





Article

Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts

Lakshminarayan M. Iyer ¹, Vivek Anantharaman ¹, Arunkumar Krishnan ², A. Maxwell Burroughs ¹
and L. Aravind ^{1,*}

¹ National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; lakshmin@mail.nih.gov (L.M.I.); ananthar@mail.nih.gov (V.A.); burrough@mail.nih.gov (A.M.B.)

² Department of Biological Sciences, Indian Institute of Science Education and Research (IISER) Berhampur, Odisha 760010, India; akrishnan@iiserbpr.ac.in

* Correspondence: aravind@mail.nih.gov

Abstract: Jumbo phages have attracted much attention by virtue of their extraordinary genome size and unusual aspects of biology. By performing a comparative genomics analysis of 224 jumbo phages, we suggest an objective inclusion criterion based on genome size distributions and present a synthetic overview of their manifold adaptations across major biological systems. By means of clustering and principal component analysis of the phyletic patterns of conserved genes, all known jumbo phages can be classified into three higher-order groups, which include both myoviral and siphoviral morphologies indicating multiple independent origins from smaller predecessors. Our study uncovers several under-appreciated or unreported aspects of the DNA replication, recombination, transcription and virion maturation systems. Leveraging sensitive sequence analysis methods, we identify novel protein-modifying enzymes that might help hijack the host-machinery. Focusing on host–virus conflicts, we detect strategies used to counter different wings of the bacterial immune system, such as cyclic nucleotide- and NAD⁺-dependent effector-activation, and prevention of superinfection during pseudolysogeny. We reconstruct the RNA-repair systems of jumbo phages that counter the consequences of RNA-targeting host effectors. These findings also suggest that several jumbo phage proteins provide a snapshot of the systems found in ancient replicons preceding the last universal ancestor of cellular life.

Keywords: DNA viruses; anti-phage systems; nicotinamide dinucleotide; DNA polymerases; DNA polymerase III; RNA polymerase; transcription; nucleotides; virus evolution



Citation: M. Iyer, L.; Anantharaman, V.; Krishnan, A.; Burroughs, A.M.; Aravind, L. Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. *Viruses* **2021**, *13*, 63. <https://doi.org/10.3390/v13010063>

Academic Editors: Julie Thomas and Lindsay Black

Received: 24 November 2020

Accepted: 31 December 2020

Published: 5 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Viral genomes span a wide size-continuum, which at the lower end includes some of the smallest natural replicons and at the higher end exceeds that of several cellular genomes. In eukaryotes, most giant viruses are unified within the diverse clade of nucleocytoplasmic large DNA viruses (NCLDVs), which includes the poxviruses, iridoviruses and asfarviruses which infect animals and several large viruses that infect microbial eukaryotes [1–3]. Large prokaryotic viruses first came to light with the discovery of the *Bacillus megatherium* phage G over 50 years ago, which measures more than 600 nanometers in length from head to tail, with a head diameter of just under 200 nm [4,5]. It also became clear early on that this virus is not just exemplary in virion size but also in terms of its genome [6]. This was followed by the discovery of several giant phages infecting a diverse array of bacteria. In the past two decades, these giant phages entered the “genomic era” as their genomes were completely sequenced, revealing numerous remarkable aspects of their biology [7–9].

Recently, the generic term jumbo phage has been applied to these giant phages, with a genome-size cutoff of 200 kb applied to include them in this category [8]. An even larger type within them, with a lower genome size cutoff of 500 kb has been termed

megaphage [10]. The availability of genomic data for over 200 jumbo phages has sparked a renewed investigation of these viruses [7]. Consequently, there has been a wealth of recent studies that structurally, biochemically and ecologically characterize these viruses. One offshoot of this work has been the realization that the jumbo phages are not an evolutionarily unitary group. The tailed bacteriophages (*Caudovirales*) belong to three major structural groups: (1) The *Podoviridae* with short tails (e.g., coliphage N4); (2) the *Siphoviridae* with long non-contractile tails (e.g., coliphage lambda); (3) *Myoviridae* with long contractile tails (e.g., coliphage T4). Jumbo phages feature both *Myoviridae* and *Siphoviridae* in their ranks suggesting the independent emergence of large genome sizes in different phage groups [7]. Within these groups, the jumbo phages show a range of head morphologies which include regular icosahedra, icosahedra with elongated central triangles and icosadeltahedra. In some cases, like *Tenacibaculum maritimum* phages PTm1 and PTm5, the heads are further decorated with an assembly of fibers that project out from the top [11]. Some also display interesting variations in the tail morphology such as the presence of a mop-like array of flexible tail fibers in the *Sphingomonas* phage PAU [12] and long whiskers (*Pectobacterium* phage CBB) or hair-like fibers (coliphage 121Q and *Agrobacterium* phage Atu_ph07) [13,14] projecting from both the head and the tail-sheath. It has been proposed, although not confirmed, that these decorations might have a specific role in enhancing host attachment, especially in the context of the dispersion of virions by aqueous flow [15].

Jumbo phages have been traditionally difficult to isolate due to their large size, which prevents both easy separation from the host via filtration and diffusion in agar to form visually detectable plaques on bacterial lawns [8]. Nevertheless, the rising interest in these viruses has resulted in the identification of jumbo phages from gammaproteobacteria, betaproteobacteria, alphaproteobacteria, zetaproteobacteria, bacteroidetes, cyanobacteria, sporulating firmicutes and actinobacteria. Gammaproteobacteria are prevalent among the easily cultivable organisms, as well as those which are medically and agriculturally relevant; hence, jumbo phages infecting gammaproteobacteria are currently overrepresented in the genomic databases. While the majority of jumbo phages have been isolated from aquatic environments, they have also been found to infect bacteria from other niches such as soil, marine sediments, animal guts and plant material. More recently, metagenomic studies have augmented traditional methods of prospecting and have helped identify a range of giant phages, including the *Prevotella* megaphages in the gut microbiome [9,10]. These observations suggest that jumbo phages are likely to be far more widely distributed and numerous than the current numbers indicate.

Jumbo phages have also sparked the interest of researchers due to several interesting aspects of their biology that have come to light from a flurry of recent studies. These include the use of multi-subunit DNA-dependent RNA polymerases (RNAPs) that are homologous to different segments of the host enzymes [16]. These homologous segments include two subunits displaying double- Ψ beta-barrel (DPBB) catalytic domains as well as separate subunits encompassing the transcript-exit clamp module and other parts of the RNAP. Several jumbo phages have two different versions of these multisubunit enzymes that are respectively packaged into the virion for early gene transcription or specialize in middle/late gene transcription [17,18]. An overlapping group of jumbo phages has been characterized as encoding a tubulin homolog, which helps form a nucleus-like compartment in the host that has been shown to protect the phage against host immune mechanisms such as the restriction-modification (R-M) and CRISPR/Cas systems by walling-off the phage replication and transcription apparatus [19,20]. Some jumbo phages have also been shown to possess their own biosynthetic capacity for NAD⁺ which is required as a substrate for phage DNA-replication and regulatory enzymes [21]. Further, some cyanophages encode their own capsular lipopolysaccharide (LPS) biosynthesis genes, which have been proposed to ward off superinfection by competing phages during periods of phage dormancy within their hosts (pseudolysogeny) [22]. Genomes of several jumbo phages also specify several diverse mechanisms that help them in their biological conflicts with their hosts such as:

(1) methyltransferases that modify their DNA to evade restriction attacks on their genome. (2) Incorporation of uracil in the genome in place of thymine. (3) Arrays of tRNAs that help overcome host defense mechanisms such as those utilizing the endoRNases that restrict translation by cleaving tRNAs [7,8].

These and other findings have been surveyed in recent reviews on jumbo phages. However, there are several outstanding questions of interest that can be addressed via comparative genomic analysis and sensitive sequence and structure analysis of the phage-encoded proteins. These include the identification of various systems that deter or compensate for host immune mechanisms. More specifically, these include the complement of phage proteins that carry out RNA repair in the face of immune attacks on the translation apparatus and the phage enzymes that might modify DNA and RNA beyond the previously described DNA methylases. The jumbo phage enzymes akin to the coliphage T4-encoded ADP ribosyltransferases that help in modifying and hijacking host systems have also not been studied closely [23]. Further, it has recently become apparent that bacteria deploy several effector systems activated by nucleotides and NAD⁺ derivatives to target viruses [24]. The complement of phage countermeasures against such systems remains poorly explored. In the current article, we carry out a systematic comparative genomic survey of jumbo phages to address these questions. Consequently, we identify diverse phage systems that are likely to form interfaces of the conflict with their hosts. We also use this information to throw light on less-appreciated aspects of various previously studied systems such as the DNA replication and repair enzymes, RNA polymerases, and different transcriptional strategies used by the jumbo phages.

Moreover, this investigation helps understand the independent, parallel growth of genome size in several viral lineages. To this end, we have tried to objectively define the genome size criteria for giant phages rather than using an arbitrary cutoff of 200 kb. The independent evolution of jumbo phages infecting diverse bacteria and also the presence of certain parallels to their eukaryotic counterparts, the NCLDVs, suggests that large genome sizes represent a potential strategy that has repeatedly emerged under natural selection independently of the type of host. This phenomenon might have general implications for the dramatic variations in genome size seen among both viruses and cellular organisms.

2. Materials and Methods

2.1. Sequence Analysis

Sequence profile searches were performed using the PSI-BLAST (RRID: SCR_001010) [25] and JACKHMMER programs (RRID: SCR_005305) [26] with the profile being built at each iteration. Clustering for both classification and purging of nearly identical sequences was performed with the BLASTCLUST program (version: 2.2.26, NCBI, USA, <ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>) (RRID: SCR_016641 with the length of the pairwise alignment (L) and measure of similarity, i.e., bit-score (S) adjusted depending on the degree of clustering required.

HMM searches were run using either HMMsearch initiated with an HMM built from an alignment or iteratively using JACKHMMER from single seeds [26]. Sequence searches were run against either the non-redundant (nr) protein database or custom databases of completely sequenced 224 jumbo phages or 10,176 *Caudovirales* available in Genbank as of 10 October 2020. Names of the phages were as provided in the taxonomy division of Entrez (<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=28883>). To obtain the proteome of the *Prevotella* megaphage Lak-B8, the nucleotide sequence deposited in Genbank was translated using Prodigal [27] using the translation table codes 11 and 15 as mentioned in the genome publication [10]. Profile-profile searches were conducted using HHpred (RRID: SCR_010276) and run against (1) HMMs derived from PDB; (2) Pfam models; (3) A custom database of alignments of diverse domains curated by the Aravind group [28]. All novel alignments that were added to this database are provided in the (Supplementary Materials S2). Multiple sequence alignments were built using Kalign (162) (RRID: SCR_011810) and Muscle with manual adjustments based on profile-profile

and structural alignments [29,30]. Secondary structures were predicted using the JPred program (RRID: SCR_016504) [31].

2.2. Structure Analysis

Structure similarity searches were performed using the DaliLite program (RRID: SCR_003047) run against the PDB database clustered at 75% sequence similarity [32]. Structure similarity trees were constructed based on Z-scores obtained from an all-vs-all search of the compared structures using average linkage clustering. Structural visualization and manipulations were performed using the PyMol (<http://www.pymol.org>) (RRID:SCR_000305) and MOL* programs (<http://molstar.org>) [33]. The structural figure panels were rendered by extracting the state PDB IDs and presenting them in the cartoon or molecular surface views with the MOL* program and colored to emphasize relevant features.

2.3. Comparative Genomics

Taxonomic lineages were obtained from the NCBI Taxonomy database (see Section 2.1). Contextual information from prokaryotic gene neighborhoods was retrieved using a Perl script to extract upstream and downstream genes of the query gene from the GenBank genome file. Their products were then clustered with BLASTCLUST to identify conserved gene-neighborhoods based on conservation between different taxa. Several additional filters were then applied to recognize valid neighborhoods for further analysis: (1) nucleotide distance constraints (generally 50 nucleotides); (2) conservation of gene directionality within the neighborhood; (3) presence in more than one phylum. Phylogenetic trees were constructed using an approximate maximum-likelihood method implemented in the FastTree 2.1 (RRID: SCR_015501) program under default parameters [34].

The phyletic patterns were used to generate two sets of vectors, namely the distribution by phage for a given protein and the complement of proteins for a given phage. These were used to compute the inter-protein or inter-phage Canberra distance [35], which is best suited for vectors with integer data in the form of presences and absences. The Canberra distance between two vectors \vec{p} and \vec{q} is defined as:

$$d(\vec{p}, \vec{q}) = \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

These distances were used to cluster the protein families and phages through agglomerative hierarchical clustering using Ward's method [36]. Ward's method takes the distance between two clusters A and B, as the amount by which the sum of squares from the center of the cluster will increase when they are merged. Ward's method then tries to keep this growth as small as possible. It tends to merge smaller clusters that are at the same distance from each other as larger ones, a behavior useful in lumping "stragglers" in terms of both phages and proteins with correlated phyletic patterns. The same vectors for phages were also used to perform principal component analysis to detect spatial clustering by dimensionality reduction. The variables were scaled to have unit variance for this analysis. Data processing (knitr and dplyr libraries), network analysis (igraph library), and visualization were performed using the R language.

A complete annotated list of all the jumbo phage protein clusters with domain architectures as classified in this study is available as Supplementary Materials (Supplementary S5). Any protein family referred to in the text may be found by searching the Supplementary table with the representative accession provided in the text.

3. Results and Discussion

3.1. The Basic Features of Genome Size and Protein Length Distributions of Giant Phages

To get a handle on the distribution of genome (proteome) sizes, we assembled a database of 10,176 *Caudovirales* with complete genomes and plotted the sizes of their predicted proteome against their genome size (Figure 1a). The proteome size ranging from 1 to 714 proteins is strongly positively and linearly correlated with genome size

($r^2 = 0.922$, $p = 2 \times 10^{-16}$) ranging from 5 kb to 551 kb. This suggests that despite the wide size range, proteins are encoded at similar densities across phage genomes. Consistent with this, a plot of the density of protein-coding genes on phage genomes shows a tight approximately normal distribution with a mean of 1.5 genes (standard deviation 0.25) per kb of the genome (Figure 1a).

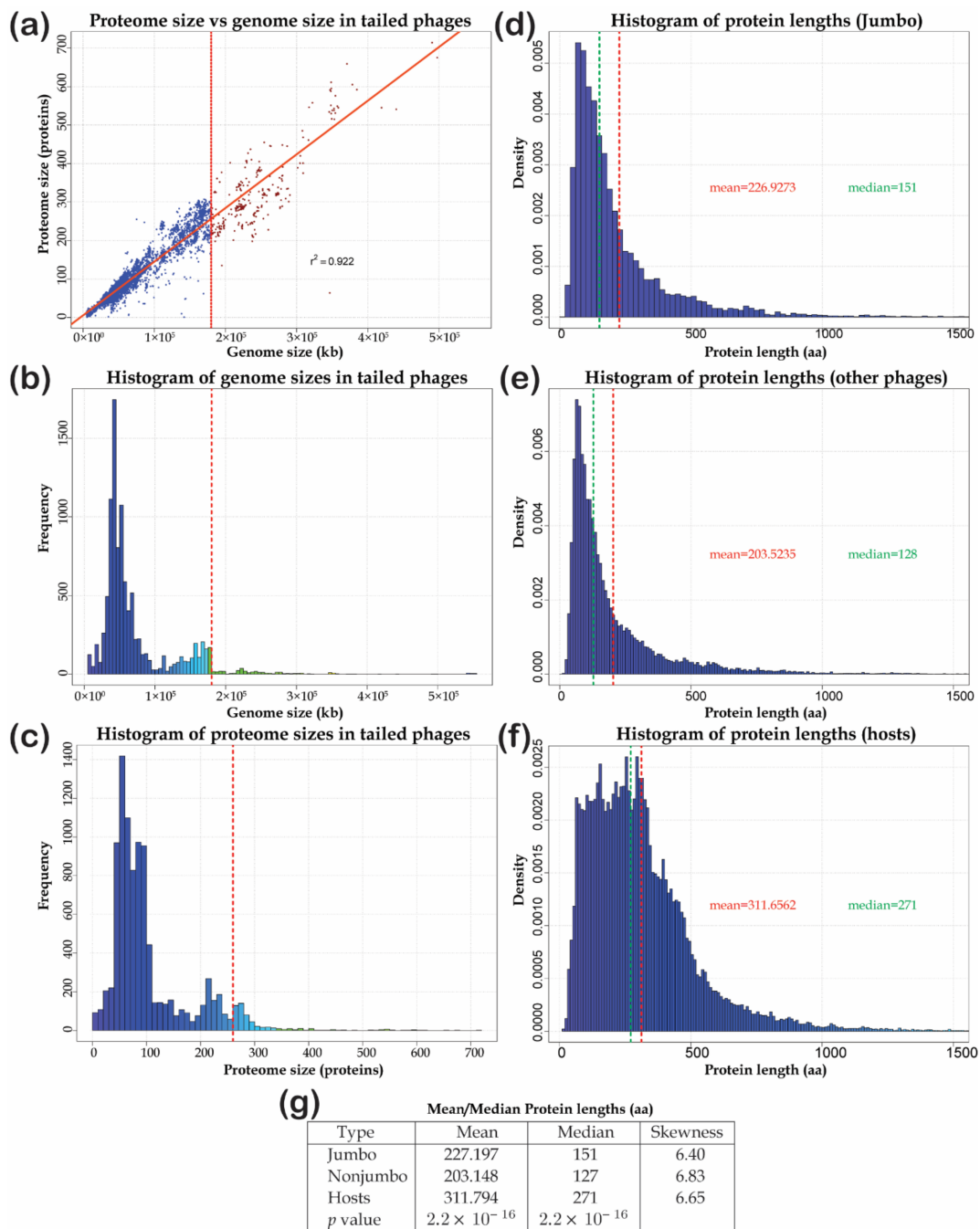


Figure 1. Bulk genome and proteome features of *Caudovirales*, jumbo phages and their hosts. Various genome and protein features of *Caudovirales* are illustrated including (a) protein density; (b) genome and (c) proteome size distributions. In (a) points are colored blue if the genome size <180 kb and red if >180 kb. Protein length distributions are illustrated for (d) jumbo phages (e) non-jumbo phages and (f) those of their hosts. The length distribution statistics are summarized in (g). The red vertical line in (a–c) shows the objective cutoff used to define the jumbo category in phages. Panels (b) and (c) use a color palette ranging from dark blue to green (R language: topo.colors function) to color each bar in the histogram.

This suggested that the histogram of the distribution of phage genome or proteome sizes would be useful to determine any objective size categories that might exist. These plots reveal a trimodal distribution with distinct valleys between the peaks dividing the phages into three size categories (Figure 1b,c). The first and most abundant category consists of the small phages with genome size less than 100 kb (modal value ~50 kb) and coding for less than 150 proteins. This is followed by the category of next abundance comprised of medium-sized phages with genome sizes less than 180 kb and typically coding for less than 250 proteins. The final category is relatively sparse with genomes larger than 180 kb with a long right tail but a current modal value of around 230 kb. While sampling bias could account for some of the differences, the modality in the distribution points to the existence of a genuine giant-phage category with genome sizes greater than 180 kb (Figure 1a,b). As compared to the cut-off of greater than 200 kb genomes employed in studies on jumbo phages, our cutoff includes 45 additional phages in the jumbo phage category with an average proteome size of 259 proteins. A recent study of phages from diverse environmental samples found diverse giant phages (with genomes greater than 200 kb) to be lodged in larger clades with genomes of at least 120 kb in length [9]. This illustrates the tendency of the jumbo phages to belong within larger clades with at least medium-sized genomes. Therefore, the above-proposed cutoff based on the size distribution helps more objectively define the genomic size category that includes the jumbo phages and is treated as the focus of our analysis. At the lower end of this category are coliphages like RB43, while the higher end features phages such as the *Prevotella* megaphage Lak-B8, *Bacillus* virus G, *Agrobacterium* hairy phage Atu_ph07 and phage SCTP-2 which infects the hyperhalophilic gammaproteobacterium *Salicola*.

A study of the distribution of the lengths of proteins in amino acids also suggests that this category can be discriminated on an average from the other phages (Figure 1d–f). The lengths of all proteins encoded by small/medium-sized phages, jumbo phages and their host genomes show similar unimodal distributions with a comparable prominent right skew (skewness = 6.4–6.8; Figure 1). However, there are differences in the median protein length with the viruses having significantly lower median protein lengths than their cellular counterparts (Kruskal–Wallis test $p = 2 \times 10^{-16}$) (Figure 1d,e). Within the phages, the jumbo phages have significantly longer median lengths of proteins (127 vs. 151, $p = 3.2 \times 10^{-13}$) than the rest. Therefore, among these DNA viruses, there is a trend for longer proteins going along with their larger genomes. Thus, while the phages with genome sizes of 180 kb or more morphologically conform to the same categories as the small/medium-sized phages, they can be defined as a distinct group encompassing the jumbo phages based on the above criteria.

3.2. Conserved Jumbo Phage Proteins Define Distinct Groups with Multiple Independent Origins

Unlike cellular genomes, there are few proteins universally conserved across jumbo phages. The terminase large subunit, the DNA-packaging protein characteristic of *Caudovirales*, comprised of a N-terminal DNA-pumping P-loop ATPase domain and a C-terminal RNase H fold endonuclease domain, is universal. The chromosome-end-processing complex comprised of the ABC ATPase (ortholog of cellular SbcC) and its nuclease partner SbcD belonging to the calcineurin-like phosphoesterase superfamily are practically universally conserved. Further, proteins with a lysozyme fold that assist phages in degrading the sugar linkages of cell walls are also nearly universal. Beyond these, all other conserved proteins show more patchy phyletic patterns across jumbo phages. In part, this arises from extreme sequence divergence, especially in the case of certain virion structural proteins. A recent study has defined phage clades using the universal terminase subunit and acknowledged that the different conserved proteins yield different tree topologies [9]. We found that the terminase large subunits from different phages can evolve at rather distinct rates; for instance, the terminase subunit of the *Prevotella* megaphage Lak-B8 is evolving at a distinctly higher rate than those from some of the other jumbo phages. In addition to the evidence for different evolutionary rates of conserved proteins, which is typical of viruses,

there are also genetic exchanges between distantly related clades. Hence, conventional phylogenetic trees based on single or concatenated protein alignments are not effective in revealing the higher-order evolutionary relationships and recombinational events shaping viral genomes. Therefore, we instead used clustering and principal component methods based on phyletic pattern vectors of conserved proteins from jumbo phages to obtain a picture of their relationships (Figure 2).

