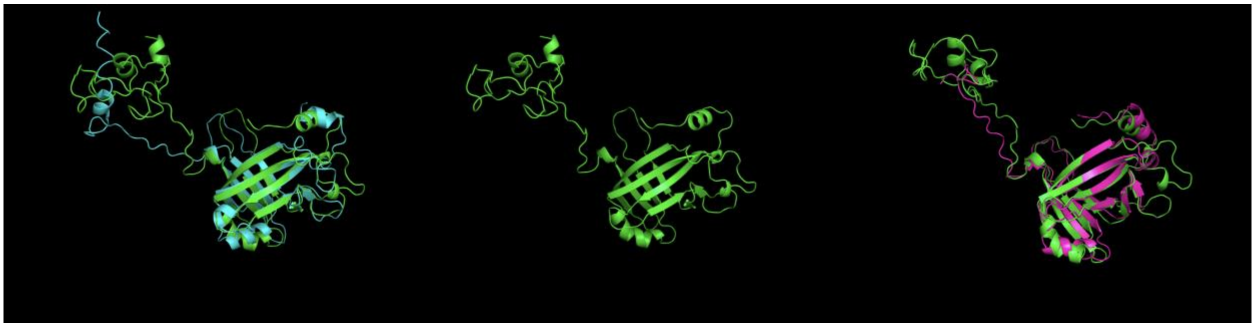


Fig.4. Lambda THN (green) is shown in the middle; Lambda THN aligned with THN of Colwellia phage 9A(cyan) – left; and Lambda THN (green) aligned with THN of Caulobacter phage Seuss (magenta) – right.

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. 5.**) forming a protein of 338 aa long (with usual TNLP in the Lambda-like phages having length of approximately 200 aa).

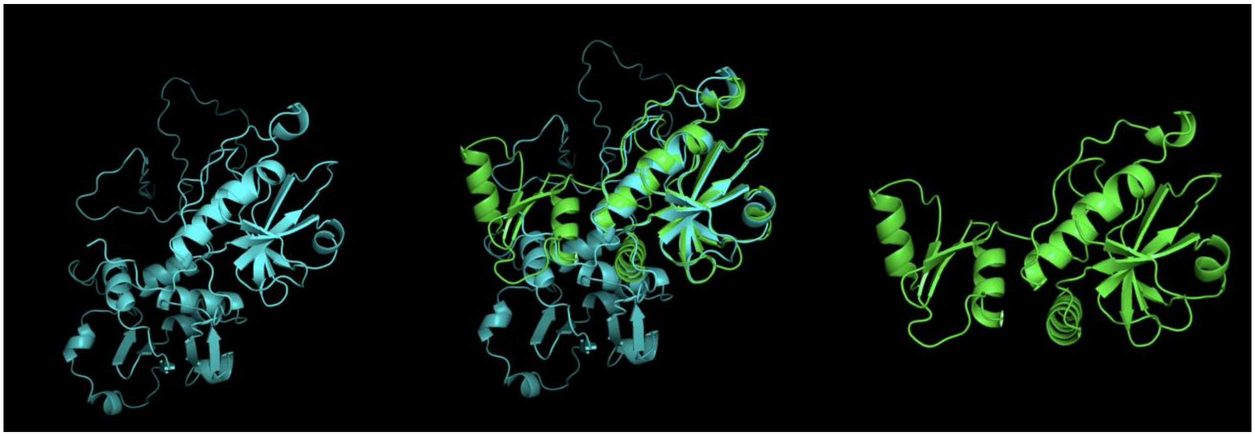


Fig.5. TNLP fused to the N-terminal end of THI of Citrobacter phage HCF1 (cyan, left); Lambda TNLP (green, right); and the alignment between Lambda TNLP and Citrobacter phage HCF1 TNLP fused to THI (in the middle).

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 exceptions, 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 several proteins. We found four genomes with confident CF HMM hits present in two separate proteins (NC_029071, NC_048150, NC_049838, and NC_049942).

In conclusion, the widespread presence, genetic stability, and deep historical study of Lambda-like phages have established them as foundational models in phage biology allowing us to employ Lambda-like organization as a benchmark for interpreting tail tip diversity across siphoviruses.

2.4.2. D3-like (Sipho-2) Tail Tips

The D3-like tail tip type defines a distinct lineage within siphoviruses, characterized by a tailored set of tail tip proteins that diverge notably from the classical Lambda-like organization. Accounting for approximately 9% of the analyzed genomes, this type exhibits a modular architecture reflective of both evolutionary divergence and functional conservation. Core components such as DTN and THN are conserved and align well with those found in the canonical D3 phage, as confirmed through profile Hidden Markov Models (HMMs). However, beyond this conserved foundation lies a more complex and dynamic evolutionary narrative. There are 43 phages in total that are D3-like.

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” during genome sequencing and assembling process.

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. 6.** 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.

