Published in final edited form as:

Biochim Biophys Acta. 2010 ; 1799(3-4): 302–318. doi:10.1016/j.bbagr.2009.10.006.

Diversity and evolution of chromatin proteins encoded by DNA viruses

Robson F. de Souza, Lakshminarayan M. Iyer, and L. Aravind*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, United States of America

Abstract

Double-stranded DNA viruses display a great variety of proteins that interact with host chromatin. Using the wealth of available genomic and functional information, we have systematically surveyed chromatin-related proteins encoded by dsDNA viruses. The distribution of viral chromatin-related proteins is primarily influenced by viral genome size and the superkingdom to which the host of the virus belongs. Smaller viruses usually encode multifunctional proteins that mediate several distinct interactions with host chromatin proteins and viral or host DNA. Larger viruses additionally encode several enzymes, which catalyze manipulations of chromosome structure, chromatin remodeling and covalent modifications of proteins and DNA. Among these viruses, it is also common to encounter transcription factors and DNA-packaging proteins such as histones and IHF/HU derived from cellular genomes, which might play a role in constituting virus-specific chromatin states. Through all size ranges a subset of domains in viral chromatin proteins appear to have been derived from those found in host proteins. Examples include the Zn-finger domains of the E6 and E7 proteins of papillomaviruses, SET-domain methyltransferases and Jumonji-related demethylases in certain nucleocytoplasmic large DNA viruses and BEN domains in poxviruses and polydnviruses. In other cases, chromatin-interacting modules, such as the LxCxE motif, appear to have been widely disseminated across distinct viral lineages, resulting in similar retinoblastoma targeting strategies. Viruses, especially those with large linear genomes, have evolved a number of mechanisms to manipulate viral chromosomes in the process of replication-associated recombination. These include topoisomerases, Rad50/SbcC-like ABC ATPases and a novel recombinase system in bacteriophages utilizing RecA and Rad52 homologs. Larger DNA viruses also encode SWI2/SNF2 and A18-like ATPases which appear to play specialized roles in transcription and recombination. Finally, it also appears that certain domains of viral provenance have given rise to key functions in eukaryotic chromatin such as a HEH domain of chromosome tethering proteins and the TET/JBP-like cytosine and thymine hydroxylases.

INTRODUCTION

In cellular life forms DNA-packaging proteins bind DNA with low sequence specificity, promote its bending and organize it into highly compacted structures. This nucleoprotein

*Corresponding author. aravind@mail.nih.gov; Tel.: +1-301-594-2445.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Supplementary material

A database of chromatin-related proteins detected in DNA viruses, along with alignments and gene-neighborhoods of particular viral chromatin proteins is available at: ftp://ftp.ncbi.nih.gov/pub/aravind/chromatin/viral_chromatin/

ensemble or chromatin has a central role in facilitating and regulating biochemical processes including DNA replication and repair, transcription and RNA processing. Evolutionary comparisons have shown that the primary DNA-packaging proteins involved in organization of chromatin are different across the three superkingdoms of life. In bacteria the primary DNA packaging proteins are members of the HU/IHF (also called DNABII) superfamily [1]. In contrast, several archaea and most eukaryotes contain histones, which form the characteristic octameric DNA compaction unit termed the nucleosome [2]. However, in some eukaryotes, such as certain dinoflagellates, bacterial type HU/IHF homologs, rather than histones, play a fundamental role in DNA packaging [3]. Likewise, in certain archaeal lineages such as Sulfolobales the histones appear to have been displaced by other chromosome packaging proteins [4]. Importantly, eukaryotic histones differ from archaeal histones in having long, low complexity tails that are enriched in positively charged residues and contact the negatively charged backbone of DNA [5]. These histones tails are substrates for a large number of chromatin modifying enzymes, which catalyze a bewildering array of covalent modifications on lysine, arginine, serine, threonine and glutamate [6, 7]. These modifications range from low molecular weight adducts such as methyl, acetyl and phosphate groups to ligation of entire protein chains such as ubiquitin and SUMO. Akin to protein modifications, DNA modifications such as methylation, momylation and more recently hydroxymethylation, amongst others, are seen to play important roles in chromatin organization [8–10].

Modifications of histones (and other chromosomal proteins) and DNA appear to act as a “code” atop that specified by the genome and are thus termed epigenetic marks [11]. Eukaryotes also display a unique proliferation of diverse “adaptor” domains, for example, the Bromo, Chromo, PHD, MYB/SANT and BMB (PWWP) domains [6]. These domains recognize modified or unmodified peptides in histone tails and other chromatin proteins. Likewise, eukaryotes are also known to possess DNA-binding proteins that specifically recognize modified DNA [12]. Thus, domains which specifically recognize such covalent modifications help in “reading” the epigenetic code and linking it to various downstream processes [11]. Supercoiling, topology and higher order arrangement of DNA in chromatin is also highly dynamic and considerably influenced by the action of multiple distinct topoisomerases [13]. Eukaryotes in particular, and to a certain degree prokaryotes, also contain other chromatin remodeling enzymes that use the free-energy of ATP hydrolysis to actively remodel DNA-protein contacts, unwind DNA or reorganize it into higher order loop-structures. Such enzymes, including Swi2/Snf2 ATPases, SMC ATPases and MORC-type ATPases, have a major role in chromosomal organization and alterations of nucleosomal positions across eukaryotes [14–16]. Proteins involved in these structural and dynamic processes of chromatin interact with other DNA-binding proteins, namely basal or general transcription factors (which recruit the RNA polymerase to a promoter) and specific transcription factors, which recognize distinctive regulatory DNA sequences associated with particular genes [17]. Transcription factors (TFs) often share DNA-binding domains with proteins involved in chromatin structure and dynamics and functionally overlap with them [6]. Thus, transcription-related protein complexes might also be considered integral components of chromatin in both eukaryotes and prokaryotes. While intimately interacting with transcription regulatory apparatus, chromatin structure and dynamics provide a distinct level of regulation with major consequences for all the cellular processes that operate on DNA [18]. This regulatory level, especially in the form of epigenetic marks, is highly developed in eukaryotes [7, 18, 19] and to lesser degree in the two prokaryotic superkingdoms [15].