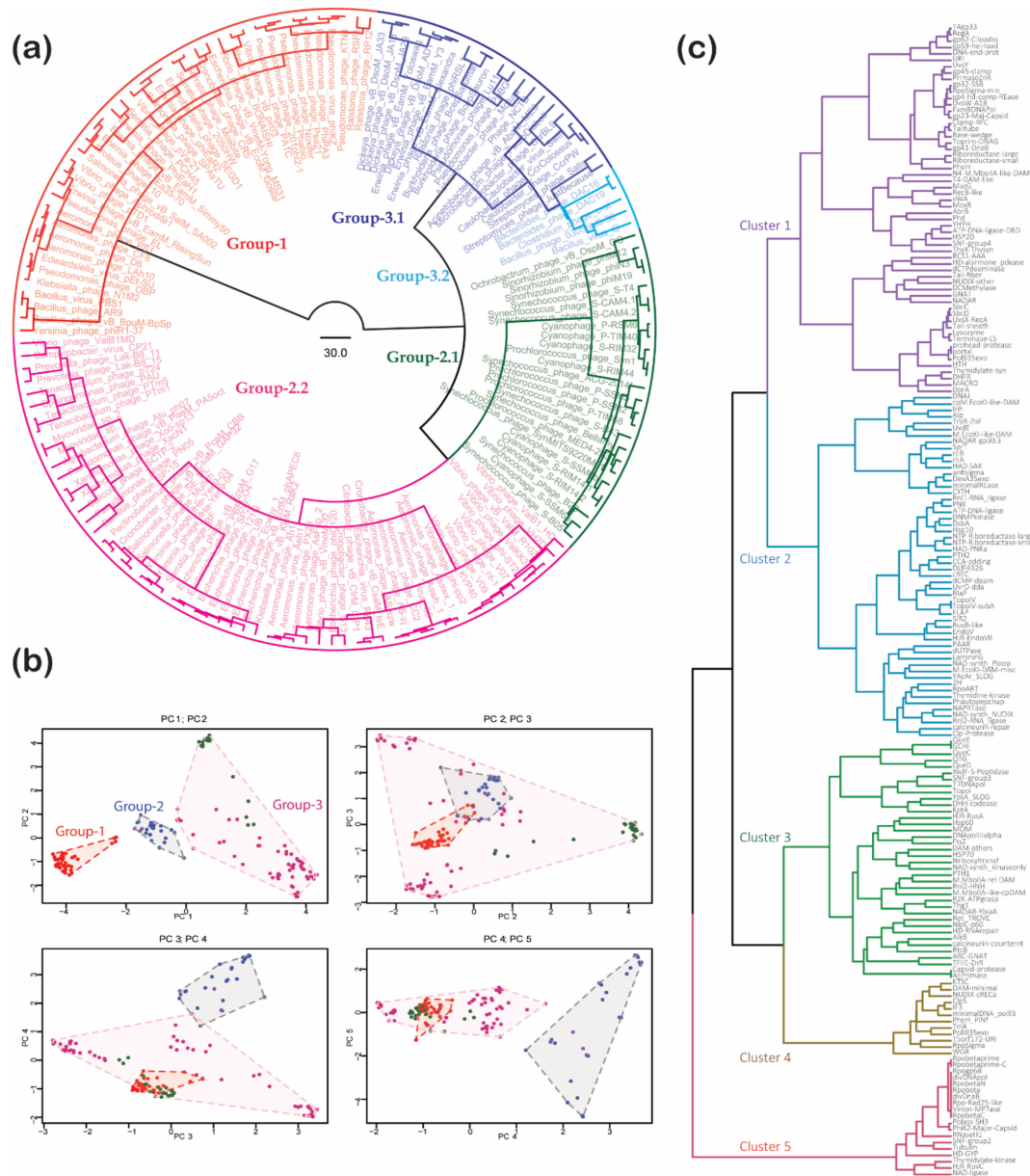


Figure 2. Phylogenetic analysis of jumbo phages and their proteins. Phyletic patterns of 113 protein families were used to (a) compute the Canberra distance for Ward’s clustering to construct a dendrogram of phages, and (b) for Principle Component Analysis for 154 jumbo phages with unique phyletic patterns. The convex hulls bounding the groups are shown on the Principal Component Analysis (PCA) plots. Members groups 1, 2.1, 2.2, 3.1 and 3.3 are colored distinctly. Organisms are colored by groups as in (a). The Canberra distance and Ward’s clustering were also used to compute the (c) protein dendrogram for 181 protein families which illustrates proteins sharing similar phyletic patterns.

To this end, we performed single-linkage clustering of all the proteins from 224 representative jumbo phages based on BLAST pairwise-alignment bit scores to identify various protein families found in them. These families were then searched using sequence profiles and hidden Markov models with the RPS-BLAST [25] and HMMSCAN (HMM3

package) [26] programs to identify conserved domains. In cases where such domains could not be identified by these methods, we carried out iterative sequence-profile, hidden Markov model searches and profile-profile searches in an attempt to identify any further conserved domains. As a result, we obtained a comprehensive collection of domain architectures for the protein families found in jumbo phages.

Using the phyletic patterns of 113 such families with wide representation in jumbo phages, we computed a distance matrix using the Canberra distance and derived a dendrogram of phages (Figure 2a) using the Ward clustering algorithm (see Methods). The same phyletic pattern matrix was also used for principal component analysis (Figure 2b) to identify potential groupings of phages by plotting pairs of the first five principal components that explain 63% of the variance.

The above procedures revealed that at the highest level the jumbo phages can be divided into three broad groups (Figure 2a). Group 1 includes the classic jumbo phages prototyped by the *Pseudomonas aeruginosa* phage PhiKZ. This group is typified by the presence of the multi-subunit “double-barrel” RNAP, an unusual DNA polymerase from a unique and divergent clade of family B DNA polymerases, a divergent version of the DnaB-helicase (see below) and a PhiKZ-type major capsid protein. This group broadly follows the PhiKZ functional model. Group 2 is characterized by a classic family B DNA polymerase, a classic OB-fold single-strand binding protein, a phage T4 UvsW/poxviral A18 type DNA helicase, the absence of an RNAP but the presence of phage-encoded sigma factors and the distinct gp23-type major capsid protein. Group 2 further splits up into two subgroups: 2.1 comprises of phages infecting both cyanobacteria (e.g., *Synechococcus* phage S-Cam4.1) and alphaproteobacteria (e.g., *Ochrobactrum* phage vB_OspM_OC); 2.2 includes phages mostly infecting gammaproteobacteria, the alphaproteobacteria (e.g., *Atu_ph07*) and bacteroidetes (e.g., the *Prevotella* megaphage LAK-B8 and *Tenacibaculum* phage PTM1). Group 3 is further split into two, of which 3.1 is defined primarily by the presence of a coliphage T7-type DNA polymerase, whereas group 3.2 (typified by the *Bacillus megatherium* phage G and *Clostridium* phage c-st) is defined by the presence of a DNA polymerase III type enzyme similar to the primary replicative enzyme of bacteria. Whereas groups 1 and 2 are entirely made up of *Myoviridae*, group 3 includes both *Myoviridae* and *Siphoviridae*. The *Myoviridae* across all groups are unified by the presence of a tail-sheath protein which forms part of the contractile tail typical of these viruses (Figure 2a). *Aeromonas* phage AP1 shows hybrid features of both groups 1 and 2.2; however, the group 1 proteins in this phage are nearly identical to other *bona fide* *Aeromonas* group 1 phages, raising the question of contamination of the genome assembly.

The characteristic proteins of each group of jumbo phages are also found in certain medium-sized phages; for instance, the group 2 phages are the jumbo phages related to the coliphage T4. Thus, group 2 phages can be seen as arising via genome expansion from coliphage T4-like predecessors. This is consistent with a recent study that found jumbo phages from environmental samples to be lodged in clades with other medium-sized phages [9]. Together, these observations indicate that there have been at least 6–7 distinct origins of jumbo phages from smaller precursors. Moreover, the presence of both *Siphoviridae* and *Myoviridae* in group 3 suggests that there has been a recombinational exchange of core machinery between viruses with distinct virion function (contractile vs. flexible tails) and morphology. This recombinational exchange is also borne out by examples such as the ATP- and NAD⁺-dependent DNA ligases shared by mutually exclusive subsets of jumbo phages (Figure 2c).

3.3. Phyletic Patterns Define Correlated and Complementary Functional Systems in Jumbo Phages

Phyletic patterns have served as a tool to objectively define potential functional linkages based on the co-occurrence, shared absences and complementary patterns of proteins. Accordingly, we clustered 181 different protein families identified in jumbo phages using the Canberra distance and Ward algorithm to detect potential function connections between them (Figure 2c). As a test of its effectiveness for detecting functional linkages, we

examined the clusters for certain known functional associations. For example, the individual components of the multi-subunit RNAP group together keeping with their forming a protein complex. Similarly, we found a grouping of dihydrofolate reductase (DHFR) and thymidylate synthase (TS) superfamilies (Figure 2c). It is known that several phages sustain their DNA replication by synthesizing their own thymidylate or a modified base through the transfer of a single carbon-atom fragment (a methyl or hydroxymethyl group) to uracil catalyzed by thymidylate synthase [37,38]. This single carbon moiety is borne on a tetrahydrofolate cofactor which in turn is produced by DHFR; thus, their grouping in the dendrogram recapitulates their functional association. In contrast, the classic TS is anticorrelated with ThyX which is a non-homologous enzyme with the same activity [39]. Thus, the majority of jumbo phages have either a member of the TS or ThyX superfamilies, indicating that they produce their own thymidylate or a modified pyrimidine.

The cluster analysis also pointed to some subtler functional connections. For example, tubulin is correlated in its phyletic pattern with the multi-subunit RNAP (Figure 2c). This suggests that the formation of the nucleus-like compartment went hand-in-hand with the acquisition of host-like transcription machinery that allowed independence from the host transcription apparatus, which is likely kept out by the sub-cellular compartment formed by the tubulin within which phage replication and transcription occurs. The same set of phages also frequently code for a member of the HSP60 superfamily of ATP-dependent chaperones (Figure 2c). This raises the possibility that the chaperone might aid in the (dis)assembly of the nucleus-like compartment in the course of the phage cycle. Curiously, the co-chaperone of HSP60, HSP10 is found in an entirely different set of phages from groups 2 and 3, in which it might recruit the host HSP60 as a partner for virion assembly. Both group 1 and group 2 phages use the prohead serine peptidase of the SH superfamily (Pfam S77); obvious homologs of these are not found in group 3 jumbo phages. However, we observed that group 3 jumbo phages show a complementary pattern to this prohead peptidase with either an unrelated XkdF family serine peptidase [40] or a divergent clade of SH superfamily peptidases [41] (AYD81192.1). This suggests that these might be the equivalent head-protein processing peptidases in this group (see Supplementary Materials S1.1–1.11 for a comprehensive collection of phyletic patterns).

Beyond these, we found several other examples of functional correlations and complementary patterns relating to proteins in DNA replication, transcription, DNA modifications and metabolism. We consider these below categorized by functional systems along with a more detailed discussion on the domain architectures of the proteins under consideration.

3.4. Major Functional Categories of Jumbo Phage Proteins

The infection cycle of jumbo phages displays the same temporal landmarks and functional modules typical of other lytic *Caudovirales* [42]. In *Myoviridae*, the first step, namely invasion, involves the contractile mechanism of the tail-sheath protein with the active injection of the DNA into the host cell via the tail-tip proteins upon engagement of the tail spike proteins with the cell-surface receptors and degradation of the peptidoglycan by lysozymes such as gp13 [43]. The flexible tails of *Siphoviridae* lack a contractile tail-sheath; here, the tape measure protein is believed to form a conduit to deliver the DNA through the host cell wall [43]. The DNA might be delivered along with the accompanying “pilot” proteins (e.g., the RNA polymerase) that help establish the virus once inside the host. The core lytic cycle is initiated by the successive transcription of early, middle and late genes mediated by different transcription factors and or RNA polymerases [42]. Among the products of the late genes, are the replication and recombination proteins on one hand and the structural and DNA-packaging proteins on the other. Together, these complete the replication of the viral DNA followed by its packaging into the head in an ATP-dependent manner and assembly of the complete virion prior to lysis of the host cell. However, there is increasing evidence that this basic lytic cycle might be paused in several jumbo phages for a “pseudolysogenic” phase depending on the environmental stress status of the host [44]. Further, right from the point of the injection of DNA to lysis of the host cell,

the virus and the host are locked in a biological conflict. On the host side, this involves the deployment of a diverse array of phage restriction mechanisms, whereas on the virus side it involves an equally diverse array of enzymatic and non-enzymatic mechanisms to hijack host machinery and counter the restriction mechanism.

Rather than discuss the functional categories of jumbo phage proteins sequentially as per the above-summarized infection cycle, we discuss them in the order of their importance in distinguishing the different groups of jumbo phages and the presence of novel biochemical features with respect to the smaller phages. Consequently, we first cover the basic DNA replication paradigms (Section 3.4.1) and recombination apparatus (Section 3.4.2) followed by the transcription apparatus (Section 3.4.3). These are followed by sections on the strategies used to hijack host systems (Section 3.4.4) and virion structure and assembly (Section 3.4.5). Since the latter of these categories generally resembles smaller *Myoviridae* and *Siphoviridae*, and has been covered in depth before [43], in this subsection we restrict ourselves to previously unreported or underappreciated findings from jumbo phages. These categories are central to distinguishing the higher-order groups of jumbo phages. These are followed by five subsections devoted to the strategies used to counter host defenses (Sections 3.4.6–3.4.10) which present several new observations on these systems.

3.4.1. Proteins that Define the Four Basic DNA Replication Paradigms in Jumbo Phages

One of the primary distinguishing features of the three higher-order groups of jumbo phages is their distinct DNA replication apparatus. All jumbo phages possess their own DNA polymerase, suggesting that they are self-sufficient with respect to their core replication apparatus (Figure 3). Jumbo phages possess one or more of five distinct types of DNA polymerases; three of them have a related core catalytic domain, termed the palm domain that adopts the “RNA-recognition-motif-like” (RRM) fold with four strands and two helices [45,46]. Two active site Mg^{2+} ions are chelated by charged residues from the end of the first strand and the β -hairpin formed by the 2nd and 3rd strands of the core RRM fold (Figure 3a). Strand 1 is followed by an insert region termed the finger module that plays a role in binding the template nucleic acid [46,47]. The remaining two jumbo phage DNA polymerases are homologous to the primary bacterial DNA polymerase III catalytic α -subunit (YP_009015632.1) with a DNA polymerase β (pol β) fold catalytic domain [48,49]. Together, these define four distinct replication paradigms which are outlined below. The group 1 phages are typified by a divergent DNA polymerase (e.g., YP_009153312.1) that has the apomorphic Mg^{2+} -chelating DXD motif in the RRM fold β -hairpin indicating that they belong to the Pol B family (Figure 3a). These polymerases are not found outside of phages and profile-profile comparisons suggest that their closest relatives are the DNA polymerases of NCLDVs like poxviruses. However, these are characterized by certain unique features that are not found to date in any other DNA polymerases, such as an unusually long multi-helical insert just N-terminal to the 2nd strand of the core RRM fold. This insert module might compensate for the absence of a classical sliding clamp in these viruses (Figure 3a). Interestingly, these are often encoded in a conserved gene-neighborhood along with an upstream gene coding for a small SH3-fold domain protein (e.g., AAL82950.1) related to those found in certain nucleic acid-binding toxins (e.g., as 5HK3; 5XE2) [50]. It remains to be seen if these might constitute a subunit of these polymerases (Figure 3a). Of note, these DNA polymerases do not have a fused 3'→5' exonuclease domain; however, such a protein is encoded elsewhere in the genome, suggesting that they might associate as a standalone protein in the replication complex. Going with this divergent DNA polymerase is a divergent version of the replication initiation helicase DnaB belonging to the RecA-ATP-synthetase superfamily of P-loop NTPases [51]. It is characterized by a modified ATP-binding Walker B motif. Nevertheless, its intact Mg^{2+} binding site indicates that it is catalytically active and that its divergence probably went together with that of the associated DNA polymerase.

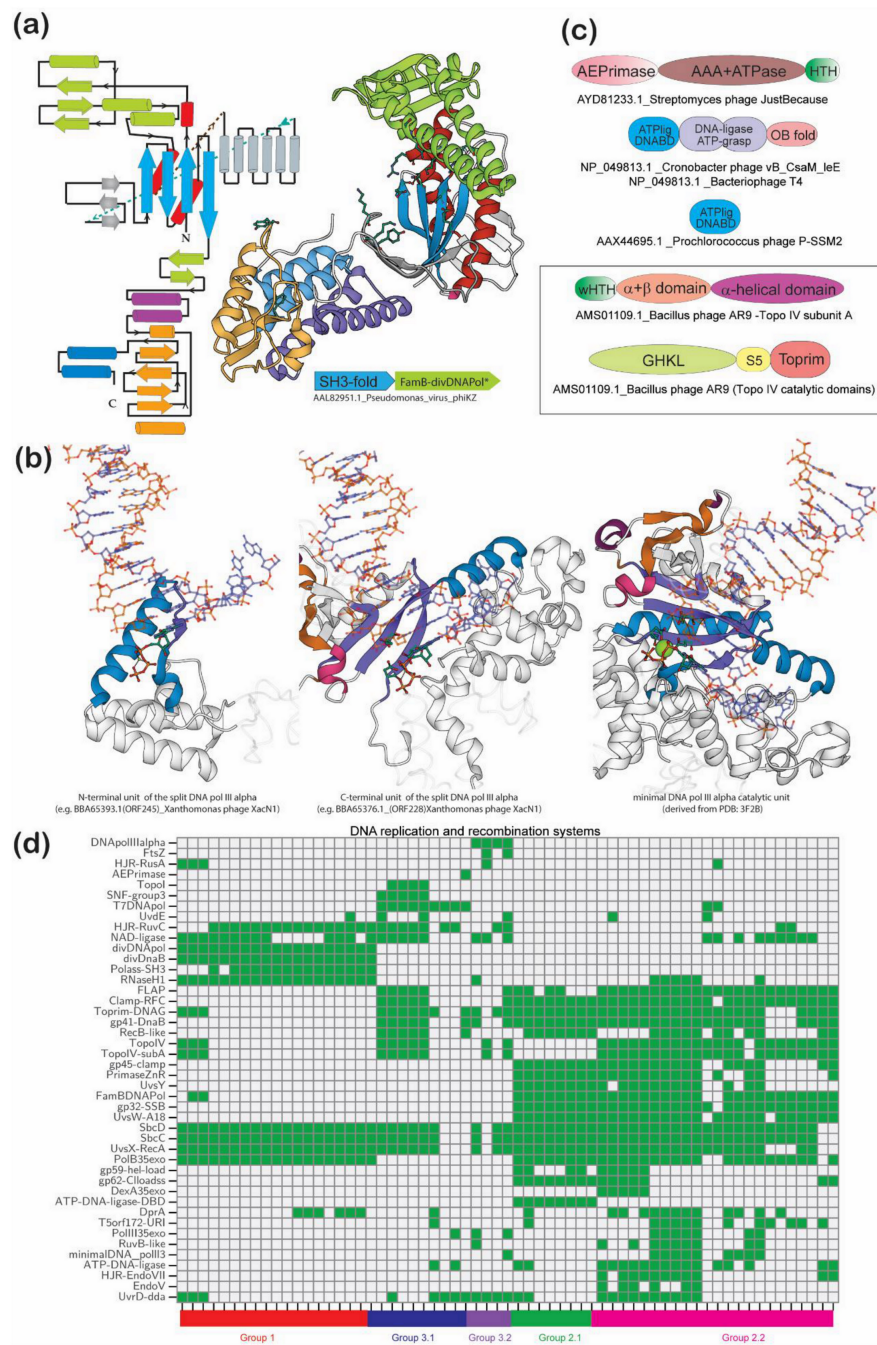


Figure 3. DNA replication and recombination systems of jumbo phages. **(a)** Predicted topology diagram, cartoon structure, and most common gene neighborhood of the divergent DNA polymerase of group 1. In the topology diagram, helices are shown as cylinders and strands as arrows. The cartoon structure was rendered using the MOL* program. Corresponding secondary structure elements are colored similarly in the topology diagram and cartoon. Genes are represented as boxed arrows with arrow-heads pointing to the 3' ends of the gene. The accession number in the label corresponds to the gene marked with a *. **(b)** Predicted components of the split DNA Polymerase III catalytic α subunit. The N-terminal unit is shown on the left and the C-terminal in the middle. **(c)** Domain architectures of some interesting proteins involved in DNA replication and recombination. Proteins are denoted by their accession numbers and phage names separated by underscores. **(d)** Phyletic vector diagram of various replication and recombination proteins in a sample of jumbo phages. The Canberra distance and Ward clustering were used to determine protein and species order in the matrix. Filled green squares indicate presence of a gene. The phage name order is the same as in Figure 4 in all vector diagrams in the main figures.

The group 2 jumbo phages follow the basic paradigm of the coliphage T4 in having a “eukaryote-like” family B DNA polymerase as the primary replicase, which is usually accompanied by a “standard” DnaB helicase (T4 gp41) closely related to the bacterial versions. Several of these phages also code for a helicase-loader (orthologs of T4 gp59) DNA-binding protein with a domain related to the eukaryotic HMG domain chromatin proteins, which recruits DnaB to replication bubbles [52,53]. These phages are also characterized by the OB fold domain SSB (T4 gp32 orthologs), which is also usually recruited to replication sites by the helicase-loader. These phages also typically possess an AAA+ ATPase of the RFC family and its partner, the helical small subunit, which together load a sliding clamp of the PCNA superfamily on DNA [54] (Figure 3d). In the group 2.2 phages, the above family B DNA polymerase is accompanied by a remarkable, previously unrecognized, phage-specific minimal version of the DNA Polymerase III catalytic α subunit (Figure 3b). Strikingly, this version is split up into separate proteins, encoded by distantly located genes, corresponding to the N-terminal NTP-binding and C-terminal Mg^{2+} - and DNA-template-binding subdomains (Figure 3b). It is conceivable that this DNA polymerase might cooperate with the more widespread family B enzyme of this group in synthesis of one of the strands or repair. The subgroup 3.1 jumbo phages show the simplest core DNA replication apparatus with the T7-like DNA polymerase as its only conserved element. The *Myoviridae* subset of these phages also displays a “standard” DnaB akin to those found in group 2 phages. A comparable *Myoviridae* subset of subgroup 3.1 also has an RFC-like DNA clamp loader but thus far lacks a detectable PCNA-like clamp homolog (Figure 3d). The subgroup 3.2 phages show a DNA polymerase III α protein with a N-terminal PHP superfamily nuclease domain. This is a likely case of non-orthologous displacement by a host-derived enzyme that they closely resemble in sequence and architecture.

Beyond these defining elements, other DNA replication components are either more sporadic in the distribution or shared by more than one group. One such is the ATP-dependent DNA ligase that is widely represented in group 2 and a subset of group 3 jumbo phages (Figure 3c,d). On the other hand, the NAD⁺-dependent DNA ligase shows a complementary pattern, being primarily present in group 1, and those members of groups 2 and 3 that lack the ATP-dependent ligase (Figure 3d). Notably, most representatives of subgroup 2.1 lack any known DNA ligase. However, they code for a previously unnoticed standalone version of the α -helical DNA-binding domain typical of ATP-dependent ligases (e.g., AAX44695.1) [55] that could potentially recruit a ligase from the host (Figure 3d). Similarly, the DnaG-like primases are present in the group 2 phages and frequently in the *Myoviridae* subset of group 3. In the former, it also tends to be accompanied by the standalone DNA-binding primase-type Zinc-ribbon (ZnR), which is fused to the DnaG Toprim domain in the cellular primases [56]. In contrast, the Siphoviridae subgroup of group 3.1 frequently exhibits the unrelated archaeo-eukaryote-type primase (AEP) fused to an AAA+ ATPase (likely functionally equivalent to the D5 AAA+-helicase domain; Figure 3c) that can catalyze an equivalent priming reaction [57]. Notably, the group 1 phages with the divergent DNA polymerase lack both the DnaG-type Toprim domain and the AEP. Nevertheless, they all have RNase H1 protein for the removal of the RNA primer. This paradoxical situation raises the possibility that these DNA polymerases might either function as “primipols” capable of both priming [57–59] and elongation or that they use their multisubunit RNA polymerases for priming. Outside of group 1, RNase H1 is seen in several subgroup 2.2 phages but not any of the others (Figure 3d). However, most subgroup 2.1 and a few 2.2 phages possess a nuclease of the 5′→3′ (Flap nuclease) superfamily (e.g., APU01440) paralleling enzymes of the same superfamily found in NCLDV such as poxviruses [2,60]. This could also function as the RNA-primer degrading nuclease in these jumbo phages.