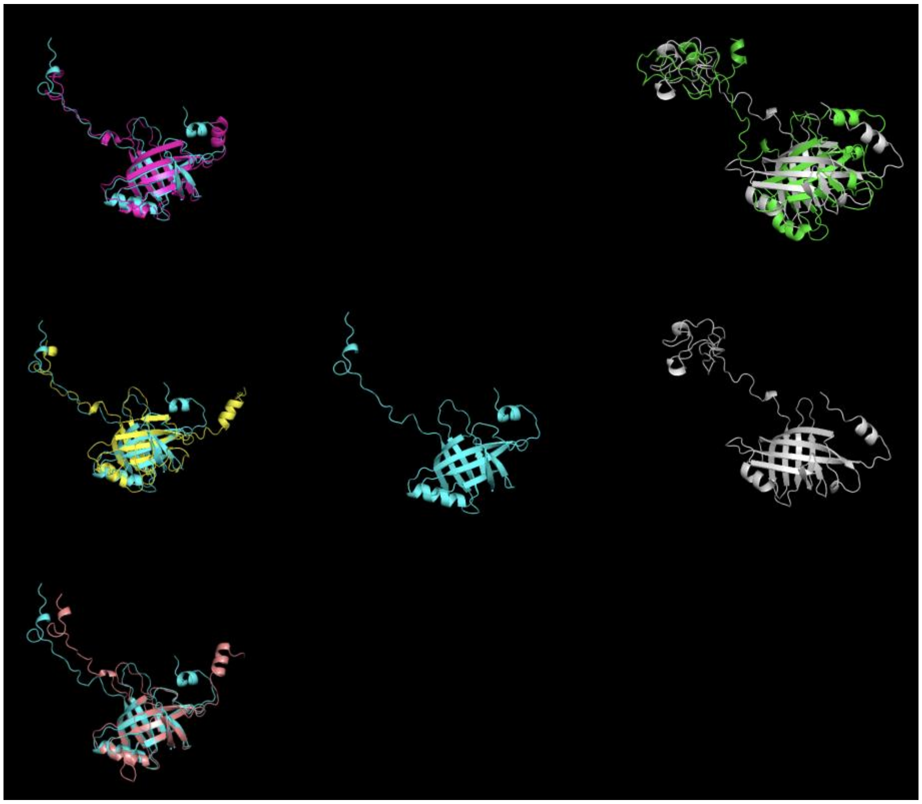


Fig. 6. THN of Pseudomonas phage D3 (cyan) is shown at the center; comparison of D3 THN (cyan) with Xp10 (magenta, top left), RcapMu (yellow, center left), and JWX (salmon, bottom left); THN of Pseudomonas phage nickie (grey, center right); nickie THN (grey) is compared with Lambda THN (green) – top right (visualized and aligned in PyMOL 3.1.1.; the structures are predicted by ESMFold).

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.7.**).

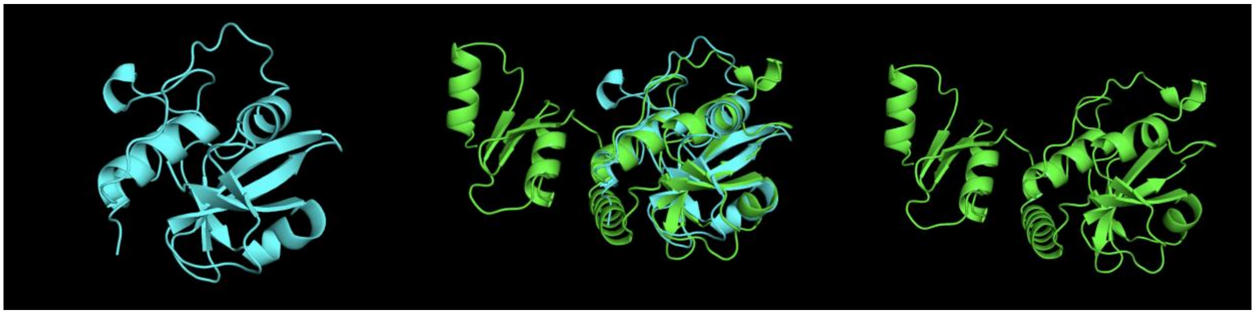


Fig.7. The comparison (in the middle) of D3-like TNLP from Pseudomonas phage nickie (cyan, left) and Lambda TNLP (green, right) (visualized and aligned in PyMOL 3.1.1.).

This significant sequence divergence renders standard Lambda gpK HMMs not suitable for detecting all homologs within this group, leading to the development of a D3-specific HMM. Yet even with this refined model, not all candidate TNLPs are captured, hinting at rapid sequence evolution paired with underlying functional conservation. Despite the challenges in detection, synteny and comparative genomic analyses consistently point to the presence of a gpK-like gene across the D3-like phages, suggesting a conserved role embedded within an adaptable framework.

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 about 100 residues long). This THI-related domain aligns with the N-terminal part of Lambda THI (**Fig. 8**.).

