Ongoing shuffling of protein fragments diversifies core viral functions to tackle bacterial resistance mechanisms

Bogna Smug
Jagiellonian University
Krzysztof Szczepaniak
Jagiellonian University
Eduardo Rocha
Institut Pasteur
Stanislaw Dunin-Horkawicz
University of Warsaw
Rafal Mostowy (✉ rafal.mostowy@uj.edu.pl)
Jagiellonian University https://orcid.org/0000-0002-4557-3748

Article

Keywords:

Posted Date: June 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3053497/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on November 28th, 2023. See the published version at https://doi.org/10.1038/s41467-023-43236-9.

Ongoing shuffling of protein fragments diversifies core viral functions to tackle bacterial resistance mechanisms

Bogna Smug[1], Krzysztof Szczepaniak[1], Eduardo P.C. Rocha[2], Stanislaw Dunin-Horkawicz[3,4] & Rafał J. Mostowy[1,†]

1) Malopolska Centre of Biotechnology, Jagiellonian University in Krakow, Poland

2) Institut Pasteur, Université Paris Cité, CNRS UMR3525, Microbial Evolutionary Genomics, Paris, France

3) Institute of Evolutionary Biology, Faculty of Biology & Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland

4) Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

† Correspondence: rafal.mostowy@uj.edu.pl

Abstract

Phages are known for their genetic modularity. Their genomes are built of independent functional modules that evolve separately and combine in various ways, making them astoundingly diverse. Multiple studies have demonstrated how genome mosaicism emerges in phage populations and facilitates adaptation to their hosts, bacteria. However, less is known about the extent of (within-)protein modularity and its impact on viral evolution. To fill this knowledge gap, here we quantified such modularity by detecting instances of protein mosaicism, defined as a homologous fragment between two otherwise unrelated proteins. We used highly sensitive homology detection to quantify protein mosaicism between pairs of 133,574 representative phage proteins and to understand its relationship with functional diversity in phage genomes. We found that diverse functional classes often shared homologous domains. This phenomenon was often linked to protein modularity, particularly in receptor-binding proteins, endolysins and DNA polymerases. We also identified multiple instances of recent diversification via exchange and gain/loss of domains in receptor-binding proteins, neck passage structures, endolysins and some members of the core replication machinery. We argue that the ongoing diversification via shuffling of protein domains associated with those functions is reflective of co-evolutionary arms race and the resulting diversifying selection to overcome multiple bacterial resistance mechanisms.

Introduction

Phages have been co-evolving with their bacterial hosts for billions of years. This process has led to an astounding ubiquity and diversity of the viruses that has only recently become evident thanks to rapid advancements of genomics and metagenomics[1]. Such diversity is not only evident at the level of DNA[2] and RNA[3], but also when considering the number of observed phage morphologies and structures[1], the number of known bacterial resistance mechanisms[4,5,6] or viral strategies to overcome them[7,8]. An important property of phage genomes that plays a major role in the emergence of genetic and phenotypic diversity is their modularity. The idea, proposed by Botstein in 1980, is that phages evolve by shuffling interchangeable functional modules, and that selection acting on those modules – rather than on entire genomes – facilitates emergence of new, mosaic genotypes that are advantageous in a given niche[9]. Multiple studies have since demonstrated that genetic mosaicism is ubiquitous in phages and results from frequent homologous and non-homologous recombination events between different viruses[10,11]. Gene flow can also occur relatively frequently between genetically unrelated phages[12]. As a result, bacteriophage population structure is better represented as a network rather than a phylogenetic tree[13,14], where modules of functionally related groups of genes have a coherent evolutionary history[15,16].

While multiple studies have focused on phage genome modularity and its evolutionary implications, relatively less is known about the extent and impact of within-protein modularity (henceforth referred to as ‘protein modularity’) on phage evolution. There are good reasons to expect that such modularity could be extensive. First, we know that some classes of phage proteins have a modular architecture. Very good examples are receptor-binding proteins (including tail fibres, tail spikes) or endolysins. Not only do these proteins exhibit remarkable modularity at both genetic and structural levels[17,18] but their modules can be experimentally shuffled to produce a viable phage virion with a modified host range[19,20,21,22]. Second, previous studies have suggested that structural phage proteins of different functions have evolved to reuse the same folds for different purpose, recombination being an important genetic mechanisms driving such evolution[23,24]. Finally, studies that looked for the presence of composite genes (fusions of different gene families) in viral genomes found this phenomenon to be extensive[25,16]. However, the true extent of protein modularity and its relationship to genetic and functional diversity in phages, remains not fully understood.

In this study, we aimed to better understand the extent of protein modularity in phages and its role in viral evolution. To this end, we analysed over 460,000 phage proteins to detect instances of 'protein mosaicism', defined as two non-homologous protein sequences sharing a homologous fragment (e.g., domain). We remain agnostic as to the exact nature of the genetic process that might have led to this observation (e.g., genetic recombination, deletion of the intergenic regions between consecutive genes, rearrangement, integration). Using a highly sensitive HMM-HMM approach to compare proteins[26], we detected instances of mosaicism between proteins assigned to various functional classes and inferred their domain architecture. We found that, while protein mosaicism is widespread, some functional groups are associated with a particularly highly mosaic composition, including tail fibres, tail spikes, endolysins and DNA polymerases.

Results

Functional annotation using protein fragments is often ambiguous

To investigate the relationship between protein diversity, function and modularity in bacteriophages, we carried out a comprehensive analysis of HMM profiles of representative phage proteins by comparing their predicted functional annotations, genetic similarity and domain architectures (see Methods and Supplementary Figure S1). Briefly, we used `mmseqs2` to cluster 462,721 predicted protein sequences in all bacteriophage genomes downloaded from NCBI RefSeq. The clustering was carried out at 95% coverage threshold to ensure that all proteins grouped within a single cluster have an identical or near-identical domain architecture. We used 133,574 representative protein sequences from the resulting clusters to search the `UniClust30` database against and converted the resulting alignments into HMM profiles (henceforth referred to as representative HMM profiles or rHMMs). To assign functions to rHMMs, we used `hhblits`[27] to search each rHMM against the PHROGs database[28] (to our knowledge the most accurate mapping to date between diverse phage proteins and manually curated functional annotations) complemented with a database of antidefence phage proteins[29].

We investigated the robustness of the PHROG functional annotations (which were additionally simplified to combine closely related biological functions; see Methods and Supplementary Table S1) by assessing how the HMM-HMM comparison parameters affected both the functional coverage of the data (i.e., proportion of representative proteins with any functional hit) and functional uniqueness (i.e., proportion of annotated representative proteins which unique functional hits). We found that pairwise coverage (both query and subject) had a much stronger effect on functional assignment than hit probability or e-value (see Supplementary Figure S2). Specifically, while changing the coverage threshold from 80% to 10% (while maintaining high probability threshold of 95%) increased the functional coverage from 19% to 34%, it also decreased functional uniqueness from 93% to 52% – meaning that at the lowest coverage threshold every second rHMM had multiple, different functional assignments. We also found that at high pairwise coverage threshold ambiguous functional assignment often reflected biological similarity (e.g., 'ribonucleoside reductase' vs. 'ribonucleotide reductase', or 'transcriptional regulator' vs. 'transcriptional activator'; see Supplementary Figure S3). By contrast, at lower sequence coverage thresholds co-occurrences between clearly different functions became more and more common and affected the majority of functions (see Supplementary Figure S3), meaning it was often impossible to confidently assign a function based on a fragment of a protein (i.e., partial match to a reference database; here PHROGs). Considering this, for further analyses we set the probability and pairwise coverage cut-offs for the PHROG annotations to 95% and 80%, respectively, while conservatively excluding all rHMMs with hits to more than a single functional class (see Methods).