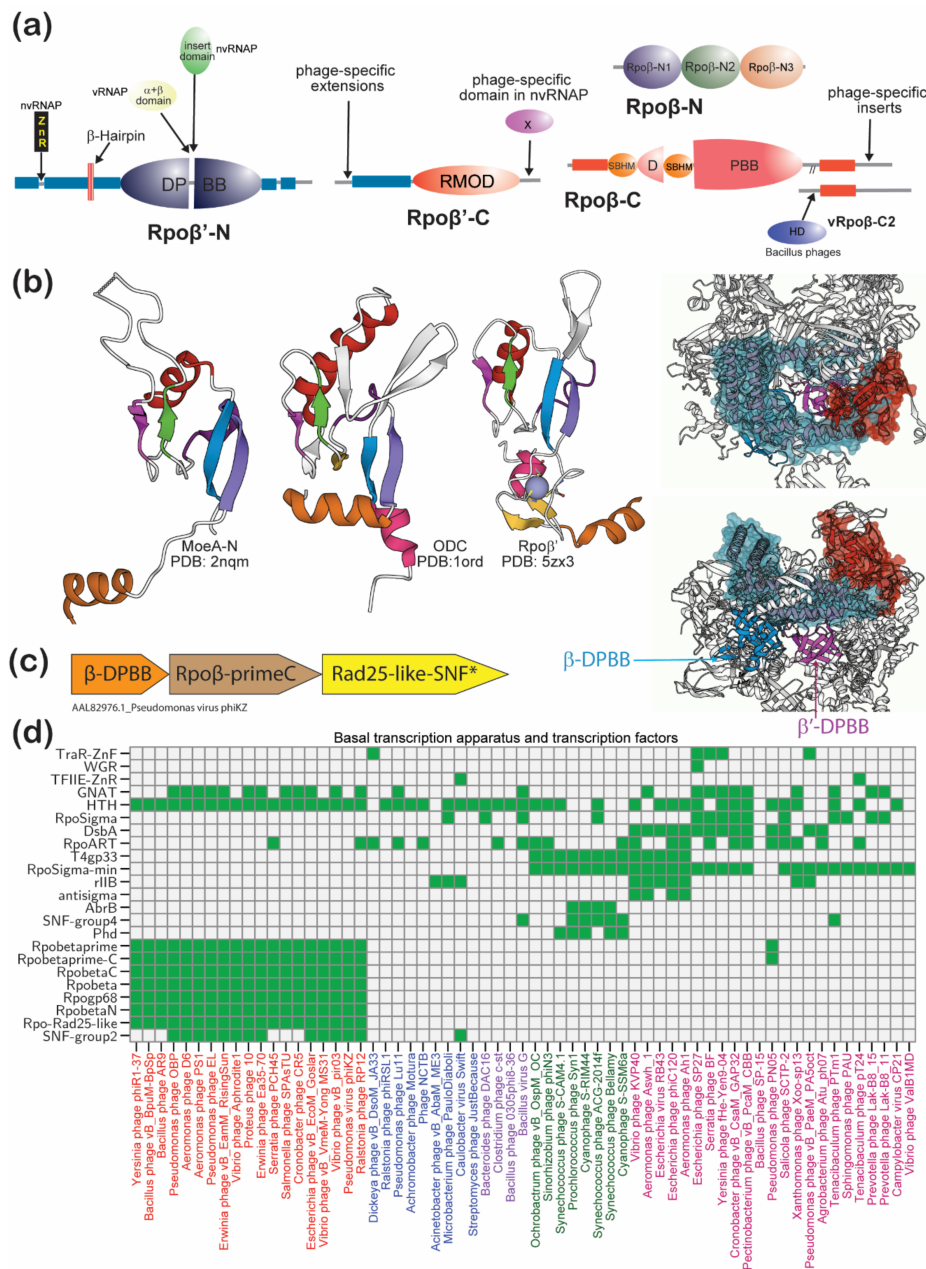


Figure 4. Basal transcriptional proteins and transcriptional factors in jumbo phages. **(a)** Domain architectures of the RNA polymerase β and β' subunits found in group 1 phages. Distinct domains are colored differently. The arrows show the distinct regions of the virion (vRNAP) and nonvirion (nvRNAP) RNA polymerase-specific insertions in the universally conserved cores. **(b)** Cartoon structures of the conserved core of the RMOD domain and its location in the transcript-exit channel of the β' subunit. On the left are MOL* cartoon renderings of the MoeA-N, the ornithine decarboxylase and the C-terminal domain of the β' subunit and on the right are two views of the RNA polymerase, with the β' subunit shaded blue and the RMOD domain brick red. **(c)** Gene neighborhood of the Rad25-like ATPase found in group 1 phages. The accession number in the label corresponds to the gene marked with a *. **(d)** Phyletic vector diagram of various proteins involved in transcription and transcriptional regulation. Filled green squares indicate presence of a gene. Phage names are colored by their corresponding groups. This phage order has been maintained for all vector diagrams in the figures.

3.4.2. The Core DNA Recombination, Topological Manipulation and Minor Repair Systems

Studies on medium-sized DNA viruses, such as coliphage T4, have shown an important role for multiple DNA recombination events, some of which are an essential aspect of the virus cycle. The dominant mode of replication in this virus is recombination-dependent

replication (RDR), which proceeds via the invasion of a duplex by a newly synthesized single strand followed by the formation of a single-strand bubble that serves as a template that is primed by the primase-helicase complex for semi-conservative DNA replication by the viral DNA-polymerase. In T4, this homologous recombination system comprised of the ATP-dependent recombinase UvsX (the homolog of RecA/Rad51), the α -helical single strand-binding protein UvsY, the UvsW DNA helicase of the A18/UvsW family and its potential DNA-binding partner, the trihelical protein UvsW.1 [61,62]. The first three components are conserved across the group 2 jumbo phages (UvsW.1 is limited to subgroup 2.2), suggesting that they utilize the RDR mechanism. Additionally, these recombination proteins also rescue stalled replication forks along with the repair of leading-strand lesions. While UvsY and UvsW are restricted to the group 2 phages, the RecA homolog is far more widespread and can be found in the *Myoviridae* from the three groups of jumbo phages (Figure 3d). This suggests that a form of RDR is likely more widely used by jumbo phages. Consistent with this proposal, we found that both groups 1 and 3 *Myoviridae* have their own distinct versions of Snf2/Swi2 superfamily-2 helicases that could take the place of UvsW in their recombination apparatus. Subgroup 2.2 and the majority of group 3 jumbo phages also share a UvrD-type SF1 DNA helicase (T4 dda) that could also provide backup for the primary helicase UvsW in recombination [63]. It is also conceivable they have recruited their own unique single-strand binding proteins in place of UvsY. In this regard, it is notable that several group 1 and 2 phages conserve a copy of the DprA/Smf protein of the SLOG superfamily that functions as a ssDNA-binding receptor in bacteria [64], which could serve as an additional single-strand binding protein (Figure 3d).

Cellular genomes possess an alternative recombination pathway (non-homologous end joining: NHEJ) that relies on short segments of identity and depends on the DNA-end processing complex comprised of the SbcC/Rad50 ABC ATPase and SbcD/Mre11 nuclease [65]. Homologs of these are also universal in *Myoviridae* jumbo phages. Like the homologous recombination enzymes, these too are seen in numerous medium-sized phages. While they have been previously termed “repair proteins” in jumbo phages [7], given that they are retained across most representatives, they should be interpreted as part of the core recombination apparatus (Figure 3d). Like their cellular counterparts, they might be required for recombination processes related to NHEJ, such as during the bridging of phage chromosomes to mediate rescue of stalled replication forks. This might be of particular importance due to the growing knowledge of host effectors that target viral DNA potentially resulting in moribund replication forks [66]. Alternatively, these could also aid in recombination for solving the end problem, wherein chromosome ends could be lost during initiation of replication by RNA primers.

An often-under-appreciated aspect of the jumbo phage recombination apparatus is the Holliday junction resolvases (HJR). Nearly all group 1 and 3 phages possess a RuvC-like Holliday junction resolvase akin to those found in the majority of bacteria and certain NCLDV such as poxviruses [67,68]. However, as in the poxviruses, this RuvC is not coupled with a RuvB-like ATPase found in the cellular recombination systems (Figure 3d). The phage T4 endo VII-like HJR [69] that is found in subgroup 2.2 and a subset of group 4 phages shows a nearly complementary pattern with RuvC. Another, complementary association for HJRs is seen with the unrelated resolvase RusA displacing RuvC in a small subset of group 1 phages typified by the *Pseudomonas* PhikZ-like phages. That still leaves several phages without a recognized HJR. We observed that several group 2 and group 3 phages possess previously uncharacterized restriction endonuclease (REase) fold domain proteins of the RecB family (e.g., AUZ94847.1) and a novel family (e.g., APU01428.1) that are reminiscent of the REase fold HJR seen in archaea [68] (Figure 3d). Group 2.2 phages also display two conserved proteins with a Uri endonuclease domain (e.g., QAY00453.1 and APU01437.1) previously implicated in HJR function [70–72]. Based on contextual connections to the recombination-related single-strand coating proteins, we previously proposed a role for one of these, the T5orf172 family, in recombination [73]. It is conceivable

that these endonucleases can perform an HJR-like role in phages lacking known HJRs or functions as alternative HJRs in specific repair or replication contexts.

A two-subunit phage-specific topoisomerase is found in a small subset of group 1, subgroup 2.2 and several group 3 jumbo phages. This topoisomerase forms a phage clade separate from other Toprim domain topoisomerases whose closest relatives are the cellular topo IV enzymes (Figure 3c). The phage versions are distinguished by the fusion into a single subunit of the two topoisomerase catalytic domains, namely the GHKL ATPase domain that drives conformational changes in the complex to mediate DNA-strand manipulation, and the Toprim domain that mediates strand breakage and rejoining [56]. Their second subunit is comprised of the mainly α -helical domain that forms a hoop around the single DNA strands that are moved during topological manipulation [74] (Figure 3c). The shared presence of this topoisomerase enzyme across otherwise distinct higher-order groups of phages points to the dissemination of these topoisomerases through lateral transfer between different viruses driven by the selective pressure of the independent growth in genome size. Some phages of subgroup 2.2, such as the *Achromobacter* phage Motura, Phage NCTB, *Ralstonia* phage phiRSL1, and *Pseudomonas* phages Lu11 and PaBG also possess a phage-specific version of the unrelated topoisomerase I with a tyrosine recombinase superfamily catalytic domain (Figure 3d, Supplementary Materials S1.3) [75].

Beyond these more widely conserved core components, there are additional sporadically distributed DNA repair proteins (Figure 3d). One of these is the backbone-cleaving DNA glycosylase typified by the coliphage T4 endonuclease V which is found in a subset of the subgroup 2.2 phages. It has been shown to repair pyrimidine dimer lesions [76]. In a similar vein, a subset of subgroup 2.2 and group 3 phages possesses a TIM-barrel fold UvdE endonuclease which has been shown to participate in the excision damaged nucleotides by cleaving DNA immediately adjacent to the 5' phosphates [77].

3.4.3. The Basal Transcription Apparatus and Transcription Factors

Medium-sized *Caudovirales* either encode their own RNAPs, which wholly or partially transcribe their genes, or rely on hijacking the host transcription apparatus for their own transcription. The RNAPs of phages belong to either of two unrelated families defined by their catalytic domains: (1) Those with the RRM fold palm domain that are distantly related to the RRM fold palm domains of the DNA polymerases. These are found in phages like T7 or N4 and are both injected during invasion and synthesized after invasion [78]. (2) The double-barrel domain RNAPs (Figure 4a). These have an active site formed at the interface of two DPBB domains, with one supplying a triad of aspartates in a DXDXD signature that binds the two Mg^{2+} ions and the second that supplies two basic residues to stabilize the phosphotransfer reaction intermediate [17,18]. This superfamily includes all cellular and killer plasmid RNAPs, the eukaryotic RNA-dependent RNA polymerases and their caudoviral (e.g., YonO) and transposon relatives, and the archaea-specific DNA polymerases [79–81]. Such enzymes are also found in all the group 1 and two group 3 jumbo phages (*Bacillus* phage 0305phi8-36 and *Clostridium* phage c-st.). Those in the group 3 phages are YonO-like enzymes with both DPBB domains in a single subunit (e.g., ABS83659.1, BAE47877.1), whereas the group 1 enzymes are closer to the multisubunit host RNAPs. The latter typically come in two paralogous copies (except possibly *Pseudomonas* phage PN05 with a single β' cognate; Supplementary Materials S1.1), one of which is packaged into the virion (vRNAP) and the other produced later for middle and late gene transcription (nvRNAP).

The jumbo phage RNAPs along with their orthologs from medium-sized phages have been extensively studied in recent works and proposed to define an ancient clade of double-barrel RNAPs [17,82]. However, as several intricacies of their architecture remain poorly described we consider those points and their significance for the evolution of this class of enzymes.

The catalytic core of cellular RNAPs consists of two subunits termed β and β' that have accreted around the two DPBB domains [79] (Figure 4a). In group 1 jumbo phages,

the equivalent of β' is split up into two distinct polypeptides. One of these encompasses the long N-terminal α -helical array domain with a characteristic β -hairpin element and the catalytic DPBB domain that chelates the Mg^{2+} ions. The cellular β' DPPBs are characterized by the presence of an insert domain just upstream of the active site residues, which in bacteria is an α -helical bundle domain and in archaeo-eukaryotes is a RAGNYA superfamily domain [80]. Notably, while the phage β' DPBB also has inserts at the same position, they are neither related to the eukaryotic nor the bacterial versions. Instead, the vRNAP and the nvRNAP each have their own unique insert domains, with a longer $\alpha + \beta$ domain seen in the former (Figure 4a). Like the bacterial RNAP β' subunit, the nvRNAPs have a Zn-ribbon inserted into their N-terminal α -helical array domain [80]. The conserved C-terminal region of β' consists of a small 2-layered $\alpha + \beta$ domain followed by a second region, which assumes a doughnut-shaped structure that allows for the exit of the mRNA (Figure 4b). This core occurs as a standalone protein in the group 1 jumbo phages and provides hints regarding its evolutionary origins. Via structural comparisons, we established that the core of the doughnut-shaped domain is defined by a hitherto unreported superfamily (the RMOD domain; Figure 4b) that unifies the basic amino acid decarboxylase C-terminal domain and the molybdopterin biosynthesis protein MoeA N-terminal domain (DALI Z-scores 5.3–6.2; Figure 4b). In the RNAPs, this domain is inserted into a Zn-cluster domain to form a more elaborate structure that further supports α -helical hairpin extensions to form the β' C-terminal module. The phage nvRNAP versions of this module are further distinguished by the fusion to a unique C-terminal domain that is not found in any cellular RNAP β' subunit (Figure 4b).

The group 1 phage cognates of the cellular β subunit are split up into two separate polypeptides in the case of the nvRNAP or three in the vRNAPs. One of these corresponds to the N-terminal part of the cellular RNAP β subunit. The remainder of the β cognate is specified in the nvRNAP by a single protein containing the catalytic DPBB domain, which contributes two conserved lysines to the active site, and its C-terminal extension. The later part occurs separately as a third protein in the vRNAPs. In group 1 *Bacillus* phages, the region homologous to the C-terminal part of the cellular RNAP β subunit is further fused to an active HD phosphohydrolase domain [16] (Figure 4a). The phage β DPBB domain shares with all cellular and killer plasmid β DPBBs the insert of a sandwich-barrel-hybrid motif (SBHM) domain, which is absent in the RDRPs (and their phage relatives like YonO and NCGL1702-like RNAPs of mobile elements) [80]. The phage proteins containing the β DPBB also have a further N-terminal SBHM domain shared with the bacterial β subunits and are distinguished by a unique N-terminal region that has so far not been found in any other RNAP (Figure 4a). The SBHM inserted into the DPBB of the β subunit (the so-called “flap”) binds the basal transcription factors, either sigma or TFIIB, in the cellular versions [80]. Notably, no sigma factor is found in any group 1 phages, raising the possibility that it interacts with some other protein (Figure 4d). In the case of the nvRNAP, this could be the enigmatic phage-specific gp68 subunit. Through secondary structural analysis, we predict that this subunit is composed of a series of α -helical hairpins reminiscent of those seen in supersecondary-structure-forming domains such as tetratricopeptide and HEAT repeats. The unique inserts and domain fusions of phage RNAP subunits, which have no parallels in the cellular versions, support the idea that these phage enzymes are early-branching and have not been derived from cellular homologs via the splitting of the β and β' subunits. Instead, they appear to be remnants of the transcription apparatus from a replicon of the pre-last universal common ancestor (LUCA) period. Hence, it is likely that such multi-subunit versions fused to give rise to the two core subunits of the cellular versions along the stem lineage leading to the LUCA.

Group 1 phages code for a Swi2/Snf2-type SF2 helicase that is specifically related to Rad25 and shows a phyletic pattern strongly correlated to the RNAP subunits (Figure 4c,d). It is also often encoded in gene-neighborhoods associated with genes for the β' C-terminal gene and the catalytic β DPBB containing subunit (Figure 4c). Hence, this helicase might

act in concert with the RNAP in transcription—one possibility is that it is a functional equivalent of the bacterial RNAP recycling SWI2/SNF2 enzyme RapA [83].

Other than group 1 and two group 3 phages, none of the remaining jumbo phages code for their own RNAPs. However, all group 2 phages code for sigma factors (Figure 4d). These sigma factors come in two types: a minimal version with just the single helix-turn-helix (HTH) domain which binds the promoter and interacts with the SBHM domain inserted into the RNAP β subunit and a more complete version close to the cellular RpoD sigma factors [80]. The minimal sigma factor seen in some group 2 phages could again be a remnant of ancient sigma factors prior to the duplication and elaboration seen in the cellular forms. The versions closer to the cellular sigma factors could have been secondarily acquired by the viruses from the cellular genomes. In addition to the *bona fide* sigma factors, the group 2 phages also possess an ortholog of the coliphage T4 late-gene transcription factor gp33 protein (Figure 4d). This protein contains a HTH domain, which recruits the RNAP by binding the SBHM domain of the β -subunit in a manner similar to the sigma and TFIIB HTH domains [84]. The sigma factors and gp33 suggest that group 2 phages adopt a strategy converse to that of the group 1 phages—they likely utilize their sigma factors to recruit the catalytic subunits of their hosts for mRNA synthesis (see below).

A neglected aspect of the jumbo phage transcription apparatus is the suite of phage-encoded transcription factors (Figure 4d). As lytic phages, it is generally believed that jumbo phages have a simple transcriptional program wherein early, middle and late genes are successively transcribed without major branching or switching for distinct alternative programs typical of lysogenic phages [85,86]. However, we found that the majority of jumbo phages (92%) code for their own specific transcription factors (TFs; median number = 2, range 1–15). While most of these have either classic HTH or the winged HTH (wHTH) DNA-binding domains, a few have the Phd and AbrB DNA-binding domains found in certain toxin-antitoxin systems [87] or the $\alpha + \beta$ WGR DNA-binding domain found in the bacterial MolR-like TFs [88] (e.g., QAY00573.1). Subgroup 3.2 phages, like *Bacillus* phage G, also show an expansion of novel Myb-type HTH TFs related to the firmicute RsfA-like TFs (Figure 4d, Supplementary Materials S1.4) [89]. At least in the case of the group 1 phages, it is possible that they deploy different TFs to work in combination with their RNAPs. The *Siphoviridae* subset of group 3 phages has an ortholog of the Zinc ribbon (ZnR) found in the archaeo-eukaryotic TFIIE (e.g., ARB13787.1) [90,91]. Given that group 3 phages lack both RNA polymerase subunits and sigma factors, it remains to be seen what role these viral TFIIE-like ZnRs might play. *Bacillus* phage G has an unprecedented lineage-specific expansion (~35 copies) of a ZnR (e.g., YP_009015343.1) that might also function as specific TFs or alternatively as structural DNA-binding proteins (Supplementary Materials S1.4). Another specific TF shared by several group 3 and group 2 phages is an ortholog of coliphage T4 rIIB protein. Early studies showed that disruption of this locus results in a more virulent phage, suggesting that it might function as a negative regulator that modulates phage transcription depending on the physiological state of the host [92,93]. In a similar vein, several group 1 (e.g., QDJ96661.1) and group 3 (e.g., ARB13751.1) phages possess repressor TFs with lambda cro/cI-type HTH domains (cHTH) that could play a role in negatively regulating certain transcriptional programs [94]. It is conceivable that the above two sets of TFs play a role in the phenomenon of pseudolysogeny reported for certain jumbo phages as a response to their hosts being under stress [44]. In principle, some of these TFs might also be used to modulate host transcription. Conversely, several group 1 phages possess predicted transcriptional activators with AraC-like HTH domains (e.g., QGZ14570.1) [95]. Subgroup 2.2 contains another potential transcriptional activator orthologous to the coliphage T4 dsDNA-binding protein DsbA that might recruit the host RNAP to late promoters [85,96]. We found that DsbA is a homolog of the *Caulobacter* GapR family of nucleoid-associated proteins [97] and is accordingly predicted to similarly form a ring around dsDNA. The above observations indicate that jumbo phages might display greater flexibility in their transcriptional programs than previously expected.

3.4.4. Alternative Mechanisms of Jumbo Phages for Hijacking Host Regulatory Machinery

Studies on the coliphages T4 and N4 and more recently the *Pseudomonas* phage phiKMV have indicated that the hijacking of the host transcriptional apparatus by phages involves a battery of non-enzymatic and enzymatic factors [85]. The non-enzymatic factors include a variety of host RNAP-interacting proteins that drive it towards preferring the phage promoters. The enzymatic mechanisms include the deployment of an array of ADP-ribosyltransferases (ARTs) such as T4 Alt, ModA and ModB that modify the RNAP α -subunit and other proteins to make it switch preference for viral templates [98] (Figure 4d). Similarly, the phiKMV-like phages deploy the GCN5-like NH_2 -group acetyltransferase (GNAT) Rac to acetylate the host RNAP α -subunit leading to its eventual cleavage, thereby drawing it away from highly active host promoters to the viral early promoters [99] (Figure 4d). We recently reported the existence of both ART and GNAT domains in the large multidomain polyvalent proteins injected by certain phages into their hosts along with their DNA that might play a similar role as the above [100]. Indeed, we also found a whole slew of such factors encoded by different jumbo phages.

Among the non-catalytic hijacking factors, we found the anti-sigma factor prototyped by T4 HTH protein AsiA in a subset of group 2 jumbo phages. AsiA interacts with the primary sigma factor of the host and inhibits transcription from both bacterial and phage early promoters to cause the RNAP to switch to middle promoters [101]. We also found orthologs of the T4 Alc/Alp protein in a subset of group 2 jumbo phages (e.g., APU01786.1) that binds both the β subunit and host primary sigma factor to shut down transcription except if the DNA template contains the modified base 5hmC which is found in viral DNA (see below) [102]. Thus, it specifically redirects the RNAP for late viral transcription (Figure 4d).

Among the enzymatic modifiers, we found numerous GNATs (e.g., QBP07143.1) across all three groups that are predicted to function like the phiKMV Rac in modifying host proteins such as RNAP subunits to favor viral transcription [99]. Some GNATs found in group 1 phages (e.g., AMR59843.1) are most closely related to the ribosomal protein GNATs [103]; hence, they might help hijack the host translational apparatus or render it refractory to ribosome-targeting effectors. A comparable peptide modification is suggested by the sporadic presence of the R2K family of ATP-grasp peptide ligases in subgroup 2.2 (e.g., AUZ95380.1), a feature shared with certain amoeba-infecting NCLDVs. Members of this family of enzymes are predicted to conjugate aminoacyl moieties to translation factors in eukaryotes and could likewise aid the viruses in modifying host proteins [104]. Similarly, jumbo phages from all three groups encode several ARTs, suggesting that ADP-ribosylation of host proteins is a common strategy for mediating outcomes that are favorable to the phage (Figure 4d). We found that ARTs are most common in group 2 and group 3 phages that do not encode their own RNAPs. Therefore, at least some of these ARTs are probably used similar to the T4 enzymes in modifying the host RNAP. Also notable is the presence of the viral versions of the polyADP-ribose polymerases or transferases (PARPs/PARTs) which are typical of eukaryotes in few jumbo phages (e.g., BBA65562.1). We previously reported such versions as being shared with bacterial polymorphic toxins and being precursors of the eukaryotic PARPs [23,105]. The phage PARPs might function similarly to the ARTs but instead of a single ADP-ribose moiety transfer longer chains of ADP-ribose polymers.

The observation that the acetylation by Rac triggers cleavage of the RNAP α -subunit points to the possible role of phage-encoded peptidases in targeting host proteins. Such a strategy has a much wider parallel across viruses, with multiple eukaryotic viruses targeting host ubiquitination [100,106–108]. Most well-conserved peptidases in jumbo phages correlate with particular virion types, suggesting that they are primarily deployed for processing viral proteins during maturation; others show a patchy pattern as might be expected of those that are deployed against the host. The most prominent of these is the ClpP serine protease (QDJ96709.1) found in some group 1 and group 2.2 phages. Several other phages instead encode their own ClpS protein (e.g., AMM43882.1), which is an adaptor that likely brings target proteins to the host ClpAP system for degradation (Figure 5a) [109]. Some group 2.2 phages also show 1–3 copies of a SprT-like metallopeptidase (e.g., APU01787.1),

whereas some group 3 phages show ArdC-like MPTases [100] both of the Zincin-like superfamily (Supplementary Materials S1.6). Their limited distribution, variability, and copy number differences between closely related phages favor the possibility of them targeting different host proteins for cleavage.

