Protein modularity in phages is extensive and associated with functions linked to core replication machinery and host tropism determinants

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

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

2) Institut Pasteur, Université de Paris, 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 in that their diversity can be viewed through the existence of evolutionary independent functional modules that occur and recombine in different combinations. Multiple studies have demonstrated how such mosaicism emerges in natural populations and facilitates adaptation to their hosts, bacteria. However, less is known about the extent of (within-)protein modularity and its impact on phage evolution. To fill this knowledge gap, here we quantified such modularity by detecting instances of 'protein mosaicism', defined as homologous fragment between two otherwise unrelated proteins. We then used highly sensitive homology detection to quantify protein mosaicism between pairs of 133,624 representative phage proteins, and to understand its relationship with genetic and functional diversity in phage genomes. We found that diverse functional classes often shared homologous fragments and domains, with multiple instances of such mosaicism having emerged relatively recently. We detected the strongest signal for protein mosaicism in receptor-binding proteins, endolysins as well as the core replication machinery, with DNA polymerases as mosaic outliers. We argue that the extensive protein modularity linked to those functional classes is reflective of the co-evolutionary interactions with bacterial hosts but with important differences in the underlying nature of diversifying selection.

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] or strategies to evade bacterial defences[4]. 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 entire genomes – facilitates emergence of the new, mosaic genotypes that are advantageous in a given niche[5]. 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[6,7]. Gene flow can also occur relatively frequently between genetically unrelated phages[8]. As a result, bacteriophage population structure is better represented as a network rather than a phylogenetic tree[9,10], where modules of functionally related groups of genes have a coherent evolutionary history[11,12].

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 phage proteins have modular 3-dimensional structures. Very good examples are receptor-binding proteins (including tail fibres, tail spikes) or endolysins. These proteins not only have a remarkable modularity at both genetic and structural levels[13,14] but their domains can be experimentally shuffled to produce a viable phage virion with a modified host range[15,16,17,18]. Second, previous studies have suggested that structural phage proteins of different functions have evolved to reuse the same folds for different purpose and that one of the genetic mechanisms of such evolution could be recombination[19,20]. Finally, studies that looked for the presence of composite genes (fusions of different gene families) in viral genomes found this phenomenon to be extensive[21,12]. However, the true extent of protein modularity, and especially its relationship to genetic and functional diversity in phages, remains not fully understood.

In this study, we aimed to better understand protein modularity and its role in evolution of phages.

To this end, we analysed over 460,000 phage proteins to detect instances of 'protein mosaicism', defined as two non-homologous protein sequences that share 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 highly sensitive homology search approach (HMM-HMM protein comparison using *HHpred*), we detected instances of protein mosaicism between proteins assigned to various functional classes and inferred their domain architecture. We found that, while protein mosaicism is extensive, some functional groups are associated with a particularly highly mosaic composition, including tail fibres, tail spikes, endolysins and DNA polymerases.

Results

Protein fragment-based functional annotation 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 next searched the `UniClust30` database using 133,624 representative protein sequences and converted the resulting alignments into HMM profiles (henceforth referred to as representative HMM profiles or rHMMs). To assign functions to representative proteins, we used `hhblits` to search each rHMM against the PHROGs database, which to our knowledge provides the most accurate mapping between diverse phage proteins and manually curated functional annotations[22].

We investigated the robustness of the PHROG functional annotations (which were additionally simplified to combine clearly synonymous 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 sequence coverage, unlike other parameters, had a very strong effect on functional assignment (see Sup-

plementary 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 low coverage threshold every second rHMM had multiple, different functional assignments. Furthermore, upon closer investigation we 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 a lower sequence coverage threshold, co-occurrences between clearly different functions became more and more common and affected the majority of functions (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 cutoffs for the PHROG annotations to 95% and 80%, respectively, while conservatively excluding all rHMMs with hits to more than a single functional class.

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 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[23]. We then used sensitive HMM-HMM comparison (`HHpred`) to detect presence of these domains in rHMMs (see Methodology and Supplementary Figure S1). Importantly, the ECOD database classifies all domains into hierarchical groups (i) that share homology and have the same topology (T-groups), (ii) that share homology but have different topologies (H-groups), and (iii) with possible homology due to evidence at the structural level only (X-groups). Hence, if two domains belong to different X-groups, their homology can be detected at neither genetic or structural level.

Figure 1 shows a distribution of ECOD domain families (H-groups) in different functional classes. (The distribution of T-groups across functional classes is also shown in a Supplementary Figure S4.) As expected, many phage proteins with different functions share homologous domains. These include well known examples of helix-turn-helix domains found in transcriptional regulator/repressor proteins, integrases, transposases or DNA-binding proteins[24]; the RIFT-related domains found in many structural proteins like tail and neck proteins[25] but also for example in RNA polymerases in the form of double-psi barrels[26]; P-loop domain-related family found in ligases, kinases or helicases[27,28,29]; or the ribonuclease H-like domain family found in many DNA processing enzymes like Holliday junction resolvases, exo- and endonucleases or transposases[30]. While some instances of domain sharing by different functional classes can simply be explained by the fact that some proteins belonging to different functional classes may indeed have the same function (e.g., generic functional class 'tail' likely contains multiple tail proteins like tail fibre or tail spike), for most cases such domain sharing very likely represents a genuine presence of the same fold in proteins carrying out different functions. Therefore, these results clearly show an important, evolutionary relationship between proteins of multiple, diverse functions.

Another important aspect of the domain-function mapping was how different domains were distributed across functions. First, some domains, particularly helix-turn-helix and P-loop domains, were not only more common than others but also genetically more diverse, and we found a strongly significant, positive correlation between domain frequency (H-groups) and diversity (see Supplementary Figure S5). This

suggests that the prevalence of these domains is not just a straightforward consequence of their genetic conservation but that some folds have likely been more evolutionarily successful than others due to their generalist nature (e.g., DNA binding or NTP hydrolysis). Second, multiple distinct domains were detected in proteins of the same function. Given that different H-groups are by definition those between which there is no signal of homology, it points to the multi-domain architecture of certain functional classes and their potentially complex, evolutionary history. We thus hypothesised that this apparent complexity is indicative of modularity and ongoing domain shuffling of proteins that belong to different functional classes.