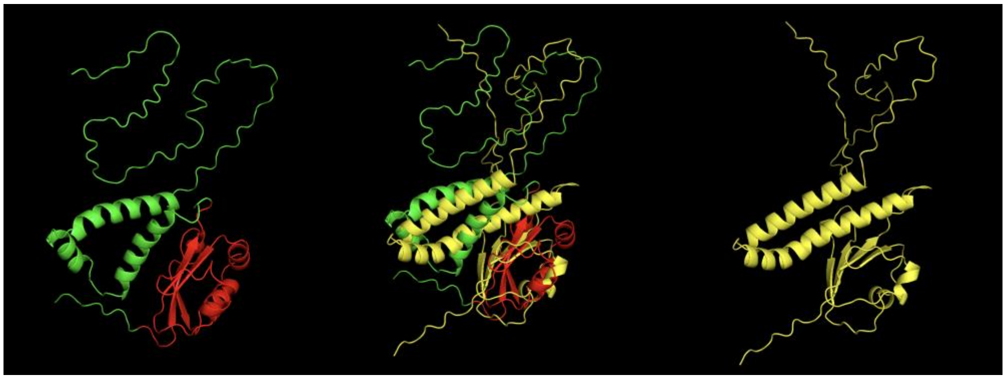


Fig.8. The comparison (center) between Lambda THI (green-red, left) and the N-terminal domain of CF from D3-like phage Pseudomonas phage nickie (yellow, right). The red color indicates the N-terminal part of THI that is similar to the THI domain of CF for D3-like phages (the structures are modeled using ESMFold).

For CF, D3-like phages share the same core domain organization with Lambda-like phages. Also, they also share similarity among the variable domains, HDII-ins1 (cyan) and HDII-ins2 (magenta) (**Fig.9**). Although subtle distinctions are evident (**Fig.10**), the continued use of cyan and magenta in the Lambda and CobraSix diagram is considered appropriate, given the close relationships that the same colors are intended to represent. The HII-ins2 domain in CobraSix is particularly notable, characterized by the presence of a non-contiguous strand that replaces one of the lambda strands

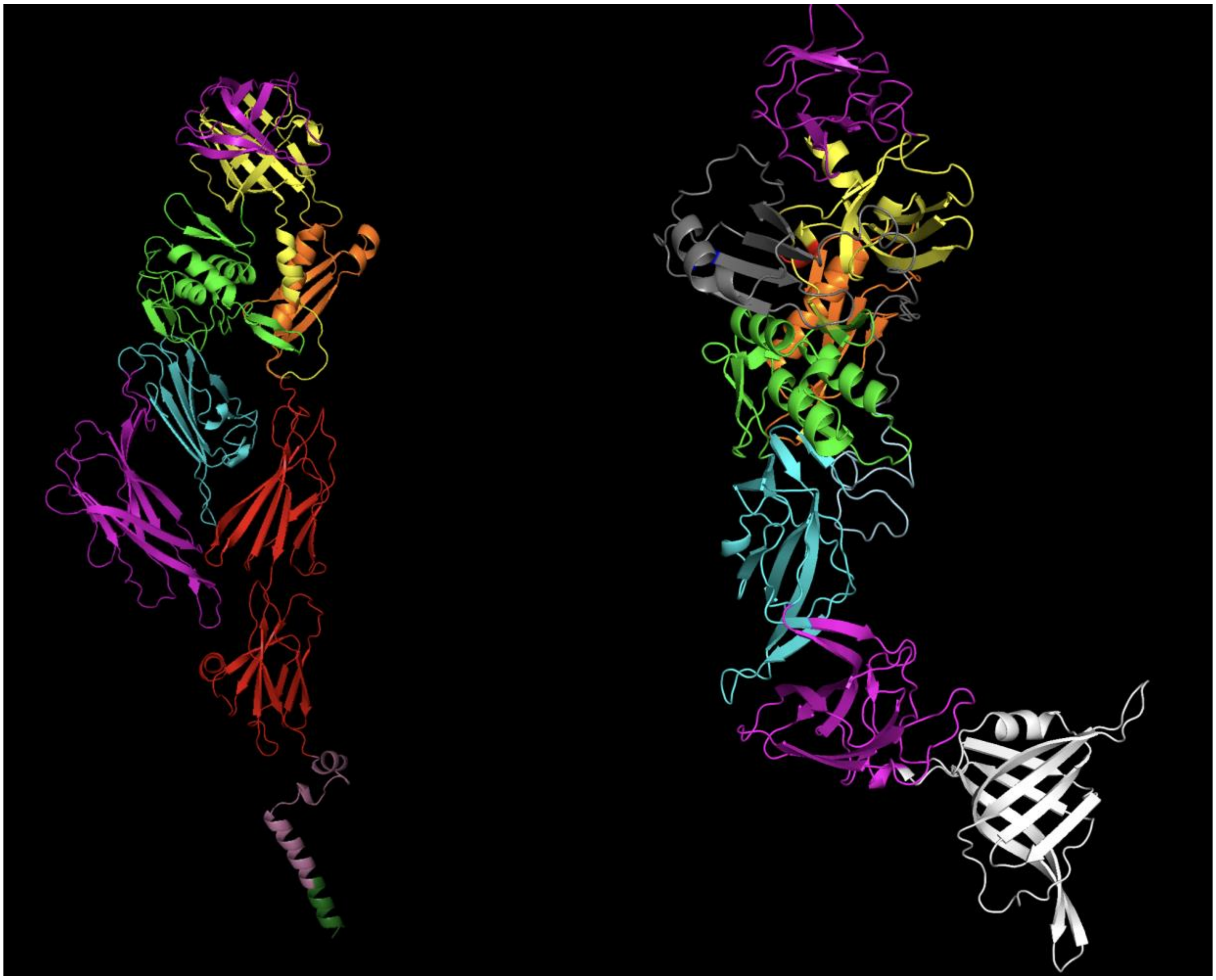
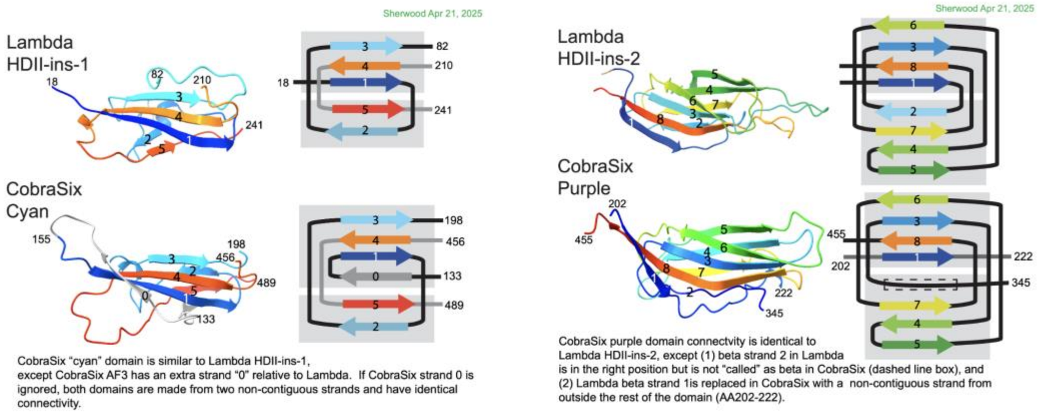