In contrast to cellular life forms, DNA viruses package their genome into externally situated protein coats (capsids) or lipid membranes situated inside such protein coats. Studies of different bacteriophages such as lambda, P22 and T4, suggest that DNA is packaged in viral

capsids as naked DNA close to the maximum possible density observed in a pure DNA crystal [20–22]. In contrast, cores of large eukaryotic poxviruses have much greater available space than in the bacteriophage capsids and DNA is packaged at lower density [23, 24]. However, even in this case the bulk of DNA in the core appears to be primarily in the form of naked strands although there might be limited linkages to proteins [23]. A similar partial linkage to a protein (conserved protein VII) has been reported in adenoviral capsids [25]. Studies on T4 DNA packaging have shown that, though positively-charged proteins of the capsid play some role in the process, majority of the charge-neutralization during viral DNA packaging comes from polyamines and monovalent metal ions included in the capsid [22]. Hence, viral DNA in capsids is packaged very differently from that of their cellular hosts. However, viral DNA, while replicating either as an episome or integrated into host DNA, is often subject to packaging similar to host chromatin.

In recent years, major advances in viral genomics have made available complete genome sequences of numerous large DNA viruses. Comparative viral genomics has gone a long way in revealing the nature of the viral proteome and previously unclear vertical and horizontal relationships between diverse dsDNA viruses [26, 27]. These studies point to a complex web of relationships in which a variety of proteins are shared between otherwise phylogenetically distinct groups of viruses as a result of extensive lateral gene exchanges between viruses and their hosts. In the past, sequences of viral proteins have been difficult to analyze due to rapid divergence relative to one and other and their cellular counterparts. Availability of numerous genome sequences and structure solution efforts have mitigated this to a certain extent and allowed recovery of distant relationships [28–31]. These studies have shown that both eukaryotic and prokaryotic viruses encode a diverse set of chromatin proteins, each of which might have functional consequences for the host or the virus. To date studies on both eukaryotic and bacterial dsDNA viruses have revealed that they encode proteins that are involved in chromatin structure and dynamics [26, 32–34]. These included various P-loop ATPases that could function as chromatin remodelers, topoisomerases, histone-modifying enzymes and DNA-binding proteins with packaging and structure-modifying potential. Experimental studies on some such virally encoded chromatin proteins have demonstrated critical roles for them in expression of host or viral genes [32–36].

In this article we attempt to systematically review virally encoded chromatin proteins from a comparative genomics perspective. In doing so we hope to bring attention to previously underappreciated viral chromatin proteins and place what is already known in a broader context. As can be seen from the above discussion, the category “chromatin proteins (CPs)” can be a bit diffuse, overlapping with other processes such as replication, recombination and transcription. In this article we stick mainly to those involved in chromatin structure and dynamics, largely refraining from detailed discussion on enzymes catalyzing DNA and RNA synthesis or mediating DNA repair. However, we do briefly consider several transcription factors and their DNA-binding domains due to their functional overlap with chromatin proteins. We begin by providing an overview of large dsDNA viral relationships and phyletic patterns of chromatin proteins encoded by them. We follow this with a summary of the various functional classes of chromatin proteins encoded by viruses and their potential significance for viral biology. Finally, we attempt to integrate this information into our current understanding of viral evolution.

A GENERAL SURVEY OF DNA VIRUSES AND PHYLETIC PATTERNS OF VIRAL CHROMATIN PROTEINS

Commonalities and differences of DNA viruses

Double stranded DNA viruses are enormously diverse in terms of virion morphology, genome size/coding capacity, genome structure and replication strategies (Fig. 1). Yet several disparate groups of viruses might display one or more shared features that include [26, 28, 30, 37]: 1) β -jellyroll domain capsid proteins; 2) DNA-packaging ATPases either of the terminase large subunit or FtsK-HerA superfamily; 3) portal proteins; 4) DNA polymerases; 5) replication-related DNA helicases belonging to the AAA+ superfamily; 6) primases either of the eukaryote-type primase superfamily or TOPRIM domain superfamily (DnaG-like). Most large DNA viruses additionally encode one or more DNA metabolism enzymes that might help in more efficiently providing precursors for DNA synthesis [26]. These features appear to have spread in viruses as a result of a combination of common origin and extensive gene exchange between disparate groups [26, 27, 38]. Beyond these core proteins, major viral groups might considerably differ from each other in their protein complements. The main morphological and genomic differences appear to mirror the three superkingdoms of cellular life: thus viruses infecting bacteria, archaea and eukaryotes do show considerable differences between each other (Fig. 1) [38, 39]. For example, the caudate morphology (tailed-bacteriophages) is primarily restricted to bacterial viruses, whereas several unusual morphologies such as bottle-shaped (ampullavirus), lemon-shaped (fusellovirus), two tailed (bicaudavirus) and hooked-filamentous forms (lipothrixvirus) are unique to archaeal viruses [39]. Some viral groups such as certain caudate phages in bacteria, poxviruses, iridoviruses, phycodnaviruses and herpesviruses are observed across phylogenetically diverse sets of host species. However, other groups such as baculoviruses and polydnaviruses appear to be restricted to certain arthropod lineages.