Figure 1. Diverse protein functions often share homologous domains. Heatmap showing groups of homologous ECOD domains (H-group names; X-axis) found in proteins with different functional classes defined by PHROGs (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 in a given functional class. Colours correspond to different PHROG categories. Only domains found in multiple functional classes (at least 3) are shown.

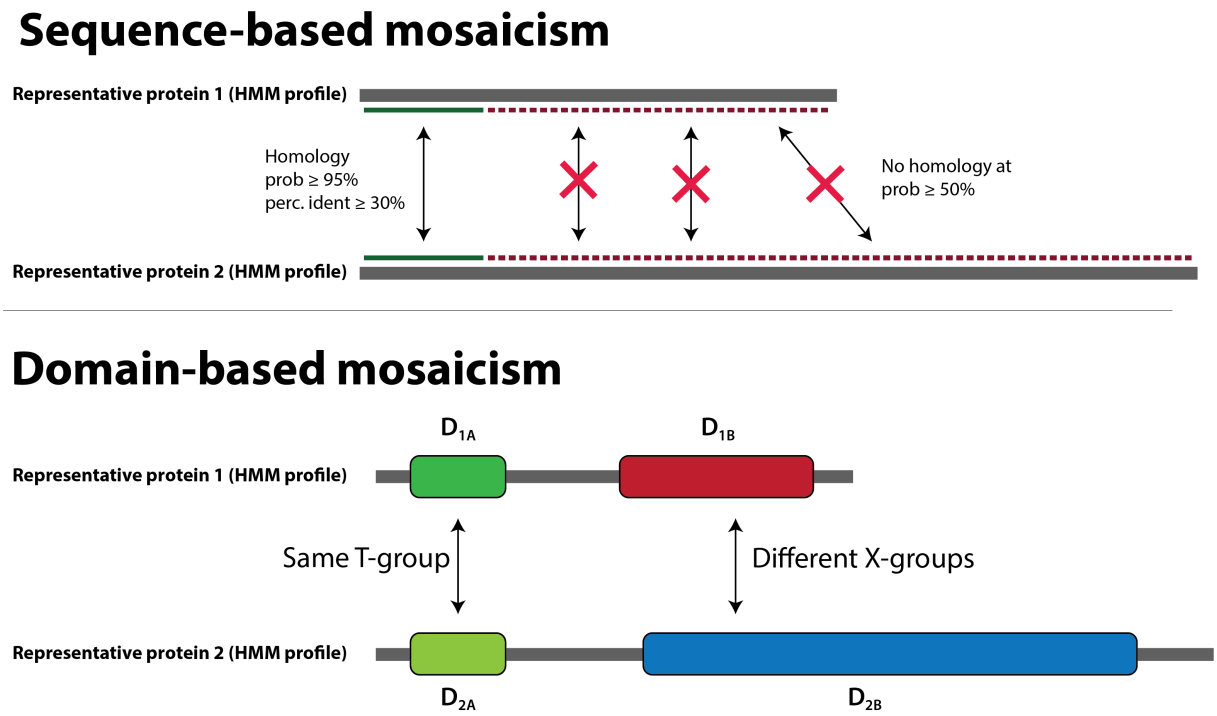

Figure 2. Two definitions of protein mosaicism. (Top) Sequence-based mosaicism for any rHMM pair was defined as presence of sequence-based homology in the background of absence of homology using their HMM profiles. First, pairwise coverage was calculated as the total aligned length using a permissive 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%) and (iii) had percentage identity of at least 30%. (Bottom) Domain-based mosaicism for any rHMM pair was defined when (i) both proteins had at two distinct ECOD domains detected, (ii) both shared a homologous domain (i.e., had domains belonging to the same T-groups), and (iii) both additionally contained non-homologous domains (i.e., belonging to different X-groups).

Protein mosaicism is associated with host range determinants and DNA replication proteins

To quantify the frequency of protein modularity in our dataset, we carried out the all-against-all comparison of 133k rHMMs and investigated the relationship between their sequence similarity and coverage (see Supplementary Figure S6). The results show that, while most protein 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. To systematically detect pairs of functional classes associated with proteins that show evidence of modularity, we defined 'protein mosaicism' using two complementary measures: sequence and domain (see Figure 2 for a visual explanation). Briefly, we reported sequence mosaicism when two proteins shared detectable similarity over less than half of their lengths but at a percentage identity (in the aligned region) of at least 30%. Domain mosaicism was reported when two proteins had each at least two domains of which at least one domain was of the same topology (same T-groups) while at least one other belonged to different structural architectures (different X-groups). The rationale for using both measures is that, while both have advantages and disadvantages, they are complimentary in overcoming different types of bias and hence capturing mosaicism in the data (see also Discussion).

We found that protein mosaicism is common in phages, with a mosaic signal in at least one protein assigned to 85 out of 95 functional classes using either criterion (sequence or domain), and in 53 out of 95 functional classes using both criteria (sequence and domain). Figure 3A shows a network of functions for which evidence of mosaicism was found between at least 20 different rHMM pairs, with different edges corresponding to signal from both tests of mosaicism. Functional groups where domain-level mosaicism was particularly common were DNA/RNA metabolism, transcription regulation, structural tail proteins and lysis proteins. In particular, functions where multiple rHMMs showed evidence of within-class mosaicism using both domain and sequence test were endolysins, tail fibers and spikes (and the generic 'tail' category), transcriptional regulators, as well as functional classes related to core nucleotide metabolism, particularly between DNA primases, DNA helicases, DNA polymerases, DNA ligases and others.

To further confirm that these functional classes were associated with protein mosaicism (and not just more common), we next calculated the odds ratio for each functional class to be over-represented in

mosaic proteins. We first subdivided all rHMMs into those with a mosaic signal (sequence or domain) and those without the signal. Then, for each functional class, all PHROG HMMs associated with this class were used to detect all rHMMs with a possible link with this class (a permissive sequence coverage threshold of 10%, instead of 80%, was used). Finally, we calculated the odds ratio that this functional class is found in a greater proportion of mosaic rHMMs than non-mosaic rHMMs. The results are shown in Figure 3B. Consistently with the trend seen in the network in Fig. 3A, we found that the functional classes over-represented within the mosaic proteins are linked to tail (e.g., tail fibre and tail spike, but also tail length tape measure or minor tail), nucleotide metabolism (endonucleases, DNA polymerases, primases and helicases) and endolysins. The proportions of different functional classes in mosaic and non-mosaic rHMM, used to calculate the odds ratio, are shown in Supplementary Figure S7 and the full list of functions with a significant over-representation in mosaic proteins is shown in Supplementary Table S2.