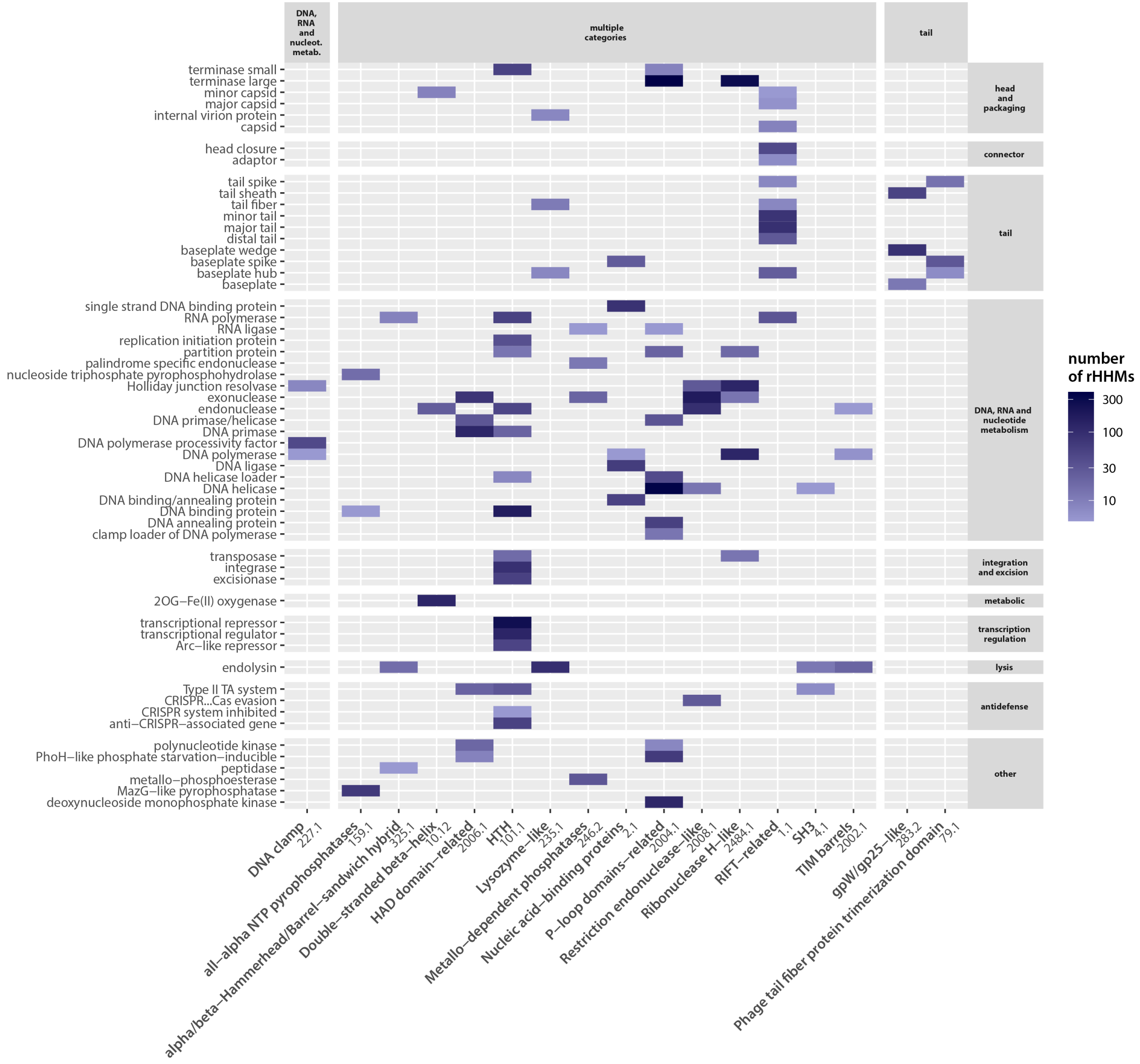

Figure 1. Diverse protein functions often share homologous domains. Heatmap showing groups of homologous ECOD domains (H-group names; x-axis) found in proteins assigned to different functional classes (y-axis). A domain was considered present in a functional class when it was present (i.e., found with a minimum 95% probability and 70% subject coverage) in at least 5 rHMMs assigned to a given functional class. The colour scale indicates number of rHMMs in which the domain was found. Only domains found in multiple functional classes (at least 3) are shown. Generic functional classes ('tail' and 'structural protein') were excluded from this visualisation. Functional classes are grouped according to their categories.

Proteins assigned to different functional classes share homologous domains

Given that for a low pairwise coverage threshold we often found rHMMs to be co-annotated by apparently distinct functional classes, we hypothesised that these functions contained rHMMs that shared homologous domains (i.e., protein structural and functional units that have been shown to have emerged from a common ancestor). To address this hypothesis, we used the Evolutionary Classification of Protein Domains (ECOD) database as it provides a comprehensive catalogue of known protein domains and their evolutionary relationships[30]. We then used sensitive HMM-HMM comparison to detect presence of these domains in rHMMs (see Methodology and Supplementary Figure S1). The ECOD database classifies all domains into T-groups (have the same topology, share evolutionary relatedness and structural homology), H-groups (have different topology but share evolutionary relatedness and structural homology) and X-groups (have different topology and no evidence of evolutionary relatedness but share structural homology). Therefore, we assumed that if two domains belong to different X-groups, they are not homologous.

Figure 1 shows the distribution of ECOD domains (H-groups) across different functional classes. (The distribution of T-groups across functional classes is also shown in a Supplementary Figure S4.) In line with previous literature, we found examples of phage proteins with different functions sharing homologous domains. These include well known examples of helix-turn-helix domains found in transcriptional regulator/repressor proteins, integrases, transposases or DNA-binding proteins[31]; the RIFT-related domains found in many structural proteins like tail and neck proteins[32] but also for example in RNA polymerases in the form of double-psi barrels[33]; P-loop domain-related family found in ligases, kinases or helicases[34,35,36]; or the ribonuclease H-like domain family found in many DNA processing enzymes like Holliday junction resolvases, exonucleases, DNA polymerases or transposases[37].

Cases of homologous domain sharing (i.e., belonging to a single ECOD H-group) between proteins assigned to different functional classes can be explained in several ways. One explanation is that such proteins may actually have the same function (e.g., baseplate and baseplate wedge). Alternatively, ancient and large domain classes that play an important, biological role (e.g., DNA binding or NTP hydrolysis) may have diverged into subfamilies specific for different functions and thus are shared by a wide range of PHROG classes. Indeed, we found a strongly significant, positive correlation between domain frequency (H-groups) and diversity (see Supplementary Figure S5), suggesting that domains that are common in nature tend to be more diverse.

Finally, domain sharing between proteins assigned to different functional classes may be the result of mosaicism, i.e., the acquisition of specialised domains for different functions. This scenario is additionally supported by the observation that distinct H-groups were detected in proteins assigned to the same functional class. For example, functional classes such as exonucleases, endonucleases, DNA polymerases or endolysins each contained as many as 4 distinct H-groups, each found in at least 5 rHMMs (see Figure 1). We thus hypothesised that this distribution is indicative of modularity and ongoing domain shuffling in functionally diverse proteins.

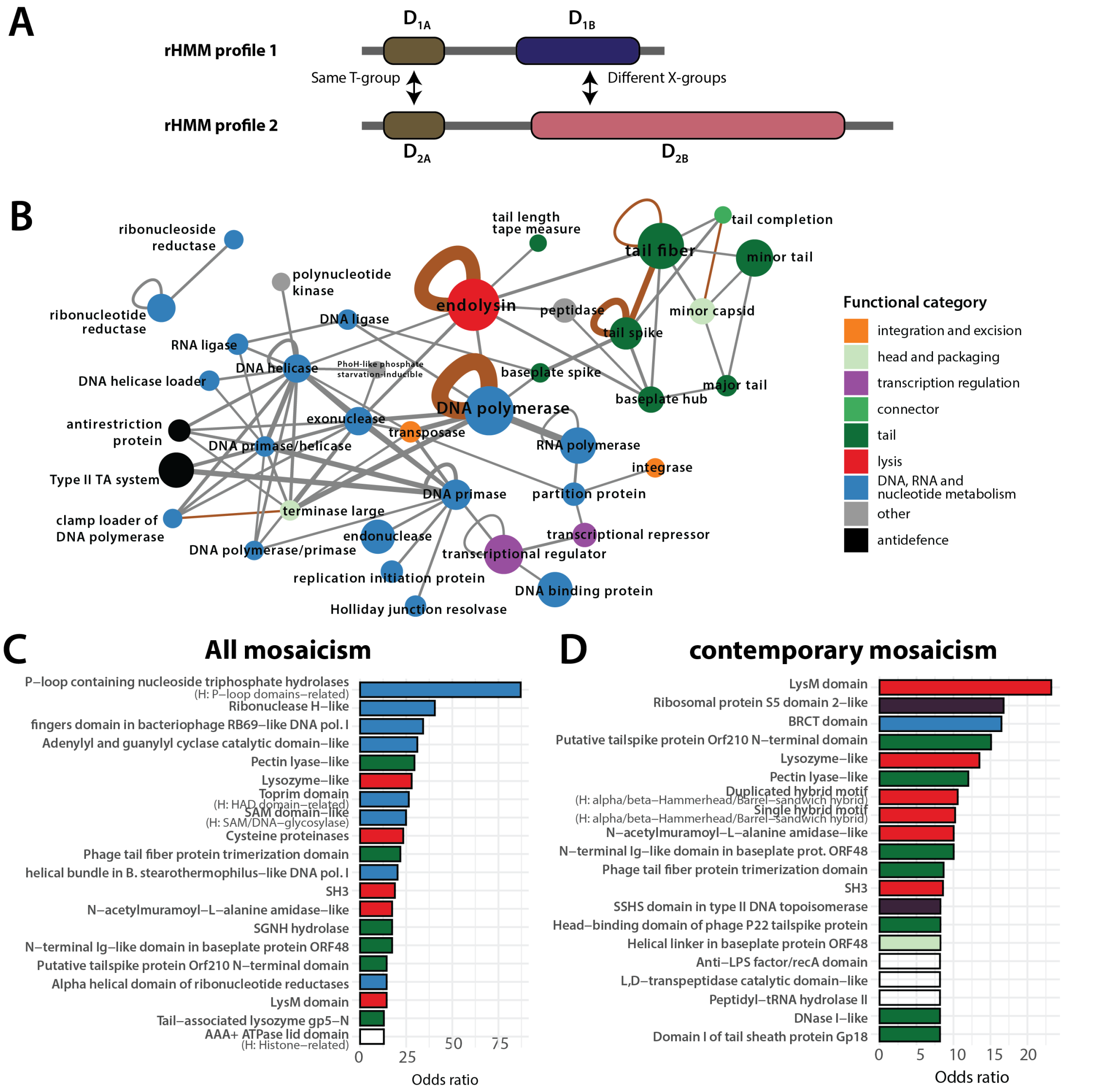

Figure 2. Map of domain mosaicism in phages. (A) Domain mosaicism for any rHMM pair was defined when (i) both proteins had at least two distinct ECOD domains detected, (ii) both shared a domain assigned to the same ECOD T-group, and (iii) both additionally contained non-homologous domains (i.e., belonging to different X-groups). (B) Mosaic network of protein functions. Each node represents a functional class and edges link functional classes where evidence of domain mosaicism was found between at least four pairs of domain architectures (i.e., unique combinations of ECOD T-groups which can be thought of structurally-equivalent proteins). Brown edges connect functions where at least one case of 'contemporary mosaicism' was found (i.e., a pair of rHMMs with the percentage identity of a shared fragment 50% or greater). Node size corresponds to the number of domain architectures in a given functional class. Edge thickness corresponds to the number of domain architecture pairs with evidence of domain mosaicism. Generic functional classes ('tail' and 'structural protein') were excluded from this visualisation. (C) Bar plot shows the odds ratio that a given domain (ECOD T-group) is found more frequently in mosaic domain architectures than non-mosaic domain architectures. Only 20 domains with the greatest odds ratio that are statistically significant are shown (Fisher's exact test; level of significance was Bonferroni-corrected for multiple testing). Colours denote the most frequent functional category in rHMMs with the given domain (when at least two categories are the most frequent then the bar is white). For each domain, corresponding H-group names are provided if different from the T-group name. (D) Same as panel C but here mosaicism is defined as 'contemporary' as for brown edges in panel B.