Distribution of DNA viral chromatin proteins

In terms of encoding chromatin proteins an important factor appears to be genome size/coding capacity (Fig. 1). Papovaviruses (papillomaviruses and polyomaviruses) and Sputnik (a satellite virus of the Mamavirus) in eukaryotes, salterproviruses, rudiviruses and fuselloviruses in archaea, and some satellite phages (e.g. bacteriophage P4 or *Lactococcus* phage bIL311), tectiviruses, picoviruses (a lineage of small tailed phages) in bacteria are amongst the smallest of dsDNA viruses with genomes less than 20kb and usually encoding no more than 15–20 proteins. These viruses typically possess a minimal DNA replication apparatus of 1–3 proteins, which might include either a DNA helicase or a helicase-primase protein or a DNA polymerase that uses a protein primer. None of these viruses are thus far known to encode any *dedicated* chromatin-modifying or remodeling enzymes. Among prokaryotic viruses in this size range there is currently little evidence for any proteins that might interact with host chromatin. However, several animal viruses in this category encode distinctive proteins that interact with host chromatin proteins. These proteins contain certain characteristic modules such as: 1) the retinoblastoma pocket-binding protein with a LXCXE module in papovaviruses [33, 40]. 2) The viral DNA-binding domain typified by the DNA-binding domain of papillomavirus E2 protein [29]. 3) The distinctive Zn-finger domains of papillomavirus E6- and E7-type proteins [41, 42].

The next category defined by genomes greater than 20kb and typically less than 50kb and encoding approximately 20–70 proteins includes dsDNA viruses such as the animal adenoviruses, archaeal lipothrixviruses and bicaudiviruses and several lineages of caudate bacteriophages (Fig. 1). Adenoviruses exhibit continuity with animal viruses in the lower size range in their chromatin-related proteins and encode proteins containing LXCXE modules (e.g. E1A protein [43]). They also possess their own distinctive virus-specific

DNA-binding domains (e.g. adenovirus E2A [44] and protein VII [45, 46]), which have chromatin related functions. In this size range, several bacteriophages and archaeal lipothrixviruses exhibit ATP-dependent enzymes of the P-loop fold (E.g. SWI2/SNF2, Fig. 1) that might have a role in chromatin remodeling, DNA manipulations during replication, and assembly of transcription related complexes [47]. Additionally, phages in this size range possess a variety of proteins that could be involved in DNA-packaging such as those with the helix-extension helix domain (HEH) [48] and homologs of the bacterial chromosomal proteins (e.g. HU/IHF family) [1]. Many of these viruses also possess a distinctive apparatus for recombination in the form of a representative of the functionally comparable single strand-binding proteins, namely Rad52, ERF1 and the classical single strand binding proteins (SSBs) [31]. Along with the above proteins most of these phages encode a specialized ABC-ATPase homologous to the Rad50 and SbcC proteins and associated nucleases of the TOPRIM and calcineurin-like phosphoesterase superfamilies (see below). Several DNA bacteriophages in this and higher size ranges display a diverse set of DNA-modifying enzymes, which generate a wide range of modified bases in DNA [9, 10, 49, 50].

Moderately sized dsDNA viruses typically have genomes in the size range from 50–150kb and usually encode 70–160 proteins (Fig. 1). This category encompasses animal viruses, including herpesviruses (though few members might exceed 150kb) and baculoviruses, and several caudate bacteriophages. These animal DNA viruses show evolutionary links to the smaller size-range viruses in terms of their chromatin proteins. Both herpesviruses and baculoviruses encode proteins with LXCXE modules, and the herpesvirus latency associated nuclear antigen (LANA) is homologous and functionally equivalent to the papilloma virus E2 protein [51, 52]. Herpesviruses also encode several other distinctive DNA-binding factors involved in chromatin related functions [32, 53–55]. They also possess RING finger domain-containing ubiquitin E3 ligases, which may play an important role in modifying the chromatin proteins deposited on the viral chromosome [56]. Several baculoviruses encode their own SWI2/SNF2 ATPases and in some cases their own homologs of histones and a polyADP ribose polymerase that could modify histones or other chromatin proteins [57, 58]. Among bacteriophages in this range there is considerable continuity with smaller phages in terms of the representation of chromatin proteins – these viruses too encode SWI2/SNF2 and A18-like ATPases, Rad50/SbcC-like ABC ATPases, SSBs proteins and homologs of bacterial chromosomal proteins (Fig. 1). However, they differ from their smaller counterparts in frequently encoding topoisomerases [13, 26]. Additionally, they might also encode their own DNA recombinases of the RecA superfamily, which are either closely related to the bacterial RecA or belong to a distinct category of phage RecA-like ATPases (this study).

