Reviewing the main types of siphophage 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, ON, 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 experimentally 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 broad 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) that infect Gram-negative bacteria exhibit an extraordinary diversity in their tail organization, 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. Despite advances in cryo-electron microscopy and genomic annotations, the classification and comparative analysis of tail tip modules remain incomplete, particularly for siphophages infecting Gram-negative hosts.

The central fiber is a critical structural component of long-tailed phages, where it plays an essential role in host recognition 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 such as lipopolysaccharides or pili. The modular nature of central fiber proteins, combined with the structural diversity of their terminal domains, 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 architecture and evolutionary diversity across phage lineages.

The current study aims to systematically explore and classify tail tip architectures across a curated dataset of 467 Gram-negative siphophages, originally 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 fusion 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 well-defined 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 the following 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. Description of Central Fiber domain organization using phage Lambda as a reference

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, each contributing to its function. 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 (**Table A.**).

Table A. 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 gpL (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), 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 on *Escherichia coli*.

Upon binding to LamB, gpJ undergoes conformational changes that facilitate DNA injection into the host. These structural rearrangements are crucial for the transition from reversible to irreversible binding, ensuring successful infection. In summary, the domain organization of gpJ integrates structural stability with functional specificity, enabling bacteriophage Lambda to effectively recognize and infect its host.

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

The Lambda-like tail tip type, the most prevalent structural class among the 467 siphovirus genomes analyzed, exhibits a remarkable degree of conservation and functional sophistication. Defined by a consistent set of distinct proteins corresponding to the principal tail tip-associated functional groups, this type exemplifies 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 components in these phages are especially noteworthy, displaying a modular domain organization. The N-terminal region of the fiber protein includes insertions such as HDII-ins-1 and HDII-ins-2, while the C-terminal end carries two fibronectin type III domains, FNIII-1 and FNIII-2, along with accessory domains like AHS, CSF,

and RBD. These structural elements collectively shape a tail tip complex tailored for precise host recognition and genome delivery.

In the prototypical member, bacteriophage Lambda, this machinery is finely orchestrated: 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. Notably, these phages are also characterized by modular genome architectures that accommodate both lytic and lysogenic lifestyles, reinforcing their adaptability.

Altogether, 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.

2.2.2. 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 phage 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.

Central to this complexity is the presence of a TNLP, typically a homolog of gpK. In D3-like phages, gpK-like proteins share only partial homology with their Lambda counterparts, primarily restricted to the C-terminal region. This significant sequence divergence renders standard Lambda gpK HMMs inadequate 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 these phages, suggesting a conserved role embedded within an adaptable framework.

Further reflecting this adaptability, phages such as CobraSix and KPP5665-2 encode genes for DTN, THN, and TNLP, yet lack a distinct gene for THI. Instead, their CF proteins exhibit N-terminal regions with sequence and structural similarity to the N-terminal portion of THI from Lambda-like phages. Interestingly, these CF proteins harbor non-core domains exclusively positioned at the N-terminal end of the core region. Many of these domains adopt “fold-back” topologies, in which the polypeptide chain loops back on itself to form anti-parallel structures - an architectural feature that may support functional plasticity and enhance host interaction dynamics.

Altogether, the D3-like tail tip organization exemplifies a lineage marked by evolutionary innovation, where structural divergence and modular design converge to accommodate host-specific pressures. This unique configuration not only highlights the plasticity of phage morphogenesis but also holds implications for understanding phage-host specificity and developing phage-based therapeutics.

2.2.3. 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. 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.

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.

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

The MP22-like tail tip type, also referred to as Sipho-4, represents a major and structurally distinct lineage within the siphophage landscape, comprising approximately 17% of the analyzed genome collection. Unlike classical Lambda-related phages, this lineage is largely absent from the Lambda Supercluster and is instead exemplified by phages such as Chi, JBD30, and certain gene transfer agents (GTAs). These phages exhibit a unique tail tip organization, characterized by divergent homologs of gpM (DTN), gpL (THN), and gpI (THI), and a conspicuous absence of the TNLP component. Structural and cryo-EM studies of Chi-like phages and GTAs have further illuminated this distinctiveness, revealing novel protein folds and modular architectures not observed in Lambda-like or D3-like counterparts.