Fig.9. CF Lambda (left) and CF D3-like phage CobraSix(right).



CobraSix "cyan" domain is similar to Lambda HDII-ins-1,
except CobraSix AF3 has an extra strand "0" relative to Lambda. If CobraSix strand 0 is ignored, both domains are made from two non-contiguous strands and have identical connectivity.

CobraSix purple domain connectvity is identical to Lambda HDII-ins-2, except (1) beta strand 2 in Lambda is in the right position but is not "called" as beta in CobraSix (dashed line box), and (2) Lambda beta strand 1is replaced in CobraSix with a non-contiguous strand from outside the rest of the domain (AA202-222).

Fig.10. Side-by-side comparison of HDII-ins-1(left) and HDII-ins-2 (right) between phage Lambda and D3-like phage CobraSix.

In summary, the D3-like tail tip type, present in approximately 9% 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.

2.4.3. PY54-like (Sipho-3) Tail Tips

The PY54-like tail tip type represents a small but distinct lineage within siphoviruses, comprising approximately 3% of the analyzed genome collection (13 phages). This group is defined by a conserved core of tail proteins - specifically DTN and THN - that exhibit strong homology to those of the Yersinia phage PY54, as revealed by profile Hidden Markov Model (HMM) analyses.

Alongside these core components, PY54-like phages encode a TNLP, homologous to the C-terminal domain (CTD) of Lambda gpK. Although sequence divergence limits full-length alignment, the conserved C-terminal similarity suggests that this gpK-like protein retains essential structural and functional roles in tail tip formation and possibly in DNA ejection.

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 two exceptions that formed singleton clusters: DTN from Rhizobium phage 16-3 (YP_002117577.1) and Paracoccus phage Shpa (YP_009593493.1). These two proteins were longer than other DTNs among the PY54-like phages on average (~200 aa).

For THN, 10 of 13 protein sequences were clustered with PY54 THN. The remaining 3 sequences were split into two clusters with one of these clusters being a singleton. For TNLP, the 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 likely fused to the C-terminus of CF based on the HMM analysis (~first 50 residues of the fused protein). Also, CFs from other phages are clustered neatly with the PY54 CF (including the potential fused region) indicating that a similar fusion event is likely relevant to other PY54-like phages.

Further supporting a modular tail tip organization, PY54-like phages demonstrate a streamlined architecture that deviates from the Lambda paradigm. Notably, these phages lack a separately encoded THI gene. However, their CF proteins appear to compensate for this absence: while they do not possess a distinct N-terminal THI-like domain, they do contain an unfolded region predicted by AF3 structural models. This region may serve a functional surrogate role in tail tip assembly or stabilization.

Interestingly, the arrangement of non-core domains within these CF proteins mirrors those observed in Lambda-like phages, pointing to partial conservation of tail tip architecture. This pattern - conservation of structural features amid genetic abbreviation - underscores the evolutionary plasticity of tail modules. The PY54-like phages, though few in number, offer a compelling case study of how siphoviral tail tips can evolve along alternative trajectories while maintaining functional integrity. Their unique configuration may hold insights into novel host-interaction strategies and the broader adaptability of phage morphogenesis.