The largest DNA-viruses typically have genomes over 150kb, extending up to 1.15mb (*Acanthamoeba* mimivirus), and encode anywhere between 160–900 proteins (Fig. 1). Among eukaryotic viruses in this range are the Shrimp White Spot Syndrome virus (WSSV), polydnviruses with multipartite genomes and the large nucleocytoplasmic DNA virus (NCLDV) clade. This latter clade unifies several distinct large eukaryotic dsDNA viruses, namely poxviruses, African Swine Fever virus (ASFV), iridoviruses, and phycodnaviruses and the mimivirus. Phylogenetic reconstructions suggest that within the NCLDV clade the mimivirus and the phycodnaviruses are unified into a single clade and they together are further united to iridoviruses to the exclusion of the other NCLDVs [26, 30]. Among bacteriophages, the largest caudate phages belong to this size range and are represented by some well-studied forms such as T4. Despite their genome size, polydnviruses are rather barren in terms of coding capacity – the recently discovered BEN domain is prevalent in most polydnviruses and proteins with this domain might perform a chromatin-related function important for these viruses [59]. Additionally, certain polydnviruses encode their own version of the histone H4. WSSV possesses a SWI2/SNF2

ATPase and also a distinct family of proteins typified by ICP11 which are believed to interact with histones [60]. Almost all NCLDV's encode SWI2/SNF2 ATPases and A18-type SFII helicases. Nearly all NCLDV's also encode one or more topoisomerases. NCLDV's frequently possess their own versions of chromosomal proteins distinct from host histones [26]. Furthermore, certain NCLDV's might encode their own histone methylases and demethylases [26, 61, 62] and also DNA-modifying enzymes [63, 64]. Iridoviruses also share with several caudate bacteriophages the DNA-binding HEH domains [26]. Thus, some NCLDV's display among the most developed repertoires of chromatin proteins that appear to function both in regulating viral gene expression and targeting host chromatin. In contrast, the largest bacteriophages with genomes comparable to NCLDV's are not distinguished by any particular adaptations in this regard and typically show the same type of chromatin proteins as those described in moderately sized bacteriophages.

In addition to chromatin proteins, across all size ranges and host specificities, dsDNA viruses might encode specific transcription factors that regulate their own gene expression as well as those of their hosts. Especially in eukaryotic viruses, many of these transcription factors belong to unique virus-specific families. However, in moderately-sized and large viruses specific transcription factors derived from cellular transcription factors are frequently encountered.

FUNCTIONAL CLASSES AND ARCHITECTURES OF DOMAINS IN VIRAL CHROMATIN PROTEINS

The above survey indicates that there is a considerable diversity of proteins performing chromatin related functions in viruses. However, majority of the better understood representatives possess conserved globular protein domains that can be classified into a few major functional classes for the purposes of a more detailed discussion: **1) Adaptor domains-** These may be defined as domains that are primarily involved in fostering physical linkages between proteins in which they are present and target proteins or nucleic acids. As per their binding specificities these domains might be further categorized as protein or nucleic acid interacting adaptor proteins. **2) DNA-binding domains of structural proteins-** This group includes domains found in proteins that bind DNA in order to package it into structures at various levels of organization (e.g. histones and other chromosomal proteins). **3) Domains found in proteins mediating ATP-dependent chromatin remodeling, manipulation of chromosomes and topoisomerases.** **4) Enzymatic domains that covalently modify chromatin components.** **5) DNA-binding domains of transcription factors.**

Adaptor domains mediating protein-protein interactions

The most commonly observed adaptor domains in viral proteins are critical in mediating interactions between viral proteins and host chromatin proteins. A widely distributed adaptor domain mediating such an interaction in eukaryotic viruses is the **LXCXE module** (Fig. 2A). The core of this module by itself is rather short and assumes a simple extended configuration which binds to a specific cleft in the conserved eukaryotic Rb protein (or its paralogs in vertebrates such as p107 and p130) [33, 40, 43, 65]. Our analysis using both PSI-BLAST [66] and the hidden Markov model-based homology detection package HMMER3 [67] showed that, rather than being the result of convergent evolution, LXCXE modules from diverse DNA viruses are likely to share a common origin. This study also showed that, beyond the LXCXE core, the module is additionally characterized by an acidic stretch at the C-terminus (Supplementary material). The LXCXE module appears to have been widely disseminated across ssDNA (certain parvoviruses and nanoviruses) and dsDNA viruses through lateral transfer and is used in both plant and animal hosts due to conservation of the

Rb protein. In the smallest DNA viruses the LXCXE module is often found fused to other domains that are related to replication of the viruses – e.g. in polyomaviruses it is part of a large protein, the T-antigen, where it is combined with the DnaJ, the E1-like origin-binding protein domain and the ATP-dependent SFIII helicase [68–71]. Likewise, in ssDNA viruses such as parvoviruses of the bocavirus and densovirus lineages it is combined with the rolling circle replicator protein (RCR) or the replicative SFIII helicase (Fig. 2A). In the papillomaviral E7 protein and adenoviral E1A protein it is combined respectively with two distinct Zn-finger domains, while in the herpesvirus it is found N-terminal of the catalytic domain of the UL97 kinase domain (Fig. 2A) [40]. Versions such as ORF virus (a poxvirus) ORF119, 8kDa protein of baculoviruses (both detected in this study) and CLINK proteins of plant ssDNA nanovirus [72] are standalone proteins with the only detectable module being LXCXE (Fig. 2A). Domain fusions of the LXCXE module to replication-related functions in several small DNA viruses suggest that binding of the Rb pocket is tightly linked to replication of the viral genome. Several studies indicate that the LXCXE module precisely interacts with the endogenous histone deacetylase (HDAC)-binding cleft of the Rb proteins (e.g. [35]) and disrupts the interaction involved in repressing E2F/DP1 bound promoters. The relief of E2F/DP1 promoters from this repression favors cell-proliferation, which in turn might favor oncogenic transformation and co-replication of the viral genome. In certain cases additional domains fused to the LXCXE domain might catalyze further modifications or interactions with other chromatin proteins (see below). An example of the former process is phosphorylation of the Rb protein by herpesviral UL97 kinases, which results in disruption of Rb interactions with HDAC and E2F and consequent progression of the cell to the S-phase [40].