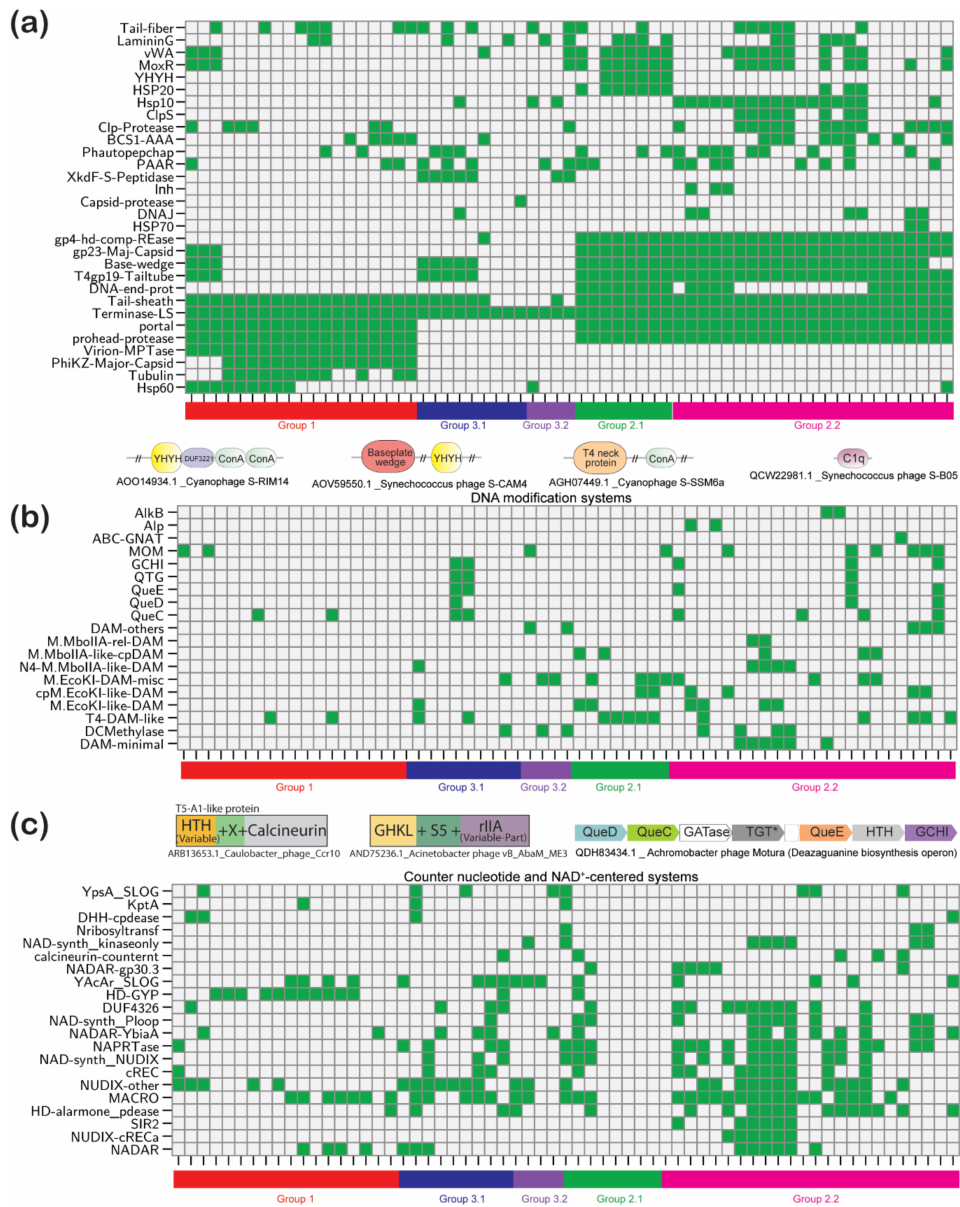


Figure 5. Virion structure and conflict systems. Vector diagrams, domain architectures and operons of (a) virion morphogenesis and chaperone proteins; (b) DNA modification proteins and (c) proteins involved in counter nucleotide and NAD⁺-centric systems. The Canberra distance and Ward clustering were used to determine protein and species order in the vector diagram. Filled green squares indicate presence of a gene. The virus name order and coloring is the same as in Figure 4. Refer to Figure 3 legend for details on domain and gene neighborhood depiction. In domain architecture representations, distinct domains are colored differently.

3.4.5. Some Notable Features of Virion Structure and Maturation

While the broad morphological features of jumbo phages and some of the conserved players in DNA packaging and virion maturation, such as the terminase and the prohead peptidases, are relatively well understood, several harder to detect or lineage-specific players in these functions remain less understood. In this section, we briefly discuss a few of these (Figure 5).

With respect to DNA-processing and packaging, the C-terminal RNase H fold endonuclease domain associated with the terminase large subunit is a universal feature across *Caudovirales* [110]. However, group 2 phages, which possess the classical terminase small subunit, are characterized by the additional virion-associated TnsA transposase-like nuclease domain with the REase fold (the so-called “head completion protein”, e.g., AMO43252.1). While this manuscript was in preparation, a report was published regarding its homolog gp4 from coliphage T4 showing that its non-specific DNase activity was required for packaging [111]. This group is also characterized by the “DNA end protector”, a HTH domain protein (e.g., AGH07633.1). This nuclease and the end protector protein likely facilitate a specific DNA-processing event required for packaging in group 2 phages. Likewise, there are several group-specific peptidases among the structural proteins that might catalyze additional processing events for virion maturation. Group 1 and 2 jumbo phages possess the portal protein (T4 gp20 ortholog) and the prohead assembly serine peptidase (e.g., AAL83076.1, coliphage T4 gp21 ortholog) that show identical phyletic patterns (Figure 5a). The latter has been demonstrated to play a role in cleavage of head proteins during assembly of the portal through which the DNA is loaded into the head [112]. Group 1 phages are distinguished by an additional conserved, uncharacterized zincin-like metallopeptidase (AMR59696.1), which could potentially catalyze other such cleavage reactions during maturation [100] keeping with reports of extensive proteolysis during maturation [113] (Figure 5a).

Some of the proteolysis events relating to the assembly and maturation of the virion are closely linked to chaperone functions that tend to have a more lineage-specific distribution. One of these is the small S74 autopeptidase domain that is found in group 2 jumbo phages (e.g., AAQ64368.1), NCLDV of the Marseillevirus clade, and in eukaryotic DNA-binding TFs such as the myelin regulatory factor [114]. This domain is characterized by conserved serine and lysine residues that are required for catalyzing autoproteolytic detachment from the larger proteins they are part of. Subsequently, the S74 domain assembles into hexamers that aid the folding of the remainder of the parent polypeptide. This domain appears to be deployed in the assembly of tail-fiber, tail-spike and other proteins that tend to have high low complexity content [114]. Other than the Hsp10 subunit of the chaperonin complex that was noted above, Hsp20 (e.g., AUZ95449.1) was previously reported in jumbo cyanophages [115] and is also found in one of the largest jumbo phages, *Atu_ph07*. We found DNAJ domain cochaperones of HSP70 to be sporadically distributed among jumbo phages mainly in group 2.2 and are predicted to recruit the host HSP70 to aid folding and assembly (Figure 5a). Interestingly, group 2.2 jumbo phages also contain a previously unrecognized MoxR AAA+ ATPase (e.g., ASV44729.1)-vWA (e.g., ASV44728.2) chaperone-cochaperone pair, wherein the vWA cochaperone is fused to a N-terminal metallopeptidase domain. A MoxR-vWA chaperone system without the peptidase domain was previously reported to be required for the assembly of the tail of the *Acidianus* two-tailed archaeal virus [116]. We propose that in group 2.2 jumbo phages it similarly plays a role in virion assembly along with the peptidolytic processing of structural proteins.

Typical of most subgroup 2.2 phages is the coliphage T4 type gp40 portal assembly chaperone (e.g., APU01407.1), which indicates that in these the assembly of the prohead occurs via recruitment of the portal protein to the membrane [117,118]. A more widely distributed mechanism of membrane-proximal assembly is indicated by the orthologs of the AAA+-ATPase chaperone/translocase BCS1 that is found in several representatives of each of the 3 major groups (Figure 5a). In eukaryotes, the heptameric BCS1 protein functions as a translocase that binds fully folded proteins in the mitochondrial matrix and translocates them across the inner membrane [119]. It is possible that the phage BCS1 proteins function similarly in membrane-proximal assembly to bind and translocate folded proteins that are packaged into the head. Different subsets of subgroup 2.2 and some group 3 phages may also possess other AAA+ chaperones such as ClpA (AUZ95251.1) or ClpX (AUZ94814.1, AEO93517.1) that could play a role as translocases of unfolded proteins or aid in virion assembly (Figure 5a).

An important aspect of the mature virion is domains associated with the fibers and surface decorations of the tail and head which play a role in adhesion via specific interactions with host peptidoglycan-associated and capsular carbohydrate molecules (see below). Notable among these are different families of lectin domains that might mediate specific interactions with carbohydrate moieties on the host surface. Several of these are related to adhesion domains that are involved in cell–cell interactions in multicellular eukaryotes like animals. Such include the FGS domain of the C-type lectin fold which is found to be diversified by error-prone reverse transcriptases in several phages [120]. While we found no FGS domains in jumbo phages, previous studies have reported another β -sandwich fold lectin domain, the discoidin domain, in certain jumbo phages [121]. We found a wider distribution of this domain across group 3 phages and sporadically in other groups. We also found certain jumbo phages to contain one or more lectin domains with the concanavalin fold that are associated with the tail fibers and other virion components. These include the Laminin G3 domains and the SPRY domain (e.g., QDH50631.1) both of which have been acquired by eukaryotes. In eukaryotes, the former plays similar roles in adhesion [122], whereas the latter has been extensively used in recognition of invasive viral molecules in the TRIM-type anti-viral immune proteins [123]. Interestingly, the cyano jumbo phage *Synechococcus* phage S-B05 has a tail protein (QCW22981.1) with a C1q domain shared with the eponymous complement proteins of the vertebrate immune systems. While a domain with a similar fold was reported in the tail knob protein of the Sf6-like tailed phages [124], the above cyanophage version is much closer to the animal homologs (Figure 5a).

Several jumbo phages code for the NlpC/P60 proteins with the papain-like peptidase fold (e.g., ARB13621.1), that are sometimes tail-associated, and have been previously characterized to hydrolyze peptide-linkages in the peptidoglycan [125]. Similarly, they might also display proteins with the enigmatic YHYH domain (e.g., AOO14285.1) that is often fused to tail/base plate component proteins or the concanavalin-like lectin domains [126] (Figure 5a). We found that this domain contains a triad of highly conserved histidines and an aspartate that is predicted to form a Zn²⁺ binding site. Given these associations, we predict that the YHYH domains might be a hitherto uncharacterized tail-associated metal-dependent hydrolase. Thus, both the NlpC/P60 and YHYH domains appear to be a key part of the phage peptidase repertoire (in addition to the tail-associated lysozymes) to breach the peptidoglycan layer during invasion.

The PAAR domain was first found in polymorphic and related toxins delivered via the type VI secretion system (T6SS) and predicted to be an integral component of that secretion system [105]. Subsequently, it was shown to sharpen the phage-tail-like structure of the T6SS injection apparatus and also recruit toxin and other effectors to the tail [127]. The majority of *Myoviridae* jumbo phages from across all three groups possess a PAAR domain protein, which unlike those from T6SS is present in a standalone form unfused to any effector domain (e.g., APU01493.1). Thus, it represents the ancestral form of the PAAR domain found in the phage tail tip that was then repurposed as a component of the cellular T6SS.

3.4.6. Conflict-Related Adaptations of Jumbo Phages

Biological conflicts between parasitic and self-sufficient replicons have shaped their genomes since the beginning of life. On the host side, this has resulted in a spectacular array of molecular immune mechanisms that limit the viral cycle by a range of different strategies. Several of these have come to light due to comparative genomic analysis and have subsequently been validated through biochemical studies [128,129]. These include disparate mechanisms such as the modification of cell-surface molecules to prevent attachment of the virus, targeting of viral nucleic acids by restriction and CRISPR/Cas systems, limiting essential metabolites such as NAD⁺ by targeted degradation, apoptotic or self-targeting mechanisms that limit viral replication and spread by targeting the ribosome or inducing cell-suicide [130]. These processes, especially those involving self-targeting or suicide, are often under tight-regulatory control by threshold-setting signals in the form of diverse nucleotides and NAD⁺ derivatives [24,131,132]. On the phage side, the counter-mechanisms

against host defense and competing viruses are less understood. However, the jumbo phages are in a privileged position as they have been recently investigated for some of these mechanisms involving “fortification” of their systems by formation of subcellular compartments [133] and CRISPR-based mechanisms to hijack host functions [9].

Our analyses reveal that they possess several other conflict systems to counter not only the host attacks but also to prevent super-infection by other viruses and plasmids. These include: (1) an array of DNA modifications; (2) RNA repair mechanisms and provisioning of tRNAs to head off host self-attacks on the translation apparatus; (3) Metabolic systems to counter host NAD⁺ restriction; (4) Systems to target nucleotides and NAD⁺ derivatives used as signals or toxins by the host immune mechanisms; (5) Inhibitory factors to prevent superinfection; (6) Host surface modification to preclude invasion of the cell by other viruses. Beyond these, there are other poorly understood phage counter-defense systems, such as those centered on the orthologs of the coliphage T4 rIIA protein [134] which combine a N-terminal GHKL superfamily ATPase domain [62,135] with a highly variable C-terminal domain (Figure 5c). We discuss some of the major players involved in each of these processes in greater detail below.

3.4.7. DNA Modifications in Jumbo Phages

DNA nucleobase modifications are among the key devices deployed across bacteriophages with DNA genomes to evade restriction attacks on their genomes by host endoDNases. These DNA modifications might also play epigenetic roles in phage-specific gene-expression and packaging of the DNA into the head [38,136,137]. Further, modification of phage DNA enables them to launch endonucleolytic attacks on the host DNA to preempt expression of host-counter phage defenses and potentially recycle host nucleotides for phage replication [138]. DNA modifications in jumbo phages came to light with the *Sphingomonas* phage PAU whose genome size was misestimated by certain electrophoretic procedures due to its extensive DNA modifications [139]. It was also observed that multiple *Bacillus* jumbo phages have almost all thymine in their DNA replaced by uracil [140,141]. Our systematic analysis of the jumbo phage genomes reveals that all three major mechanisms of DNA modifications can be seen in different lineages: (1) production of pre-modified nucleotides, like UTP or hydroxymethyl CTP, for DNA synthesis; (2) in situ modification by specialized DNA modification; (3) production of a pre-modified base followed by its incorporation in DNA via a transglycosylation reaction.

In terms of the phages using pre-modified nucleotides for DNA synthesis, multiple negative determinants are required for using uracil in place of thymine, namely the absence of a thymidylate synthase or a Thy1, and a dUTPase (Supplementary Materials S1.1, S1.5). This unique situation is seen in the *Bacillus* jumbo phages like PBS1 that use uracil in place of thymine [16]. Phages with specialized thymidylate synthase superfamily enzymes for synthesis of 5 hmC can generate that base in a manner similar to the synthesis of thymine from uracil [38]. Other than the dUTPase, there are also MazG-like and CYTH superfamily triphosphatases scattered across all groups of jumbo phages, which could potentially play a comparable role in degrading endogenous NTPases with unmodified nucleobases (Supplementary Materials S1.1, S1.11). A subset of these phages often has a second determinant in the form of an ortholog of the T4 Alc protein that helps redirecting the host RNA polymerase solely towards templates that contain 5 hmC [102] (Figure 5b). This is primarily seen in a subset of group 2.2 phages. The absence of TET/JBP pyrimidine hydroxylases suggests that all 5 hmC in jumbo phages characterized to date is synthesized at the level of free nucleotide and not generated in situ in DNA [136].

We extend the previous observations of DNA N⁶ adenine methylases (DAMs) and DNA 5-cytosine methylases (DCMs) in jumbo phages to define several previously unrecognized exemplars of these [137] (Figure 5b). The DAMs can be classified into three higher-order clades and the related cytosine N4-methylases; representatives of all these are found in different jumbo phages. The most common are the coliphage T4-like DAMs (Clade 3) which are found in representatives of all three higher-order groups of jumbo

phages (e.g., AUZ94846.1). The next most prevalent are the M.EcoKI-like DAMs (Clade 2), a circularly permuted version of it and other smaller related clades are found in some group 2 and 3 jumbo phages (e.g., APU01582.1). Similarly, some group 2 and 3 phages also possess the M.MboIIA/M.MunI-like DAMs (Clade 1) with or without a circular permutation (e.g., AUZ95158.1, AQW88654.1). A subset (AAX44419.1) of these methylases shows the characteristic motifs that indicate them as being cytosine N4-methylases [137] and are found sporadically in members of groups 2 and 3. A distinct family of DAMs from subgroup 2.2 (AMM43637.1), which are much smaller than the DAMs with DNA substrates, are closest to the Clade 3 DAMs defined by the PCIF1 enzymes, which were recently shown to methylate adenines at the 5' ends of eukaryotic mRNAs [137,142]. Hence, these phage enzymes might also be comparable RNA adenine methylases that play a role in post-transcription regulation. Lastly, DCMs are also found in group 2 and group 3 phages. A small number of phages display an AlkB-like 2OGFeDO (e.g., AXQ68776.1) which acts on methylated adenines either in the context of DNA repair or resetting of the methyl marks [143,144]. Its sporadic and limited presence suggests that rather than DNA repair it might be directed at resetting methyl marks either in the phage as part of an epigenetic regulatory process or in host DNA to target protective host DNA adenine methylation (Figure 5b).

Other than these well-known modifications, we also identified enzymes for other larger DNA-modifications that have previously not been reported in jumbo phages. Several group 2 phages code for enzymes of the Mom family of the GNAT superfamily that are prototyped by the Mom enzyme of phage Mu. This enzyme catalyzes the transfer of an acyl group from an acylated coA moiety, such as glycyl coA, to the N⁶ atom of adenine (Momylation) [145,146]. Since then, several momylation systems are encoded by both phage and bacterial genomes [136,145]. The jumbo phage Mom domains are typically fused to REase fold endoDNase domains that are predicted to function as the restriction component (Figure 5b). This implies that Momylation protects the phage DNA via modified adenines, while the associated endonuclease domain attacks the DNA of the host or superinfecting phages. We previously identified a further member of the GNAT superfamily coupled with an ABC ATPase domain that was predicted to carry out a yet uncharacterized acyl modification of a DNA nucleobase [136]. This DNA modification system is found in a small subset of group 2 jumbo phages such as the *Tenacibaculum* phage pT24 (e.g., QAX98348.1) pointing to hitherto unrecognized modifications among the jumbo phages (Figure 5b). Similarly, a small number of jumbo phages such as the *Achromobacter* phage Motura contain a system for the synthesis of a deazaguanine base like queuine or archaeosine, and a transglycosylase enzyme for their incorporation into DNA in place of guanine [136] (Figure 5c). In some jumbo phages, only a subset of enzymes of this system is observed (Figure 5b). However, given that several bacteria possess such DNA modification systems, it is possible that the phages complement their deazaguanine production and incorporation systems with the host enzymes [136].

DNA modification enzymes are most common in group 2 and 3 jumbo phages (Figure 5b). These might contain more than one such system—for example, phage vB_PaeM_PA5oct contains five momylating enzymes and a deazaguanine synthesis/incorporation system (Supplementary Materials S1.11 and S5). Similarly, multiple phages might possess more than one type of DNA methylase. This multiplicity possibly helps them overcome more than one host restriction system. The group 1 phages generally possess fewer DNA modification systems, keeping with the proposal that the nucleus-like compartment formed by them limits host restriction attacks [19,20] (Figure 5b). However, those that do carry such enzymes could utilize them for protecting viral DNA immediately after invasion.

3.4.8. Counter-Nucleotide and NAD⁺-Centered Systems

A key discovery in recent years is the pervasive role of cyclic and linear (oligo)nucleotides, nucleotide- and NAD⁺ derivatives as signals in immune mechanisms across the three superkingdoms of life [24,131,147–151]. Such signaling molecules are produced by dedicated signal-generating nucleotide synthetases, the Ter-biosynthetic system (see below), and

NAD⁺-processing enzymes and induce a diverse array of effectors that target the viral or host macromolecules (in a suicidal or dormancy-inducing response) [24,131,152]. The systems which are united under the umbrella of these signaling mechanisms include the eukaryotic and prokaryotic SMODs (e.g., interferon-induced oligoadenylate synthetase and cyclic guanylate-adenylate systems), the prokaryotic type-I and type-III CRISPR/Cas, the SLOG-TIR and alarmone-based immune systems [131]. There is increasing evidence, especially in eukaryotic viruses, that there are several viral counter-mechanisms that degrade the nucleotide/nucleotide-derived signals to block these immune processes [153,154]. The analysis of the polyvalent proteins injected by certain tailed phages along with their DNA suggested that such signal-targeting enzymes might be widespread even among bacterial viruses [100]. Here, we present the identification of several such proteins across jumbo phages that provide prototypes for understanding such counter-nucleotide defenses in the viral world. These further enmesh with various viral NAD⁺-centered systems.

We found at least six distinct families of phosphodiesterases (PDease) that we predict to be the mainstay of counter-nucleotide defenses in jumbo phages, which are likely directed at different types of nucleotides (Figure 5c). Across the three jumbo phage groups, there are HD-GYP enzymes (e.g., QDB70412.1) of the HD superfamily of phosphoesterases that target cyclic di- and oligo-nucleotides [155–158] and are likely to serve as a defense against the SMODS-based and type I/III CRISPR systems that use such nucleotides as signals. Another member of the HD superfamily, HD-alarmone PDease from group 2 and 3 jumbo phages (e.g., BBI90576.1), likely targets the alarmone [159], a nucleotide used as the signal by the stringent response system and the recently described related host immune systems [131,160]. PDEases of the DHH superfamily, which have been shown to act on diverse cyclic linkages, such as 3'-5', 2'-3' and cyclic di-nucleotides [161–163], are also found in a small number of group 2 and 3 phages (e.g., ALN97901.1) and possibly counter signals of SMODS and type I/III CRISPR systems [164]. A similar activity is predicted for the coliphage T5 A1 orthologs [165], which are phosphodiesterases of the calcineurin-like superfamily [48], featured by a set of group 2 and 3 phages. Members of the 2H superfamily of phosphoesterases found in group 2.2 and some group 3 jumbo phages (e.g., QDH83605.1) are specialists of 2'3' cyclic phosphates that are formed from the cleavage of tRNA and NAD⁺-dependent RNA processing (Figures 5c and 6a) [166,167]. These could signal the attack on the translation apparatus and potentially induce other defense systems. Phage enzymes targeting such cyclic phosphates could play an additional role in RNA repair (see below).

NAD⁺ is used as a substrate by enzymes of the ART, SIR2, ADP-ribosyl cyclase (ARC), TIR and SLOG superfamilies to produce a variety of ADPr derivatives, which are either conjugated to other macromolecules (ADP ribosylation), constitute soluble signals such as ADPr 1''2'' cyclic phosphate or cADPr with a cyclic diphosphate linkage or potential toxic metabolites such as ADPr 1''phosphate [24,131,168–174] (Figure 6a). Soluble ADPr-derived signals are central to the counter-phage defense systems that feature domains belonging to the above superfamilies. Such host-generated soluble ADPr-derivative signals (e.g., cADPr) and toxins (ADPr 1''P) are potentially degraded by the Macro (e.g., APU01542.1) and NADAR (e.g., QAY00367.1) enzymatic domains, which are displayed by several jumbo phages across all three groups and shared with unrelated families of eukaryotic RNA viruses [23,175,176]. A subset of these might also function in NAD⁺-dependent RNA repair (see below). Through the analysis of conserved gene-neighborhood or operonic linkages, we uncovered a divergent version of the Receiver (Rec) domain that is linked to different NAD⁺/ADPr-processing enzymes (Figure 6b–d). Unlike conventional Rec domains, this domain is never found in two-component Histidine Kinase signaling systems [177]. Instead, it is found only in association with NAD⁺-processing systems that generate or recognize cyclic nucleotides (e.g., 2'-3' cyclic ends of RNA and WYL domains that recognize cyclic nucleotides [178,179]; Figure 6b). Known Rec domains process phosphoester linkages by means of catalytic aspartate residues that are also conserved in these divergent Rec domains [180]. Accordingly, we posit that these are cADPr or 2'-3' cyclic nucleotide-processing enzymes and term them the cRec (cyclic-phosphate processing Rec) domains.