Analogously, we next investigated an association between the frequency of specific protein domains and protein mosaicism. Topologies which were most commonly shared between rHMMs that show evidence of mosaicism (see Supplementary Figure S8) were linked to nucleotide metabolism (e.g., P-loop containing nucleoside triphosphate hydrolases or 'winged' helix-turn-helix), to structural tail proteins or depolymerases (e.g., RIFT-related tail proteins or pectin lyase-like) or lysins (e.g., lysozyme-like). To assess which of these domains were in fact over-represented in rHMMs with evidence of mosaicism (only those rHMMs where any ECOD domain was found were considered), we calculated the frequency of ECOD domains (T-groups) found in mosaic and non-mosaic rHMMs and calculated the odds ratio for a given topology to be over-represented in mosaic rHMMs (Figure 3C). Topologies with the greatest odds ratio were those often found in DNA/RNA metabolism proteins (particularly ribonucleotide reductase, DNA polymerase, DNA primase and RNA ligase) and tail spike proteins. (The full list of domains with a significant over-representation in mosaic proteins is shown in Supplementary Table S3.) Overall, these results highlight a few functional classes, namely receptor-binding proteins, endolysins and DNA/RNA replication proteins, as those showing the strongest signs of mosaicism.

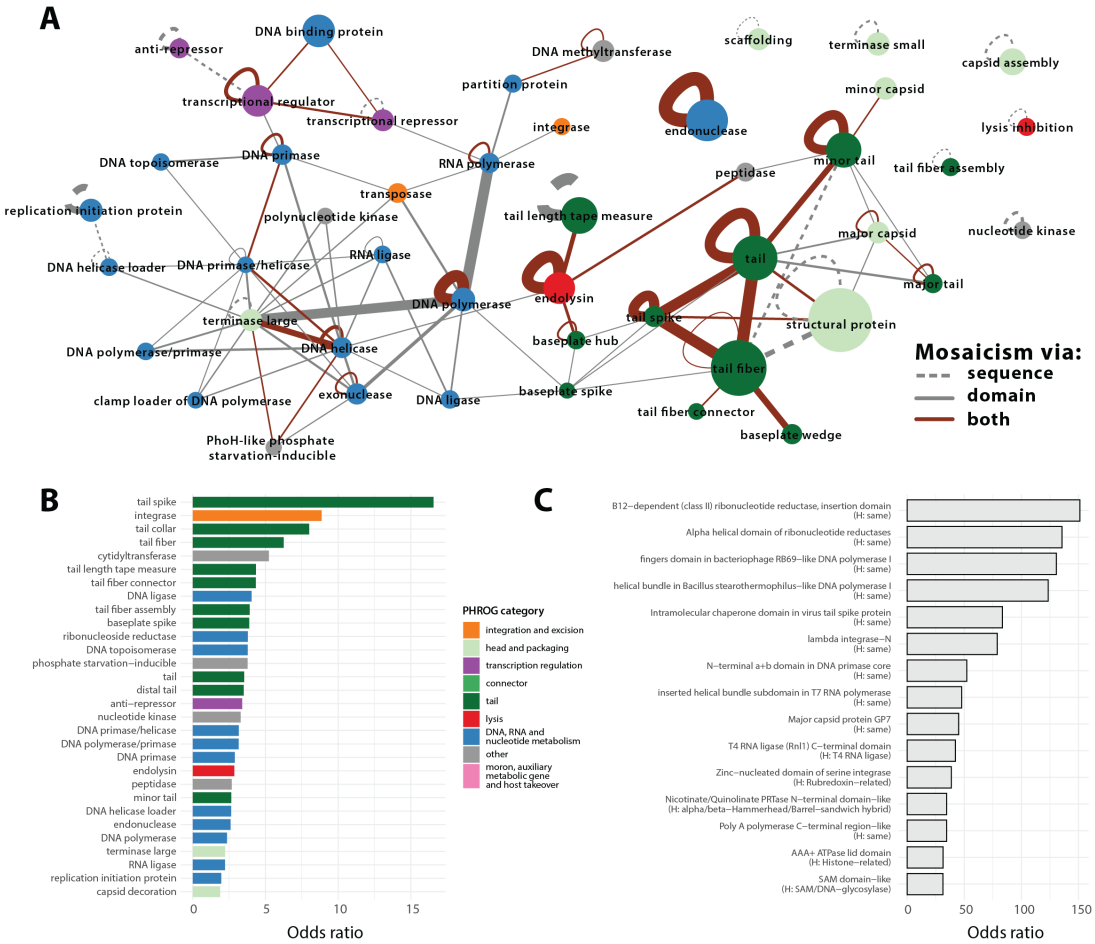

Figure 3. Protein mosaicism is common in phages. (A) Mosaic network of protein functions. Each node represents a functional class and edges link functional classes where evidence of mosaicism was found between at least 20 different rHMM pairs. The type of line corresponds to the level of evidence: either sequence-only (dashed), domain-only (solid) or both (red); line thickness corresponds to the number of rHMM pairs with either level of evidence of mosaicism (sequence or domain). (B) Bar plot shows the odds ratio between the proportion of mosaic and non-mosaic rHMMs where a given domain (ECOD T-groups) was found. Mosaicism was considered as either level of evidence (sequence or domain). The topologies displayed are sorted by the calculated odds ratio, provided that their over-representation in mosaic rHMMs was statistically significant (Fisher's exact test; level of significance was Bonferroni corrected for multiple testing). (C) Bar plot shows the odds ratio between the proportion of mosaic and non-mosaic rHMMs with a hit to a given functional class at 10% coverage. Results are again sorted by the magnitude of the effect (odds ratio) provided that the over-representation was statistically significant (same as in panel B).