Protein modularity is most often linked to replication, lysis and structural proteins

To explicitly detect protein modularity in our data, we defined domain mosaicism (see Figure 2A) as a protein (rHMM) pair with (i) at least two domains (different X-groups) detected in each of them and (ii) at least one domain of the same topology (same T-groups) and (iii) at least one other domain of different structural architectures (different X-groups). This definition thus excluded pairs of proteins where the shared domains may have been homologous but belonged to divergent subfamilies specific for different functions. We found evidence of such domain mosaicism in 45 out of 101 functional classes (assuming at least 3 mosaic rHMMs per functional class). Figure 2B shows a map of domain mosaicism visualised as a network with nodes representing functional classes and edges linking those classes that contained rHMMs with evidence of domain mosaicism. We found that functional categories where domain-level mosaicism was common were DNA/RNA metabolism (e.g., RNA and DNA polymerases, DNA ligases, helicases, exo- and endonucleases, DNA binding proteins), transcription regulation, structural tail proteins (tail fibre, tail spike and baseplate proteins) and endolysins. Three functional classes with the most examples of within-class domain mosaicism were DNA polymerases, endolysins and tail spikes.

To examine the relationship between protein function and domain mosaicism independently of assignment to functional classes, we next investigated which domain architectures are statistically associated with mosaicsm. To this end, we calculated the odds ratio for each domain (ECOD T-group) to be over-represented in proteins with evidence of domain mosaicism. Specifically, we first considered only rHMMs with significant hits to at least a single domain. Then, for a given domain, we calculated the number of all domain architectures (i.e., unique combinations of ECOD T-groups) with and without that domain and the number of all domain architectures with and without evidence of domain mosiacism. Finally, we calculated the odds ratio that this domain is found more frequently in mosaic domain architectures than non-mosaic domain architectures (see Methods). Results, shown in Figure 2C, are consistent with the network in panel B. Domains with the greatest odds ratio of being over-represented in mosaic proteins typically fall into three categories: (i) domains occurring in proteins associated with DNA/RNA metabolism, particularly in DNA polymerases, DNA primases, DNA helicases, exonucleases, ribonucleotide reductases and Holliday junction resolvases, e.g., P-loop containing nucleoside triphosphate hydrolase, Ribonuclease H-like, adenylyl and guanylyl cyclase catalytic, toprim or SAM-like domains; (ii) domains occurring in endolysins, e.g., lysozyme-like, SAM-like, cysteine proteinases or SH3; and (iii) domains occurring in receptor-binding proteins, e.g., pectin lysase-like, tail fiber trimerization domain, SGNH hydrolases or tail-associated lysozymes and Ig-domains.

The existence of domain-level mosaicism in phages is not a new phenomenon as some functions analysed here have been previously linked with mosaic domain architectures[38,39]. We thus next enquired which cases of domain mosaicism are ancient (i.e., represent ancestral domain shuffling underlying functional diversification) and which cases of domain mosaicism are contemporary (i.e., are the result of a relatively recent emergence of protein modularity). This issue was partially addressed using ECOD T-groups instead of H-groups to assign shared domains in protein pairs. However, to investigate this problem further, we first looked into the sequence similarity distribution of all mosaic pairs of rHMMs and found that only 9% of them shared fragments with a percentage identity of 10% or greater. Then, we reanalysed the data using a definition of 'contemporary mosaicism' by requiring that the shared protein fragments have an amino-acid percentage identity level of 50% or greater. We found that four of the functional classes fulfil that criterion (Figure 2B, brown edges): DNA polymerase, tail spikes, endolysins and tail fibers. Finally, using the domain-based approach (Figure 2D), we found that domains (ECOD T-groups) significantly over-represented in proteins showing evidence of contemporary mosaicism are most often linked to receptor-binding proteins and baseplate proteins (e.g., putative tailspike protein N-terminal domain, pectin lyase-like, N-terminal Ig-like domain, head-binding domain of phage P22 tailspike, helical linker in baseplate protein) and to endolysins (e.g., LysM domain, SH3, amidase-like, lysozyme-like etc.); we also found signal to domains that are typically associated with replication proteins (e.g., SSHS domain in type II DNA topoisomerase or BRCT domain) or antidefence proteins (ribosomal protein S5 2-like).

Overall, these results show that (i) domain mosaicism is common in phage proteins and associated with DNA/RNA replication, lysis and structural proteins, (ii) while most of that mosaicism appears to be due to ancient domain shuffling or specialisation, we see clear examples of contemporary mosaicism particularly in receptor-binding proteins and endolysins, and (iii) there are also rare and intriguing cases of recently emerged mosaicism associated with other functions.

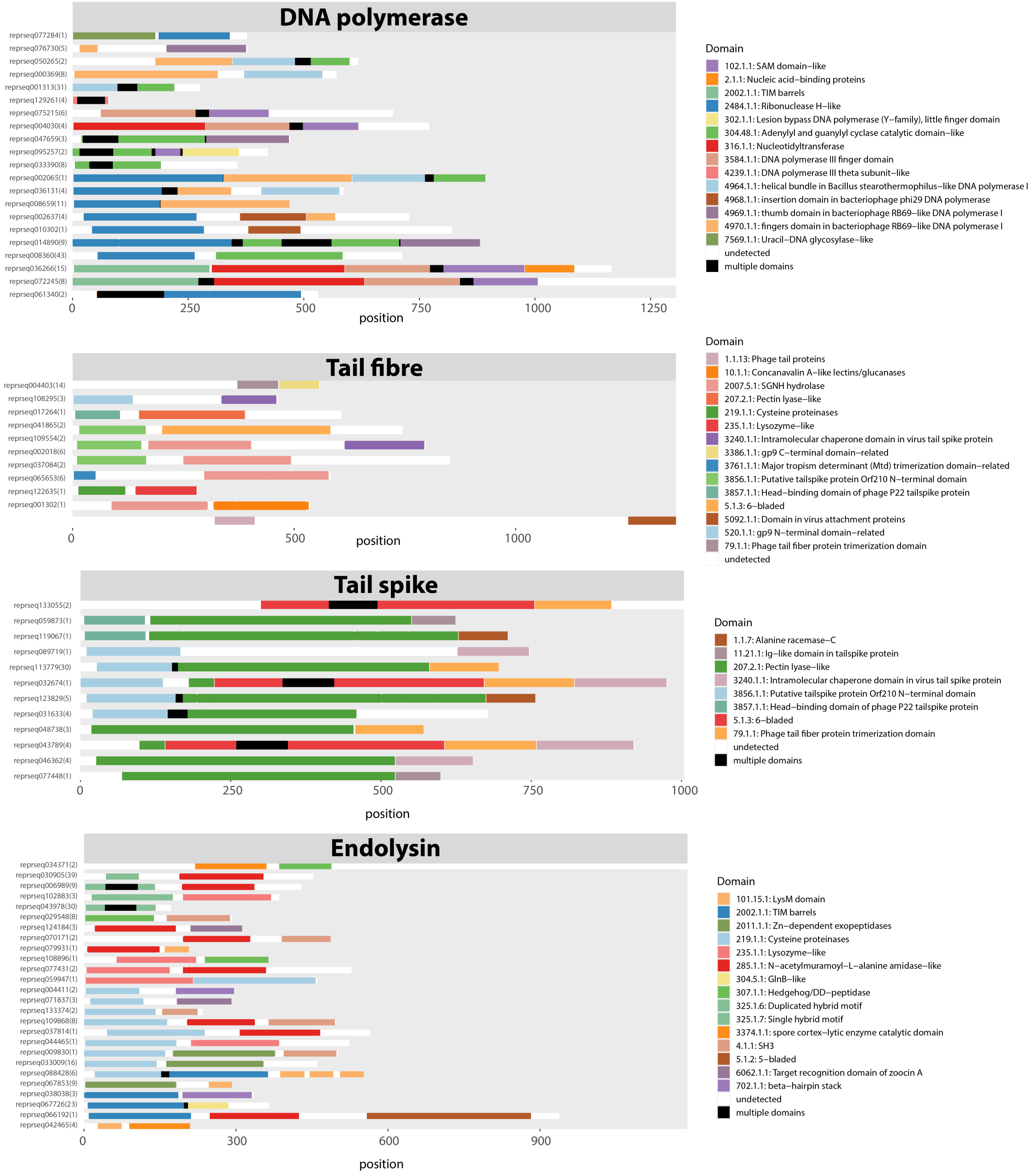

Figure 3. Visualisation of domain architectures for functional classes exhibiting the highest levels of domain mosaicism. Each line shows one chosen rHMM (abbreviated `reprseqXXXXXX`) per each domain architecture, with the number of protein sequences having this domain architecture displayed in bracket, and the ECOD domains (T-groups) found within that rHMM. Colours denote the ECOD T-groups, with black denoting multiple domains found in this region and white denoting absence of ECOD hits. ECOD T-groups are in the following format: $X_{\text{id}}.H_{\text{id}}.T_{\text{id}}$. Only domain architectures with at least two different T-groups are shown.

Protein modularity hotspots

To better understand the nature of protein modularity, we next looked into the specific domain architectures of the three mosaicism-outliers: DNA polymerase, tail fibre, tail spike and endolysin. We also developed a Shiny webserver that allows users to interactively look up and visualise domain architectures in all functional classes used in this analysis as well as to connect specific domains shared with proteins of other functions: `bognasmug.shinyapps.io/PhageDomainArchitectureLookup`.

DNA polymerase and other replication proteins

As far as individual functional classes are concerned, DNA polymerases are the clear mosaicism outlier in the 'DNA, RNA and nucleotide metabolism' category. The representative domain architectures of all of those found in DNA polymerases are shown in Figure 3. For comparison, in Supplementary Figure S6 we show an overview of the domain architectures for representative members of all families of DNA polymerases known to occur in bacterial or viral genomes (A, B, C, X and Y)[40,41] detected using `HHpred`[42] with ECOD as the database.