More recently another potential adaptor domain, which is found in both viral and cellular chromatin proteins, namely the BEN (BANP/SMAR1, E5R and NAC1) domain was identified [59] (Fig. 2B). In cellular proteins, this domain is found typically associated with a variety of other protein domains involved in interactions with chromatin proteins such as the POZ, C2H2-Zn finger and C4DM domains. The version in the human protein SMAR1 has been implicated in interactions with histone deacetylases HDAC3 and HDAC4 [73]. It appears to have been independently acquired by chordopoxviruses and polydnviruses. In the former it is typified by the vaccinia virus protein E5R, which is an abundant early component of the virosome and these proteins typically have 3–4 tandem repeats of the BEN domain. Given that the virosome is the center for replication of the poxviral genome it is conceivable that the BEN domain proteins are required for the organization of the chromatin in the virosome [59]. They could also be potentially involved in the process by which poxviruses recruit host chromatin proteins to the cytoplasmic viral replication complexes [74]. However, it should be noted that vertebrates possess an endogenous protein closely related to the viral E5R-like proteins (e.g. human KIAA1553). Hence, it is conceivable that the poxviral E5R-like proteins also play a role in interfering with host chromatin functions by mimicking KIAA1553 orthologs. Consistent with this the *Mollusum contagiosum* virus shows a recent displacement of its E5R by a version acquired from its host [59]. In polydnviruses BEN domain proteins might be encoded by multiple genes from several DNA circles (e.g. in *Cotesia congregata* bracovirus) and might similarly help in organizing viral chromatin. Some of these versions are fused to an RNaseT2 domain suggesting an additional role in processing chromatin-associated transcripts (Fig. 2B). A possible role in interactions with host chromatin is also suggested by the observation that a viral BEN domain encoding gene is one of those transferred to the host wasp's genome [59].

Other potential adaptor domains that might mediate protein-protein interactions pertinent to chromatin are more restricted in their distribution. One of the better studied of such domains is the Zn-finger domain found C-terminal to the LXCXE module in the papillomavirus E7 proteins. This domain has been implicated in interaction with Rb (with low affinity), the

histone acetyltransferase CBP/p300 and the cyclin-dependent kinase inhibitor p21 (CIP1) [33, 75, 76]. The NMR structure of this domain has been determined and it was claimed that it represented a unique fold [76]. However, structural analysis of this E7 domain revealed that it is a derived form of the PHD finger specifically related to a class of cellular PHD fingers typified by the Pygopus protein [77] (Fig. 3). However, in E7 the N-terminal portion and one of the pair of Zn-chelating elements of the ancestral fold have been lost (Fig. 3). This observation indicates that the E7 Zn-finger domain might at least in part interact like the ancestral PHD finger from which it has been derived. Another adaptor module unique to papillomaviruses is the Zn-binding domain found in their highly conserved E6 protein [42]. E6 proteins from avian and chelonian papillomaviruses have a single copy, whereas those from mammals show a duplication of this Zn-binding domain [78].

Analysis of the sequence alignment and NMR structure of this domain reveals that it contains a derived version of the treble clef fold in which a helical insert has emerged immediately after the N-terminal Zn-chelating flap of the fold (Fig. 3) [79]. Sequence analysis also indicates that another viral adaptor domain, the Zn-binding domain of the adenoviral E1A protein (called CR3 region of E1A [33]) shows a conservation pattern similar to the A20 Zn-finger [80] and is likely to adopt a treble clef fold (Supplementary material). Cellular treble clef fold domains such as the PHD and ZZ mediate specific interactions in the context of chromatin, whereas other such as the RING, UBP and A20 finger mediate interactions in the context of ubiquitination [6, 79]. Roles of the viral treble clef domains are consistent with their cellular counterparts – papillomavirus E6 recruits a ubiquitin E3 ligase to degrade p53 and also interacts with other chromatin proteins such as the histone acetyltransferases CBP and p300 and also the telomerase [81]. While E6 has also been implicated in several other interactions their significance in viral pathogenesis remains unclear. The predicted adenoviral E1A CR3 treble clef domain has been shown to be a major factor in regulating viral transcription by comprising an activation domain that interacts with the host mediator complex protein Med23 [33]. It would be interesting to test if it might also mediate ubiquitin related interactions as suggested by its predicted structure.

Adaptor domains mediating DNA-protein interactions

Viral adaptor domains, which mediate specific interactions between DNA and viral or host chromatin proteins, are best known from animal viruses. While such proteins are likely to be far more widespread, rapid divergence of these proteins and a paucity of studies in other viral systems have limited the information on them elsewhere in the viral world. One domain which appears to be widespread across animal DNA viruses is the domain with the RNA-recognition motif (RRM)-like fold present in certain key DNA binding proteins such as E2 from papillomaviruses and LANA/EBNA1 from vertebrate herpesviruses [29, 51, 52, 82]. In both these viral groups this domain binds DNA sequences associated with the viral replication origins, tethers the viral chromosome to host mitotic chromosomes and allows its efficient partitioning and persistence in host cells. In this regard they are archetypal adaptors of this class that bind viral DNA via the RRM-fold domain and interact with host nucleosomes via the poorly conserved interaction module in the remainder of the protein. In the case of the E2 protein this interaction occurs via the BRD4 protein which binds acetylated histones through its bromodomains and the E2 protein by the C-terminal helical Br-C domain [83]. In the case of LANA, interactions have been observed with paralogous bromodomain proteins such as BRD2 and the acetyltransferase CBP [84, 85]. In addition to recruitment of cellular replication factors (e.g. the origin recognition complex and MCM3) to the viral genome the adaptor role of these proteins is also critical to regulate viral gene expression. The E2 protein's interaction with BRD4 is required for trans-activation of papillomavirus promoters, whereas LANA's interaction with the CBP/p300 proteins is involved in repression of certain promoters during latency which is observed in

herpesviruses [86]. Further, LANA has also been implicated as an adaptor that recruits DNA methylases (especially DNMT3a, the *de novo* methylase) to host genes to facilitate *de novo* methylation of promoters and their repression [82, 87]. To date LANA/EBNA1 are not detectable in the ostreid herpesvirus which infects molluscs. However, we found a virus related to the ostreid herpesvirus inserted into the chromosome of its host amphioxus (*Branchiostoma*), which suggests that even this lineage of herpesviruses is likely to have a comparable protein that allows it to associate with host chromosomes (Supplementary material).