Tail spikes, endolysins and DNA polymerases are mosaicism outliers

Since our results suggest that domain shuffling is an important process that shapes diversification of proteins associated with some biological functions, we hypothesised that some functional classes may be more diverse and exhibit a stronger degree of mosaicism than others. To address this, we defined two measures of genetic diversity for each functional class: genetic and structural. Genetic diversity was measured by the number of protein families, where protein family was defined as a cluster of similar rHMMs (see Methods). Structural diversity was measured by the number of unique domain architectures (ECOD T-groups). Functional classes with a poor ECOD coverage (less than 20%; see Supplementary Figures S9 and S10) were removed from further analyses with the rationale that for those functions there was a strong bias against the domain-based measure of structural diversity. Interestingly, some of the removed functional classes were also amongst the most diverse when considering the number of protein families (e.g., tail length tape measure or lysins; see Supplementary Figure S9.). This suggests that their diversity may have emerged as a result of shuffling of sub-domain fragments or yet unidentified domains.

The relationship between the two measures of diversity for the remaining 72 functional classes is shown in Figure 4A. We found a positive correlation between the two measures (Spearman's rank test; $\rho = 0.61$, $p < 10^{-7}$). We then subdivided the functional classes into those of high and low diversity. Specifically, functional classes that were in top 50th percentile of both genetic and structural diversity were considered of high diversity, while classes that were in the bottom 50th percentile of both genetic and structural diversity were considered of low diversity; remaining classes were discarded. We found that the proteins belonging to the high diversity functional classes had on average a greater number of topologies (1.22 vs. 1.59) and were longer (219 vs. 334 aa; one-sided Wilcoxon rank sum test, $p < 10^{-4}$ and $p < 0.02$ respectively) compared to the proteins belonging to the low diversity functional classes.

We also found that the diverse functional classes had on average a greater proportion of proteins that exhibited a mosaic signal (sequence or domain; Wilcoxon rank sum test, $p = 0.001$). This is expected as diversifying selection is likely to promote mosaicism via protein fragment shuffling while a greater diversity allows for a stronger signal to detect mosaicism. Hence to minimise such bias in comparing the levels of mosaicism between different functions, we focused on the high diversity functional classes. Figure 4B shows the 30 most diverse functional classes, their average number of topologies per protein and results of both measures of mosaicism (sequence and domain). First, several functional classes were outliers in terms of mosaicism (both sequence and domain), particularly DNA polymerases, endolysins and tail

spikes. Second, some functional classes showed predominantly sequence-based evidence for mosaicism, e.g., tail fibre, baseplate hub or peptidases. Some of the functional classes for which there was evidence of sequence-based mosaicism also had a low average number of domains per protein (e.g., endonucleases or capsid decoration proteins) or generally had a low ECOD coverage (e.g., replication initiation protein). This suggests that proteins associated with these functions might have mosaic domain architectures that we could not detect in this study or that the sequence-based approach detects the signal of sub-domain recombination which the domain-based mosaicism approach cannot. Finally, several functions – and in particular receptor-binding proteins – remained modular by sequence when requiring a high sequence similarity in the homologous region (50% or greater).

To better understand how the measures in Figure 4B change when considering the between-class comparisons instead of the within-class comparison, we applied them to PHROG functional categories – nine related groups of functional classes that are typically co-located in viral genomes ('head and packaging', 'connector', 'tail', 'DNA, RNA and nucleotide metabolism', 'integration and excision', 'lysis', 'transcription regulation', 'moron, auxiliary metabolic gene and host takeover' and 'other'; see Figure 4C). While the strongest signal for domain-based mosaicism was obtained for the 'DNA, RNA and nucleotide metabolism' category, followed by 'tail' and 'lysis', by far the greatest proportion of proteins where sequence-based mosaicism was found were proteins from the 'tail' and 'lysis' categories. To further investigate the nature and prevalence of recently-emerged protein mosaicism, we searched for pairs of functional classes with multiple instances of sequence-based mosaicism when requiring sequence similarity of 70% or 90% (see Supplementary Figure S11). Interestingly, we found evidence of such recent mosaicism not only in receptor-binding proteins and endolysins but also in several other classes like DNA polymerase, replication initiation protein or DNA methyltransferases. We then searched for ECOD domains associated with such recent mosaicism, namely domains overlapping with high-similarity protein fragments in otherwise unrelated proteins (see Supplementary Figure S12). Accordingly, we found multiple instances of highly similar fragments overlapping with domains associated with receptor-binding proteins (e.g., 'Putative tailspike protein N-terminal domain', 'Pectin lyase-like', 'N-terminal Ig-like domain in baseplate protein') and core replication proteins ('P-loop containing nucleoside triphosphate hydrolases' and 'Riboncluease H-like'). These results suggest that the emergence of protein mosaicism is an ongoing evolutionary phenomenon.

To better understand the nature of protein modularity, as well as patterns seen in functional classes

linked to tail and lysis, we next focused on the three mosaicism-outliers: DNA polymerase, endolysin, tail spike.