Our results point to a few notable observations. First, we have recovered domain architectures of not only families A and B, which are well known to occur in phages such as T4 and T7, but also of families C and Y (c.f., Supplementary Figure S6) which are characteristic of bacteria. Second, we have identified other domain architectures that are variants of the above. For example, instead of the four domains typical of family A (ECOD X-groups 2484, 4970, 4964 and 304), we found rHMMs that contained only the first three (2484, 4970, 4964) and two (2484, 4970). Such rHMMs with unusual domain architectures represented clusters with multiple protein sequences, suggesting multiple occurrences of such architectures in the analysed genomes. Finally, the comparison of these domain architectures points to clear cases of mosaicism, such as the insertion domain of bacteriophage ϕ29 found alongside the exonuclease (ECOD X-group 2484) and/or the finger domain (ECOD X-group 4970). It also highlights that conserved folds found in DNA polymerases are reused in various combinations, but also in combination with domains present in other proteins belonging to the 'DNA, RNA and nucleotide metabolism' category (see Supplementary Figure S7).

Receptor-binding proteins and other tail proteins

Receptor-binding proteins, like tail fibers and tail spikes, are often described in the literature as consisting of three domains: a conserved N-terminal which binds to the tail structure (e.g., to the baseplate), a variable and host-dependent C-terminal which binds to the receptor at the bacterial surface, and the central domain which contains enzymes (hydrolases) that help the phage penetrate layers of surface sugars like the capsular polysaccharide[43]. Our results show clear evidence for the emergence of modularity via shuffling of all of these domains (see Figure 3). First, we find N- and C-terminal domains in multiple arrangements. For example, the C-terminal 'Alanine racemase-C' domain (ECOD T-group 1.1.7) is found in tail spikes in combination with either the 'Head-binding domain of phage P22 tailspike protein' domain (3857.1.1) or with the 'Putative tailspike protein Orf210 N-terminal' domain (3856.1.1), providing an excellent example of mosaicism. Second, we found co-existence of various enzymatic domains within the same protein in different combinations. For example, the endosialidase domain was found to co-occur with the 'SGNH hydrolase' domain (2007.5.1) in a tail fibre protein as well as with the 'Pectin-lyase like' domain (207.2.1) in a tail spike. Finally, some domains present in receptor-binding proteins were also found to occur in other functional classes. A good example here is the 'tail fibre trimerization domain' domain (79.1.1) which is also found in baseplate spikes in combination with other domains like lysozyme (see also Supplementary Figure S8). Overall, these results suggest that domains found in receptor-binding proteins can not only be shuffled in different combinations, but that multiple enzymatic domains can co-occur in the same protein.

Endolysins

Endolysins are classically described as having catalytic domains (lysozymes, muramidases, amidases, endopeptidases, etc.) and/or cell wall-binding domain; and they may be observed in multiple combinations[18]. Here we find both types of domains co-occurring in various combinations (see Figure 3). For example, the catalytic domain Cysteine proteinases (ECOD T-group 219.1.1) is found in combination with either SH3 domain (4.1.1) or target recognition domain of zoocin A domain (6062.1.1). Second, we found the presence of multiple catalytic domains within the same proteins. This includes co-occurrence of exopeptidases (2011.1.1) and Cysteine proteinases (219.1.1), or co-occurrence of the spore cortex-lytic enzyme (3374.1.1) and Hedgehog/DD-peptidase (307.1.1). These domain architectures are in line with those previously described for endolysins of mycobacteriophages, where apart from multiple instances of co-occurrence between the peptidase-like N-terminal and a cell wall-binding C-terminal there were also central domains with amidases, glycoside hydrolases and lytic transglycosylases[44]. Interestingly, a domain-based network of this diversity in all lysis genes shows a much more inter-connected network than for replication and tail proteins, suggesting that the endolysin domains likely co-occur many out of all possible combinations (see Supplementary Figure S9).

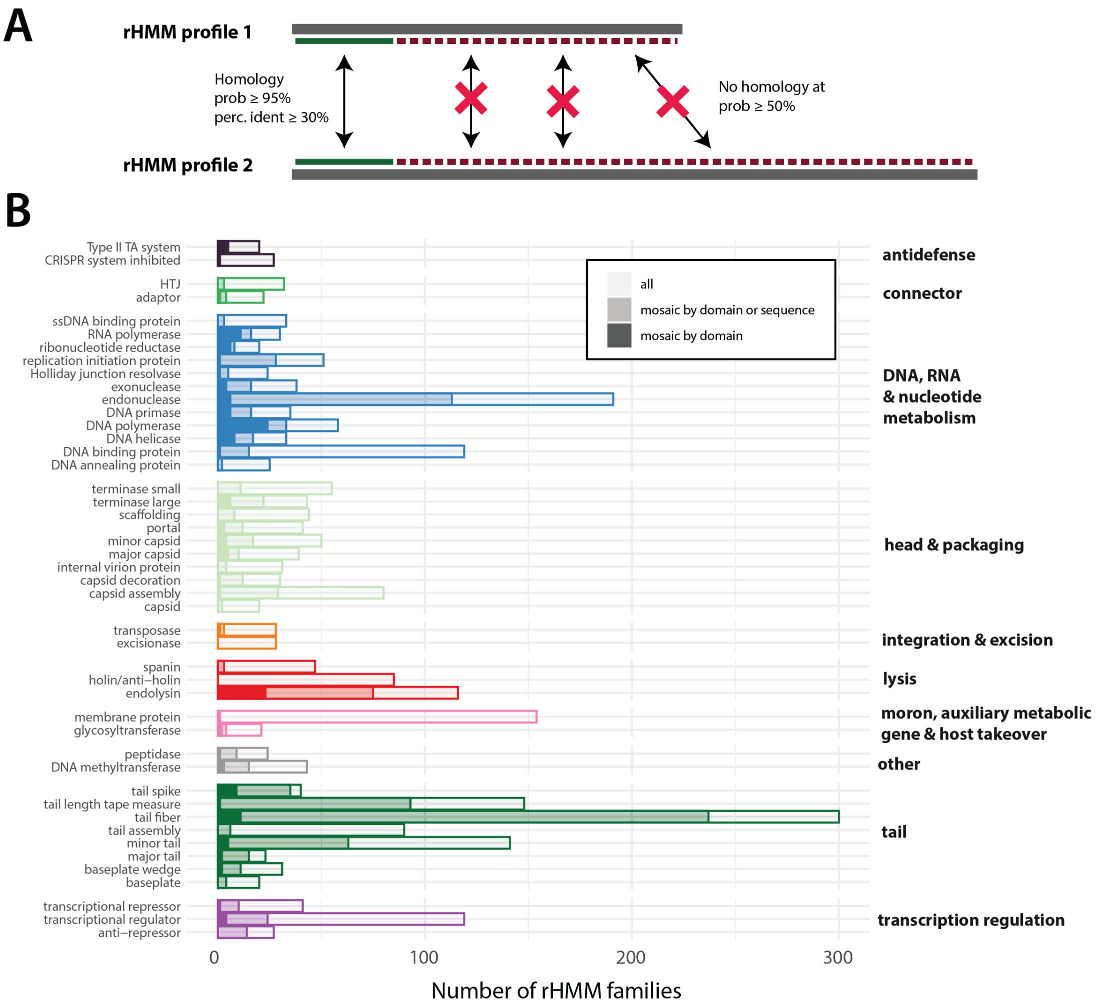

Figure 4. Sequence mosaicism. (A) Sequence mosaicism for any rHMM pair was defined as presence of sequence-based homology in the background of absence of homology using their HMM profiles. First, pairwise coverage was calculated as the total aligned length using a permissive probability hit threshold of 50%. Mosaicism was claimed when the aligned fragment (i) constituted a short proportion of the length of both proteins (both query and sequence coverage $\leqslant$50%), (ii) was detected with high probability ($p \geqslant$95%), (iii) had percentage identity of at least 30% and (iv) was of length $\geqslant$ 50aa. (B) Number of rHMM families with signal of mosaicism detected by domain alone (opaque) vs. by domain or sequence (moderate) vs. number of all rHMM families (transparent) in different functional classes. Colours are equivalent to Figure 2B. Only functional classes with genetic diversity of at least 20 families are shown.

Sequence-based insight points to extensive mosaicism beyond domain analysis

Our measure of domain mosaicism (c.f., Figure 2A) is a robust approach to detect protein modularity as it uses evolutionary information stored in the ECOD database to distinguish the lack of local sequence similarity due to possible divergence (same X-groups) from that due to lack of common ancestry (different X-groups); it also ensures that the units of mosaicism, i.e., domains, are evolutionarily-meaningful as they can fold independently and hence can be horizontally shuffled. However, this approach has two important limitations.

The first limitation is that the domain mosaicism approach relies on the assumption that all functional classes have a comparable coverage in the ECOD domain database, which may not be true. Indeed, our analysis of such coverage (see Supplementary Figure S10) shows that while some functional classes – for example those belonging to the 'DNA/RNA nucleotide metabolism' – are relatively well annotated by ECOD, other functional classes (e.g., tail completion, head scaffolding, spanin, holin/anti-holin, nucleotide kinase or tail length tape measure) seem to be strongly under-represented in domain databases. An interesting example are tail fibres which rarely exhibit hits to more than a single ECOD domain in spite of being known as long and multi-domain proteins. Furthermore, while we saw a strong and significant correlation between structural diversity (number of unique domain architectures detected by ECOD at the T-group level) and genetic diversity (measured by the number of protein families, where protein family was defined as a cluster of similar rHMMs; see Methods) in different functional classes, some classes – including tail length tape measure protein, membrane proteins, head-tail joining proteins or ssDNA binding proteins – had a disproportionately low structural diversity compared to genetic diversity (see Supplementary Figure S11).

The second limitation of the domain mosaicism approach is that it relies on a highly restrictive definition of mosaicism – it requires that each protein in a mosaic pair have two structurally unrelated domains detected (different X-groups). This requirement might miss many cases of mosaicism where a domain is undetected or when mosaicism occurs at the sub-domain level[45]. To gauge the potential extent of such bias, we carried out the all-against-all comparison of 134k rHMMs using hhblits[27] and investigated the relationship between their sequence similarity and coverage of all pairs (see Supplementary Figure S12). The results show that, while most rHMM pairs align at high coverage, reflecting their likely homology over the majority of the sequence length, there is a substantial fraction of pairs that overlap by a fragment that constitutes a short proportion of their length, indicating possible modularity at the domain or sub-domain level. This relationship was qualitatively identical when we subtracted all rHMM pairs where we detected domain mosaicism with ECOD, suggesting that the domain mosaicism approach may miss potentially interesting cases of protein modularity.