2.4.4. MP22-like (Sipho-4) Tail Tips

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 and Salmonella phages. MP22-like phages often display variable modular arrangements that include fused or multi-domain tail components, indicative of structural streamlining or functional innovation. Unlike other common tail tip types, this tail type is not observed among the lamboid phages and is exemplified by phages such as Chi, JBD30, and certain gene transfer agents (GTAs).

Within the MP22-like phages, the DTN protein is unusually large ~560 amino acids and contains two gpV-like beta-sandwich domains rather than one, followed by a C-terminal beta-barrel domain. This unique arrangement results in a DTN hexameric ring that effectively behaves as a trimeric structure - marking a departure from the canonical hexameric assemblies observed in Lambda-like, D3-like, and T5-like tail tips. Meanwhile, the THN protein also presents a distinct modular organization, incorporating a Lambda-like HDI domain, a central iron-binding motif, and a distinctive C-terminal OB-fold domain.

For DTN, DTN proteins among the MP22-like phages were split into two main clusters by MMseq2 (0 0): more than 75% of DTN clustered together with MP22 including Chi and JBD30 and less than one quarter of DTNs clustered together with 10 large Caulobacter phages having genome length above 200 Kbp. The first cluster had longer DTNs on average than the second cluster. Moreover, the arrangements of tail tip associated functional components were different in these clusters. Cluster 1 (larger cluster with MP22, Chi, and JBD30) had DTN-MP22, THN, THI, and CF as separate proteins (no apparent TNLP or a spare ORF between THN and THI). Cluster 2 (smaller cluster with large Caulobacter phages) had DTN-GTA, THN, TNLP as separate proteins and THI likely fused to CF as exemplified by the joint HMM CF and THI hits to the fused protein in Alphaproteobacteria phage PhiJL001 (NC_006938.1).

The CF domain arrangement houses a series of conserved domains - HDII, HDIII, and HDIV - followed by two FNIII domains. Strikingly, these FNIII domains are positioned similarly to those in the CF proteins of Lambda-like (Sipho-1), PY54-like (Sipho-3), and T5-like (Sipho-5) phages, establishing a bridge of structural similarity despite broader organizational divergence.

Altogether, the MP22-like tail tip type embodies a distinct evolutionary pathway in siphophage tail morphogenesis. Its unique modular configuration and innovative domain arrangement point to a lineage finely tuned for specific host interactions - offering new insights into the diversity and adaptability of phage infection mechanisms.

2.4.5. T5-like (Sipho-5) Tail Tips

The T5-like tail tip type on 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 ejection 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 84 phages have DTN protein sequences that are clustered neatly with T5 DTN and their lengths are very similar as well (~200 aa). The remaining phage required further analysis due to no apparent DTN candidate detected based on both synteny and HMM analysis. The results of the further analysis indicate that 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 detected by the HMM hits to the resulting fused protein by both CF HMMs and THN HMMs with low Evalues. The other T5-like phages 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 separate TNLP and THI proteins. However, the C-terminus of TM in T5 has a muralytic activity (Boulanger et al., 2008) that is likely compensate for this.

At the domain level, CFP in T5-like phages retains the three universal "core" domains - HDII, HDIII, and HDIV - found across most siphoviral CFs. However, uniquely among all tail tip types, the T5-like CFP also includes the HDI domain (a homolog of the N-terminal domain of lambda gpL) as its N-terminal region. This incorporation of HDI into the CFP, rather than its expression as a separate THN protein, is a defining structural hallmark of the T5-like type phages.

Together, these features - fused THN-CF architecture, absence of TNLP and THI, and distinctive CFP domain composition - underscore the unique evolutionary trajectory of the T5-like group. Their complex tail tip machinery exemplifies the architectural plasticity of phage design and supports their classification as a structurally and functionally distinct lineage within siphophages.

2.4.6. Roaming the Unknowns

In this section, we describe the further analysis of the unclassified phages with the main focus on exploring tail tip organization of phages infecting Proteobacteria. There were 16 phages of Proteobacteria marked as Unknown that could not be confidently assigned to the main tail tip types due to a variety of factors, including highly divergent sequences, genome mis-annotation, issues with sequencing and/or assembling, or other factors.

There were three phages that were misannotated in the metadata as siphophages, two of them were podophages, NC_047742.1 and NC_048113.1, and one was a myophage NC_019929.1. These phages were removed and not included in the subsequent analysis.

There was one likely T5-like phage, Aeromonas phage AhSzw-1 (NC_047950.1) with PY54-THN HMM hit with high E-value. Considering that in T5-like phages THN is fused to CF, we compared the protein with the HMM hit to the fused THN-CF protein of T5.

There were two likely Lambda-like phages with a missing ORF for THN (these ORFS were not detected in their GenBank files). In Escherichia phage ECP1 (NC_049926.1), there was a gap between DTN and TNLP that was large enough to house a THN ORF. In Escherichia phage vB_Eco_mar001J1 (NC_048206.1), the DTN ORF was located in the beginning of the GenBank file and the TNLP was located at the end indicating the anticipated THN ORF was likely "lost in translation" during genome sequencing and/or assembling process. Other tail tip associated proteins including DTN, TNLP, THI, and CF were present and similar to Lambda based on the results of the HMM analysis and synteny.