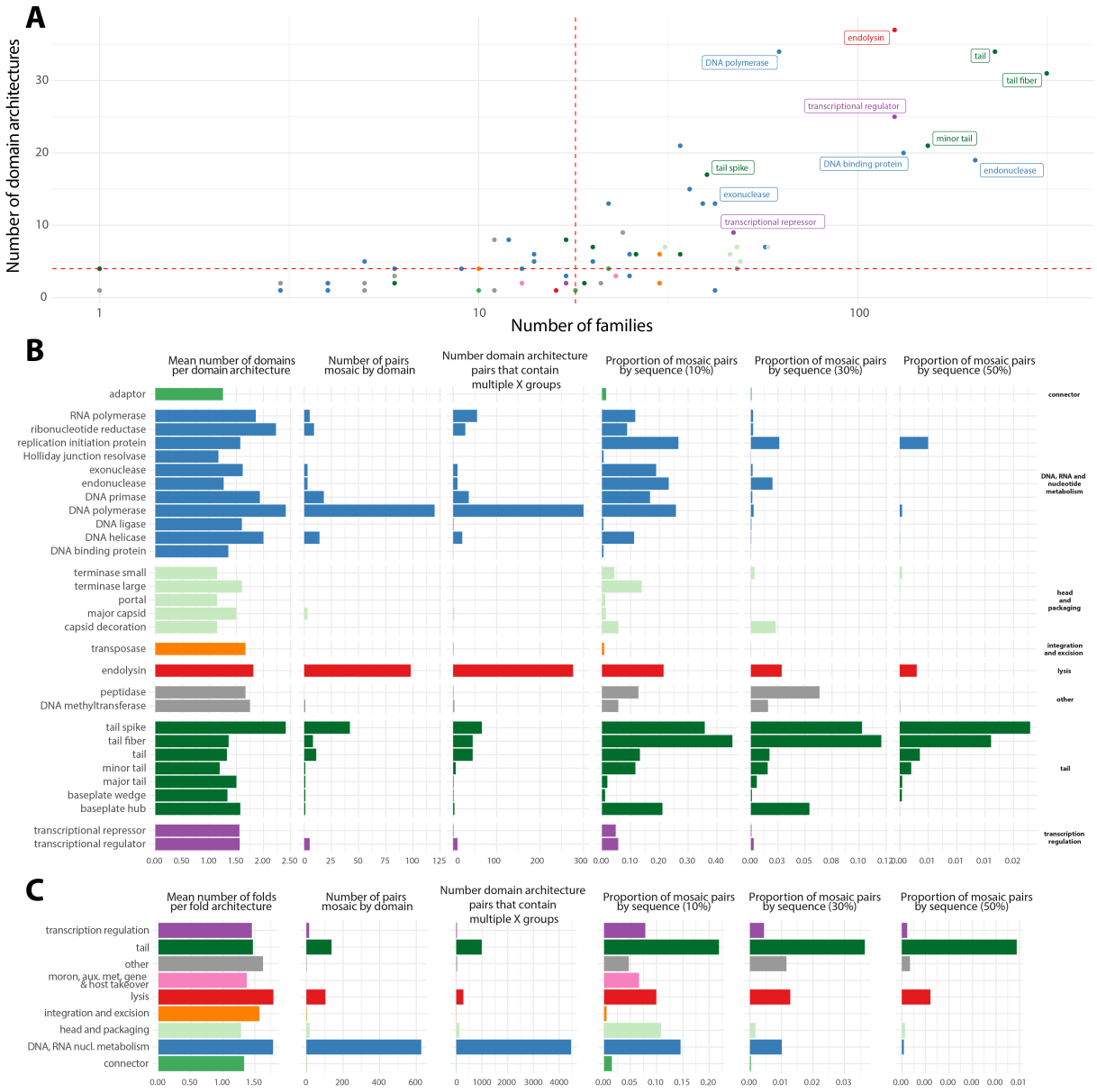

Figure 4. Mosaicism is associated with functions related to host tropism and nucleotide metabolism. (A) Diversity of all functional classes as measured by the number of protein families (X-axis) and the number of domain architectures (i.e., unique combination of ECOD T-groups; Y-axis). Colours correspond to different PHROG functional categories (same as in Figure 1). Red dashed line denotes the median values for the two measures of diversity. Functional classes above the values indicated by both dashed lines are analysed in the next panel. (B) Various measures obtained for the most diverse functional classes, including mean number of domains per architecture (column 1), number of domain architecture pairs that are mosaic according to the domain mosaicism test (column 2), number of domain architecture pairs that contain multiple X groups (column 3), and proportion of rHMM pairs that are mosaic to all pairs, according to the sequence mosaicism test, for three different percentage identity thresholds: 10%, 30% and 50% (columns 4-6). (C) The same measures applied per PHROG category instead of individual functional classes.

Specific examples of protein modularity

To better understand the biological nature of the detected protein mosaicism and how it generates protein modularity, we looked into the specific domain architectures of the highly mosaic functional classes. While below we focus on three classes and categories they belong to (core replication machinery, tail and lysis), we have developed a Shiny webserver that allows for interactive domain architecture lookup in each of 96 the functional classes and a look-up of 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 5. For comparison, in Supplementary Figure S13 we show an outline 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)[31,32] using `HHpred` and ECOD. Our results point to a few notable observations. First, we have recovered the exact or near-exact topological combinations of not only families A and B, which are well known to occur in phages like T4 and T7, but also of families C and Y (c.f., Supplementary Figure S13). Second, we have found several other domain architectures which are close variations in terms of presence/absence of the topologies mentioned 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). Likewise, we also found an example of family B architecture (ECOD X-groups 2002, 316, 3584, 102) with an extra nucleic acid-binding protein fold (T-group 2.1.1), found in family C, towards the C-terminal. In most of these cases, rHMMs in unusual domain architectures represented clusters with multiple protein sequences, suggesting multiple occurrences of such architectures in the analysed phage population. Finally, the comparison of the similarities and differences between these domain architectures points to clear cases of mosaicism. Examples include (i) sharing of the exonuclease domain between A and B families (the 'Ribonuclease H-like' fold) in the background of other, non-homologous, polymerase-specific folds; (ii) the 'Adenylyl and guanylyl cyclase catalytic domain' fold that is commonly found in families A and B was also present in family Y with other, unrelated domains; or (iii) the insertion domain in bacteriophage ϕ29 fold found alongside the exonuclease (2484) and/or the finger domain fold

(4970). It is also worth noting that all topologies found in *polX*, member of the family X (Supplementary Figure S13), were recovered in other combinations in the dataset analysed here. Overall, the recovered architectures point to a considerable amount of modularity via structural flexibility to reuse conserved folds found in DNA polymerases in various combinations.

To better understand these domain architectures in the context of other proteins in functional classes belonging to the 'DNA, RNA and nucleotide metabolism' category, we visualised it as a network where each node is a domain and an edge connects domains that co-occur in the same protein (see Supplementary Figure S14). Such network shows a very high degree of connectedness between multiple functional classes via nodes of high centrality (i.e., star-like nodes connecting multiple nodes that do not directly connect). Such connections are often indicative of modularity as cooccurrence of domains AB and BC but not AC points to likely mosaicism (provided that A and C are unrelated). These central nodes were predominantly highly conserved domain families that often co-occur in proteins of the replicative and transcription machinery (c.f., Figure 1). Upon a closer look, these domains, which often occurred in proteins of multiple functions, tend to co-occur with more specialist domains, which is indicative of domain mosaicism. Examples include the P-loop related topology (ECOD T-group 2004.1.1) which occurs in clamp loaders of DNA polymerase, DNA helicase and RNA ligase; the toprim topology (ECOD T-group 2006.1.3) which occurs in DNA topoisomerase, DNA primase or DNA primase/helicase; or the winged topology (ECOD T-group 101.1.2 which occurs in DNA primase and DNA topoisomerase) – in all these cases these topologies co-occured with other, unrelated domains. The existence of such a modular domain architecture in DNA/RNA replication proteins has been known for some time and is generally thought to have emerged over long evolutionary time scales[33]. However, given a few examples of recent mosaicism in proteins associated with core replication (see Supplementary Figures S11 and S12), it is possible that some prevalent and highly conserved domains may play an important role in the ongoing emergence of mosaicism of these proteins (see Discussion).