Given the above, we next defined sequence mosaicism when two rHMMs shared detectable similarity over less than half of their length but with a percentage identity (in the aligned region) of at least 30% (see Figure 4A). We then calculated the number of rHMM families with signal of mosaicism detected by domain vs. by domain or sequence in different functional classes. Results are shown in Figure 4B. We saw that, on average, the proportion of rHMM families with the signal of sequence mosaicism in a given functional class was greater than the proportion of rHMM families with the signal of domain mosaicism. This result was in line with our expectations since, as explained above, the sequence mosaicism test is the less conservative one. Notably, however, in some functional classes the proportion of families with the signal of sequence mosaicism alone was disproportionately high. Examples include functional classes such as replication initiation protein, endonuclease, DNA binding protein, capsid assembly, endolysin, tail spike, tail length tape measure, tail fibre, minor tail or transcriptional repression/regulation. This suggests that these – and potentially some other – functional classes may harbour an under-explored reservoir of protein modularity.

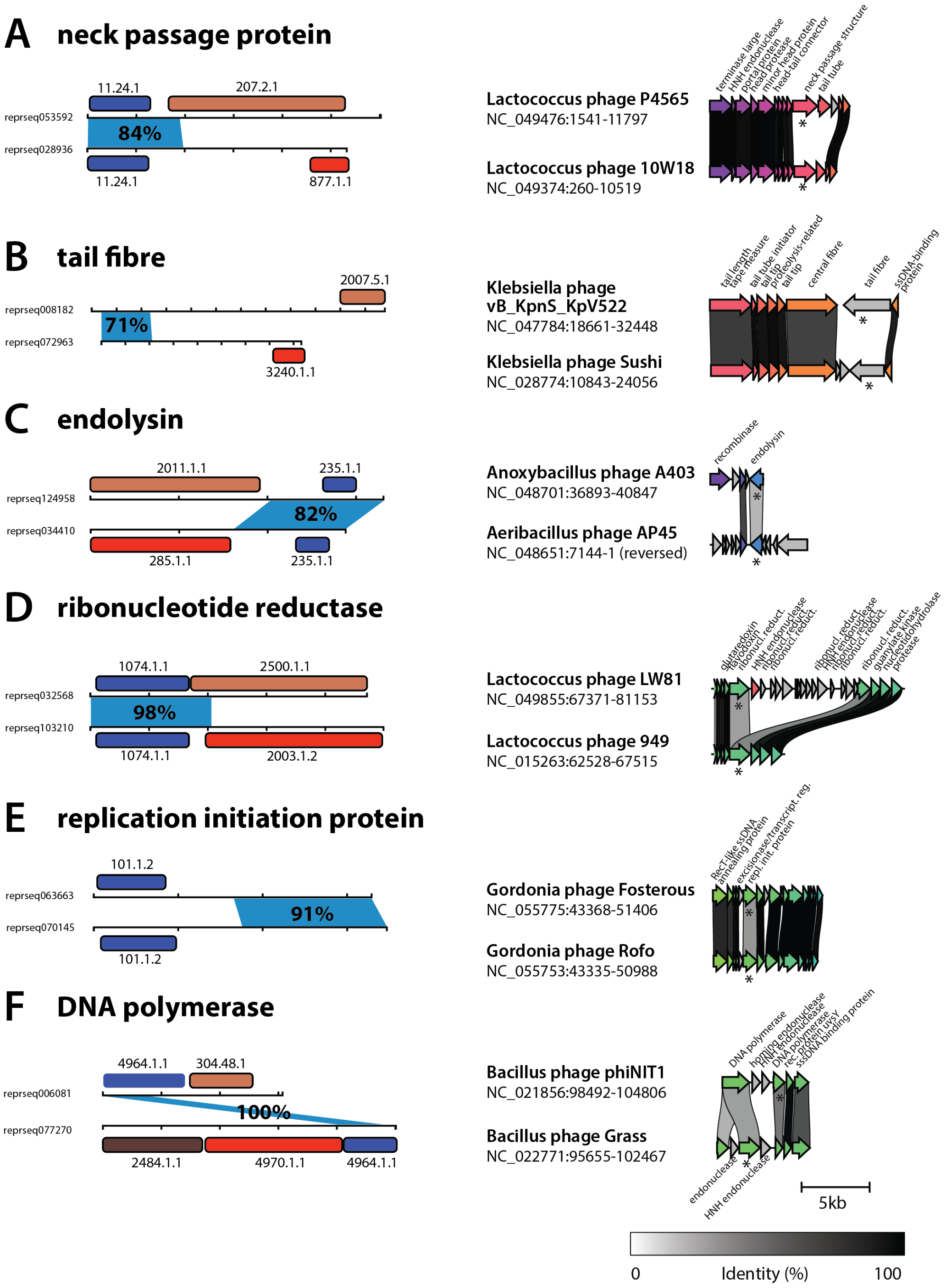

Figure 5. Protein modularity caught red-handed. Six panels (A-F) show representative examples of recently emerged protein modularity. (Left) Horizontal lines denote the protein sequence length (abbreviated `reprseqXXXXXX`) with dashes every 100aa. Blue shows the region with high genetic similarity and information about the percentage identity of that region (amino-acid level). Boxes show detected ECOD domains with their ECOD IDs; homologous domains in a pair have the same colour. (Right). Genomic comparison using `clinker`[46] of the regions where the two proteins were found, with the corresponding names of the phage and genome coordinates. The upper genome fragment corresponds to the upper protein, and stars show the location of the two proteins. Only proteins with informative functional hits (NCBI Genbank) are labelled. Links are drawn between genes with percentage identity of at least 30% across the full length, with the level of identity represented by the scale at the bottom of the figure.

Emergence of protein modularity is an ongoing evolutionary process

Our results so far show that domain mosaicism often underlies functional diversity in phages, and that its emergence in some functional classes (RBPs, endolysins and potentially other functions) may be contemporary. We also found that many functional classes exhibit strong signal of sequence mosaicism that is not detected by the ECOD-based analysis. Given these observations, we hypothesised that there may be cases of recently emerged protein modularity resulting from an ongoing evolution and diversification that would only be detected using the sequence mosaicism test. To find putative pairs of functional classes that might have undergone recent diversification, we created a network of functional classes with rHMMs that exhibit a signal of a recent sequence mosaicism (percentage identity $\geqslant$70% and $\geqslant$90%; see Supplementary Figure S13). While, as expected, multiple links were found between proteins classified as tail fibres, tail spikes and endolysins, connections between proteins of other functions were also identified, including replication proteins, neck proteins and anti-defence proteins. Additionally, almost all of the functional classes linked with rHMMs with an 'unknown' status, implying that they had a mosaic signal with a protein the function of which was uncertain.

To examine whether these cases represented genuine, ongoing emergence of protein modularity and are not false-positives, we used all levels of information available to us to (pairwise comparison, HHpred domain detection, genomic context) to examine dozens of these pairs in detail. As a result, we provide representative cases of ongoing emergence of protein modularity in six different functional classes: neck passage protein, tail fibre, endolysin, ribonucleotide reductase, replication initiation protein and DNA polymerase (see Figure 5). These examples demonstrate a variety of mechanisms and biological contexts in which protein diversification emerges at the domain level.

As shown in Figure 5A, one mechanism that mediates protein diversification is exchange of domains. This can be best seen using the example of neck passage structure proteins. These proteins have been previously identified as a diversity hotspot in *Lactococcus* phages[47] and some are known to carry carbohydrate-binding domains[48]. The provided example shows an exchange of a non-homologous C-terminal receptor-binding domain and pectin lyase-like domain while preserving the near-identical N-terminal in two closely related phages. An analogous example are two tail fibre proteins (Figure 5B), found in closely related *Klebsiella* phages, with a very similar N-terminal and two, non-homologous receptor-binding C-terminal domains – a phenomenon very well known to occur in phages infecting bacteria with extensive surface polysaccharide diversity[17]. A similar but converse example are two fragment-sharing endolysins found in two otherwise unrelated genomes of *Anoxybacillus* and *Aeribacillus* phages (Figure 5C). The said endolysins contain a highly similar C-terminal (lysozyme) and two unrelated N-terminal domains (exopeptidase and amidase).

On the other hand, we observed multiple different mechanisms driving protein diversification in core replication proteins. One was domain exchange between ribonucleotide reductases in *Lactococcus* phages (Figure 5D). The two closely related genomes both carry a ribonucleotide reductase protein with an identical N-terminal domain and unrelated C-terminals: ten stranded beta/alpha barrel domain (ECOD 2500.1.1) and FAD/NAD(P)-binding domain (2003.1.2). Interestingly, one of the genomes has the other C-terminal domain in another protein that is located downstream from a genetic island that contains other ribonucleotide reductases and endoluclease domains. This suggests that diversification of the discussed protein was linked to the insertion/deletion of a new domain, possibly together with the mentioned genetic island.

Another example of a protein diversification mechanism was found in two replication initiation proteins present in two closely related genomes of *Gordonia* phages (Figure 5E). The two proteins share a near-identical C-terminal regions but with no detectable ECOD domains hits; they also both have hits to the winged helix-turn-helix domain (ECOD 101.1.2) but with no detectable similarity at the sequence level. While the homolgous N-terminal could potentially be explained by strong diversifying selection, the high similarity between the two phage genomes (ANI = 97%, coverage = 89%) suggests that the most likely explanation is a domain exchange via recombination into its distantly related variant.

Last but not least, we investigated the underlying mechanism of diversification of DNA polymerases. Intertestingly, this mechanism is quite different from the ones above and involves shuffling (i.e., gain or loss) of domains, as shown in Figure 5F. Two proteins, found in related genomes of *Bacillus* phages, share an identical sequence that we identified as a helical bundle in DNA polymerase I. Investigation of other proteins in the neighbouring genetic region revealed that the two genomes contain the same set of DNA polymerase domains at high percentage identity but split into different open reading frames due to presence and absence of several endonucleases between those domains. This suggests that diversification of replication regions, including DNA polymerases, in phages may often occur via gain and loss of domains.