The linearized ADPr generated by the Macro, NADAR and cRec domains or by NADase action of host NAD⁺-targeting effectors (see below) are likely to be processed further in certain group 2 and 3 phages by the Nudix domain (e.g., QDH83582.1), which specifically cleaves nucleotide diphosphate-X linkages, to release AMP [181,182] (Figure 6a,c).

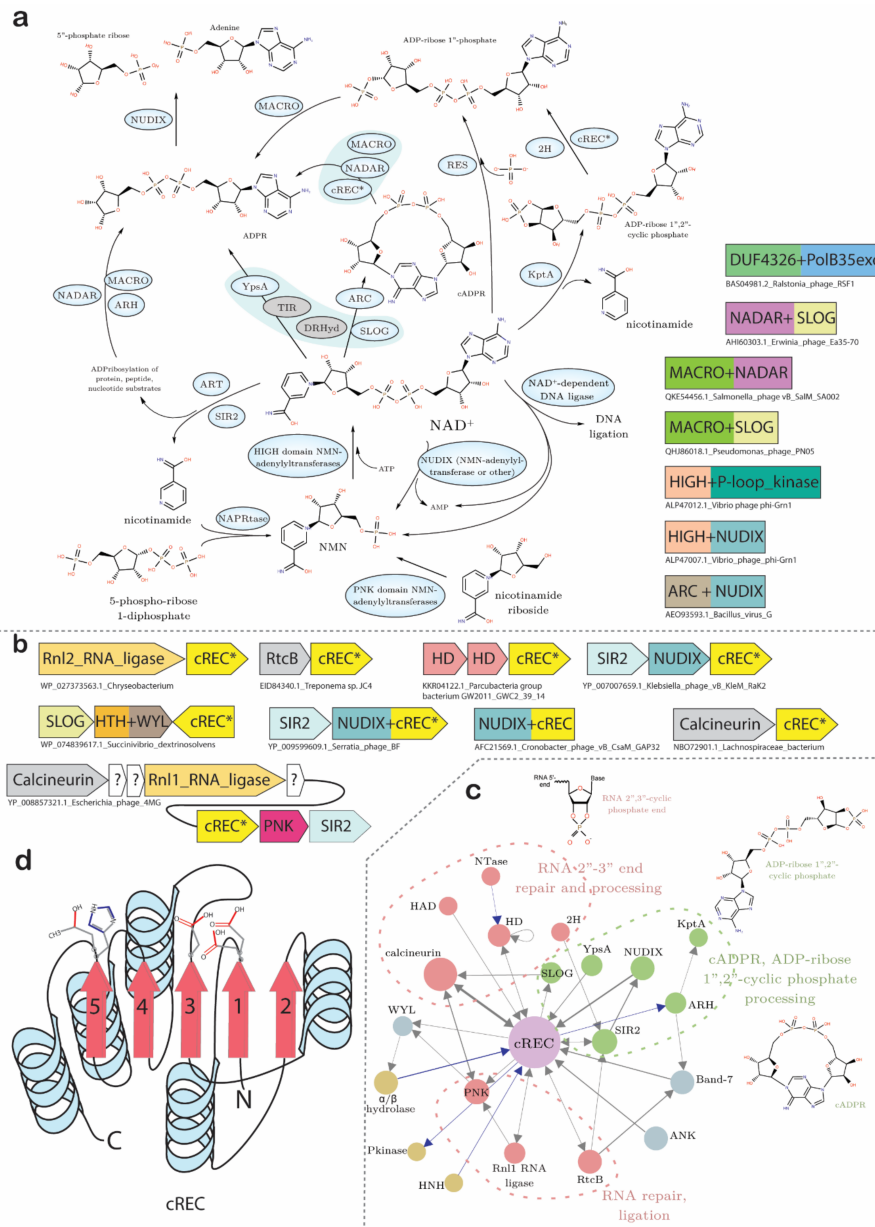


Figure 6. NAD⁺-centric conflict systems. **(a)** Biochemical action of various proteins involved in the NAD⁺-centric conflict systems and examples of interesting domain architectures of these proteins. **(b)** Gene neighborhoods of the cREC-containing systems and predicted topology of the cREC domain. Genes are represented as block arrows with arrow-heads pointing to the 3' ends of the gene. Neighborhoods and architectures are denoted by their accession numbers and phage names separated by underscores. The accession number in the label corresponds to the gene marked with a *. **(c)** Contextual network of the cREC domain derived from gene neighborhoods and domain architectures. Gene neighborhood associations are depicted with black lines with the arrowhead pointing to the 3' gene and domain associations with blue lines with the arrowhead pointing to the C-terminal domain. Distinct domains in architecture representations are colored differently. Nodes in the same functional category are colored similarly and grouped. **(d)** The topology of the cREC domain was predicted based on the know Receiver domains. The conserved residues were then superimposed onto the topology based on the cognate positions of conserved residues observed in classical Receiver domains.

On the phage side, too various systems such as replication (NAD⁺-dependent DNA ligases) and RNA repair [105,176,183] use NAD⁺ as an essential substrate. Moreover, ADPr conjugates to proteins catalyzed by the ARTs (see above) are also a key part of the phage control of the host machinery. Some jumbo phages possess SIR2 (e.g., QAY00582.1) and YAcAr family SLOG domains (e.g., QBP07259.1). We also detected novel members of the ARC superfamily in certain jumbo phages belonging to groups 2 and 3 (e.g., AEO93593.1). All these domains are known or predicted to utilize NAD⁺ [131,184] to produce ADPr-derivative signals, such as cADPr. These domains are frequently fused to MACRO, NADAR, or Nudix domains in the same polypeptide (Figure 6), which can degrade these signals, suggesting that they could act as toggles deployed by the jumbo phages to control host behavior using ADPr-derivative signals. Our contextual analysis of domain architectures also revealed the hitherto enigmatic domain DUF4326 (Figures 5c and 6a) found in several group 2 and 3 jumbo phages as well as amoeba-infecting NCLDVs to show several independent fusions to Macro, SLOG and NADAR domains on one hand and DNA replication or R-M components on the other. Accordingly, we predict it to be a potential enzymatic domain that bridges NAD⁺-utilizing signaling systems with DNA-modifications that interface with replication and R-M systems (AMB and LA, in preparation).

Host defense systems counter the viral requirement for NAD⁺ by the deployment of NADase effectors, such as the TIR, DrHyd and SIR2 domain proteins [24,131,168–173]. Given that NAD⁺ limitation can cripple all the above-mentioned systems, it is not surprising that several jumbo phages have been previously reported to carry their own NAD⁺-synthesis system [21]. The current analysis reveals that this system exists in certain representatives of each of the three major groups of jumbo phages (Figure 6a). The most expanded version found in phages such as *Vibrio* phage phi-Grn1 which has 3 genes: a nicotinamide phosphoribosyltransferase (NAPRTase; e.g., ALP46980.1) [185] and two distinct nicotinamide-nucleotide adenylyltransferase genes. The first enzyme synthesizes nicotinamide mononucleotide NMN from nicotinamide and 5-phospho-ribose 1-diphosphate. Both versions of the adenylyltransferases contain a HIGH superfamily NTase domain that adenylates NMN. However, they are differentiated by fusions of the HIGH NTase domain to either a Nudix domain (e.g., ALP47007.1), which can hydrolyze NAD⁺ to NMN [186], thereby regulating NAD⁺ levels or a P-loop kinase domain (e.g., ALP47012.1), which phosphorylates nicotinamide riboside to generate NMN [187]. Some jumbo phages contain more abbreviated versions of these systems with a single NAPRTase-NTase pair or just the former enzyme.

3.4.9. RNA Repair and RNA-Based Regulatory Systems

Bacteria possess several “second-line” immune mechanisms that are brought into play when the primary invader-targeting restriction mechanisms fail [24,131,188]. Typically, these restrict phage replication by targeting their own translation system components, such as tRNAs and ribosomal proteins. A widespread mode of attack is in situ cleavage of the tRNA anticodon loops at the ribosome to “jam” translation [189–191]. While these measures might negatively impact the host fitness, in the net they prove beneficial because they can limit the spread of the virus to kin cells (inclusive fitness) and the host could potentially tide over periods of dormancy while stopping the viral cycle. Starting with the work on the coliphage T4 [192,193], it became clear that there are several RNA repair/ribosome rescue processes by which phages reverse such attacks on the translation system [178]. The best-described of these is the presence of phage tRNAs in the genomes of disparate jumbo phages [9,194–196], which could substitute for the cleaved host tRNAs to allow translation of viral proteins to continue. We found that group 1, and subgroup 2.1 jumbo phages on an average possess only 4.5 and 7.1 tRNAs, respectively, per genome. However, subgroups 2.2 and 3.1 possess a mean number of 18 and 22 tRNA per genome, respectively (Figure 7a). This suggests that the reliance on self-encoded tRNAs widely differs between phages. More generally, we found RNA repair systems to be more common in group 2 and

3 than group 1 phages (Figure 7a). This difference presumably arises from the protection offered by their nucleus-like compartment.

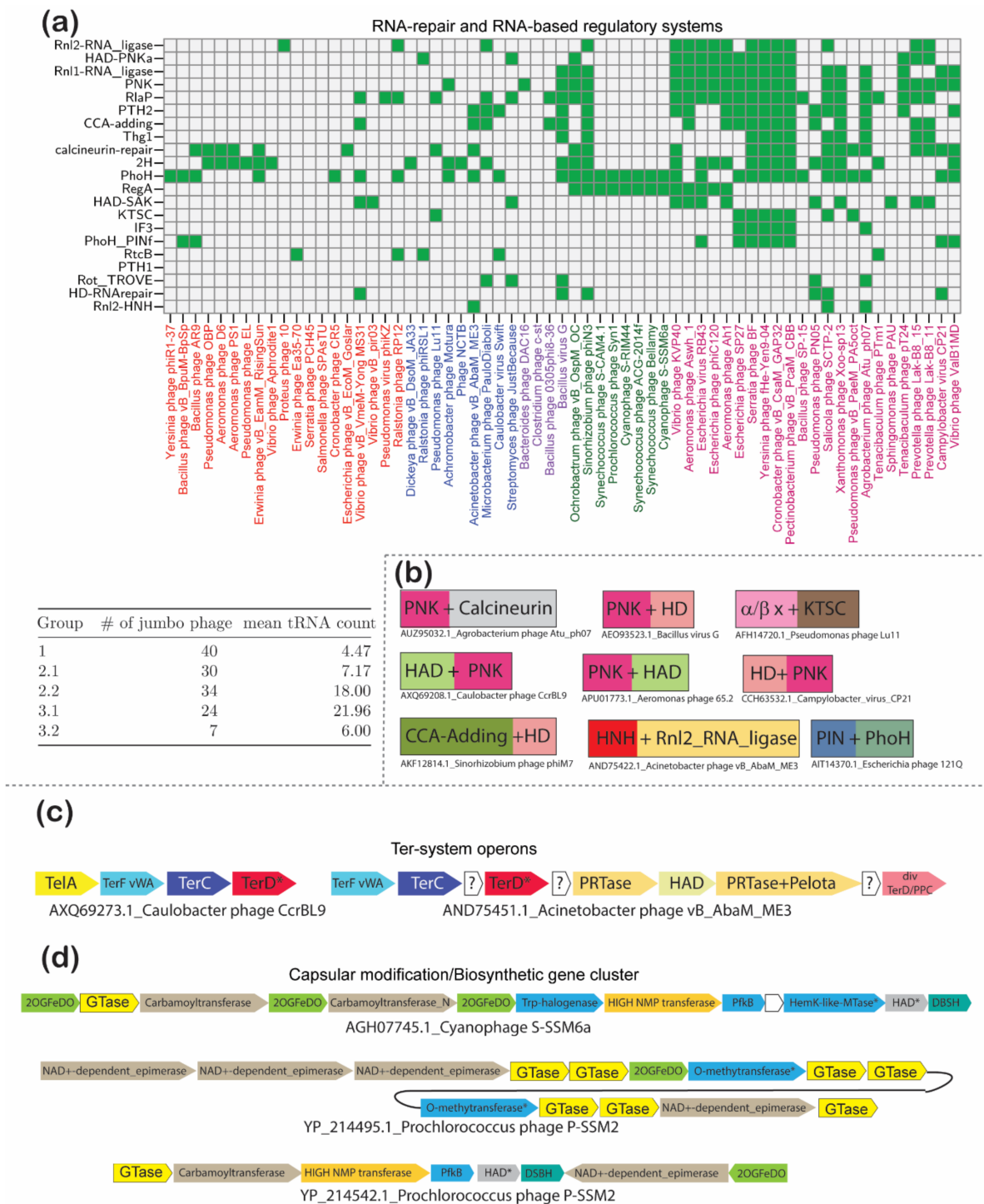


Figure 7. RNA-repair and RNA-based regulatory systems. **(a)** Phyletic vector diagrams and **(b)** domain architectures of jumbo proteins predicted to be involved in RNA-repair and RNA-based conflict and regulation. Also shown in **(a)** is the mean tRNA count for the different groups of jumbo phages. Also illustrated are the gene neighborhoods of **(c)** Ter-system and **(d)** capsular polysaccharide modification systems. GTase: glycosyltransferase. Genes are represented as boxed arrows with arrow-heads pointing to the 3' ends of the gene. Neighborhoods and architectures are denoted by their accession numbers and phage names separated by underscores. The accession number in the label corresponds to the gene marked with a *. In domain architectures, distinct domains are colored differently.

In all three jumbo phage groups, the tRNAs are supplemented by two classes of RNA ligases that ligate cleaved tRNAs, which belong to the RtcB-like and the ATP-grasp folds, respectively [178] (Figure 7a). The RtcB ligases are less common in jumbo phages and ligate RNAs both with 3' phosphate and 2'-3' cyclic phosphate ends, which are generated by the action of the metal-independent RNases [197,198] (Figure 7a). In contrast, the ATP-grasp ligases [199,200], which are more common in jumbo phages, require RNA ends with 5' phosphates and free 3' hydroxyls as substrates [201]. The majority of phage ATP-grasp ligases belong to one of two families, Rnl1 and Rnl2, with some rare divergent versions, like those from actinophages (e.g., AYD81305.1), which are outside of these families. Certain members of the Rnl2 family are fused to N-terminal HNH endonuclease domains [178] that might be used to launch attacks on superinfecting phage or host DNA even as the ligase repairs damaged RNA. A given phage might have any combination of RtcB, Rnl1 and Rnl2; e.g., the *Burkholderia* phages BcepSaruman and BcepSauron have all three of them, whereas *Xanthomonas* phage XacN1 has three paralogous versions of the Rnl2 family (Figure 7a, Supplementary Materials S1.8 and S5). Hence, these RNA ligases are probably not redundant and have specialized roles in repairing RNA cleaved by different kinds of RNase effectors. A small subset of group 3 and 2 phages additionally feature a ROT/TROVE RNA-binding protein, which, in cellular organisms constitutes a ribonucleoprotein platform for RNA repair along with a tRNA-like partner [178,202] (Figure 7a).

As several RNase effectors, especially the metal-independent RNases, generate RNA ends that are not directly usable by ATP-grasp ligases, they function along with polynucleotide kinases which phosphorylate the 5' ends, and phosphoesterases which hydrolyze 3' phosphates and make them amenable to ligation [178]. In jumbo phages, such enzymes are generally congruent with the presence of ATP-grasp RNA ligases (Figure 7a). The clearing of cyclic 2'-3' ends for ligation by ATP-grasp RNA ligases is likely performed in part by the above-mentioned 2H phosphoesterases that are specialist enzymes for such linkages [166,167]. Additionally, as noted above, the cRec phosphoesterase domain is likely to perform such a role in a subset of the jumbo phages. In support of this proposal, we found several operonic linkages of the cRec domain to the different families of RNA ligases in both phage and cellular genomes (Figure 6b). After the cyclic ends are resolved, the linear phosphates are hydrolyzed by linear phosphatases. In jumbo phages, these are typically members of the HAD superfamily and these belong to either of two families: PKN-HAD (e.g., QAX98520.1) usually fused to a N-terminal polynucleotide kinase domain and the so-called "Swiss-army-knife" (e.g., AUE22617.1) family which [178] tends to occur as a standalone domain (Figure 7b). In a small number of jumbo phages, this phosphatase activity appears to be also catalyzed by calcineurin-like and HD-like domains. In few group 1 and 3 jumbo phages, the 2' phosphates emerging from the action of metal-independent RNases might also be "cleaned up" by the action of the KptA family of ARTs (e.g., AUG85871.1) that transfer the phosphate to NAD⁺-generating a 1''-2'' cyclic phosphorylated ADPr [23,203], which is further processed by 2H and Macro domains [204] (Figure 7).

Other than ligating RNAs targeted by endoRNases, jumbo phages deploy one or more template-independent and template-dependent nucleotidyltransferases (NTases) to restore missing bases from RNA ends. The most prevalent of these is the recently described RlaP (RNA ligase-associating Pol β) NTase family that functions alongside the RNA ligases [178]. Its distribution in jumbo phages closely mirrors the RNA ligases supporting the functional cooperation between these enzymes (Figure 7a). RNase effectors deployed by host immune mechanisms, in addition to endonucleolytic action, might exonucleolytically cleave off a further nucleotide from the end (e.g., RloC RNase action on the wobble base after cleaving the anticodon loop [205]). It is predicted that the RlaP NTases restore such lost nucleotides before ligation [178]. The next most common is the CCA-adding enzyme (also of the Pol β superfamily) found in several group 2 and 3 jumbo phages, which restores the CCA trinucleotide or parts thereof in a template-independent manner at the 3' end of tRNAs [206]. In a similar vein, nucleotide restoration at the 5' ends (especially in the histidinyl tRNA) is catalyzed by the unusual 3'-5' polymerase Thg1 in

a template-dependent or independent manner [207]. This enzyme is only sporadically present in a few group 2 and 3 jumbo phages (Figure 7a).

Cloven tRNAs at the ribosome pose an additional challenge—they are linked to the incomplete polypeptides through a peptidyl tRNA linkage. Hence, hydrolysis of such peptidyl tRNA linkages is an important step in the recycling process to rescue jammed ribosomes [208]. Across group 2 and 3 jumbo phages, we found the peptidyl tRNA hydrolase (PTH2, e.g., APU01446.1) [195] which can potentially hydrolyze such linkages during ribosome rescue (Figure 7a). Interestingly, PTH2 is the characteristic peptidyl tRNA hydrolase of the archaeo-eukaryotic lineage [209] rather than that of host bacteria, which instead usually contain Pth1; thus, it could be the remnant of the translation apparatus of an ancient replicon that contributed to both archaeo-eukaryotic and phage genomes. In contrast, a single group 2 jumbo phage identified from metagenomic data with a bacterial-type Pth1 (AXH72886.1) [195]. The translation initiation factor IF3 gene (e.g., AUZ94813.1) and the lysyl tRNA synthetase KTSC RNA-binding domain protein (e.g., AZU98290.1) seen in few group 2.2 phages could also play a role in ribosome rescue or supporting phage translation (Figure 7a).

A further RNA-related function found across several group 2 and 3 phages is the PhoH protein, which contains an ATPase domain related to the N-terminal domain of SF1 helicases [210]. PhoH proteins have been shown to function as RNA helicases and at least some of these are sequence-specific. Both in jumbo phages and cellular genomes, they might be coupled to a N-terminal PIN endoRNase domain with the 5'→3' nuclease fold [87]. This combination has been shown to function in cellular genomes as a toxin-antitoxin system that potentially facilitates dormancy in response to certain environmental stresses [210,211]. Its role in the jumbo phages remains enigmatic; among other roles, these could help in the clearing of jammed ribosomes via its helicase action and endonucleolytic action of the associated PIN domains when present. Alternatively, it could function in post-transcriptionally regulating certain genes by operating on their RNAs. This latter function is related to the proposed role in post-transcription gene regulation for the orthologs of the coliphage T4 RegA proteins with a RRM fold RNA-binding domain [212] that are found across group 2.1 jumbo phages (Figure 7a). The presence of a conserved histidine in these proteins raises the possibility that RegA might function as a metal-independent RNase (Supplementary Materials S1.8 and S5).

3.4.10. Pseudolysogeny and Adaptations for Preventing Superinfection in Jumbo Phages

In a subset of jumbo phages, specialized tubulin/FtsZ-like proteins facilitate their chromosome segregation synchronously with host cell division during pseudolysogeny [213,214]. There is evidence that at least some of the jumbo phages without tubulin homologs (e.g., cyanophages) are also capable of pseudolysogeny during which they alter host-biochemistry, such as the photosynthetic apparatus. In addition to attacks from the host immune systems, jumbo phages also experience heightened conflict during pseudolysogeny with other phages and plasmids that might invade the cells they are residing in. Other than deploying the R-M systems [213,214] and CRISPR systems [213,214] encoded in the phage genomes, several jumbo phages have evolved at least two alternative strategies for these conflicts.

The first strategy involves the deployment of intracellular inhibitory factors that prevent the establishment of rival phages. We observed that several jumbo phages possess one or more proteins from the Ter system (Supplementary Materials S5), a functionally linked network of proteins found across bacteria that is involved in resistance to phages and certain xenobiotic substances [152]. The Ter system is also encoded by some plasmids which use it as a “fertility inhibitory factor”, i.e., preventing other plasmids from parasitically utilizing their transfer apparatus [215]. The Ter system includes several functionally disparate components: (1) biosynthesis of a nucleoside/nucleotide derivatives (nucleobase phosphoribosyltransferases: PRTases and HAD fold phosphoesterases), which evidently signal the detection of the incoming threat; (2) ligand-binding domains such as the TerB

and TerD proteins that might sense the signal; (3) TM components that might respond directly by changing membrane structure and transport (TerC); (4) components forming sub-membrane structures (e.g., TelA and the TerF vWA domain protein) which prevent entry of the phage or xenobiotic [152]. We found a relatively complete Ter system with signal biosynthesis PRTases and HAD enzymes in the group 2.2 jumbo phage, *Acinetobacter* phage vB_AbaM_ME3 (Figure 7c). More abbreviated Ter operons with TerD, TerC, TerF and TelA are seen in the group 3 *Caulobacter* jumbo phages. As standalone genes, TelA is also found in the gammaproteobacterial group 2.2 jumbo phages, TerB in group 1 *Erwinia* phages (e.g., ANZ49651) and TerD in actinobacterial group 3 phages. These observations suggest that the jumbo phages (and certain other related smaller tailed phages), might utilize the Ter system in their competition with superinfecting phages. Additionally, these genes might also confer a degree of xenobiotic resistance to the host, thereby allowing the phage to tide periods of pseudolysogeny.

The second mechanism is to modify the host biochemistry. This includes modification of molecules such as the capsules and peptidoglycan to prevent to counter attachment of rival phages [22]. Bacteria with Gram-negative cell walls are characterized by a complex glycolipid component of the outer membranes, the lipopolysaccharide (LPS), comprised of less-variable lipid A and oligosaccharide components and the highly variable outward-facing polysaccharide (O antigen) component. In contrast, Gram-positive bacteria have their own surface structures such as teichoic acid and teichuronopeptide shells. Systems modifying surface molecules have previously been found as part of the cellular defense against viruses [130] and have recently been reported to be coupled to intracellular counter-viral defenses to provide a two-pronged mode of immunity [24]. While these systems have been long recognized in lysogenic *Caudovirales*, the presence of such systems in jumbo phages first came to light with the report of the so-called LPS biosynthesis system in group 2.1 cyanophages [7]. Our analysis points out that such systems might more widespread and sometimes more elaborate than originally reported.

The *Synechococcus* phage P-SSM2, the prototypical group 2.1 jumbo phage with such systems, has a major 15-gene cluster. Such large clusters are primarily seen in cyanophages [7] (Figure 7d), with smaller complements of modifying enzymes in some of the other group 2 and 3 phages [22]. The cyanophages display complete complements of sugar-production and modification enzymes, whereas the other jumbo phages have more restricted sets of enzymes. The complete clusters contain 1–4 paralogous genes coding for a NAD⁺-dependent epimerase/dehydratase that are key enzymes in the production of modified sugars (Figure 7d). They may also feature the double-stranded β -helix (cupin) fold sugar isomerases [216]. The presence of PfkB family sugar kinases [217] in the cyanophages suggests that these catalyze the production of phospho-sugars (Figure 7d). These act as the substrates for the production of NDP-linked sugars by the HIGH-superfamily nucleotidyl-transferases that add NMP to phosphosugars (e.g., QAY00630.1, AGN12290.1) [218], which have a wider distribution beyond the cyanophages. The mainstay of these clusters is one or more paralogous glycosyltransferase (GTase) genes belonging to one of at least eight distinct GTase families [219] that synthesize oligosaccharides using NDP-sugars as substrates (Figure 7d). Their sporadic presence and polymorphisms seen in the cyanophage gene-clusters even between related viruses suggest that their products are primarily involved in the production of diversified outer polysaccharide complements (O antigen in the case of LPS) that might aid in preempting attachment of other phages.