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 enzymatic domains (hydrolases) that help the phage penetrate layers

of surface sugars, for example the polysaccharide capsule[34]. Our results show clear evidence for the emergence of modularity via shuffling of all of these domains (see Figure 5). First, we find N-terminals and C-terminals in multiple combinations. 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) at the N-terminus or with the 'Putative tailspike protein Orf210 N-terminal domain' topology (3856.1.1), providing an excellent example of mosaicism. The latter domain, apart from being common in tail spikes, can also be frequently found in tail fibers. Interestingly, the combination of the 'Putative tailspike protein Orf210 N-terminal domain' fold and the 'Intramolecular chaperone domain in virus tail spike protein' fold (3240.1.1) at the C-terminal, a known endosialidase fold, is found in proteins associated with both tail spike and tail fibre classes. Second, we found co-existence of various enzymatic domains within the same protein in different combinations. For example, the mentioned endosialidase fold was found to co-occur with the 'SGNH hydrolase' fold (2007.5.1) in a tail fibre protein as well as with the 'Pectin-lyase like' fold (207.2.1) in a tail spike, while the 'SGNH hydrolase' fold was found to co-occur with the 'Concanavalin A-like lectins/glucanases' fold that is a marker of diversity-generating retroelements[35]. 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' fold (79.1.1) which is also found in baseplate spikes in combination with other domains like lysozyme (see also Supplementary Figure S15). Overall, these results suggest that receptor-binding domains, including enzymatic ones, can also be shuffled in different combinations, including multiple occurrences of enzymatic domains.

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[14]. The domain architectures found here contain both types of domains (see Figure 5). Specifically, the catalytic domains found include the 'N-acetylmuramoyl-L-alanine amidase-like' topology (ECOD T-group 285.1.1), a topology with structural similarity to amidase2; Cysteine proteinases (219.1.1), a topology that includes multiple peptidase and CHAP domains; and the 'Lysozyme-like' topology (235.1.1). These topologies are found in various combinations with the cell wall-binding domains that include the lysM domain (101.15.1), SH3 domain (4.1.1), or the '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); co-occurrence of the 'spore cortex-lytic enzyme' topology (3374.1.1) and Hedgehog/DD-peptidase (307.1.1); or co-occurrence of Hedgehog/DD-peptidase (307.1.1) and 'Lysozyme-like' (235.1.1). 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[36]. 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 S16).

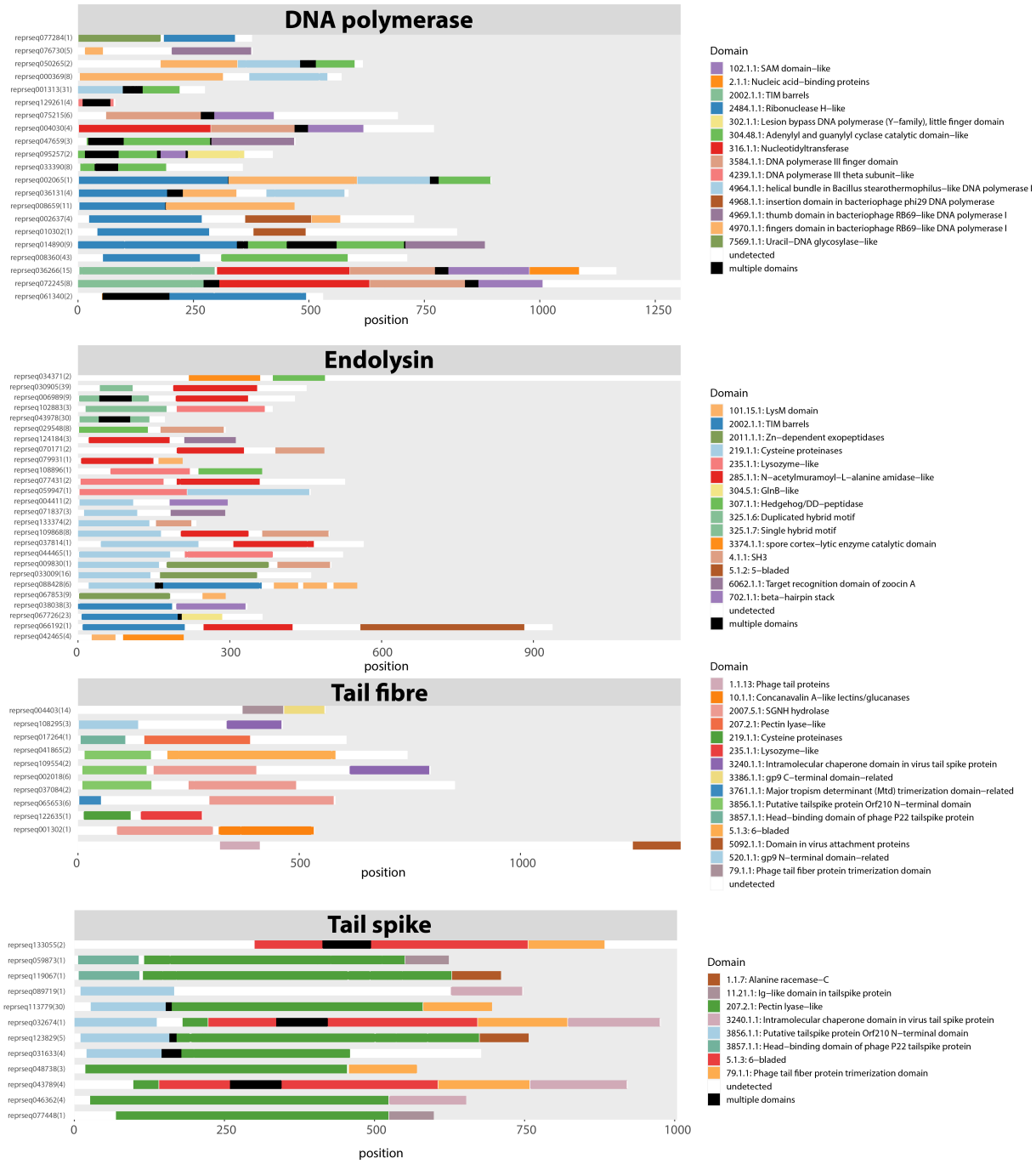

Figure 5. Visualisation of domain architectures for functional classes exhibiting the highest levels of domain mosaicism. Each line shows one chosen rHMM (abbreviated `reprseqXXXXXX`) per each domain architecture (with the number of protein sequences having this domain architecture displayed in bracket) 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: $X_{id}.H_{id}.T_{id}$.