Discussion

In this study we have systematically analysed the relationship between genetic diversity, functional diversity and protein modularity in phages using 134k HMM profiles of representative phage proteins (rHMMs) and comparing them to each other and to the ECOD domain database using a sensitive homology search via HMM-HMM comparison. Our results demonstrate that domain conservation in phage proteins is extensive, often linking proteins with different functions, and that these domains often co-occur in multiple combinations. This is consistent with our knowledge of how phages evolve and their remarkable ability to not only alter their protein sequence through rapid evolution but also to recycle existing folds in novel biological contexts[23,24,49]. Indeed, our findings show that such domain shuffling, that is known to be often recombination-driven[50], not only links different functions but also underlies genetic diversity within multiple functional classes, notably related to tail proteins, lysins, and the core replication machinery. It also shows how emergence of protein modularity is an important mechamism of ongoing diversification in phage populations.

Modularity in receptor-binding proteins (tail fibre, tail spike) as well as in endolysins has been extensively studied before, though to our knowledge it has not been systematically quantified and compared to other functional classes. Both receptor-binding proteins and lysins can play an important role in host range determination[51], and previous studies have repeatedly demonstrated their rapid evolution in face of adaptation to new hosts, particularly in receptor-binding proteins[18,52,53]. It is therefore not surprising that these proteins would have evolved a LEGO-like, modular architecture that facilitates rapid structural alterations to aid viral adaptation. There are nevertheless important differences between the two groups in terms of how such modularity has been and continues to be shaped by evolution. While receptor-binding proteins and endolysins are both specific in that they contain enzymes that recognise and hydrolyse specific sugar moieties, the diversity of the sugar repertoire on which they act can be quite different. Receptor-binding proteins often use surface polysaccharides as the primary receptor, notably capsular polysaccharides and LPS, which due to their rapid evolution can often vary considerably, even between two bacterial isolates of the same lineage[54]. This means that phages are under selective pressure to rapidly adapt to new hosts that may bear completely different surface receptors than their close relatives. A good example is *Klebsiella pneumoniae*, which is known to often exchange polysaccharide synthesis loci with other bacterial lineages[55] while its phages are known for not only extensive modularity of receptor-binding proteins[17] but also existence of phages with complex tails with a broad host spectrum[56]. In line with this, we found clear evidence of the emergence of recent mosaicism within tail fibers and tail spikes.

Endolysins, on the other hand, target the peptidoglycan of their bacterial hosts. While there is a considerable diversity of peptidoglycans in bacteria[57], its diversity does not vary as dramatically between different lineages of the same species as it can be the case with surface receptors. Consequently, one would expect a weaker diversifying selection acting within phages that infect closely related bacteria and a stronger one for those phages that infect distantly related hosts. In line with this reasoning, Oechslin and colleagues recently found that the fitness costs of endolysin exchange between phages increased for viruses infecting different bacterial strains or species[18]. However, they also found evidence of recombination-driven exchange of endolysins between virulent phages infecting the same host and the prophages carried by this host, pointing to the likely importance of recombination in driving the evolution of endolysins. This is consistent with the previous reports that domain shuffling is an important driver of endolysin diversity in Mycobacteriophages[44], and with our results showing that emergence of protein modularity in endolysins is often a contemporary and ongoing phenomenon.

Another major group for which there was evidence of extensive protein modularity and mosaicism were core replication proteins, particularly DNA polymerases. This result may seem counter-intuitive as core replication proteins are known to contain highly conserved sites due to the very precise way in which they process and metabolise DNA/RNA. But the DNA replication machinery is known to be highly diverse across the tree of life[58], including in viruses[39], and this diversity is known to have been evolving since the existence of the last common universal ancestor (LUCA) with evidence for the importance of recombination and domain shuffling in this process[38]. It can be thus expected that the much of the protein modularity that we detect in this study is ancient and predates the emergence of bacteria and phages. However, there are a few arguments to suggest that such mosaicism has been emerging, and continuously emerges, during co-evolution between bacteria and phages. First, the scale of diversity of (and mosaicism in) some core replication proteins, for example DNA polymerases or endonucleases, suggests that maintaining such diversity must have been beneficial for phages. Second, previous studies have reported modularity of DNA polymerases[59] as well as plasticity and modularity of the DNA replication machinery as a whole[60] in T4-like phages. The authors argued that such flexibility gives these viruses an edge in adapting to their diverse bacterial hosts[60]. Finally, our data points to clear examples of recently-emerged protein modularity in DNA polymerases, ribonucleotide reductases and replication initiation proteins. This suggests that core replication proteins continue to evolve in the process of bacteria-phage co-evolution.

One possible and potentially important driver of the diversity of core replication proteins in phages could be bacterial defence systems. There is a growing body of literature describing bacterial defence systems that target phage replication machinery to prevent viral infection and their spread in bacterial populations. One example is the DarTG toxin that was recently shown to ADP-ribosylate phage DNA to prevent phage DNA polymerase from replicating viral DNA, and escape mutations in DNA polymerase allowed the phage to process the modified DNA[61]. Another example is the Nhi, a bacterial nuclease-helicase that competes with the phage DNA polymerase for the 3' end of DNA to prevent phage replication[62]. A recent study by Stokar-Avihail and colleagues systematically investigated molecular mechanisms of phage escape from 19 different phage-defence systems in bacteria and found that such escape was often linked to mutations in core replication proteins including DNA polymerase, DNA primase-helicase, ribonucleotide reductase or SSB proteins[8]. The authors speculate that, from the evolutionary point of view, it makes sense for the bacterial defences to target essential components of the viral core replication machinery as an escape mutation would likely induce greater fitness cost for the virus. We thus think that the observed diversity and mosaicism observed within and between the proteins associated with core nucleotide metabolism reflects the ongoing co-evolutionary arms race between bacterial phage-defence systems and phages co-adapting to new bacterial defences. Given that mutations can often bear a high fitness cost, recombination of existing folds could be a viable evolutionary mechanism of adaptation to move across the steep fitness landscape.

Altogether, our results can be viewed as one approach to identify evolutionary hotspots in phage genomes. Bacteria employ a wide range of, often highly genetically diverse, strategies to resist infection by phages and mobile genetic elements. Variation in how bacteria protect themselves over time, space and phylogeny means that no single strategy – or even a combination or strategies – can universally work for either side[63]. This is the type of scenario where one expects balancing selection (e.g., negative frequency-dependent selection) to maintain diversity of such strategies[64], and where genetic innovation can be evolutionarily favoured[65]. In line with this thinking, we would thus expect protein modularity, and its ongoing emergence, to become associated over time with functions that are essential in overcoming different bacterial resistance mechanisms that determine host range such as host entry, lysis and evasion of multiple bacterial defence systems. It is also tempting to speculate whether such co-evolutionary dynamics might have played an important role in evolution of the core replication machinery through occasional production of lasting functional innovation.

As mentioned before, each of our two approaches to detect protein modularity has strengths and weaknesses. While the domain approach is a robust approach to detect protein modularity, we showed that it is bound to often miss genuine emergence of modularity (e.g., horizontal swaps of homologous domains, sub-domain recombination or domain gain/loss), especially in proteins that are under-represented in domain databases. On the other hand, while sequence mosaicism is likely to identify these problematic cases, it can result in false-positive cases of modularity, for example stemming from highly variable rates of evolution in different areas of the protein. One good example are tail length tape measure proteins. They often exhibited mosaic signal by sequence but we were not able to confirm any genuine cases of recent emergence of modularity either due to the presence of long and repetitive coiled coil regions or due to occurrence of frequent splits of very long, near-identical ORFs into multiple ones. Ideally, the most robust approach to assess the role of emergence of protein modularity in phage evolution should use all three levels of information: sequence, domain and structure.

One important caveat of this work is that our conclusions are bound to be more robust for functions assigned to protein clusters which are more diverse and relatively common, like the structural proteins or core replication proteins. This of course stems from a common problem in studies of horizontal gene transfer and genetic recombination in that on the one hand recombination promotes the emergence of diversity but on the other hand greater diversity provides a stronger signal to detect composite sequences. Given the limitations discussed above, it is to be expected that a great reservoir of mosaicism exists beyond what was reported in this study, namely in (a) proteins of unknown functions, (b) proteins which are under-represented in domain databases and (c) in less frequent, accessory proteins that themselves could have emerged as a result of domain shuffling and diversifying selection acting on the phage pangenome. We thus expect that our results are only the tip of the iceberg which is the true extent of protein modularity in phage populations.

Methodology

Data

We downloaded all complete bacteriophage genomes from NCBI Virus in January 2022 using the following criteria: virus = bacteriophage, genome completeness = complete, sequence type = RefSeq, yielding 4,548 complete genomes. We then detected open reading frames in those genomes using the approach based on `MultiPhate2`[66]. This resulted in 462,721 predicted protein sequences which were clustered with `mmseqs2`[67] using the following parameters: minimum sequence identity = 0.3, sensitivity = 7, coverage = 0.95, yielding 133,624 clusters.

HMM profile construction

For each of the clusters, a representative protein sequence was taken as the one suggested by `mmseqs2` (i.e., sequence with the most alignments above the special or default thresholds with other sequences of the database and these matched sequences). Of those, 50 included more than 10 unknown characters and were thus excluded from further analysis. Each of the remaining 133,574 representative sequences was then used as a starting point to build a hidden Markov model (HMM) profile for each of the clusters. The profile was built by aligning the UniClust30 database[68] against each representative sequence with `hhblits` with the following parameters: minimum probability = 90%, minimum sequence identity with master sequence = 10%, minimum coverage with master sequence = 30%, and other parameters set to default[27]. The resulting profiles are referred to as rHMMs (HMM profiles of representative proteins) throughout this work. See also Supplementary Figure S1 for a visual outline of the methodology.

All-by-all profile-profile comparison

All 133,574 rHMMs were compared to each other using `hhblits` with the following parameters: minimum probability = 50% and other parameters set to default. Then, for every pair, we calculated query and subject coverage as the total number of residues in the aligned sequence regions by the length of the query and subject sequence, respectively. Unless stated otherwise, only hits with a minimum probability of $p \geqslant 0.95$ were considered. To assign rHMMs into protein families, we did as follows. First, we only considered all pairs of rHMMs with a pairwise coverage $\text{cov} = \min(\text{qcov},\text{scov}) \geqslant 0.8$. For each pair, we then calculated a weighted score of $p \times \text{cov}$, which was used as a weight of a undirected network. Finally, we used a Markov clustering algorithm (MCL)[69] with an inflation factor `--I 2` to cluster rHMMs into 72,078 families.