Also, there were three Shiga-toxin converting phages of Escherichia coli (NC_004914.3, NC_004913.3, and NC_008464.1) with the confident hits by the CF-Lam HMMs and lack of other defining clues from the HMM hits and synteny analysis.

In addition, there was one Caulobacter phage Sansa (NC_047756.1) with confident HMM hits for TM and CF proteins with three ORFs separating them and lack of other type-defining clues. In the dataset, there were 12 Caulobacter phages in total. Among those, 10 phages with large genomes (above 200 Kbp) were classified as MP22-like and 1 phage, Caulobacter phage Seuss (NC_047757.1) that was similar in length and GC-content to phage Sansa. Their TM proteins were confidently annotated by the HMM and synteny analysis and similar in length ~1800 aa. Phage Seuss was classified as Lambda-like type based on the HMM and synteny analysis. However, the synteny patterns of Seuss and Sansa did not match.

There were several other phages that did not fall into the main tail tip types. Importantly, Pseudomonas phage F-10 and Stenotrophomonas phage S1 that have been previously identified as distinct morphotypes that expand the developed tail tip framework beyond the mainstream types among siphophages, F-10 type was discovered by Dr. Davidson and S1-type by Dr. Lenskaia. Further exploration and detailed characterization of minor siphophage tail tip types are outside the scope of this study.

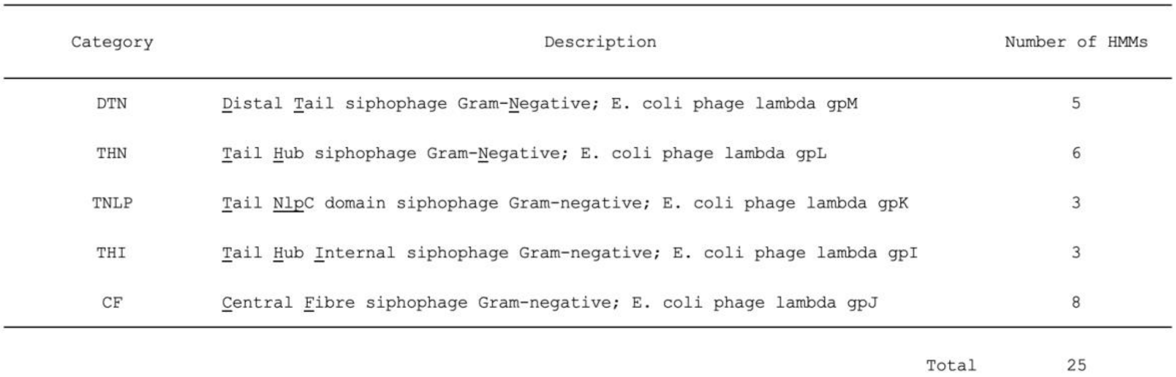
3. Materials and methods

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 B.**). These categories include as follows: 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 3. 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.11.**). 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.

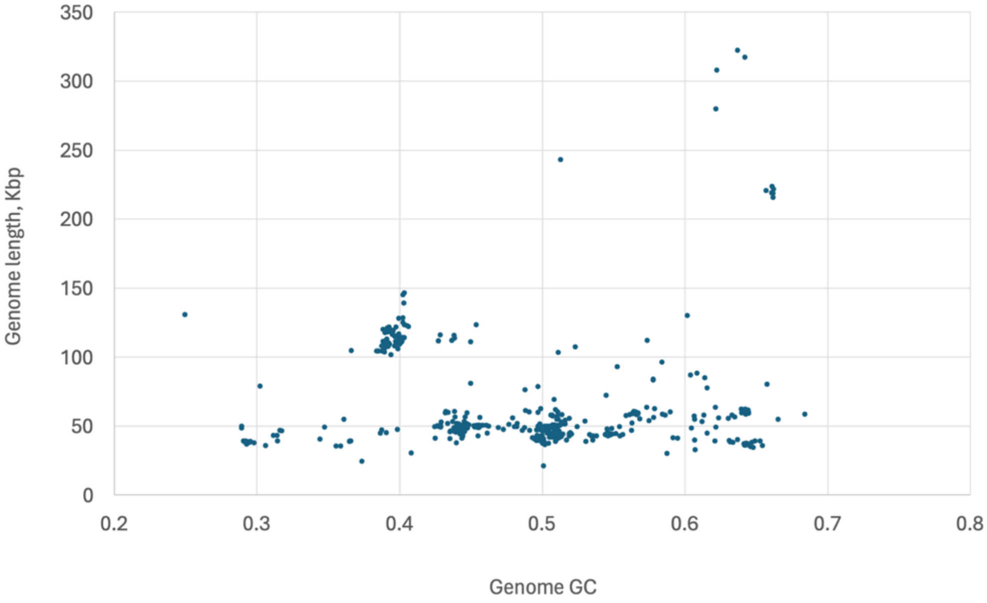
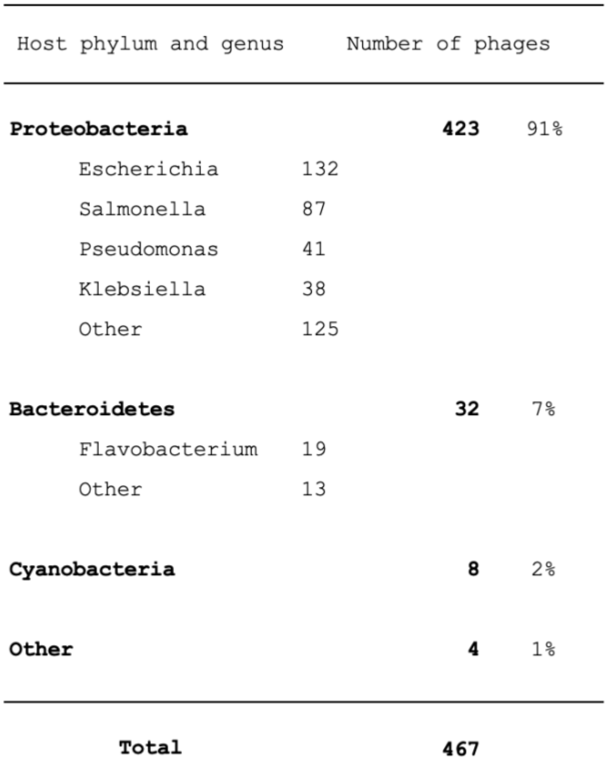