MP22-like tail tips are typically found in small, virulent siphophages infecting Enterobacteriales, including Escherichia and Salmonella. Within this group, the DTN protein (e.g., gp26) is unusually large at ~550 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 the Sipho-1, -2, and -3 tail tips. Meanwhile, the THN protein (gp27) presents a modular organization, incorporating a Lambda-like HDI domain, a central iron-binding motif, and a distinctive C-terminal OB-fold domain.

The THI protein (gp28), although relatively small, occupies a strategic position within the upper lumen of the tail tip, paralleling the location of its counterpart in Lambda-like phages and potentially playing a role in DNA ejection or tail stabilization. The central fiber (CF), encoded by

gp30, reinforces the structural innovation of this group. It houses a series of conserved domains - HDII-ins-2, 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 and PY54-like (Sipho-3) phages, establishing a bridge of structural similarity despite broader organizational divergence. Among known structures, the Sipho-4 CF shows the closest architectural resemblance to that of phage FSL_SP-016.

Altogether, the MP22-like tail tip type embodies a distinct evolutionary pathway in siphoviral tail morphogenesis. Its unique domain configurations, absence of TNLP, and conserved yet innovatively arranged CF components point to a lineage finely tuned for specific host interactions - offering new insights into the diversity and adaptability of phage infection mechanisms.

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

The T5-like tail tip type, also referred to as Sipho-5, defines a distinct and relatively prevalent structural lineage among siphoviruses with large genomes (>100 Kbp), representing approximately 18% of the analyzed genome collection. While these phages exhibit some homology to the PY54-like group - particularly a Distal Tail Needle (DTN) protein that clusters within the PY54 HMM family - their overall tail tip architecture diverges markedly. Most notably, the Tail Hub Needle (THN), homologous to gpL, is fused to the Central Fiber Protein (CFP), forming a chimeric structural module that is entirely unique among the five major tail tip types identified. This fusion likely reflects an evolutionary innovation designed to accommodate the functional demands of large-genome packaging and the more intricate infection mechanisms associated with these complex phages.

The prototypical member of this group, phage T5, is emblematic of this complexity, with a well-characterized two-step DNA injection system. In T5, pre-early genes are transcribed and expressed before the full genome is introduced into the host, illustrating a sophisticated regulatory strategy likely enabled by the structural innovations in its tail machinery. Consistent with this complexity, T5-like phages lack a distinct gpK-like protein (TNLP), suggesting either a functional

replacement by other domains or a streamlined adaptation in their tail tip module. This absence further differentiates them from the other structural types - Sipho-1 through -4.

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 Sipho-5 type.

Together, these features - fused THN-CF architecture, absence of TNLP, 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 the siphovirus superfamily.

2.3. Structural Comparison of Central Fiber Proteins Reveals Functional and Evolutionary Diversity in siphophages

2.3.1. Phi80 Central Fiber: Structural Homology and Receptor Specificity

The central fiber protein of phage phi80 shows remarkable structural and functional similarity to that of phage T1. Like gp26 of T1, the phi80 fiber harbors an N-terminal domain (NTD) that adopts an HK97-like fold, presumed to anchor the fiber to the tail structure. The C-terminal region is structurally homologous to that of T1 and also targets the FhuA receptor, a ferrichrome transporter in *Escherichia coli*. This shared receptor usage underscores a conserved infection strategy among FhuA-targeting phages. Notably, the phi80 central fiber lacks a distinct tail needle, instead relying on its elongated fiber structure to mediate host recognition and initial contact. Although there is no high-resolution structure yet for the full-length phi80 fiber, homology modeling and functional studies support its role as a modular adhesin, combining receptor-binding and potential fiber maturation domains.