Functional annotation

To assign each rHMM into a functional category, we used the Prokaryotic Virus Remote Homologous Groups database (PHROGs; version 4)[28]. Every rHMM was compared with the PHROGs HMM profile database using `hhblits`. We used functional classes as defined by PHROGs, but we additionally simplified and merged the names referring to closely related biological functions (e.g., *RusA-like Holliday junction resolvase* and *RuvC-like Holliday junction resolvase* became *Holliday junction resolvase*; *Dda-like helicase* and *DnaB-like replicative helicase* became *DNA helicase*; *head-tail adaptor Ad1* and *head-tail adaptor Ad2* became *adaptor*, etc.). The exact mapping of used functional categories onto PHROGs is provided in Supplementary Table S1. Only functional classes that (i) were assigned to PHROGs with the total number of at least 500 sequences and (ii) were found in at least 20 rHMMs were considered (unless stated otherwise). Additionally, every rHMM was compared with a database of antidefence proteins[29] using hhblits, and those that had hits to PHROGs and some specific antidefence function were assigned the specific antidefense class. Functional classes were assigned as those with hits to a known class at 80% coverage and 95% probability hit threshold. Finally, rHMMs with hits to more than a single functional class were discarded unless they only co-occurred with generic classes like tail or structural protein.

Domain detection

To detect domains in rHMMs, we used the Evolutionary Classification of Protein Domains[30] database (ECOD, version from 13.01.2022). Each of the 133,574 rHMMs was compared to the `HHpred` version of the ECOD database using `hhblits` with with a minimum probability of 20% and otherwise default parameters. Domains were considered as those hits to rHMMs with probability $p \geqslant 0.95$ and subject coverage scov $\geqslant 0.7$.

Detection of mosaic protein pairs

To look for potential mosaicism between rHMMs at the domain level (cf., 2A), we searched for pairs of rHMMs that shared a domain of the same topology (i.e., fold; ECOD T-groups), detected at the 95%

probability threshold, while each containing domains that belonged to different ECOD X-groups (i.e., there is absence of evidence of homology between these domains at both sequence and structural level). To look for potential mosaicism between proteins at the sequence level (cf., Figure 4B), we compared all rHMMs with each other at the permissive probability threshold of 50% to account for potential distant homology between the two sequences. Again, the query and subject coverage were calculated as the total number of residues in the aligned sequence regions by their respective lengths. The pair of rHMMs was considered mosaic if it was found to share a similar genetic fragment (probability $p \geqslant 0.95$ percentage identity $p_{id} \geqslant 0.3$) in the background of the absence of homology at the permissive probability threshold: max(scov,qcov) $\leqslant 0.5$. We only considered rHMM pairs with a minimum aligned fragment length of 50aa.

Odds ratio to be over-represented in proteins with evidence of mosaicism

Each domain architecture was classified as mosaic (i.e., having evidence of mosaicism) or non-mosaic (no evidence of mosaicism). Then for each topology (ECOD T group) we calculated the number mosaic domain architectures including this topology (m_t), number of mosaic domain architectures not including this topology (m_{nt}), number of non-mosaic domain architectures including this topology (n_t) and number of non-mosaic domain architectures not including this topology (n_{nt}). Then the odds ratio was calculated as:

$$\mathrm{OR} = \frac{m_t/m_{nt}}{n_t/n_{nt}}.$$

Reproducibility

Code used to generate the data and figures is publicly available at:

- `https://github.com/bioinf-mcb/phage-protein-modularity-data`
- `https://github.com/bioinf-mcb/phage-protein-modularity-figures`

Mapping table between rHMMs and the NCBI database is available via:

- `https://figshare.com/projects/Protein_modularity_in_phages/156350`.

Domain architecture lookup in different functional classes is available at:

- `https://bognasmug.shinyapps.io/PhageDomainArchitectureLookup`.

Acknowledgements

This work was financed by the Polish National Agency for Academic Exchange (RM), the Polish National Science Centre OPUS Grant (grant 2020/37/B/NZ8/03492) and the EMBO Installation Grant (RM). SDH was supported by First TEAM programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund (grant POIR.04.04.00-00-5CF1/18-00). We thank Virkam Alva, Carina Büttner, Alan R. Davidson, Mart Krupovic, Andrei Lupas, Joana Pereira and Eugen Pfeifer for helpful discussions and/or comments related to this work which was instrumental in finalising this manuscript.

References

[1] Dion, M. B., Oechslin, F., and Moineau, S. (03, 2020) Phage diversity, genomics and phylogeny. *Nat Rev Microbiol,* **18**(3), 125–138.

[2] Ofir, G. and Sorek, R. (03, 2018) Contemporary Phage Biology: From Classic Models to New Insights. *Cell,* **172**(6), 1260–1270.

[3] Neri, U., Wolf, Y. I., Roux, S., Camargo, A. P., Lee, B., Kazlauskas, D., Chen, I. M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., Bryant, D. A., Bhaya, D., Krupovic, M., Dolja, V. V., Kyrpides, N. C., Koonin, E. V., Gophna, U., Narrowe, A. B., Probst, A. J., Sczyrba, A., Kohler, A., Séguin, A., Shade, A., Campbell, B. J., Lindahl, B. D., Reese, B. K., Roque, B. M., DeRito, C., Averill, C., Cullen, D., Beck, D. A. C., Walsh, D. A., Ward, D. M., Wu, D., Eloe-Fadrosh, E., Brodie, E. L., Young, E. B., Lilleskov, E. A., Castillo, F. J., Martin, F. M., LeCleir, G. R., Attwood, G. T., Cadillo-Quiroz, H., Simon, H. M., Hewson, I., Grigoriev, I. V., Tiedje, J. M., Jansson, J. K., Lee, J., VanderGheynst, J. S., Dangl, J., Bowman, J. S., Blanchard, J. L., Bowen, J. L., Xu, J., Banfield, J. F., Deming, J. W., Kostka, J. E., Gladden, J. M., Rapp, J. Z., Sharpe, J., McMahon, K. D., Treseder, K. K., Bidle, K. D., Wrighton, K. C., Thamatrakoln, K., Nusslein, K., Meredith, L. K., Ramirez, L., Buee, M., Huntemann, M., Kalyuzhnaya, M. G., Waldrop, M. P., Sullivan, M. B., Schrenk, M. O., Hess, M., Vega, M. A., O'Malley, M. A., Medina, M., Gilbert, N. E., Delherbe, N., Mason, O. U., Dijkstra, P., Chuckran, P. F., Baldrian, P., Constant, P., Stepanauskas, R., Daly, R. A., Lamendella, R., Gruninger, R. J., McKay, R. M., Hylander, S., Lebeis, S. L., Esser, S. P., Acinas, S. G., Wilhelm, S. S., Singer, S. W., Tringe, S. S., Woyke, T., Reddy, T. B. K., Bell, T. H., Mock, T., McAllister, T., Thiel, V., Denef, V. J., Liu, W. T., Martens-Habbena, W., Allen Liu, X. J., Cooper, Z. S., and Wang, Z. (Oct, 2022) Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell,* **185**(21), 4023–4037.

[4] Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (Mar, 2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science,* **359**(6379).

[5] Egido, J. E., Costa, A. R., Aparicio-Maldonado, C., Haas, P. J., and Brouns, S. J. J. (Feb, 2022) Mechanisms and clinical importance of bacteriophage resistance. *FEMS Microbiol Rev,* **46**(1).

[6] Tesson, F., é, A., Mordret, E., Touchon, M., res, C., Cury, J., and Bernheim, A. (May, 2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun,* **13**(1), 2561.

[7] Samson, J. E., Magadán, A. H., Sabri, M., and Moineau, S. (Oct, 2013) Revenge of the phages: defeating bacterial defences. *Nat Rev Microbiol,* **11**(10), 675–687.

[8] Stokar-Avihail, A., Fedorenko, T., r, J., Garb, J., Leavitt, A., Millman, A., Shulman, G., Wojtania, N., Melamed, S., Amitai, G., and Sorek, R. (Apr, 2023) Discovery of phage determinants that confer sensitivity to bacterial immune systems. *Cell,* **186**(9), 1863–1876.

[9] Botstein, D. (1980) A theory of modular evolution for bacteriophages. *Ann N Y Acad Sci,* **354**, 484–490.

[10] Hatfull, G. F. and Hendrix, R. W. (Oct, 2011) Bacteriophages and their genomes. *Curr Opin Virol,* **1**(4), 298–303.

[11] Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (Mar, 1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A,* **96**(5), 2192–2197.

[12] Moura de Sousa, J. A., Pfeifer, E., Touchon, M., and Rocha, E. P. C. (05, 2021) Causes and Consequences of Bacteriophage Diversification via Genetic Exchanges across Lifestyles and Bacterial Taxa. *Mol Biol Evol,* **38**(6), 2497–2512.

[13] Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (Apr, 2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol,* **25**(4), 762–777.

[14] Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D., and Sullivan, M. B. (Jun, 2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol,* **37**(6), 632–639.

[15] Lima-Mendez, G., Toussaint, A., and Leplae, R. (Oct, 2011) A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Res Microbiol,* **162**(8), 737–746.

[16] Iranzo, J., Krupovic, M., and Koonin, E. V. (08, 2016) The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio,* **7**(4).

[17] Latka, A., Leiman, P. G., Drulis-Kawa, Z., and Briers, Y. (2019) Phages. *Front Microbiol,* **10**, 2649.

[18] Oechslin, F., Zhu, X., Dion, M. B., Shi, R., and Moineau, S. (08, 2022) Phage endolysins are adapted to specific hosts and are evolutionarily dynamic. *PLoS Biol,* **20**(8), e3001740.

[19] Dunne, M., Rupf, B., Tala, M., Qabrati, X., Ernst, P., Shen, Y., Sumrall, E., Heeb, L., Plückthun, A., Loessner, M. J., and Kilcher, S. (Oct, 2019) Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. *Cell Rep,* **29**(5), 1336–1350.