Discussion

In this study we have systematically analysed the relationship between genetic diversity, functional diversity and domain architecture of over 133k HMM profiles of representative phage proteins (rHMMs) using a sensitive homology search via HMM-HMM comparison (`HHpred`). Our results demonstrate that the phage protein domain conservation 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[19,20,37]. Indeed, our findings show that such domain shuffling, that is known to be often recombination-driven[38], 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.

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[39], and previous studies have repeatedly demonstrated their rapid evolution in the face of adaptation to new hosts, particularly in receptor-binding proteins[14,40,41]. 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[42]. 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[43] while its phages are known for not only extensive modularity of receptor-binding proteins[13] but also existence of phages with complex tails with a broad host spectrum[44]. In line with this, we found that tail spikes and tail fibres were those functional classes where

we found clear evidence of the emergence of recent mosaicism.

Endolysins, on the other hand, target the peptidoglycan of their bacterial hosts. While there is a considerable diversity of peptidoglycans in bacteria[45], 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[14]. However, they also found evidence of recombination-driven exchange of endolysins between virulent phages infecting the 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[36].

Another major group for which there was evidence of extensive protein 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 DNA and RNA. But the DNA replication machinery is known to be highly diverse across the tree of life[46], including in viruses[47], 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 in this process[33]. It can be thus expected that the mosaicism that we detect in this study is ancient – in part at least – 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 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[48] as well as plasticity and modularity of the DNA replication machinery as a whole[49] in T4-like phages. The authors argued that such modularity gives these viruses the possibility to adapt to change by redesigning the interactions between their components[49]. This implies that modular protein evolution must have played a role in diversification of specific phage lineages and families. Finally, our data points to clear examples of recently-emerged mosaicism in highly conserved and common protein domains like P-loop hydrolases or the Ribonuclease H-like fold. This supports the idea of continuously emerging modularity of core

replication proteins and points to an interesting hypothesis that the highly conserved domains may act as genetic linkers to facilitate domain shuffling via homologous recombination. Overall, all of this suggests that core replication proteins, like DNA polymerases, continue to evolve in the process of bacteria-phage co-evolution.

One potentially important driver of the diversity of core replication proteins in phages, like DNA polymerases, 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[50]. Another example is the Nhi, a bacterial nuclease-helicase that competes with the page DNA polymerase for the 3' end of DNA to prevent phage replication[51]. 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, primase-helicase or SSB proteins[52]. The authors speculated 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. Indeed, in some proteins assigned to 'replication initiation protein', 'ribonucleotide reductase' and 'DNA polymerase' classes, we have seen individual examples of a relatively high protein sequence similarity ($\geqslant$50%) in a distant background (no evidence of homology at 50% hit probability), suggesting that it may indicate recent domain exchanges. This was characteristic of other proteins known for their ability to rapidly evolve via mosaicism like receptor-binding proteins or endolysins.

Each of our two approaches to detect mosaicism has strengths and weaknesses. Mosaicism via domain composition is the more robust approach to detect protein modularity as it uses structural information to refute the possibility that the absence of sequence similarity is due to the loss of evidence for homology

of one domain. Another advantage of the domain-based approach is that it ensures that the "units" of mosaicism are evolutionarily-meaningful as they can fold independently and hence can be horizontally shuffled. However, that approach only works on multi-domain proteins and is strongly dependent on the completeness of the domain databases like ECOD. On the other hand, the sequence-based approach can potentially detect mosaicism that will be missed via domain composition analysis (e.g., sub-domain mosaicism), but may not be able to distinguish genuine domain-shuffling from highly variable rates of evolution in different areas of the protein. One good example are tail length tape measure proteins which consist of conserved domains and long coiled coils which are repetitive and difficult to compare via sequence-similarity searches. (However in this case we could not verify this hypothesis using a domain-based approach due to the near-zero ECOD coverage of tail length tape measure proteins.) Nevertheless, for many of the functional classes defined here we obtained support for mosaicism using both, complementary measures, which makes a convincing case for the importance of modularity in the evolution of these functions.

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`[53]. This resulted in 462,721 predicted protein sequences which were clustered with `mmseqs2`[54] 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). This representative sequence 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[55] 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[56]. 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 133,624 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 cov = min(qcov,scov) $\geqslant 0.8$. For each pair, we then calculated a weighted score of $p \times$ cov, which was used as a weight of an undirected network. Finally,

we used a Markov clustering algorithm (MCL)[57] with an inflation factor `--I 2` to cluster rHMMs into 74,562 families.

Functional annotation

To assign each rHMM into a functional category, we used the Prokaryotic Virus Remote Homologous Groups database (PHROGs; version 4)[22]. 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 500 sequences and (ii) were found in at least 20 rHMMs were considered (unless stated otherwise). Functional classes were assigned as those with hits to a known class at 80% coverage and 95% probability hit threshold. Unless stated otherwise, rHMMs with hits to more than a single functional class were discarded.

Domain detection

To detect domains in rHMMs, we used the Evolutionary Classification of Protein Domains[23] database (ECOD, version from 13.01.2022). Each of the 133,624 rHMMs was compared to the `HHpred` version of the ECOD database using `hhblits` 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 proteins at the sequence level, 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

were 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. To look for potential mosaicism between rHMMs at the domain level, 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). See Figure 2 for a schematic drawing for both measures of mosaicism.