Fig. 11. Distribution of 467 phage sequences.

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 4**.) 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, Cyanobacteria, and even Thermophilic lineages, reflecting the ecological breadth and adaptive versatility of Gram-negative siphophages.

Table 4. 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 and conclusion

This study introduces a unified framework for exploring and understanding siphophage tail tip architectures in phages infecting Gram-negative bacteria, based on the CF modular organization and tail tip structures. By analyzing main structural types - corresponding to Lambda-like (Sipho-1), D3-like (Sipho-2), PY54-like (Sipho-3), MP22-like (Sipho-4), and T5-like (Sipho-5) - we reveal a layered landscape of structural variation and conservation that underpins the functional logic of phage-host interaction. Our domain-based annotation strategy, combining profile Hidden Markov Models (HMMs) with synteny analysis and structural predictions, allowed us to uncover deep evolutionary relationships and detect divergent and novel domain architectures across diverse of phage genomes.

Among the 423 phages infecting Proteobacteria of 467 Gram-negative siphophages in our dataset, we were able to classify over 96% into defined tail tip types using our computational approach based on HMM profile hits and synteny analysis. The remaining 16 of 423 phages -

fewer than 4% - could not be confidently assigned due to a variety of factors, including highly divergent sequences, genome mis-annotation, issues with sequencing and/or assembling, or other factors. These unclassified phages likely represent outliers or potentially novel structural variants that warrant further investigation. Nonetheless, the high rate of classification demonstrates the robustness and comprehensiveness of our current framework for tail tip type identification among Proteobacteria-infecting siphophages.

Key among our findings is exploring the variability and versatility among CF domain combinations and their impact on functional roles the Central Fiber and other tail tip components. These structures not only diversify functional targeting strategies but also exemplify the modular, mosaic nature of phage evolution. Our framework brings conceptual clarity to the patchwork of tail tip architectures by demonstrating that the arrangement and connectivity of domains - not just their presence - are central to function and evolutionary lineage.

For example, the absence of a separate THI in D3-like (Sipho-2), PY54-like (Sipho-3), and MP22-like (Sipho-4) phages correlates with major architectural shifts, including the relocation of N-terminal accessory folds to the CF and the fusion of structural roles into fewer gene products. This suggests a broader principle of domain economy in tail architecture, where evolutionary innovation often involves fusion, duplication, or truncation rather than de novo generation of structural motifs.

Importantly, the use of profile HMMs enabled the identification of highly divergent homologs, revealing lineage-specific elaborations and adaptations even within canonical tail types. However, this method remains constrained by the scope of current domain models and reference annotations. Highly novel, rare, or horizontally transferred elements may still elude detection, particularly in under-sampled environments or phage lineages. We propose that future work integrate structural prediction (e.g., AlphaFold), virion proteomics, and in situ cryo-EM analysis to validate and extend our bioinformatic predictions. Such approaches would not only refine existing classifications but also open up new vistas for experimental phage biology.

One emerging implication of this work is that CF modularity provides a latent combinatorial code for host specificity and adaptation. As shown in modular swaps between Lambda-like phages (Sipho-1) or between MP22-like CFs from Chi and related enterophages, structural shuffling of terminal or linker domains can result in altered host ranges or virulence profiles. This underscores the value of domain-level resolution not just for taxonomy, but also for synthetic biology and therapeutic applications of phages. Understanding the structure - function relationships in these tail architectures will be critical for engineering phages with predictable and tunable infectivity.