Studies on herpesviruses have uncovered a rich diversity of lineage-specific DNA-binding adaptor proteins with several roles in assembling particular chromatin states on viral and cellular promoters [32, 82]. One of these, RTA/BRFL1/ORF50 [88], contains a conserved N-terminal DNA-binding domain that is unique to herpesviruses with a 6-helical bundle fold (supplementary material). In certain herpesviruses infecting New World monkeys (e.g. Saimiriine herpesvirus 2) this domain is further combined with a downstream DNA-binding AT-hook motif [89]. These proteins bind various viral promoters and via their C-terminal LXXLL motif recruit the histone acetyltransferase CBP to transactivate gene expression specific to the lytic phase of the herpesviral lifecycle [90]. RTA might also interact with other cellular transcription factors such as OCT1 to recruit activating chromatin complexes to viral promoters [91]. Like RTA, another herpesvirus protein, VP16, plays an important role as a DNA-binding adaptor that allows activation via a series of transcription factors (e.g. Oct1) that bind viral promoters [92]. It has a unique structure with two domains, one centered on two large helices forming an antiparallel coiled-coil and a C-terminal $\alpha+\beta$ domain with 4 strands. Both domains cooperate to form multiple interaction interfaces, such as a basic interface for DNA and other for cellular proteins. Among its cellular partners are CBP and the chromatin remodeling SWI2/SNF2 ATPases BRG1 and BRM which help in keeping the viral genome in an euchromatin state (typified by modifications like H3K9Ac, H3K14Ac, H3K4me2/me3) for active transcription of early genes upon infection [32]. It also forms a complex with the host protein HCF that allows it to recruit host transcription factors like Oct1 to viral promoters. One of the early herpesviral genes induced upon infection, ICP8 [53, 54], encodes a protein with an OB-fold DNA-binding domain that is related to the phage SSB proteins [54, 93, 94]. The herpesviral version is however decorated by several α -helical insertions that form additional interfaces for interactions with host chromatin proteins. Like its phage counterparts, ICP8 binds ssDNA, is important for recombination and replication and is a key adaptor in recruiting the remodeling ATPases BRG1, BRM and hSNF2H, and recombination and replication proteins to viral chromatin [95]. It also appears to be required for transcription regulation due to its role in recruiting cellular chromatin opening complexes [32].

Conserved domains in viral DNA structuring proteins

Evidence from diverse dsDNA viruses indicates that DNA in virions is packed at considerably high density without being associated with nucleosome-like structures. However, paradoxically, the above discussion shows that several medium to large-sized eukaryotic viruses and a wide range of bacteriophages encode their own homologs of DNA-packaging proteins such as histones, IHF/HU, MC1 and *Chlamydia* HC1 [96]. To date there is no evidence for these proteins being packaged into virions [23, 24]. This suggests that they might provide an alternative packaging for the viral DNA inside the cell. Studies on other DNA viruses indicate that they have several mechanisms to ensure that they are either coated by histones modified to maintain an active chromatin state (adenoviruses and herpesviruses) or have regions of DNA free of nucleosomes (polyomaviruses) [32, 33]. This observation suggests that several large DNA viruses could use their self-encoded DNA-packaging proteins to maintain favorable chromatin states for their transcription and

replication machinery. It is also noted that in some viruses, such as members of the NCLDV clade, viral DNA is maintained separate from the host DNA in distinct bodies such as the virosome or virus factories [74]. In such cases presence of virally encoded DNA-packaging proteins probably provides a means to allow specialized packaging of DNA in these bodies.

In some cases, viral DNA-packaging proteins belong to the same family as those used by their hosts (e.g. histone-fold proteins). However, in other cases, like IHF/HU in ASFV, MC1 in phycodnaviruses (including Mimivirus) and *Chlamydia* HC1 in the *Thermus* phage phiYS40 (gi: 118197626, YS40_006), the viral DNA-packaging protein is unrelated to that of the host (Fig. 1). Instead they appear to have been derived from the DNA-packaging protein of an entirely different superkingdom (e.g. IHF/HU from bacteria and MC1 from archaea) or from a different phylogenetically lineage (e.g. HC1 is typical of chlamydiae and bacteroidetes rather than *Thermus*) [26]. The histones of baculoviruses (e.g. the H3/H4 fusion protein observed in the *Heliothis zea* virus 1 protein HZV1gp001 of the nudivirus lineage) and polydnaviruses (H4 of the bracovirus lineage), are most similar to the corresponding insect histones. Similarly, the histone H2B of the herpesvirus related to the ostreid herpesvirus integrated into the amphioxus genome (gi: 260792362, BRAFLDRAFT_62176, supplementary material) is most similar to the amphioxus version. These examples suggest that the histones found in these viruses are recent derivation from their hosts. However, in these cases the histone tails of the viral sequences have markedly diverged from the host versions. This suggests that these viral histones are probably refractory to modifications typical of host histones and thereby allow maintenance of chromatin states favorable to viral processes. Likewise, the frequent adoption of chromosomal proteins distinct from that of the host might also help the viral chromatin to be unaffected by host chromatin remodeling enzymes, thereby circumventing host heterochromatinization based defenses. Presence of other chromosomal proteins, like those with HMG domains [97], indicates that in some viruses (e.g. in iridoviruses and phycodnaviruses) these domains could potentially mediate via DNA-bending further higher order structuring of chromatin. A distinct virally encoded histone-related function is performed by the ICP11 family of WSSV [60]. The globular domain conserved in these proteins has a RRM-like fold reminiscent of the comparable domain in the papillomavirus E2 and herpesviral LANA/EBNA1 proteins, but in this case it interacts with host histones to inhibit their interactions with DNA. This might represent a novel mechanism to modulate the assembly of nucleosomes and the deposition of condensed chromatin on host and viral chromosomes [60].