2.3.2. ES18 Central Fiber and the S74 Protease-Chaperone Paradigm

Phage ES18 represents an intriguing variant in the family of FhuA-binding siphophages. Its central fiber protein contains a well-characterized S74-type protease-chaperone domain at the C-terminus, which is thought to be autocatalytically cleaved after assisting in proper folding or multimerization of the fiber. This maturation mechanism mirrors that observed in phage K1H and may be more widespread than previously appreciated. The presence of this domain suggests that ES18 may have evolved under selective pressures to optimize fiber assembly and functionality, potentially enhancing its infectivity. The dual-domain nature of the ES18 fiber - anchoring and maturation - emphasizes the modularity of these structures and points to a broader evolutionary theme of domain swapping and specialization in phage tail fiber evolution.

2.3.3. T5 Central Fiber: A Possible ES18-Like Configuration

Although the tail tip architecture of phage T5 has been classically considered distinct from FhuA-binding phages like T1, phi80, and ES18, emerging evidence suggests potential structural convergence. Preliminary sequence comparisons reveal that the T5 tail fiber protein shares limited homology with ES18 in its C-terminal domain, raising the possibility that T5 may also possess a protease-chaperone mechanism for fiber maturation. T5 does not use FhuA as a receptor but relies on FhuA-related proteins in some hosts, possibly explaining this partial conservation. Further comparative modeling and biochemical validation are needed to determine whether T5 indeed employs a cleavage-mediated assembly strategy akin to that of ES18 and K1H.

2.3.4. K1H: A Model for Chaperone-Guided Fiber Maturation

Phage K1H has served as a prototype for understanding chaperone-assisted tail fiber assembly. Its central fiber protein includes a well-studied C-terminal S74-type protease domain that is essential for correct folding and trimerization of the fiber prior to autocatalytic cleavage. Structural studies have shown that this domain is dispensable in the mature virion but required for proper expression and assembly. K1H's example provides a useful framework for interpreting similar domain architectures in other phages, including T1, phi80, and ES18, suggesting a convergent evolutionary strategy to manage the complexity of large, multidomain fiber proteins.

These findings highlight the structural and functional variability of central fiber proteins across Gram-negative siphophages. Despite differences in receptor specificity and domain

composition, a recurring theme emerges: the use of modular architectures - anchoring domains, long coiled-coils, and self-cleaving protease-chaperones - to build versatile and effective host-recognition appendages. This diversity not only underscores the adaptability of phages to different bacterial surfaces but also offers insights into the evolutionary pressures shaping their infection machinery.

2.4. Structural Features of T1-like Central Fibers

In exploring the tail tip organization of Gram-negative siphophages, we investigated the structural characteristics of the central fiber protein (CFP) in T1-like phages, focusing in particular on gp26 of phage T1. Our results support the presence of a long coiled-coil structure predicted by structural modeling, prior literature, and our own analyses. While direct experimental evidence remains limited, there are some indications of such a coiled-coil structure at the distal end of the tail, although the image quality is insufficient to resolve discrete domains. The presence of “brushy schmutz” at the tail terminus may reflect this coiled-coil, though further high-resolution structural work is needed to confirm it.

AlphaFold3 (AF3) modeling of the T1 gp26 protein supports a modular architecture. The N-terminal domain (NTD) displays a fold consistent with the HK97 NTD - a feature commonly associated with anchoring to tail structures - suggesting its role in binding or stabilizing the coiled-coil. Interestingly, the C-terminal half of gp26 is closely related to that of the fiber protein in phage phi80, and both are known to use the FhuA outer membrane protein as their host receptor. This structural and functional homology supports the hypothesis that these phages have co-evolved tail fiber modules optimized for similar receptor targeting strategies.

Furthermore, both T1 and phi80 harbor a “Ctd-x” C-terminal domain, which we propose may represent an unidentified protease-chaperone module. This domain likely serves a self-cleaving function during fiber maturation, analogous to the “S74 protease” C-terminal domain described in the tail fiber of phage K1H. Intriguingly, this S74-type domain is also found in ES18 fibers, which use the same receptor, suggesting a convergent strategy among distinct phage groups. These findings raise the possibility that phage T5 may also share structural and functional similarities with this group.