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). BS acknowledges funding by the French Government Scholarship (BGF). 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 Vikram Alva, Carina Büttner and other members of Plevka Lab, Alan R. Davidson, Bas E. Dutilh, Mart Krupovic, Andrei Lupas as well as Joana Pereira and other members of Torsten Schwede's group for helpful discussions and/or comments related to this work that were instrumental in finalising this manuscript.

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`.

References

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

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

[3] Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell. 2022 Oct;185(21):4023–4037.

[4] Samson JE, Magadán AH, Sabri M, Moineau S. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol. 2013 Oct;11(10):675–687.

[5] Botstein D. A theory of modular evolution for bacteriophages. Ann N Y Acad Sci. 1980;354:484–490.

[6] Hatfull GF, Hendrix RW. Bacteriophages and their genomes. Curr Opin Virol. 2011 Oct;1(4):298–303.

[7] Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A. 1999 Mar;96(5):2192–2197.

[8] Moura de Sousa JA, Pfeifer E, Touchon M, Rocha EPC. Causes and Consequences of Bacteriophage Diversification via Genetic Exchanges across Lifestyles and Bacterial Taxa. Mol Biol Evol. 2021 05;38(6):2497–2512.

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

[10] Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019 Jun;37(6):632–639.

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

[12] Iranzo J, Krupovic M, Koonin EV. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. mBio. 2016 08;7(4).

[13] Latka A, Leiman PG, Drulis-Kawa Z, Briers Y. Phages. Front Microbiol. 2019;10:2649.

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

[15] Dunne M, Rupf B, Tala M, Qabrati X, Ernst P, Shen Y, et al. Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Rep. 2019 Oct;29(5):1336–1350.

[16] Latka A, Lemire S, Grimon D, Dams D, Maciejewska B, Lu T, et al. Phages Switches Their Capsule Serotype Specificity. mBio. 2021 05;12(3).

[17] Gerstmans H, Grimon D, Gutiérrez D, Lood C, Rodríguez A, van Noort V, et al. A VersaTile-driven platform for rapid hit-to-lead development of engineered lysins. Sci Adv. 2020 06;6(23):eaaz1136.

[18] Dunne M, Prokhorov NS, Loessner MJ, Leiman PG. Reprogramming bacteriophage host range: design principles and strategies for engineering receptor binding proteins. Curr Opin Biotechnol. 2021 04;68:272–281.

[19] Cardarelli L, Pell LG, Neudecker P, Pirani N, Liu A, Baker LA, et al. Phages have adapted the same protein fold to fulfill multiple functions in virion assembly. Proc Natl Acad Sci U S A. 2010 Aug;107(32):14384–14389.

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

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

[22] Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021 Sep;3(3):lqab067.

[23] Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014 Dec;10(12):e1003926.

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

[25] Alva V, Koretke KK, Coles M, Lupas AN. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. Curr Opin Struct Biol. 2008 Jun;18(3):358–365.

[26] Iyer LM, Koonin EV, Aravind L. 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. 2003 Jan;3:1.

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

[28] Medvedev KE, Kinch LN, Grishin NV. Functional and evolutionary analysis of viral proteins containing a Rossmann-like fold. Protein Sci. 2018 08;27(8):1450–1463.

[29] Vyas P, Trofimyuk O, Longo LM, Deshmukh FK, Sharon M, Tawfik DS. Helicase-like functions in phosphate loop containing beta-alpha polypeptides. Proc Natl Acad Sci U S A. 2021 Apr;118(16).

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

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

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

[33] Leipe DD, Koonin EV, Aravind L. 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. 2004 Oct;343(1):1–28.

[34] Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, Lavigne R, et al. Targeting mechanisms of tailed bacteriophages. Nat Rev Microbiol. 2018 Dec;16(12):760–773.

[35] Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res. 2018 Jan;46(1):11–24.

[36] Payne KM, Hatfull GF. Mycobacteriophage endolysins: diverse and modular enzymes with multiple catalytic activities. PLoS One. 2012;7(3):e34052.

[37] Fraser JS, Yu Z, Maxwell KL, Davidson AR. Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J Mol Biol. 2006 Jun;359(2):496–507.

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

[39] de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends Microbiol. 2019 01;27(1):51–63.

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

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

[42] Holt KE, Lassalle F, Wyres KL, Wick R, Mostowy RJ. Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. ISME J. 2020 Jul;14(7):1713–1730.

[43] Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae. Nat Rev Microbiol. 2020 Jun;18(6):344–359.

[44] Ouyang R, Costa AR, Cassidy CK, Otwinowska A, Williams VCJ, Latka A, et al. High-resolution reconstruction of a Jumbo-bacteriophage infecting capsulated bacteria using hyperbranched tail fibers. Nat Commun. 2022 Nov;13(1):7241.

[45] Turner RD, Vollmer W, Foster SJ. Different walls for rods and balls: the diversity of peptidoglycan. Mol Microbiol. 2014 Mar;91(5):862–874.

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

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

[48] Wang CC, Yeh LS, Karam JD. Modular organization of T4 DNA polymerase. Evidence from phylogenetics. J Biol Chem. 1995 Nov;270(44):26558–26564.

[49] Petrov VM, Nolan JM, Bertrand C, Levy D, Desplats C, Krisch HM, et al. Plasticity of the gene functions for DNA replication in the T4-like phages. J Mol Biol. 2006 Aug;361(1):46–68.

[50] LeRoux M, Srikant S, Teodoro GIC, Zhang T, Littlehale ML, Doron S, et al. The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. Nat Microbiol. 2022 07;7(7):1028–1040.

[51] Bari SMN, Chou-Zheng L, Howell O, Hossain M, Hill CM, Boyle TA, et al. A unique mode of nucleic acid immunity performed by a multifunctional bacterial enzyme. Cell Host Microbe. 2022 04;30(4):570–582.

[52] Stokar-Avihail A, Fedorenko T, Garb J, Leavitt A, Millman A, Shulman G, et al. Discovery of phage determinants that confer sensitivity to bacterial immune systems. bioRxiv. 2022;p. 2022.08.27.505566.

[53] Ecale Zhou CL, Kimbrel J, Edwards R, McNair K, Souza BA, Malfatti S. MultiPhATE2: code for functional annotation and comparison of phage genomes. G3 (Bethesda). 2021 05;11(5).

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

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

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

[57] Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002 Apr;30(7):1575–1584.