Though DNA in virions is largely naked, there is evidence in several viruses for key interactions with certain proteins that condense and tether DNA inside capsids [22, 25]. In adenoviruses such a role has been demonstrated for the protein VII which condenses DNA inside the capsid and upon infection remains attached to the viral genome as the early DNA-packaging protein. It is eventually displaced by the E1A protein, which recruits host factors to establish a euchromatin-like state [25]. Protein VII is widely conserved in adenoviruses and contains a characteristic N-terminal all β -strand domain with a 'GWG' signature followed by a positively charged poorly conserved region, resembling histone tails, which probably non-specifically contacts DNA (supplementary material). Interestingly, several bacteriophages of firmicutes and iridoviruses encode small proteins that almost entirely consists of a domain with a HEH fold [28, 48]. These bacteriophage proteins are related to the HEH fold domains, SAP and LEM, which are found in eukaryotic cellular proteins [48]. These latter domains are specifically involved in tethering and recruiting various catalytic domains to matrix/scaffold attachment sites on DNA or tethering chromosomes to the nuclear membrane [48]. In bacteriophages, the HEH domain proteins are consistently associated in a conserved gene neighborhood with the major capsid protein and are occasionally fused to it (Fig. 4). Earlier studies had indicated that genes encoding proteins

involved in packaging DNA into capsids (e.g. the terminase subunits and the capsid proteins) show a statistically significant tendency to occur either close to the start or the center of the viral chromosome [28]. An identical trend was seen in the case of genes encoding phage SAP domain proteins (Fig. 4), strongly suggesting a role in DNA packaging in capsids. We propose that like their eukaryotic counterparts they might be involved in tethering the viral chromosome at certain points to the capsid. It is also possible that, as in the case of the adenoviral protein VII, the HEH domain proteins additionally participate in viral DNA packing within cells.

Domains found in proteins mediating ATP-dependent chromatin remodeling, manipulation of chromosomes and topoisomerases

Widely distributed viral P-loop ATPases—All cellular life forms possess an orthologous set of proteins typified by Rad50/SbcC, an ABC ATPase of the SMC superfamily, and its nuclease partner Mre11/SbcD of the calcineurin-like phosphoesterase fold, suggesting that this complex was already present in the last universal common ancestor of all life forms [98]. Several bacteriophages and some NCLDV possess homologous ATPase and nuclease partners (e.g. Phage T4 gp46 and gp47 respectively) [99]. This family of ABC ATPases is characterized by the presence of a variable-sized coiled-coil insert within the P-loop ATPase domain, with which it dimerizes to form a filamentous structure with two ATPase heads [100]. Studies on Rad50 have indicated that these dimeric filaments are critical in tethering distantly located dsDNA units (e.g. ends of two separate chromosomes), with each ATPase head contacting one of the units that are being bridged [100]. This bridging function of Rad50 appears to be critical in bringing chromosomes together for non-homologous end joining and a comparable role has been reported for SbcC and gp46 in processing dsDNA ends for processing during recombination [101, 102]. Our analysis shows that the majority of viruses (70%) with SbcC/gp46-like proteins possess linear chromosomes. This suggests that it might have a role in bridging chromosome ends during the process of recombination-dependent-replication, which is a mechanism to solve the end replication problem of linear viral genomes [99]. Consistent with this, analysis of genomic contexts of SbcC/Gp46 homologs in various bacteriophages shows a strong linkage, in addition to the gene for the Mre11/SbcD-like nuclease subunit (gp47), to various replication proteins such as the DNA clamp, clamp-loader ATPase, primase and certain Holliday junction resolvases (Fig. 4). Unlike their cellular counterparts, the coiled-coil insert in phage SbcC homologs is much shorter. This might be reflective of the shorter lengths of viral chromosomes and their denser packing in viral replication centers compared to cellular chromosomes. Some viruses, such as the caudate phage P2, possess a related ABC ATPase with a fused TOPRIM domain nuclease (e.g. P2 OLD protein) [103, 104] and it might play a comparable role to the classical Rad50/SbcC-like proteins in these phages.