Taken together, the structural features of the T1-like tail tip - specifically the predicted coiled-coil, HK97-like NTD, and C-terminal maturation domains - illustrate the evolutionary modularity and functional specialization in siphophage central fiber proteins. These insights not only refine our understanding of tail tip diversity but also highlight common molecular solutions adopted across different phages for receptor engagement and structural assembly.

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: 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 B. 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.1.**). 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.

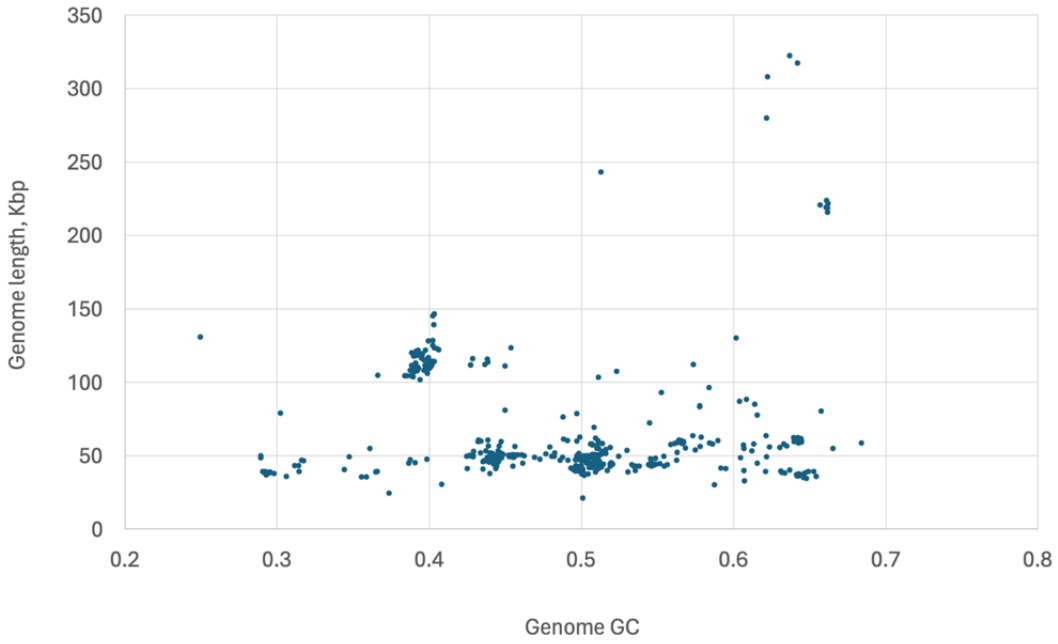

Fig. 1. Distribution of 467 phage sequences.

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 C**.) 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 C. 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 hundreds of phage genomes.

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 (formerly TTC4) in D3-like (Sipho-2) and PY54-like (Sipho-3) 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 Sipho-1a and Sipho-1b phages 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 Central Fiber 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 decorated with variable domains that determine host specificity, infection mechanics, and evolutionary plasticity.

The classification scheme we present not only resolves long-standing ambiguities surrounding phage types like D3 and MP22 but also establishes a reference scaffold for mapping new and hybrid variants. For a long time, phages like Pseudomonas phage D3 had tail tip structures that did match the well-characterized Lambda-like paradigm. Their gene organization was unusual, with fusions and truncations of known tail proteins (like gpL and gpI), and their Central Fiber had a strange combination of domains. Consequently, researchers could not place them into a known tail tip category, and it was not clear if they had equivalents of known tail proteins (like gpJ, gpL, 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, and scientists could not determine whether they were Lambda-like variants or a distinct structural group. For many years, researchers lacked a clear framework or terminology to describe and compare these phages. The ambiguity limited the ability to map their evolution, functions, or use them in synthetic applications.

Our study help resolve these ambiguities 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 lays 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 is more than a structural curiosity - it is a record 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.