We observed that cyanophages might feature a second gene cluster with 8–12 genes coding for sugar- and amino acid-modification enzymes. For instance, jumbo cyanophages like S-SSM6a contain a 10 gene cluster that is predicted to specify a biosynthetic system producing a halogenated amino acid and sugar-derived secondary metabolite (Figure 7d). The core of this system shared with actinobacterial secondary metabolite biosynthetic systems features carbamoyl transferases (e.g., AGH07737.1, AGH07739.1) related to those found in beta-lactam antibiotic synthesis and a HemK-like amino acid methylase (e.g., AGH07745.1). The cyanophage gene-cluster also codes for an amino acid halogenase (AGH07741.1), sev-

eral hydroxylating enzymes of the 2OGFeDO (e.g., AGH07740.1) and cupin-like DSBH (e.g., AGH07747.1) families and sugar-dephosphorylating HAD superfamily phosphatases [220]. Variations on this theme might include additional O-methyltransferases (e.g., YP_214498.1, YP_214495.1; Figure 7d). This system functions either in the further decoration of capsular polysaccharides or in producing a secondary metabolite and might serve as an antibiotic or toxin to improve host competitiveness during pseudolysogeny.

3.5. Evolutionary Considerations

Jumbo phages provide a window into understanding key evolutionary questions: (1) What was the form of the proteins performing key functions encoded by early replicons? (2) What were the evolutionary events that shaped the domain architectures of proteins associated with information transmission through the “central dogma” prior to the last common universal ancestor (LUCA) of cellular life? (3) How did the early replicons accrete genes to become larger replicons with increasing degrees of self-sufficiency?

Since the availability of the first viral genome sequences, it has been discussed as to whether they carry any features of early replicons beyond those that can be inferred from cellular life forms or if they represent different stages of the “degeneration” of cellular life [221]. The jumbo phages on the bacterial side and the NCLDV on the eukaryotic side have been at the center of this discussion as their large genomes straddle the size range between the smallest cellular genomes and medium-sized viral genomes [2,3]. In the past three decades, the availability of degenerate cellular genomes, such as *Mycoplasma genitalium*, the bacteroidetes *Sulcia muelleri* and the gammaproteobacterium *Baumannia cicadellinicola*, has emphatically indicated that they display gene complements and phylogenetic relationships quite unlike the phage proteins [222,223]. First, despite extreme degeneration, these bacteria retain sufficiently large complements of “core proteins” (i.e., those associated with central dogma information flow: replication-, transcription- and translation-related proteins). Second, these proteins show high sequence similarity with those of other bacteria, usually sufficient to precisely position them on the bacterial phylogenetic tree. Third, the conserved proteins show mostly congruent domain architectures to their bacterial counterparts.

In sharp contrast, jumbo phages do not show any tendency to conserve relatively complete complements of core systems. Instead, they possess a patchwork of such in so far as they help their successful replication. Thus, jumbo phages despite their size, might entirely lack an RNAP and usually do not code for a conserved aminoacyl tRNA synthetase complement, which is found even in the highly degenerate bacterial genomes. Further, the core systems that are present can be markedly divergent from all their host counterparts—the divergent DnaB and family B DNA polymerases typical of the group 1 jumbo phages are a case-in-point. This has also been demonstrated for the multisubunit RNA polymerases [16–18]. Likewise, the jumbo phage topoisomerases form distinct branches from the cellular versions and show a separation of the DNA-encircling and enzymatic components (Figure 3c). Several jumbo phages also possess the archaeo-eukaryotic version of the peptidyl tRNA hydrolase (PTH2) rather than the version commonly found in their hosts [224]. These observations suggest that the jumbo phages (and the related smaller phages), rather than being degenerations of cellular systems, have evolved from distinct pre-LUCA replicons. Hence, it is possible that they preserve some of the features of these replicons closer to the ancestral form, which were subsequently consolidated in somewhat distinct configurations in the conserved systems of cellular genomes.

Thus, they provide a snapshot of the diverse “experiments” that occurred in the central-dogma-related systems of the early replicons. For instance, it may be inferred that the double-barrel RNAPs were more fluid in the early replicons and comprised of multiple separate subunits, each performing a specific role, that came together to constitute the active RNAP [79,80]. In this regard, the minimal sigma factor and the gp33 TF of group 2 phages provide a possible picture of the ancestral state of the HTH TF that recruited the RNAP to specific promoters. Likewise, the minimal version of the DNA polymerase III module (see

below) provides a model of the ancestral versions of this replication enzyme. Thus, versions similar to these phage versions likely duplicated or fused with other domains to give rise to their more complex cellular cognates. Further, the two distinct representatives of the FtsZ/Tubulin family from jumbo phages also suggest that the diversity of functions offered by these self-organizing NTPase proteins (e.g., chromosome segregation and subcellular compartment formation) had already emerged early in evolution and were subsequently recapitulated in the eukaryotic centrosome and primary cilium [225,226].

The minimal version of the cellular DNA polymerase III catalytic module, which we uncovered in this study, maps to more or less just the core catalytic nucleotidyltransferase domain and template DNA strand binding domain (Figure 3c). It is further split into two standalone proteins that are likely to reconstitute complete catalytic enzymes. Notably, the jumbo phages with this version do not contain any of the other domains of DNA polymerase III, instead possess a family B DNA polymerase, like in the archaeo-eukaryotic lineage. Similarly, the C-terminal zinc ribbon of TFIIIE, a TF typical of the archaeo-eukaryotic lineage, occurs in jumbo phages that depend on the bacterial RNAP [90,91]. Thus, conserved components of the replication, transcription and translation apparatus that occur separately either in the bacterial or the archaeo-eukaryotic branch of cellular life are juxtaposed with components of the other branch in jumbo phage systems. This suggests that in the pre-LUCA period there was potentially a greater “mixing and matching” of the protein domains performing different central-dogma functions, some of which has survived in the extant phages. It also suggests that the replicons grew by accretion of cooperating genetic elements that provided different core components. Those that acquired different, more complete complements of components leading to self-sufficiency branched off as the two lineages of cellular life. Some of the rest adopted different degrees of parasitic existence and survived as the viral lineages.

This brings us to the final question of what drove the multiple independent increases in genome size resulting in the jumbo phages and whether it informs us in any way about the origins of cellular genomes? A popular proposal is the ratchet model wherein mutations resulting in bigger heads favor growth of genome size through acquisition of new genes. However, the random loss of the genes after they have been fixed would be selected against thereby ratcheting up genome size [139]. While this provides a simple baseline model, which has the advantage of being agnostic to specific genetic features, there are several indications that the underlying landscape of pre-adaptations might play a larger role than implied by this model. The systematic survey of 10,176 complete phage genomes indicates that whereas *Siphoviridae* are present at 2.7 times the frequency of *Myoviridae*, among jumbo phages *Myoviridae* are almost nine times more frequent than *Siphoviridae*. Further, there is no case so far of other relatively common *Caudovirales* groups, like *Podoviridae* or *Autographiviridae*, ever evolving to the jumbo state. This suggests that there is a distinct pre-adaptation among *Myoviridae*, perhaps relating to their DNA injection machinery, that predisposes them to successfully operate at larger genome sizes.

At least 6–7 independent emergences of jumbo phages can be inferred among *Caudovirales* that use the terminase-portal mechanism of DNA packaging [139,227,228]. Strikingly, to date, no prokaryotic virus with lipid membranes internal to capsid, which use the A32/FtsK/HerA-like packaging ATPases, i.e., tecti/cortico-like viruses, are known to have expanded to large sizes [2,229]. In contrast, in eukaryotes, the viruses that accreted on such a tecti-like core gave rise to a wide array of small viruses (e.g., adenoviruses and adomaviruses) and plasmids infecting the mitochondrial descendants of alphaproteobacteria on one hand and the large NCLDV with genomes reaching sizes comparable to the jumbo phages on the other [2]. Thus, their evolutionary trajectory mirrored *Caudovirales* in terms of occupying the entire spectrum of genome sizes. From among the eukaryotic viruses that accreted around a *Caudovirales*-like core arose the medium-sized herpesviruses that remained at genome sizes comparable to medium-sized precursors of the jumbo phages. These patterns suggest that in addition to the viral pre-adaptations, the host superkingdom has also been a determinant of which viruses expand in genome size.

Several jumbo phages carry several alternatives to countering host defenses and rival phages: for example, they might carry more than one type of RNA ligase and tRNAs or multiple enzymes for NAD⁺ synthesis or potentially alternative enzymes for capsular modification. This parallels the existence of alternative biosynthetic and repair pathways in cellular genomes, suggesting that there has been a certain selection for robustness via functional backups, over and beyond what exists in medium or smaller tailed phages. A parallel might also be struck with cellular genomes wherein within the same clade one might encounter relatively small and large genomes as sister groups—e.g., the gammaproteobacterium *Escherichia coli* has around 4350 genes while the related *Hemophilus influenzae* has only 1800 genes with extensive variation within each genus [230]. A part of this difference arises from the larger genomes coding for a much wider range of metabolic and immune functions. A part of the difference in genome size also arises from the accumulation of selfish genetic elements in the genome. This tendency appears in phage genomes that grow to around the coliphage T4 size [231] and increases as they expand to the size of jumbo phages. These are particularly visible in the form of self-splicing introns (e.g., in the DNA polymerase gene of the *Bacillus* phage G) or inteins (e.g., in the ribonucleotide reductase gene of *Clostridium* phage c-st) in essential phage genes similar to what has been reported for host genomes [232].

Thus, the differences in cellular genome sizes can be seen as occupying a spectrum similar to what in ecology has been described as the r-K selection spectrum—those with smaller genomes are selected by higher replication rates (r), whereas those with larger genomes replicate more slowly but have an entire array of features that make them stronger competitors [132,233]. Similarly, the caudoviral genomes may be seen as being selected along a similar spectrum with fast replication on one end and a stronger capacity to weather host immune attacks and metabolic restrictions on the other. This is borne out by the above observations that jumbo phages are enriched in a diverse array of strategies, both to counter host defenses and withstand physiological limitations arising from environmental conditions. In light of this, the repeated emergence of jumbo phages and their persistence across diverse bacterial groups might reflect an evolutionarily stable strategy in the face of hosts with well-developed immune mechanisms that face a diversity of environmental conditions.

4. Conclusions

Using 224 jumbo phages infecting diverse bacteria, including the recently discovered mega phages from *Prevotella*, we provide a comparative genomics overview along with in-depth sequence and structure analysis to provide a synthetic overview of the functionalities encoded by these viruses. Although these phages have evolved from distinct starting points within *Caudovirales*, as reflected in their DNA replication, transcription, and structural systems, they have converged to large genome sizes. This has also been paralleled by similar acquisitions and innovations with respect to metabolic and counter-immunity adaptations, several of which are described here for the first time. We hope that this analysis and the accompanying supporting material provide a framework for future investigations into these viruses.

Supplementary Materials: The following are available online at <https://www.mdpi.com/1999-4915/13/1/63/s1>, Supplementary S1.1–1.11: Phyletic vector diagrams for various jumbo phage proteins or protein domain families. Supplementary S2: Multiple Sequence Alignments of various protein domain families described in the text. Supplementary S3: Counts of tRNA genes per genome in representative jumbo phages. Supplementary S4: Relative gene positions of the 100 most frequent protein domains in a set of representative jumbo phage genomes. Supplementary S5: Sequence clusters of all proteins of the jumbo phage genomes analyzed in this study. Also provided is an excel file with complete annotation of all sequence clusters.

Author Contributions: Conceptualization, L.A.; methodology, V.A., L.M.I., L.A.; software, V.A., L.A.; validation, L.M.I., A.M.B., L.A.; formal analysis, L.M.I., V.A., A.K., A.M.B., L.A.; investigation, L.M.I., V.A., A.M.B., L.A.; resources, V.A., L.M.I.; data curation, L.M.I., V.A., A.K., A.M.B., L.A.; writing—original draft preparation, L.A.; writing—review and editing, L.M.I., A.M.B.; visualization, L.M.I., V.A., A.K., A.M.B.; supervision, L.A.; project administration, L.A.; funding acquisition, L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Iyer, L.M.; Aravind, L.; Koonin, E.V. Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* **2001**, *75*, 11720–11734. [[CrossRef](#)]
- Iyer, L.M.; Balaji, S.; Koonin, E.V.; Aravind, L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* **2006**, *117*, 156–184. [[CrossRef](#)] [[PubMed](#)]
- Koonin, E.V.; Yutin, N. Evolution of the large Nucleocytoplasmic DNA viruses of Eukaryotes and convergent origins of viral gigantism. *Adv. Virus Res.* **2019**, *103*, 167–202.
- Donelli, G.; Guglielmi, F.; Paoletti, L. Structure and physico-chemical properties of bacteriophage G. I. Arrangement of protein subunits and contraction process of tail sheath. *J. Mol. Biol.* **1972**, *71*, 113–125. [[CrossRef](#)]
- González, B.; Monroe, L.; Li, K.; Yan, R.; Wright, E.; Walter, T.; Kihara, D.; Weintraub, S.T.; Thomas, J.A.; Serwer, P.; et al. Phage G structure at 6.1AA resolution, condensed DNA, and host identity revision to a lysinibacillus. *J. Mol. Biol.* **2020**, *432*, 4139–4153. [[CrossRef](#)] [[PubMed](#)]
- Hatfull, G.F.; Hendrix, R.W. Bacteriophages and their genomes. *Curr. Opin. Virol.* **2011**, *1*, 298–303. [[CrossRef](#)]
- Yunker, I.T.; Duffy, C. Jumbo Phages. In *Reference Module in Life Sciences*; Elsevier: Amsterdam, The Netherlands, 2020.
- Yuan, Y.; Gao, M. Jumbo bacteriophages: An overview. *Front. Microbiol.* **2017**, *8*, 403. [[CrossRef](#)] [[PubMed](#)]
- Al-Shayeb, B.; Sachdeva, R.; Chen, L.-X.; Ward, F.; Munk, P.; Devoto, A.; Castelle, C.J.; Olm, M.R.; Bouma-Gregson, K.; Amano, Y.; et al. Clades of huge phages from across Earth's ecosystems. *Nature* **2020**, *578*, 425–431. [[CrossRef](#)]
- Devoto, A.E.; Santini, J.M.; Olm, M.R.; Anantharaman, K.; Munk, P.; Tung, J.; Archie, E.A.; Turnbaugh, P.J.; Seed, K.D.; Blekhman, R.; et al. Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat. Microbiol.* **2019**, *4*, 693–700. [[CrossRef](#)]
- Kawato, Y.; Istiqomah, I.; Gaafar, A.Y.; Hanaoka, M.; Ishimaru, K.; Yasuike, M.; Nishiki, I.; Nakamura, Y.; Fujiwara, A.; Nakai, T. A novel jumbo Tenacibaculum maritimum lytic phage with head-fiber-like appendages. *Arch. Virol.* **2020**, *165*, 303–311. [[CrossRef](#)]
- Ackermann, H.W.; Auclair, P.; Basavarajappa, S.; Konjin, H.P.; Savanurmah, C. Bacteriophages from Bombyx mori. *Arch. Virol.* **1994**, *137*, 185–190. [[CrossRef](#)] [[PubMed](#)]
- Buttimer, C.; Hendrix, H.; Oliveira, H.; Casey, A.; Neve, H.; McAuliffe, O.; Ross, R.P.; Hill, C.; Noben, J.P.; O'Mahony, J.; et al. Things are getting hairy: Enterobacteria bacteriophage vB_PcaM_CBB. *Front. Microbiol.* **2017**, *8*, 44. [[CrossRef](#)]
- Attai, H.; Boon, M.; Phillips, K.; Noben, J.P.; Lavigne, R.; Brown, P.J.B. Larger than life: Isolation and genomic characterization of a jumbo phage that infects the bacterial plant pathogen, agrobacterium tumefaciens. *Front. Microbiol.* **2018**, *9*, 1861. [[CrossRef](#)] [[PubMed](#)]
- Buttimer, C.; Born, Y.; Lucid, A.; Loessner, M.J.; Fieseler, L.; Coffey, A. Erwinia amylovora phage vB_EamM_Y3 represents another lineage of hairy Myoviridae. *Res. Microbiol.* **2018**, *169*, 505–514. [[CrossRef](#)] [[PubMed](#)]
- Lavysh, D.; Sokolova, M.; Minakhin, L.; Fvina, M.; Artamonova, T.; Kozyavkin, S.; Makarova, K.S.; Koonin, E.V.; Severinov, K. The genome of AR9, a giant transducing Bacillus phage encoding two multisubunit RNA polymerases. *Virology* **2016**, *495*, 185–196. [[CrossRef](#)]
- Sokolova, M.L.; Misovets, I.V.; Severinov, K. Multisubunit RNA polymerases of jumbo bacteriophages. *Viruses* **2020**, *12*, 1064. [[CrossRef](#)]
- Yakunina, M.; Artamonova, T.; Borukhov, S.; Makarova, K.S.; Severinov, K.; Minakhin, L. A non-canonical multisubunit RNA polymerase encoded by a giant bacteriophage. *Nucleic Acids Res.* **2015**, *43*, 10411–10420. [[CrossRef](#)]
- Malone, L.M.; Warring, S.L.; Jackson, S.A.; Warnecke, C.; Gardner, P.P.; Gumy, L.F.; Fineran, P.C. A jumbo phage that forms a nucleus-like structure evades CRISPR-Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nat. Microbiol.* **2020**, *5*, 48–55. [[CrossRef](#)]

20. Mendoza, S.D.; Nieweglowska, E.S.; Govindarajan, S.; Leon, L.M.; Berry, J.D.; Tiwari, A.; Chaikerasitak, V.; Pogliano, J.; Agard, D.A.; Bondy-Denomy, J. A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **2020**, *577*, 244–248. [[CrossRef](#)]
21. Lee, J.Y.; Li, Z.; Miller, E.S. Vibrio phage KVP40 encodes a functional NAD⁺ salvage pathway. *J. Bacteriol.* **2017**, *199*, 9. [[CrossRef](#)]
22. Bertani, B.; Ruiz, N. Function and biogenesis of lipopolysaccharides. *EcoSal Plus* **2018**, *8*, 1. [[CrossRef](#)] [[PubMed](#)]
23. Aravind, L.; Zhang, D.; de Souza, R.F.; Anand, S.; Iyer, L.M. The natural history of ADP-ribosyltransferases and the ADP-ribosylation system. *Curr. Top. Microbiol. Immunol.* **2015**, *384*, 3–32. [[PubMed](#)]
24. Burroughs, A.M.; Aravind, L. Identification of uncharacterized components of prokaryotic immune systems and their diverse eukaryotic reformulations. *J. Bacteriol.* **2020**, *2020*. [[CrossRef](#)] [[PubMed](#)]
25. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [[CrossRef](#)] [[PubMed](#)]
26. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **2009**, *23*, 205–211.
27. Hyatt, D.; LoCasio, P.F.; Hauser, L.J.; Uberbacher, E.C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **2012**, *28*, 2223–2230. [[CrossRef](#)]
28. Soding, J.; Biegert, A.; Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005**, *33*, W244–W248. [[CrossRef](#)]
29. Lassmann, T.; Frings, O.; Sonnhammer, E.L. Kalign2: High-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res.* **2009**, *37*, 858–865. [[CrossRef](#)]
30. Edgar, R.C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **2004**, *5*, 113. [[CrossRef](#)]
31. Cole, C.; Barber, J.D.; Barton, G.J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **2008**, *36*, W197–W201. [[CrossRef](#)]
32. Holm, L.; Kaariainen, S.; Rosenstrom, P.; Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **2008**, *24*, 2780–2781. [[CrossRef](#)] [[PubMed](#)]
33. Schrodinger LLC. *The PyMOL Molecular Graphics System*, version 1.8; Schrodinger: New York, NY, USA, 2015.
34. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **2010**, *5*, e9490. [[CrossRef](#)] [[PubMed](#)]
35. Lance, G.N.; Williams, W.T. Computer programs for hierarchical polythetic classification (“similarity analyses”). *Comput. J.* **1966**, *9*, 60–64. [[CrossRef](#)]
36. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: New York, NY, USA, 1990.
37. Asare, P.T.; Jeong, T.Y.; Ryu, S.; Klumpp, J.; Loessner, M.J.; Merrill, B.D.; Kim, K.P. Putative type 1 thymidylate synthase and dihydrofolate reductase as signature genes of a novel Bastille-like group of phages in the subfamily Spounavirinae. *BMC Genom.* **2015**, *16*, 582. [[CrossRef](#)] [[PubMed](#)]
38. Weigele, P.; Raleigh, E.A. Biosynthesis and function of modified bases in bacteria and their viruses. *Chem. Rev.* **2016**, *116*, 12655–12687. [[CrossRef](#)]
39. Leduc, D.; Graziani, S.; Meslet-Cladiere, L.; Sodolescu, A.; Liebl, U.; Myllykallio, H. Two distinct pathways for thymidylate (dTMP) synthesis in (hyper)thermophilic Bacteria and Archaea. *Biochem. Soc. Trans.* **2004**, *32*, 231–235. [[CrossRef](#)]
40. Duda, R.L.; Martincic, K.; Hendrix, R.W. Genetic basis of bacteriophage HK97 prohead assembly. *J. Mol. Biol.* **1995**, *247*, 636–647. [[CrossRef](#)]
41. Chen, P.; Tsuge, H.; Almassy, R.J.; Gribskov, C.L.; Katoh, S.; Vanderpool, D.L.; Margosiak, S.A.; Pinko, C.; Matthews, D.A.; Kan, C.C. Structure of the human cytomegalovirus protease catalytic domain reveals a novel serine protease fold and catalytic triad. *Cell* **1996**, *86*, 835–843. [[CrossRef](#)]
42. Miller, E.S.; Kutter, E.; Mosig, G.; Arisaka, F.; Kunisawa, T.; Ruger, W. Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **2003**, *67*, 86–156. [[CrossRef](#)]
43. Fokine, A.; Rossmann, M.G. Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage* **2014**, *4*, e28281. [[CrossRef](#)]
44. Los, M.; Wegrzyn, G. Pseudolysogeny. *Adv. Virus Res.* **2012**, *82*, 339–349. [[PubMed](#)]
45. Iyer, L.M.; Abhiman, S.; Aravind, L. A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases. *Biol. Direct* **2008**, *3*, 39. [[CrossRef](#)] [[PubMed](#)]
46. Delarue, M.; Poch, O.; Tordo, N.; Moras, D.; Argos, P. An attempt to unify the structure of polymerases. *Protein Eng.* **1990**, *3*, 461–467. [[CrossRef](#)] [[PubMed](#)]
47. Aravind, L.; Mazumder, R.; Vasudevan, S.; Koonin, E.V. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **2002**, *12*, 392–399. [[CrossRef](#)]
48. Aravind, L.; Koonin, E.V. Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucleic Acids Res.* **1998**, *26*, 3746–3752. [[CrossRef](#)]
49. Lamers, M.H.; Georgescu, R.E.; Lee, S.G.; O’Donnell, M.; Kuriyan, J. Crystal structure of the catalytic alpha subunit of E. coli replicative DNA polymerase III. *Cell* **2006**, *126*, 881–892. [[CrossRef](#)]