By contextualizing major siphophage tail tip architectures and analyzing their mosaic organization, we provide a blueprint for understanding the diversity, function, and evolution of these critical infection machines in phages targeting Gram-negative bacteria. Our framework reframes phage tail tip biology through the lens of modularity, where conserved structural cores are combined with variable domains that determine host specificity, infection mechanics, and evolutionary plasticity.

The classification scheme we present not only resolves long-standing puzzles surrounding phage types like D3 and MP22 but also establishes a reference scaffold for mapping new and hybrid variants. Their gene organization was unusual, with fusions and truncations of known tail proteins (like gpL and gpI), and their CF had a strange combination of domains. Consequently, it was difficult to assign them to a known tail category, and it was not clear if they had equivalents of known tail proteins like gpK or gpI. This made comparative and functional analysis difficult. Also, MP22-like phages (like Enterobacteria phage Chi or MP22) had tail tips that resembled Lambda in some ways, but making it unclear if they represent Lambda-like variants or a distinct structural group. For many years, researchers lacked a clear framework and terminology to describe and compare these phages.

Our study help resolve these puzzles by proposing a consistent classification system (e.g., Sipho-1 to Sipho-5), identifying conserved and variable components of tail tips using domain architecture, and clarifying how even divergent structures still follow modular principles. While acknowledging the possible unexplored diversity beyond the current siphophage data, our work serves as a foundational “living atlas” of phage tail tips - one that can be continuously expanded and refined as more genomes and experimental data become available.

In conclusion, the patchwork of CF domain architectures and other mosaic patterns are more than structural curiosities – they are records of evolutionary innovation and ecological adaptation. By decoding this elaborate tapestry, we take a significant step toward a systematic, functional taxonomy of phage infection machinery, with broad implications for microbiology, evolutionary biology, and the therapeutic deployment of phages.

References

Bárdy, P., Füzik, T., Hrebík, D., Pantůček, R., Thomas Beatty, J., & Plevka, P. (2020). Structure and mechanism of DNA delivery of a gene transfer agent. *Nature Communications, 11*(1), 3034. 10.1038/s41467-020-16669-9

Boulanger, P., Jacquot, P., Plançon, L., Chami, M., Engel, A., Parquet, C., Herbeuval, C., & Letellier, L. (2008). Phage T5 straight tail fiber is a multifunctional protein acting as a tape measure and carrying fusogenic and muralytic activities. *The Journal of Biological Chemistry, 283*(20), 13556–13564. 10.1074/jbc.M800052200

Casjens, S. R., & Hendrix, R. W. (1974). Locations and amounts of the major structural proteins in bacteriophage lambda. *Journal of Molecular Biology, 88*(2), 535–545. 10.1016/0022-2836(74)90500-2

Casjens, S. R., & Molineux, I. J. (2012). Short Noncontractile Tail Machines: Adsorption and DNA Delivery by Podoviruses. In M. G. Rossmann, & V. B. Rao (Eds.), *Viral Molecular Machines* (pp. 143–179). Springer US.

Chen, Y., Xiao, H., Zhou, J., Peng, Z., Peng, Y., Song, J., Zheng, J., & Liu, H. (2025). The In Situ Structure of T-Series T1 Reveals a Conserved Lambda-Like Tail Tip. *Viruses, 17*(3), 351. 10.3390/v17030351

Davidson, A. R., Cardarelli, L., Pell, L. G., Radford, D. R., & Maxwell, K. L. (2012). Long Noncontractile Tail Machines of Bacteriophages. In M. G. Rossmann, & V. B. Rao (Eds.), *Viral Molecular Machines* (pp. 115–142). Springer US.

Leiman, P. G., & Shneider, M. M. (2012). Contractile Tail Machines of Bacteriophages. In M. G. Rossmann, & V. B. Rao (Eds.), *Viral Molecular Machines* (pp. 93–114). Springer US.

Linares, R., Arnaud, C., Degroux, S., Schoehn, G., & Breyton, C. (2020). Structure, function and assembly of the long, flexible tail of siphophages. *Current Opinion in Virology, 45*, 34–42. 10.1016/j.coviro.2020.06.010

Linares, R., Arnaud, C., Effantin, G., Darnault, C., Epalle, N. H., Boeri Erba, E., Schoehn, G., & Breyton, C. (2023). Structural basis of bacteriophage T5 infection trigger and E. coli cell wall perforation. *Science Advances, 9*(12), eade9674. 10.1126/sciadv.ade9674

Sonani, R. R., Esteves, N. C., Scharf, B. E., & Egelman, E. H. (2024). Cryo-EM structure of flagellotropic bacteriophage Chi. *Structure, 32*(7), 856–865.e3. 10.1016/j.str.2024.03.011

Tsui, L., & Hendrix, R. W. (1983). Proteolytic processing of phage λ tail protein gpH: timing of the cleavage. *Virology, 125*(2), 257–264. 10.1016/0042-6822(83)90199-X

Wang, C., Duan, J., Gu, Z., Ge, X., Zeng, J., & Wang, J. (2024). Architecture of the bacteriophage lambda tail. *Structure, 32*(1), 35–46.e3. 10.1016/j.str.2023.10.006