[20] Latka, A., Lemire, S., Grimon, D., Dams, D., Maciejewska, B., Lu, T., Drulis-Kawa, Z., and Briers, Y. (05, 2021) Phages Switches Their Capsule Serotype Specificity. *mBio,* **12**(3).

[21] Gerstmans, H., Grimon, D., Gutiérrez, D., Lood, C., Rodríguez, A., van Noort, V., Lammertyn, J., Lavigne, R., and Briers, Y. (06, 2020) A VersaTile-driven platform for rapid hit-to-lead development of engineered lysins. *Sci Adv,* **6**(23), eaaz1136.

[22] Dunne, M., Prokhorov, N. S., Loessner, M. J., and Leiman, P. G. (04, 2021) Reprogramming bacteriophage host range: design principles and strategies for engineering receptor binding proteins. *Curr Opin Biotechnol,* **68**, 272–281.

[23] Cardarelli, L., Pell, L. G., Neudecker, P., Pirani, N., Liu, A., Baker, L. A., Rubinstein, J. L., Maxwell, K. L., and Davidson, A. R. (Aug, 2010) Phages have adapted the same protein fold to fulfill multiple functions in virion assembly. *Proc Natl Acad Sci U S A,* **107**(32), 14384–14389.

[24] Veesler, D. and Cambillau, C. (Sep, 2011) A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiol Mol Biol Rev,* **75**(3), 423–433.

[25] Jachiet, P. A., Colson, P., Lopez, P., and Bapteste, E. (Aug, 2014) Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. *Genome Biol Evol,* **6**(9), 2195–2205.

[26] Soeding, J. (Apr, 2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics,* **21**(7), 951–960.

[27] Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (Sep, 2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics,* **20**(1), 473.

[28] Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M. A., and Enault, F. (Sep, 2021) PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform,* **3**(3), lqab067.

[29] Samuel, B. and Burstein, D. (2023) A diverse repertoire of anti-defense systems is encoded in the leading region of plasmids. *bioRxiv,* pp. 2023–02.

[30] Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B. H., and Grishin, N. V. (Dec, 2014) ECOD: an evolutionary classification of protein domains. *PLoS Comput Biol,* **10**(12), e1003926.

[31] Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., and Iyer, L. M. (Apr, 2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev,* **29**(2), 231–262.

[32] Alva, V., Koretke, K. K., Coles, M., and Lupas, A. N. (Jun, 2008) Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Curr Opin Struct Biol,* **18**(3), 358–365.

[33] Iyer, L. M., Koonin, E. V., and Aravind, L. (Jan, 2003) Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct Biol,* **3**, 1.

[34] Bork, P. and Koonin, E. V. (Dec, 1994) A P-loop-like motif in a widespread ATP pyrophosphatase domain: implications for the evolution of sequence motifs and enzyme activity. *Proteins,* **20**(4), 347–355.

[35] Medvedev, K. E., Kinch, L. N., and Grishin, N. V. (08, 2018) Functional and evolutionary analysis of viral proteins containing a Rossmann-like fold. *Protein Sci,* **27**(8), 1450–1463.

[36] Vyas, P., Trofimyuk, O., Longo, L. M., Deshmukh, F. K., Sharon, M., and Tawfik, D. S. (Apr, 2021) Helicase-like functions in phosphate loop containing beta-alpha polypeptides. *Proc Natl Acad Sci U S A,* **118**(16).

[37] Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., and Bujnicki, J. M. (Apr, 2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res,* **42**(7), 4160–4179.

[38] Leipe, D. D., Koonin, E. V., and Aravind, L. (Oct, 2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol,* **343**(1), 1–28.

[39] Kazlauskas, D., Krupovic, M., and Venclovas, C. (06, 2016) The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res,* **44**(10), 4551–4564.

[40] Filee, J., Forterre, P., Sen-Lin, T., and Laurent, J. (Jun, 2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol,* **54**(6), 763–773.

[41] Morcinek-Orlowska, J., Zdrojewska, K., and Wegrzyn, A. (Jan, 2022) Bacteriophage-Encoded DNA Polymerases-Beyond the Traditional View of Polymerase Activities. *Int J Mol Sci,* **23**(2).

[42] Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., bler, J., Lozajic, M., Gabler, F., Soeding, J., Lupas, A. N., and Alva, V. (Jul, 2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol,* **430**(15), 2237–2243.

[43] Nobrega, F. L., Vlot, M., de Jonge, P. A., Dreesens, L. L., Beaumont, H. J. E., Lavigne, R., Dutilh, B. E., and Brouns, S. J. J. (Dec, 2018) Targeting mechanisms of tailed bacteriophages. *Nat Rev Microbiol,* **16**(12), 760–773.

[44] Payne, K. M. and Hatfull, G. F. (2012) Mycobacteriophage endolysins: diverse and modular enzymes with multiple catalytic activities. *PLoS One,* **7**(3), e34052.

[45] Kolodny, R., Nepomnyachiy, S., Tawfik, D. S., and Ben-Tal, N. (May, 2021) Bridging Themes: Short Protein Segments Found in Different Architectures. *Mol Biol Evol,* **38**(6), 2191–2208.

[46] Gilchrist, C. L. M. and Chooi, Y. H. (Aug, 2021) clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics,* **37**(16), 2473–2475.

[47] Murphy, J., Bottacini, F., Mahony, J., Kelleher, P., Neve, H., Zomer, A., Nauta, A., and van Sinderen, D. (Feb, 2016) Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci Rep,* **6**, 21345.

[48] Hayes, S., Mahony, J., Vincentelli, R., Ramond, L., Nauta, A., van Sinderen, D., and Cambillau, C. (Jul, 2019) Ubiquitous Carbohydrate Binding Modules Decorate 936 Lactococcal Siphophage Virions. *Viruses,* **11**(7).

[49] Fraser, J. S., Yu, Z., Maxwell, K. L., and Davidson, A. R. (Jun, 2006) Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J Mol Biol,* **359**(2), 496–507.

[50] Jayaraman, V., o, S., a, L., and Laurino, P. (Jul, 2022) Mechanisms of protein evolution. *Protein Sci,* **31**(7), e4362.

[51] de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J., and Dutilh, B. E. (01, 2019) Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends Microbiol,* **27**(1), 51–63.

[52] De Sordi, L., Khanna, V., and Debarbieux, L. (Dec, 2017) The Gut Microbiota Facilitates Drifts in the Genetic Diversity and Infectivity of Bacterial Viruses. *Cell Host Microbe,* **22**(6), 801–808.

[53] Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (01, 2020) A continuous evolution system for contracting the host range of bacteriophage T7. *Sci Rep,* **10**(1), 307.

[54] Holt, K. E., Lassalle, F., Wyres, K. L., Wick, R., and Mostowy, R. J. (Jul, 2020) Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. *ISME J,* **14**(7), 1713–1730.

[55] Wyres, K. L., Lam, M. M. C., and Holt, K. E. (Jun, 2020) Population genomics of Klebsiella pneumoniae. *Nat Rev Microbiol,* **18**(6), 344–359.

[56] Ouyang, R., Costa, A. R., Cassidy, C. K., Otwinowska, A., Williams, V. C. J., Latka, A., Stansfeld, P. J., Drulis-Kawa, Z., Briers, Y., Pelt, D. M., Brouns, S. J. J., and Briegel, A. (Nov, 2022) High-resolution reconstruction of a Jumbo-bacteriophage infecting capsulated bacteria using hyper-branched tail fibers. *Nat Commun,* **13**(1), 7241.

[57] Turner, R. D., Vollmer, W., and Foster, S. J. (Mar, 2014) Different walls for rods and balls: the diversity of peptidoglycan. *Mol Microbiol,* **91**(5), 862–874.

[58] Koonin, E. V., Krupovic, M., Ishino, S., and Ishino, Y. (06, 2020) The replication machinery of LUCA: common origin of DNA replication and transcription. *BMC Biol,* **18**(1), 61.

[59] Wang, C. C., Yeh, L. S., and Karam, J. D. (Nov, 1995) Modular organization of T4 DNA polymerase. Evidence from phylogenetics. *J Biol Chem,* **270**(44), 26558–26564.

[60] Petrov, V. M., Nolan, J. M., Bertrand, C., Levy, D., Desplats, C., Krisch, H. M., and Karam, J. D. (Aug, 2006) Plasticity of the gene functions for DNA replication in the T4-like phages. *J Mol Biol,* **361**(1), 46–68.

[61] LeRoux, M., Srikant, S., Teodoro, G. I. C., Zhang, T., Littlehale, M. L., Doron, S., Badiee, M., Leung, A. K. L., Sorek, R., and Laub, M. T. (07, 2022) The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat Microbiol,* **7**(7), 1028–1040.

[62] Bari, S. M. N., Chou-Zheng, L., Howell, O., Hossain, M., Hill, C. M., Boyle, T. A., Cater, K., Dandu, V. S., Thomas, A., Aslan, B., and Hatoum-Aslan, A. (04, 2022) A unique mode of nucleic acid immunity performed by a multifunctional bacterial enzyme. *Cell Host Microbe,* **30**(4), 570–582.

[63] Bernheim, A. and Sorek, R. (Feb, 2020) The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol,* **18**(2), 113–119.

[64] Rocha, E. P. C. and Bikard, D. (Jan, 2022) Microbial defenses against mobile genetic elements and viruses: Who defends whom from what?. *PLoS Biol,* **20**(1), e3001514.

[65] Ebert, D. and Fields, P. D. (Dec, 2020) Host-parasite co-evolution and its genomic signature. *Nat Rev Genet,* **21**(12), 754–768.

[66] Ecale Zhou, C. L., Kimbrel, J., Edwards, R., McNair, K., Souza, B. A., and Malfatti, S. (05, 2021) MultiPhATE2: code for functional annotation and comparison of phage genomes. *G3 (Bethesda),* **11**(5).

[67] Steinegger, M. and Söding, J. (11, 2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol,* **35**(11), 1026–1028.

[68] Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Soeding, J., and Steinegger, M. (Jan, 2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res,* **45**(D1), D170–D176.

[69] Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (Apr, 2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res,* **30**(7), 1575–1584.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytables.xlsx
- supplementarymaterial.pdf