50. Ahn, D.H.; Lee, K.Y.; Lee, S.J.; Park, S.J.; Yoon, H.J.; Kim, S.J.; Lee, B.J. Structural analyses of the MazEF4 toxin-antitoxin pair in *Mycobacterium tuberculosis* provide evidence for a unique extracellular death factor. *J. Biol. Chem.* **2017**, *292*, 18832–18847. [[CrossRef](#)]
51. Leipe, D.D.; Aravind, L.; Grishin, N.V.; Koonin, E.V. The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome Res.* **2000**, *10*, 5–16.
52. Dudas, K.C.; Kreuzer, K.N. Bacteriophage T4 helicase loader protein gp59 functions as gatekeeper in origin-dependent replication in vivo. *J. Biol. Chem.* **2005**, *280*, 21561–21569. [[CrossRef](#)]
53. Bleuit, J.S.; Xu, H.; Ma, Y.; Wang, T.; Liu, J.; Morrical, S.W. Mediator proteins orchestrate enzyme-ssDNA assembly during T4 recombination-dependent DNA replication and repair. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 8298–8305. [[CrossRef](#)]
54. Oakley, A.J. Dynamics of open DNA sliding clamps. *PLoS ONE* **2016**, *11*, e0154899. [[CrossRef](#)] [[PubMed](#)]
55. Shi, K.; Bohl, T.E.; Park, J.; Zasada, A.; Malik, S.; Banerjee, S.; Tran, V.; Li, N.; Yin, Z.; Kurniawan, F.; et al. T4 DNA ligase structure reveals a prototypical ATP-dependent ligase with a unique mode of sliding clamp interaction. *Nucleic Acids Res.* **2018**, *46*, 10474–10488. [[CrossRef](#)] [[PubMed](#)]
56. Aravind, L.; Leipe, D.D.; Koonin, E.V. Toprim—A conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res.* **1998**, *26*, 4205–4213. [[CrossRef](#)] [[PubMed](#)]
57. Iyer, L.M.; Koonin, E.V.; Leipe, D.D.; Aravind, L. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: Structural insights and new members. *Nucleic Acids Res.* **2005**, *33*, 3875–3896. [[CrossRef](#)]
58. Lipps, G.; Weinzierl, A.O.; von Scheven, G.; Buchen, C.; Cramer, P. Structure of a bifunctional DNA primase-polymerase. *Nat. Struct. Mol. Biol.* **2004**, *11*, 157–162. [[CrossRef](#)]
59. Rudd, S.G.; Bianchi, J.; Doherty, A.J. PrimPol—A new polymerase on the block. *Mol. Cell. Oncol.* **2014**, *1*, e960754. [[CrossRef](#)]
60. Senkevich, T.G.; Koonin, E.V.; Moss, B. Predicted poxvirus FEN1-like nuclease required for homologous recombination, double-strand break repair and full-size genome formation. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 17921–17926. [[CrossRef](#)]
61. Barry, J.; Wong, M.L.; Alberts, B. In vitro reconstitution of DNA replication initiated by genetic recombination: A T4 bacteriophage model for a type of DNA synthesis important for all cells. *Mol. Biol. Cell* **2019**, *30*, 146–159. [[CrossRef](#)]
62. De Souza, R.F.; Iyer, L.M.; Aravind, L. Diversity and evolution of chromatin proteins encoded by DNA viruses. *Biochim. Biophys. Acta* **2010**, *1799*, 302–318. [[CrossRef](#)]
63. He, X.; Byrd, A.K.; Yun, M.K.; Pemble, C.W.T.; Harrison, D.; Yeruva, L.; Dahl, C.; Kreuzer, K.N.; Raney, K.D.; White, S.W. The T4 phage SF1B helicase Dda is structurally optimized to perform DNA strand separation. *Structure* **2012**, *20*, 1189–1200. [[CrossRef](#)]
64. Mortier-Barriere, I.; Velten, M.; Dupaigne, P.; Mirouze, N.; Pietrement, O.; McGovern, S.; Fichant, G.; Martin, B.; Noirot, P.; Le Cam, E.; et al. A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA. *Cell* **2007**, *130*, 824–836. [[CrossRef](#)] [[PubMed](#)]
65. Chang, H.H.Y.; Pannunzio, N.R.; Adachi, N.; Lieber, M.R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 495–506. [[CrossRef](#)] [[PubMed](#)]
66. Rostol, J.T.; Marraffini, L. (Ph)ighting phages: How bacteria resist their parasites. *Cell Host Microbe* **2019**, *25*, 184–194. [[CrossRef](#)] [[PubMed](#)]
67. Garcia, A.D.; Aravind, L.; Koonin, E.V.; Moss, B. Bacterial-type DNA holliday junction resolvases in eukaryotic viruses. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8926–8931. [[CrossRef](#)] [[PubMed](#)]
68. Aravind, L.; Makarova, K.S.; Koonin, E.V. Suervey and summary: Holliday junction resolvases and related nucleases: Identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.* **2000**, *28*, 3417–3432. [[CrossRef](#)]
69. Biertumpfel, C.; Yang, W.; Suck, D. Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature* **2007**, *449*, 616–620. [[CrossRef](#)] [[PubMed](#)]
70. Aravind, L.; Koonin, E.V. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res.* **2001**, *11*, 1365–1374. [[CrossRef](#)]
71. Fricke, W.M.; Brill, S.J. Slx1-Slx4 is a second structure-specific endonuclease functionally redundant with Sgs1-Top3. *Genes Dev.* **2003**, *17*, 1768–1778. [[CrossRef](#)]
72. Hoogenboom, W.S.; Boonen, R.; Knipscheer, P. The role of SLX4 and its associated nucleases in DNA interstrand crosslink repair. *Nucleic Acids Res.* **2019**, *47*, 2377–2388. [[CrossRef](#)]
73. Iyer, L.M.; Koonin, E.V.; Aravind, L. Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. *BMC Genom.* **2002**, *3*, 8. [[CrossRef](#)]
74. Schoeffler, A.J.; Berger, J.M. DNA topoisomerases: Harnessing and constraining energy to govern chromosome topology. *Q. Rev. Biophys.* **2008**, *41*, 41–101. [[CrossRef](#)] [[PubMed](#)]
75. Champoux, J.J. DNA topoisomerases: Structure, function, and mechanism. *Annu. Rev. Biochem.* **2001**, *70*, 369–413. [[CrossRef](#)]
76. Piersen, C.E.; McCullough, A.K.; Lloyd, R.S. AP lyases and dRPases: Commonality of mechanism. *Mutat. Res.* **2000**, *459*, 43–53. [[CrossRef](#)]
77. Paspaleva, K.; Thomassen, E.; Pannu, N.S.; Iwai, S.; Moolenaar, G.F.; Goosen, N.; Abrahams, J.P. Crystal structure of the DNA repair enzyme ultraviolet damage endonuclease. *Structure* **2007**, *15*, 1316–1324. [[CrossRef](#)] [[PubMed](#)]
78. Murakami, K.S.; Davydova, E.K.; Rothman-Denes, L.B. X-ray crystal structure of the polymerase domain of the bacteriophage N4 virion RNA polymerase. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5046–5051. [[CrossRef](#)]

79. Iyer, L.M.; Koonin, E.V.; 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**, *3*, 1. [[CrossRef](#)] [[PubMed](#)]
80. Iyer, L.M.; Aravind, L. Insights from the architecture of the bacterial transcription apparatus. *J. Struct. Biol.* **2012**, *179*, 299–319. [[CrossRef](#)]
81. Sauguet, L.; Raia, P.; Henneke, G.; Delarue, M. Shared active site architecture between archaeal PolD and multi-subunit RNA polymerases revealed by X-ray crystallography. *Nat. Commun.* **2016**, *7*, 12227. [[CrossRef](#)]
82. Thomas, J.A.; Benítez Quintana, A.D.; Bosch, M.A.; Coll De Peña, A.; Aguilera, E.; Coulibaly, A.; Wu, W.; Osier, M.V.; Hudson, A.O.; Weintraub, S.T.; et al. Identification of essential genes in the Salmonella phage SPN3US reveals novel insights into giant phage head structure and assembly. *J. Virol.* **2016**, *90*, 10284–10298. [[CrossRef](#)]
83. Shaw, G.; Gan, J.; Zhou, Y.N.; Zhi, H.; Subburaman, P.; Zhang, R.; Joachimiak, A.; Jin, D.J.; Ji, X. Structure of RapA, a Swi2/Snf2 protein that recycles RNA polymerase during transcription. *Structure* **2008**, *16*, 1417–1427. [[CrossRef](#)]
84. Twist, K.A.; Campbell, E.A.; Deighan, P.; Nechaev, S.; Jain, V.; Geiduschek, E.P.; Hochschild, A.; Darst, S.A. Crystal structure of the bacteriophage T4 late-transcription coactivator gp33 with the beta-subunit flap domain of Escherichia coli RNA polymerase. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 19961–19966. [[CrossRef](#)] [[PubMed](#)]
85. Hinton, D.M. Transcriptional control in the prereplicative phase of T4 development. *Virol. J.* **2010**, *7*, 289. [[CrossRef](#)] [[PubMed](#)]
86. Ptashne, M. Regulation of transcription: From lambda to eukaryotes. *Trends Biochem. Sci.* **2005**, *30*, 275–279. [[CrossRef](#)] [[PubMed](#)]
87. Anantharaman, V.; Aravind, L. New connections in the prokaryotic toxin-antitoxin network: Relationship with the eukaryotic nonsense-mediated RNA decay system. *Genome Biol.* **2003**, *4*, R81.
88. Lee, J.H.; Wendt, J.C.; Shanmugam, K.T. Identification of a new gene, molR, essential for utilization of molybdate by *Escherichia coli*. *J. Bacteriol.* **1990**, *172*, 2079–2087. [[CrossRef](#)]
89. Wang, S.T.; Setlow, B.; Conlon, E.M.; Lyon, J.L.; Imamura, D.; Sato, T.; Setlow, P.; Losick, R.; Eichenberger, P. The forespore line of gene expression in *Bacillus subtilis*. *J. Mol. Biol.* **2006**, *358*, 16–37. [[CrossRef](#)]
90. Aravind, L.; Koonin, E.V. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* **1999**, *27*, 4658–4670. [[CrossRef](#)]
91. Plaschka, C.; Hantsche, M.; Dienemann, C.; Burzinski, C.; Plitzko, J.; Cramer, P. Transcription initiation complex structures elucidate DNA opening. *Nature* **2016**, *533*, 353–358. [[CrossRef](#)]
92. Young, K.K.; Edlin, G.J.; Wilson, G.G. Genetic analysis of bacteriophage T4 transducing bacteriophages. *J. Virol.* **1982**, *41*, 345–347. [[CrossRef](#)]
93. Paddison, P.; Abedon, S.T.; Dressman, H.K.; Gailbreath, K.; Tracy, J.; Mosser, E.; Neitzel, J.; Guttman, B.; Kutter, E. The roles of the bacteriophage T4 r genes in lysis inhibition and fine-structure genetics: A new perspective. *Genetics* **1998**, *148*, 1539–1550.
94. Johnson, A.; Meyer, B.J.; Ptashne, M. Mechanism of action of the cro protein of bacteriophage lambda. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 1783–1787. [[CrossRef](#)]
95. Aravind, L.; Anantharaman, V.; Balaji, S.; Babu, M.M.; Iyer, L.M. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiol. Rev.* **2005**, *29*, 231–262. [[CrossRef](#)] [[PubMed](#)]
96. Sieber, P.; Lindemann, A.; Boehm, M.; Seidel, G.; Herzing, U.; van der Heusen, P.; Müller, R.; Rüger, W.; Jaenicke, R.; Rösch, P. Overexpression and structural characterization of the phage T4 protein DsbA. *Biol. Chem.* **1998**, *379*, 51–58. [[CrossRef](#)] [[PubMed](#)]
97. Tarry, M.J.; Harmel, C.; Taylor, J.A.; Marczyński, G.T.; Schmeing, T.M. Structures of GapR reveal a central channel which could accommodate B-DNA. *Sci. Rep.* **2019**, *9*, 16679. [[CrossRef](#)]
98. Depping, R.; Lohaus, C.; Meyer, H.E.; Ruger, W. The mono-ADP-ribosyltransferases Alt and ModB of bacteriophage T4: Target proteins identified. *Biochem. Biophys. Res. Commun.* **2005**, *335*, 1217–1223. [[CrossRef](#)] [[PubMed](#)]
99. Ceysens, P.J.; De Smet, J.; Wagemans, J.; Akulenko, N.; Klimuk, E.; Hedge, S.; Voet, M.; Hendrix, H.; Paeshuyse, J.; Landuyt, B.; et al. The phage-encoded N-Acetyltransferase Rac mediates inactivation of *Pseudomonas aeruginosa* transcription by cleavage of the RNA polymerase alpha subunit. *Viruses* **2020**, *12*, 976. [[CrossRef](#)] [[PubMed](#)]
100. Iyer, L.M.; Burroughs, A.M.; Anand, S.; de Souza, R.F.; Aravind, L. Polyvalent proteins, a pervasive theme in the intergenomic biological conflicts of bacteriophages and conjugative elements. *J. Bacteriol.* **2017**, *199*, 15. [[CrossRef](#)]
101. Gilmore, J.M.; Bieber Urbauer, R.J.; Minakhin, L.; Akoyev, V.; Zolkiewski, M.; Severinov, K.; Urbauer, J.L. Determinants of affinity and activity of the anti-sigma factor AsiA. *Biochemistry* **2010**, *49*, 6143–6154. [[CrossRef](#)]
102. Kashlev, M.; Nudler, E.; Goldfarb, A.; White, T.; Kutter, E. Bacteriophage T4 Alc protein: A transcription termination factor sensing local modification of DNA. *Cell* **1993**, *75*, 147–154. [[CrossRef](#)]
103. Favrot, L.; Blanchard, J.S.; Vergnolle, O. Bacterial GCN5-Related N-Acetyltransferases: From resistance to regulation. *Biochemistry* **2016**, *55*, 989–1002. [[CrossRef](#)]
104. Burroughs, A.M.; Zhang, D.; Aravind, L. The eukaryotic translation initiation regulator CDC123 defines a divergent clade of ATP-grasp enzymes with a predicted role in novel protein modifications. *Biol. Direct* **2015**, *10*, 21. [[CrossRef](#)] [[PubMed](#)]
105. Zhang, D.; de Souza, R.F.; Anantharaman, V.; Iyer, L.M.; Aravind, L. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol. Direct* **2012**, *7*, 18. [[CrossRef](#)] [[PubMed](#)]
106. Kumari, P.; Kumar, H. Viral deubiquitinases: Role in evasion of anti-viral innate immunity. *Crit. Rev. Microbiol.* **2018**, *44*, 304–317. [[CrossRef](#)] [[PubMed](#)]

107. Lindner, H.A. Deubiquitination in virus infection. *Virology* **2007**, *362*, 245–256. [[CrossRef](#)] [[PubMed](#)]
108. Snyder, L. Phage-exclusion enzymes: A bonanza of biochemical and cell biology reagents? *Mol. Microbiol.* **1995**, *15*, 415–420. [[CrossRef](#)]
109. Dougan, D.A.; Micevski, D.; Truscott, K.N. The N-end rule pathway: From recognition by N-recognins, to destruction by AAA+proteases. *Biochim. Biophys. Acta* **2012**, *1823*, 83–91. [[CrossRef](#)]
110. Burroughs, A.M.; Iyer, L.M.; Aravind, L. Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems. *Genome Dyn.* **2007**, *3*, 48–65.
111. Benler, S.; Hung, S.-H.; Vander Griend, J.A.; Peters, G.A.; Rohwer, F.; Segall, A.M. Gp4 is a nuclease required for morphogenesis of T4-like bacteriophages. *Virology* **2020**, *543*, 7–12. [[CrossRef](#)]
112. Sun, L.; Zhang, X.; Gao, S.; Rao, P.A.; Padilla-Sanchez, V.; Chen, Z.; Sun, S.; Xiang, Y.; Subramaniam, S.; Rao, V.B.; et al. Cryo-EM structure of the bacteriophage T4 portal protein assembly at near-atomic resolution. *Nat. Commun.* **2015**, *6*, 7548. [[CrossRef](#)]
113. Thomas, J.A.; Weintraub, S.T.; Wu, W.; Winkler, D.C.; Cheng, N.; Steven, A.C.; Black, L.W. Extensive proteolysis of head and inner body proteins by a morphogenetic protease in the giant *Pseudomonas aeruginosa* phage phiKZ. *Mol. Microbiol.* **2012**, *84*, 324–339. [[CrossRef](#)]
114. Schwarzer, D.; Stummeyer, K.; Gerardy-Schahn, R.; Muhlenhoff, M. Characterization of a novel intramolecular chaperone domain conserved in endosialidases and other bacteriophage tail spike and fiber proteins. *J. Biol. Chem.* **2007**, *282*, 2821–2831. [[CrossRef](#)] [[PubMed](#)]
115. Sullivan, M.B.; Huang, K.H.; Ignacio-Espinoza, J.C.; Berlin, A.M.; Kelly, L.; Weigele, P.R.; DeFrancesco, A.S.; Kern, S.E.; Thompson, L.R.; Young, S.; et al. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **2010**, *12*, 3035–3056. [[CrossRef](#)]
116. Scheele, U.; Erdmann, S.; Ungewickell, E.J.; Felisberto-Rodrigues, C.; Ortiz-Lombardia, M.; Garrett, R.A. Chaperone role for proteins p618 and p892 in the extracellular tail development of Acidianus two-tailed virus. *J. Virol.* **2011**, *85*, 4812–4821. [[CrossRef](#)] [[PubMed](#)]
117. Marusich, E.I.; Kurochkina, L.P.; Mesyanzhinov, V.V. Chaperones in bacteriophage T4 assembly. *Biochemistry* **1998**, *63*, 399–406. [[PubMed](#)]
118. Michaud, G.; Zachary, A.; Rao, V.B.; Black, L.W. Membrane-associated assembly of a phage T4 DNA entrance vertex structure studied with expression vectors. *J. Mol. Biol.* **1989**, *209*, 667–681. [[CrossRef](#)]
119. Tang, W.K.; Borgnia, M.J.; Hsu, A.L.; Esser, L.; Fox, T.; de Val, N.; Xia, D. Structures of AAA protein translocase Bcs1 suggest translocation mechanism of a folded protein. *Nat. Struct. Mol. Biol.* **2020**, *27*, 202–209. [[CrossRef](#)]
120. Medhekar, B.; Miller, J.F. Diversity-generating retroelements. *Curr. Opin. Microbiol.* **2007**, *10*, 388–395. [[CrossRef](#)]
121. Day, A.; Ahn, J.; Salmond, G.P.C. Jumbo bacteriophages are represented within an increasing diversity of environmental viruses infecting the emerging phytopathogen, *Dickeya solani*. *Front. Microbiol.* **2018**, *9*, 2169. [[CrossRef](#)]
122. Beckmann, G.; Hanke, J.; Bork, P.; Reich, J.G. Merging extracellular domains: Fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins. *J. Mol. Biol.* **1998**, *275*, 725–730. [[CrossRef](#)]
123. Williams, F.P.; Haubrich, K.; Perez-Borrajerro, C.; Hennig, J. Emerging RNA-binding roles in the TRIM family of ubiquitin ligases. *Biol. Chem.* **2019**, *400*, 1443–1464. [[CrossRef](#)]
124. Bhardwaj, A.; Molineux, I.J.; Casjens, S.R.; Cingolani, G. Atomic structure of bacteriophage Sf6 tail needle knob. *J. Biol. Chem.* **2011**, *286*, 30867–30877. [[CrossRef](#)] [[PubMed](#)]
125. Anantharaman, V.; Aravind, L. Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biol.* **2003**, *4*, R11.
126. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [[CrossRef](#)] [[PubMed](#)]
127. Shneider, M.M.; Buth, S.A.; Ho, B.T.; Basler, M.; Mekalanos, J.J.; Leiman, P.G. PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature* **2013**, *500*, 350–353. [[CrossRef](#)]
128. Aravind, L.; Anantharaman, V.; Zhang, D.; de Souza, R.F.; Iyer, L.M. Gene flow and biological conflict systems in the origin and evolution of eukaryotes. *Front. Cell. Infect. Microbiol.* **2012**, *2*, 89. [[CrossRef](#)]
129. Lavender, P.; Kelly, A.; Hendy, E.; McErlean, P. CRISPR-based reagents to study the influence of the epigenome on gene expression. *Clin. Exp. Immunol.* **2018**, *194*, 9–16. [[CrossRef](#)]
130. Seed, K.D. Battling phages: How bacteria defend against viral attack. *PLoS Pathog.* **2015**, *11*, e1004847. [[CrossRef](#)]
131. Burroughs, A.M.; Zhang, D.; Schaffer, D.E.; Iyer, L.M.; Aravind, L. Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids Res.* **2015**, *43*, 10633–10654. [[CrossRef](#)]
132. Kaur, G.; Burroughs, A.M.; Iyer, L.M.; Aravind, L. Highly regulated, diversifying NTP-dependent biological conflict systems with implications for the emergence of multicellularity. *eLife* **2020**, *9*, e52696. [[CrossRef](#)]
133. Guan, J.; Bondy-Denomy, J. Intracellular organization by jumbo bacteriophages. *J. Bacteriol.* **2020**. [[CrossRef](#)]
134. Edgar, R.S.; Feynman, R.P.; Klein, S.; Lielausis, I.; Steinberg, C.M. Mapping experiments with r mutants of bacteriophage T4D. *Genetics* **1962**, *47*, 179–186. [[CrossRef](#)]
135. Inoue, N.; Hess, K.D.; Moreadith, R.W.; Richardson, L.L.; Handel, M.A.; Watson, M.L.; Zinn, A.R. New gene family defined by MORC, a nuclear protein required for mouse spermatogenesis. *Hum. Mol. Genet.* **1999**, *8*, 1201–1207. [[CrossRef](#)]

136. Iyer, L.M.; Zhang, D.; Burroughs, A.M.; Aravind, L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res.* **2013**, *41*, 7635–7655. [[CrossRef](#)] [[PubMed](#)]
137. Iyer, L.M.; Zhang, D.; Aravind, L. Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. *Bioessays* **2016**, *38*, 27–40. [[CrossRef](#)] [[PubMed](#)]
138. Lobočka, M.B.; Rose, D.J.; Plunkett, G., 3rd; Rusin, M.; Samojedny, A.; Lehnher, H.; Yarmolinsky, M.B.; Blattner, F.R. Genome of bacteriophage P1. *J. Bacteriol.* **2004**, *186*, 7032–7068. [[CrossRef](#)] [[PubMed](#)]
139. Hua, J.; Huet, A.; Lopez, C.A.; Toropova, K.; Pope, W.H.; Duda, R.L.; Hendrix, R.W.; Conway, J.F. Capsids and genomes of jumbo-sized bacteriophages reveal the evolutionary reach of the HK97 fold. *mBio* **2017**, *8*, 5. [[CrossRef](#)] [[PubMed](#)]
140. Uchiyama, J.; Takemura-Uchiyama, I.; Sakaguchi, Y.; Gamoh, K.; Kato, S.; Daibata, M.; Ujihara, T.; Misawa, N.; Matsuzaki, S. Intragenus generalized transduction in *Staphylococcus* spp. by a novel giant phage. *ISME J.* **2014**, *8*, 1949–1952. [[CrossRef](#)]
141. Lopez, P.; Espinosa, M.; Piechowska, M.; Shugar, D. Influence of bacteriophage PBS1 and phi W-14 deoxyribonucleic acids on homologous deoxyribonucleic acid uptake and transformation in competent *Bacillus subtilis*. *J. Bacteriol.* **1980**, *143*, 50–58. [[CrossRef](#)]
142. Akichika, S.; Hirano, S.; Shichino, Y.; Suzuki, T.; Nishimasu, H.; Ishitani, R.; Sugita, A.; Hirose, Y.; Iwasaki, S.; Nureki, O.; et al. Cap-specific terminal N (6)-methylation of RNA by an RNA polymerase II-associated methyltransferase. *Science* **2019**, *363*, 6423. [[CrossRef](#)]
143. Iyer, L.M.; Abhiman, S.; Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **2011**, *101*, 25–104.
144. Fedeles, B.I.; Singh, V.; Delaney, J.C.; Li, D.; Essigmann, J.M. The AlkB family of Fe(II)/alpha-Ketoglutarate-dependent dioxygenases: Repairing nucleic acid alkylation damage and beyond. *J. Biol. Chem.* **2015**, *290*, 20734–20742. [[CrossRef](#)] [[PubMed](#)]
145. Iyer, L.M.; Tahiliani, M.; Rao, A.; Aravind, L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* **2009**, *8*, 1698–1710. [[CrossRef](#)] [[PubMed](#)]
146. Kaminska, K.H.; Bujnicki, J.M. Bacteriophage Mu Mom protein responsible for DNA modification is a new member of the acyltransferase superfamily. *Cell Cycle* **2008**, *7*, 120–121. [[CrossRef](#)]
147. Zhang, X.; Shi, H.; Wu, J.; Zhang, X.; Sun, L.; Chen, C.; Chen, Z.J. Cyclic GMP-AMP containing mixed phosphodiester linkages is an endogenous high-affinity ligand for STING. *Mol. Cell* **2013**, *51*, 226–235. [[CrossRef](#)] [[PubMed](#)]
148. Whiteley, A.T.; Eaglesham, J.B.; de Oliveira Mann, C.C.; Morehouse, B.R.; Lowey, B.; Nieminen, E.A.; Danilchanka, O.; King, D.S.; Lee, A.S.Y.; Mekalanos, J.J.; et al. Bacterial cGAS-like enzymes synthesize diverse nucleotide signals. *Nature* **2019**, *567*, 194–199. [[CrossRef](#)]
149. Severin, G.B.; Ramliden, M.S.; Hawver, L.A.; Wang, K.; Pell, M.E.; Kieninger, A.K.; Khataokar, A.; O'Hara, B.J.; Behrmann, L.V.; Neiditch, M.B.; et al. Direct activation of a phospholipase by cyclic GMP-AMP in El Tor *Vibrio cholerae*. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E6048–E6055. [[CrossRef](#)]
150. Niewoehner, O.; Garcia-Doval, C.; Rostol, J.T.; Berk, C.; Schwede, F.; Bigler, L.; Hall, J.; Marraffini, L.A.; Jinek, M. Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **2017**, *548*, 543–548. [[CrossRef](#)]
151. Hornung, V.; Hartmann, R.; Ablasser, A.; Hopfner, K.P. OAS proteins and cGAS: Unifying concepts in sensing and responding to cytosolic nucleic acids. *Nat. Rev. Immunol.* **2014**, *14*, 521–528. [[CrossRef](#)]
152. Anantharaman, V.; Iyer, L.M.; Aravind, L. Ter-dependent stress response systems: Novel pathways related to metal sensing, production of a nucleoside-like metabolite, and DNA-processing. *Mol. Biosyst.* **2012**, *8*, 3142–3165. [[CrossRef](#)]
153. Eaglesham, J.B.; Pan, Y.; Kupper, T.S.; Kranzusch, P.J. Viral and metazoan poxins are cGAMP-specific nucleases that restrict cGAS-STING signalling. *Nature* **2019**, *566*, 259–263. [[CrossRef](#)]
154. Zhang, R.; Jha, B.K.; Ogden, K.M.; Dong, B.; Zhao, L.; Elliott, R.; Patton, J.T.; Silverman, R.H.; Weiss, S.R. Homologous 2',5'-phosphodiesterases from disparate RNA viruses antagonize antiviral innate immunity. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13114–13119. [[CrossRef](#)]
155. Aravind, L.; Koonin, E.V. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **1998**, *23*, 469–472. [[CrossRef](#)]
156. Galperin, M.Y.; Natale, D.A.; Aravind, L.; Koonin, E.V. A specialized version of the HD hydrolase domain implicated in signal transduction. *J. Mol. Microbiol. Biotechnol.* **1999**, *1*, 303–305. [[PubMed](#)]
157. Skotnicka, D.; Smaldone, G.T.; Petters, T.; Trampari, E.; Liang, J.; Kaeffer, V.; Malone, J.G.; Singer, M.; Sogaard-Andersen, L. A minimal threshold of c-di-GMP is essential for fruiting body formation and sporulation in *Myxococcus xanthus*. *PLoS Genet.* **2016**, *12*, e1006080. [[CrossRef](#)] [[PubMed](#)]
158. Wright, T.A.; Jiang, L.; Park, J.J.; Anderson, W.A.; Chen, G.; Hallberg, Z.F.; Nan, B.; Hammond, M.C. Second messengers and divergent HD-GYP phosphodiesterases regulate 3',3'-cGAMP signaling. *Mol. Microbiol.* **2020**, *113*, 222–236. [[CrossRef](#)] [[PubMed](#)]
159. Hogg, T.; Mechold, U.; Malke, H.; Cashel, M.; Hilgenfeld, R. Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response. *Cell* **2004**, *117*, 57–68. [[CrossRef](#)]
160. Steinchen, W.; Zegarra, V.; Bange, G. (p)ppGpp: Magic modulators of bacterial physiology and metabolism. *Front. Microbiol.* **2020**, *11*, 2072. [[CrossRef](#)]

161. Rao, F.; Qi, Y.; Murugan, E.; Pasunooti, S.; Ji, Q. 2',3'-cAMP hydrolysis by metal-dependent phosphodiesterases containing DHH, EAL, and HD domains is non-specific: Implications for PDE screening. *Biochem. Biophys. Res. Commun.* **2010**, *398*, 500–505. [[CrossRef](#)]
162. Rao, F.; See, R.Y.; Zhang, D.; Toh, D.C.; Ji, Q.; Liang, Z.X. YybT is a signaling protein that contains a cyclic dinucleotide phosphodiesterase domain and a GGDEF domain with ATPase activity. *J. Biol. Chem.* **2010**, *285*, 473–482. [[CrossRef](#)]
163. Huynh, T.N.; Woodward, J.J. Too much of a good thing: Regulated depletion of c-di-AMP in the bacterial cytoplasm. *Curr. Opin. Microbiol.* **2016**, *30*, 22–29. [[CrossRef](#)]
164. Zhao, R.; Yang, Y.; Zheng, F.; Zeng, Z.; Feng, W.; Jin, X.; Wang, J.; Yang, K.; Liang, Y.X.; She, Q.; et al. A membrane-associated DHH-DHHA1 nuclease degrades type III CRISPR second messenger. *Cell Rep.* **2020**, *32*, 108133. [[CrossRef](#)] [[PubMed](#)]
165. Benzinger, R.; McCorquodale, D.J. Transfection of Escherichia coli spheroplasts. VI. Transfection of nonpermissive spheroplasts by T5 and BF23 bacteriophage DNA carrying amber mutations in DNA transfer genes. *J. Virol.* **1975**, *16*, 1–4. [[CrossRef](#)] [[PubMed](#)]
166. Mazumder, R.; Iyer, L.M.; Vasudevan, S.; Aravind, L. Detection of novel members, structure-function analysis and evolutionary classification of the 2H phosphodiesterase superfamily. *Nucleic Acids Res.* **2002**, *30*, 5229–5243. [[CrossRef](#)] [[PubMed](#)]
167. Banerjee, A.; Goldgur, Y.; Schwer, B.; Shuman, S. Atomic structures of the RNA end-healing 5'-OH kinase and 2',3'-cyclic phosphodiesterase domains of fungal tRNA ligase: Conformational switches in the kinase upon binding of the GTP phosphate donor. *Nucleic Acids Res.* **2019**, *47*, 11826–11838. [[CrossRef](#)] [[PubMed](#)]
168. Smith, B.C.; Denu, J.M. Sir2 protein deacetylases: Evidence for chemical intermediates and functions of a conserved histidine. *Biochemistry* **2006**, *45*, 272–282. [[CrossRef](#)] [[PubMed](#)]
169. Essuman, K.; Summers, D.W.; Sasaki, Y.; Mao, X.; Yim, A.K.Y.; DiAntonio, A.; Milbrandt, J. TIR domain proteins are an ancient family of NAD(+)-consuming enzymes. *Curr. Biol.* **2018**, *28*, 421–430.e4. [[CrossRef](#)]
170. Wan, L.; Essuman, K.; Anderson, R.G.; Sasaki, Y.; Monteiro, F.; Chung, E.H.; Osborne Nishimura, E.; DiAntonio, A.; Milbrandt, J.; Dangl, J.L.; et al. TIR domains of plant immune receptors are NAD(+)-cleaving enzymes that promote cell death. *Science* **2019**, *365*, 799–803. [[CrossRef](#)]
171. Essuman, K.; Summers, D.W.; Sasaki, Y.; Mao, X.; DiAntonio, A.; Milbrandt, J. The SARM1 toll/interleukin-1 receptor domain possesses intrinsic NAD(+) cleavage activity that promotes pathological axonal degeneration. *Neuron* **2017**, *93*, 1334–1343.e5. [[CrossRef](#)]
172. Samanovic, M.I.; Tu, S.; Novak, O.; Iyer, L.M.; McAllister, F.E.; Aravind, L.; Gygi, S.P.; Hubbard, S.R.; Strnad, M.; Darwin, K.H. Proteasomal control of cytokinin synthesis protects *Mycobacterium tuberculosis* against nitric oxide. *Mol. Cell* **2015**, *57*, 984–994. [[CrossRef](#)]
173. Anantharaman, V.; Aravind, L. Analysis of DBC1 and its homologs suggests a potential mechanism for regulation of sirtuin domain deacetylases by NAD metabolites. *Cell Cycle* **2008**, *7*, 1467–1472. [[CrossRef](#)]
174. Freire, D.M.; Gutierrez, C.; Garza-Garcia, A.; Grabowska, A.D.; Sala, A.J.; Ariyachakun, K.; Panikova, T.; Beckham, K.S.H.; Colom, A.; Pogenberg, V.; et al. An NAD(+) phosphorylase toxin triggers *Mycobacterium tuberculosis* cell death. *Mol. Cell* **2019**, *73*, 1282–1291.e8. [[CrossRef](#)] [[PubMed](#)]
175. Rack, J.G.; Perina, D.; Ahel, I. Macrod domains: Structure, function, evolution, and catalytic activities. *Annu. Rev. Biochem.* **2016**, *85*, 431–454. [[CrossRef](#)] [[PubMed](#)]
176. De Souza, R.F.; Aravind, L. Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. *Mol. Biosyst.* **2012**, *8*, 1661–1677. [[CrossRef](#)] [[PubMed](#)]
177. Pao, G.M.; Saier Jr, M.H. Response regulators of bacterial signal transduction systems: Selective domain shuffling during evolution. *J. Mol. Evol.* **1995**, *40*, 136–154. [[CrossRef](#)]
178. Burroughs, A.M.; Aravind, L. RNA damage in biological conflicts and the diversity of responding RNA repair systems. *Nucleic Acids Res.* **2016**, *44*, 8525–8555. [[CrossRef](#)] [[PubMed](#)]
179. Makarova, K.S.; Anantharaman, V.; Grishin, N.V.; Koonin, E.V.; Aravind, L. CARF and WYL domains: Ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **2014**, *5*, 102. [[CrossRef](#)]
180. Bourret, R.B. Receiver domain structure and function in response regulator proteins. *Curr. Opin. Microbiol.* **2010**, *13*, 142–149. [[CrossRef](#)]
181. Gabelli, S.B.; Bianchet, M.A.; Bessman, M.J.; Amzel, L.M. The structure of ADP-ribose pyrophosphatase reveals the structural basis for the versatility of the Nudix family. *Nat. Struct. Biol.* **2001**, *8*, 467–472. [[CrossRef](#)]
182. Mildvan, A.S.; Xia, Z.; Azurmendi, H.F.; Saraswat, V.; Legler, P.M.; Massiah, M.A.; Gabelli, S.B.; Bianchet, M.A.; Kang, L.W.; Amzel, L.M. Structures and mechanisms of Nudix hydrolases. *Arch. Biochem. Biophys.* **2005**, *433*, 129–143. [[CrossRef](#)]
183. Skjerning, R.B.; Senissar, M.; Winther, K.S.; Gerdes, K.; Brodersen, D.E. The RES domain toxins of RES-Xre toxin-antitoxin modules induce cell stasis by degrading NAD⁺. *Mol. Microbiol.* **2019**, *111*, 221–236. [[CrossRef](#)]
184. Hawse, W.F.; Wolberger, C. Structure-based mechanism of ADP-ribosylation by sirtuins. *J. Biol. Chem.* **2009**, *284*, 33654–33661. [[CrossRef](#)] [[PubMed](#)]
185. Dulyaninova, N.G.; Podlepa, E.M.; Touloukhonova, L.V.; Bykhovskiy, V.Y. Salvage pathway for NAD biosynthesis in *Brevibacterium ammoniagenes*: Regulatory properties of triphosphate-dependent nicotinate phosphoribosyltransferase. *Biochim. Biophys. Acta* **2000**, *1478*, 211–220. [[CrossRef](#)]

186. Sorci, L.; Martynowski, D.; Rodionov, D.A.; Eyobo, Y.; Zogaj, X.; Klose, K.E.; Nikolaev, E.V.; Magni, G.; Zhang, H.; Osterman, A.L. Nicotinamide mononucleotide synthetase is the key enzyme for an alternative route of NAD biosynthesis in *Francisella tularensis*. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3083–3088. [[CrossRef](#)]
187. Grose, J.H.; Bergthorsson, U.; Roth, J.R. Regulation of NAD synthesis by the trifunctional NadR protein of *Salmonella enterica*. *J. Bacteriol.* **2005**, *187*, 2774–2782. [[CrossRef](#)]
188. Makarova, K.S.; Anantharaman, V.; Aravind, L.; Koonin, E.V. Live virus-free or die: Coupling of antiviral immunity and programmed suicide or dormancy in prokaryotes. *Biol. Direct* **2012**, *7*, 40. [[CrossRef](#)] [[PubMed](#)]
189. Winther, K.S.; Gerdes, K. Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7403–7407. [[CrossRef](#)]
190. Bitton, L.; Klaiman, D.; Kaufmann, G. Phage T4-induced DNA breaks activate a tRNA repair-defying anticodon nuclease. *Mol. Microbiol.* **2015**, *97*, 898–910. [[CrossRef](#)] [[PubMed](#)]
191. Tomita, K.; Ogawa, T.; Uozumi, T.; Watanabe, K.; Masaki, H. A cytotoxic ribonuclease which specifically cleaves four isoaccepting arginine tRNAs at their anticodon loops. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8278–8283. [[CrossRef](#)]
192. Silber, R.; Malathi, V.G.; Hurwitz, J. Purification and properties of bacteriophage T4-induced RNA ligase. *Proc. Natl. Acad. Sci. USA* **1972**, *69*, 3009–3013. [[CrossRef](#)]
193. Walker, G.C.; Uhlenbeck, O.C.; Bedows, E.; Gumport, R.I. T4-induced RNA ligase joins single-stranded oligoribonucleotides. *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 122–126. [[CrossRef](#)]
194. Yoshikawa, G.; Askora, A.; Blanc-Mathieu, R.; Kawasaki, T.; Li, Y.; Nakano, M.; Ogata, H.; Yamada, T. *Xanthomonas citri* jumbo phage XacN1 exhibits a wide host range and high complement of tRNA genes. *Sci. Rep.* **2018**, *8*, 4486. [[CrossRef](#)] [[PubMed](#)]
195. Simoliunas, E.; Kaliniene, L.; Truncaite, L.; Zajackauskaite, A.; Staniulis, J.; Kaupinis, A.; Ger, M.; Valius, M.; Meskys, R. Klebsiella phage vB_KleM-RaK2—A giant singleton virus of the family Myoviridae. *PLoS ONE* **2013**, *8*, e60717. [[CrossRef](#)] [[PubMed](#)]
196. Monson, R.; Foulds, I.; Foweraker, J.; Welch, M.; Salmond, G.P.C. The *Pseudomonas aeruginosa* generalized transducing phage phiPA3 is a new member of the phiKZ-like group of ‘jumbo’ phages, and infects model laboratory strains and clinical isolates from cystic fibrosis patients. *Microbiology* **2011**, *157*, 859–867. [[CrossRef](#)] [[PubMed](#)]
197. Tanaka, N.; Shuman, S. RtcB is the RNA ligase component of an *Escherichia coli* RNA repair operon. *J. Biol. Chem.* **2011**, *286*, 7727–7731. [[CrossRef](#)] [[PubMed](#)]
198. Chakravarty, A.K.; Shuman, S. RNA 3'-phosphate cyclase (RtcA) catalyzes ligase-like adenylation of DNA and RNA 5'-monophosphate ends. *J. Biol. Chem.* **2011**, *286*, 4117–4122. [[CrossRef](#)] [[PubMed](#)]
199. Iyer, L.M.; Abhiman, S.; Maxwell Burroughs, A.; Aravind, L. Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: Synthesis of novel metabolites and peptide modifications of proteins. *Mol. Biosyst.* **2009**, *5*, 1636–1660. [[CrossRef](#)]
200. Shuman, S.; Schwer, B. RNA capping enzyme and DNA ligase: A superfamily of covalent nucleotidyl transferases. *Mol. Microbiol.* **1995**, *17*, 405–410. [[CrossRef](#)]
201. Shuman, S.; Lima, C.D. The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. *Curr. Opin. Struct. Biol.* **2004**, *14*, 757–764. [[CrossRef](#)]
202. Sim, S.; Hughes, K.; Chen, X.; Wolin, S.L. The bacterial Ro60 protein and its noncoding Y RNA regulators. *Annu. Rev. Microbiol.* **2020**, *74*, 387–407. [[CrossRef](#)]
203. Spinelli, S.L.; Kierzek, R.; Turner, D.H.; Phizicky, E.M. Transient ADP-ribosylation of a 2'-phosphate implicated in its removal from ligated tRNA during splicing in yeast. *J. Biol. Chem.* **1999**, *274*, 2637–2644. [[CrossRef](#)]
204. Shull, N.P.; Spinelli, S.L.; Phizicky, E.M. A highly specific phosphatase that acts on ADP-ribose 1'-phosphate, a metabolite of tRNA splicing in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **2005**, *33*, 650–660. [[CrossRef](#)] [[PubMed](#)]
205. Klaiman, D.; Steinfels-Kohn, E.; Krutkina, E.; Davidov, E.; Kaufmann, G. The wobble nucleotide-excising anticodon nuclease RloC is governed by the zinc-hook and DNA-dependent ATPase of its Rad50-like region. *Nucleic Acids Res.* **2012**, *40*, 8568–8578. [[CrossRef](#)] [[PubMed](#)]
206. Yamashita, S.; Takeshita, D.; Tomita, K. Translocation and rotation of tRNA during template-independent RNA polymerization by tRNA nucleotidyltransferase. *Structure* **2014**, *22*, 315–325. [[CrossRef](#)] [[PubMed](#)]
207. Gu, W.; Jackman, J.E.; Lohan, A.J.; Gray, M.W.; Phizicky, E.M. tRNA^{His} maturation: An essential yeast protein catalyzes addition of a guanine nucleotide to the 5' end of tRNA^{His}. *Genes Dev.* **2003**, *17*, 2889–2901. [[CrossRef](#)]
208. Burroughs, A.M.; Aravind, L. The origin and evolution of release factors: Implications for translation termination, ribosome rescue, and quality control pathways. *Int. J. Mol. Sci.* **2019**, *20*, 1981. [[CrossRef](#)]
209. Burroughs, A.M.; Glasner, M.E.; Barry, K.P.; Taylor, E.A.; Aravind, L. Oxidative opening of the aromatic ring: Tracing the natural history of a large superfamily of dioxygenase domains and their relatives. *J. Biol. Chem.* **2019**, *294*, 10211–10235. [[CrossRef](#)] [[PubMed](#)]
210. Andrews, E.S.V.; Arcus, V.L. PhoH2 proteins couple RNA helicase and RNase activities. *Protein Sci.* **2020**, *29*, 883–892. [[CrossRef](#)]
211. Andrews, E.S.; Arcus, V.L. The mycobacterial PhoH2 proteins are type II toxin antitoxins coupled to RNA helicase domains. *Tuberculosis* **2015**, *95*, 385–394. [[CrossRef](#)]
212. Sengupta, T.K.; Gordon, J.; Spicer, E.K. RegA proteins from phage T4 and RB69 have conserved helix-loop groove RNA binding motifs but different RNA binding specificities. *Nucleic Acids Res.* **2001**, *29*, 1175–1184. [[CrossRef](#)]

213. Aylett, C.H.; Izore, T.; Amos, L.A.; Lowe, J. Structure of the tubulin/FtsZ-like protein TubZ from Pseudomonas bacteriophage PhiKZ. *J. Mol. Biol.* **2013**, *425*, 2164–2173. [[CrossRef](#)]
214. Oliva, M.A.; Martin-Galiano, A.J.; Sakaguchi, Y.; Andreu, J.M. Tubulin homolog TubZ in a phage-encoded partition system. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7711–7716. [[CrossRef](#)]
215. Fong, S.T.; Stanisich, V.A. Location and characterization of two functions on RP1 that inhibit the fertility of the IncW plasmid R388. *J. Gen. Microbiol.* **1989**, *135*, 499–502. [[CrossRef](#)] [[PubMed](#)]
216. Uberto, R.; Moomaw, E.W. Protein similarity networks reveal relationships among sequence, structure, and function within the Cupin superfamily. *PLoS ONE* **2013**, *8*, e74477. [[CrossRef](#)] [[PubMed](#)]
217. Sigrell, J.A.; Cameron, A.D.; Jones, T.A.; Mowbray, S.L. Structure of Escherichia coli ribokinase in complex with ribose and dinucleotide determined to 1.8 Å resolution: Insights into a new family of kinase structures. *Structure* **1998**, *6*, 183–193. [[CrossRef](#)]
218. Aravind, L.; Anantharaman, V.; Koonin, E.V. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETPF, photolyase, and PP-ATPase nucleotide-binding domains: Implications for protein evolution in the RNA. *Proteins* **2002**, *48*, 1–14. [[CrossRef](#)]
219. Breton, C.; Fournel-Gigleux, S.; Palcic, M.M. Recent structures, evolution and mechanisms of glycosyltransferases. *Curr. Opin. Struct. Biol.* **2012**, *22*, 540–549. [[CrossRef](#)]
220. Burroughs, A.M.; Allen, K.N.; Dunaway-Mariano, D.; Aravind, L. Evolutionary genomics of the HAD superfamily: Understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J. Mol. Biol.* **2006**, *361*, 1003–1034. [[CrossRef](#)]
221. Nasir, A.; Romero-Severson, E.; Claverie, J.M. Investigating the concept and origin of viruses. *Trends Microbiol.* **2020**, *28*, 959–967. [[CrossRef](#)]
222. McCutcheon, J.P.; McDonald, B.R.; Moran, N.A. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15394–15399. [[CrossRef](#)]
223. Koonin, E.V. How many genes can make a cell: The minimal-gene-set concept. *Annu. Rev. Genom. Hum. Genet.* **2000**, *1*, 99–116. [[CrossRef](#)]
224. Powers, R.; Mirkovic, N.; Goldsmith-Fischman, S.; Acton, T.B.; Chiang, Y.; Huang, Y.J.; Ma, L.; Rajan, P.K.; Cort, J.R.; Kennedy, M.A.; et al. Solution structure of Archaeoglobus fulgidis peptidyl-tRNA hydrolase (Pth2) provides evidence for an extensive conserved family of Pth2 enzymes in archaea, bacteria, and eukaryotes. *Protein Sci.* **2005**, *14*, 2849–2861. [[CrossRef](#)] [[PubMed](#)]
225. Pedersen, L.B.; Schroder, J.M.; Satir, P.; Christensen, S.T. The ciliary cytoskeleton. *Compr. Physiol.* **2012**, *2*, 779–803. [[PubMed](#)]
226. Chaaban, S.; Brouhard, G.J. A microtubule bestiary: Structural diversity in tubulin polymers. *Mol. Biol. Cell* **2017**, *28*, 2924–2931. [[CrossRef](#)] [[PubMed](#)]
227. Rao, V.B.; Feiss, M. Mechanisms of DNA packaging by large double-stranded DNA viruses. *Annu. Rev. Virol.* **2015**, *2*, 351–378. [[CrossRef](#)] [[PubMed](#)]
228. Black, L.W.; Rao, V.B. Structure, assembly, and DNA packaging of the bacteriophage T4 head. *Adv. Virus Res.* **2012**, *82*, 119–153. [[PubMed](#)]
229. Iyer, L.M.; Makarova, K.S.; Koonin, E.V.; Aravind, L. Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: Implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res.* **2004**, *32*, 5260–5279. [[CrossRef](#)] [[PubMed](#)]
230. Mushegian, A.R.; Koonin, E.V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 10268–10273. [[CrossRef](#)] [[PubMed](#)]
231. Edgell, D.R.; Gibb, E.A.; Belfort, M. Mobile DNA elements in T4 and related phages. *Virol. J.* **2010**, *7*, 290. [[CrossRef](#)] [[PubMed](#)]
232. Stoddard, B.; Belfort, M. Social networking between mobile introns and their host genes. *Mol. Microbiol.* **2010**, *78*, 1–4. [[CrossRef](#)]
233. Smith, J.M. *Evolutionary Genetics*; Oxford University Press: Oxford, UK, 1998.