The most widespread P-loop ATPase modules encoded by viruses are families of SFII helicases. Versions pertaining to chromatin related functions belong to two major families: 1) SWI2/SNF2 ATPases, which are distinguished by a characteristic motif N-terminal to the core SFII helicase module and 2) A18/UvsW ATPases typified by the A18R helicase of the vaccinia virus and the UvsW helicase of phage T4. In eukaryotes most SWI2/SNF2 ATPases form the core of several paralogous chromatin remodeling complexes which mediate displacement of nucleosomes, while a few function as helicase promoting duplex annealing [6, 14, 105]. In bacteria, the role SWI2/SNF2 ATPases is much less clear, but certain versions, like RapA, have been implicated in recycling the RNA polymerase complex stalled at transcription termination sites [106]. Two sub-types of SWI2/SNF2 ATPases are observed in dsDNA viruses, namely one group related to their cellular counterparts and another virus-specific group typified by the vaccinia virus D6R and D11L proteins [26]. Functional studies on D11L from vaccinia indicate that it is involved in

transcription termination [107] suggesting that at least a subset of these virus-specific SWI2/SNF2 ATPases are likely to play a role comparable to the bacterial RapA proteins. This observation suggests that a role in transcription termination could have potentially been an ancestral feature of the SWI2/SNF2 family. Gene context analysis indicates that in certain bacteriophages such as *Pseudomonas* phage 201phi2-1, the gene for the SWI2/SNF2 ATPase is linked to those for RNA polymerase subunits (Fig. 4). Hence, it is likely that these versions also play a role similar to the bacterial RapA protein in polymerase recycling upon termination. In contrast, the poxviral D6R protein has been shown to heterodimerize with the NCLDV-specific transcription factor (vaccinia A8L/variola A7L) and play a role in initiating transcription by the viral RNA polymerase [108, 109]. This function is generally more in line with certain eukaryotic cellular SWI2/SNF2 ATPases and the D6R protein might help in opening up viral chromatin to facilitate transcription. Given the conservation of D6R and its functional partner A8L throughout NCLDVs, it is likely that such a function was ancestral to the NCLDVs. In other bacteriophages SWI2/SNF2 ATPases are combined in conserved gene neighborhoods with genes encoding a distinctive nuclease of the restriction endonuclease fold and certain replication proteins (Fig. 4) [110], suggesting that these versions might function as conventional helicases in replication or associated recombination events.

The A18/UvsW family is one of the most widely distributed families of SFII helicase modules in DNA viruses and representatives from both bacteriophages and NCLDVs have been functionally characterized (Fig. 1). Vaccinia virus A18R is a negative regulator of transcription elongation and its helicase action is required to release the transcript from DNA upon transcription termination [111]. An ortholog of A18R is conserved throughout the NCLDV clade [26]. The phage T4 UvsW protein is a branch-specific helicase that is required for invasion of single strands and migrations of branches, during the replication-associated recombination process observed in these phages [99]. These observations suggest that the A18/UvsW helicases are likely to function as conventional helicases that promote strand invasion or annealing, both in DNA-DNA and DNA-RNA duplexes. Analysis of conserved gene-neighborhoods in phages indicates that A18/UvsW genes are in several cases combined with primases and D5-like helicases involved in replication (Fig. 4). In these cases they might help in releasing the RNA primer just as transcripts in poxviruses. However, in most other phages they are linked to a variety of genes for recombination related proteins, suggesting that they function similar to T4 UvsW. These observations indicate that there was possibly a functional shift that occurred in the NCLDV clade, wherein the A18 helicases were adapted for transcription related roles.

Phage-specific ATP-dependent recombinases and their partners—The importance of recombination in replication is consistent with the presence of proteins in several of bacteriophages for the structuring and coating of viral chromosomes during this process. These proteins function, either in conjunction with the host ATP-dependent recombinase RecA, or with a viral homolog of RecA such as UvsX of T4. These proteins bind ssDNA, similar to the functionally comparable OB fold protein SSB, and act as single strand annealing proteins (SSAPs) [93, 94]. A number of these proteins are known from diverse bacteriophages, namely RecT (from an enterobacterial prophage), phage λ Red β , phage P22 ERF, viral homologs of the eukaryotic SSAP Rad52 and T7 gp2.5 [31]. Structure determination of these proteins revealed that gp2.5 and SSB adopt an OB-fold whereas the Rad52 adopts a dsRBD fold [93, 94]. Analysis of the predicted secondary structure of RecT/Red β and ERF reveals that they possess the same conserved $\alpha\beta 3\alpha\alpha$ core as that seen in Rad52 [112], suggesting that they too adopt the dsRBD fold and have diverged from a common ancestor. In contrast to the widespread presence of these SSAPs, homologs of RecA itself were only known from a relatively small group of phages [113]. A more detailed analysis of ATPases encoded by phages (this study) showed that a larger number of phages

infecting firmicutes, actinobacteria, proteobacteria, planctomycetes and chlamydiae possess a novel conserved P-loop ATPase typified by the Sak4 protein of the *Lactococcus* phage phi31 [114]. The Sak4-like ATPase shows conserved signatures, such as a catalytic glutamate in the second core strand of the P-loop domain and an arginine-finger from a C-terminal β -hairpin, which are typical of the RecA superfamily [115]. Sak4-like ATPases are, however, distinct from classical RecA proteins in containing a rather compact and minimal version of the ATPase domain. Analysis of the gene contexts of these Sak4 ATPases reveals that they are consistently associated with genes for the OB-fold SSB (Fig.4) or are functionally associated with SSAPs such as ERF [114]. A previously uncharacterized C2C2 Zn-finger protein is encoded by a Sak4 gene neighbor especially in cases lacking a neighboring SSB or ERF gene and might be a distinct type of SSB or SSAP (Fig.4, supplementary material). Sak4 ATPases are also occasionally fused in certain firmicute phages to the RecB nuclease domain that displays a restriction endonuclease fold. Further, genes for Sak4 ATPases might also show linkages to genes encoding A18/UvsW family helicases and are commonly embedded in a larger gene neighborhood encoding viral replication proteins (Fig.4). Both UvsX and Sak4 like ATPase are predominantly (82%) found in phages with linear chromosomes. These observations strongly suggest that the Sak4 proteins along with the products of the neighbor genes might constitute a widespread ATP-dependent recombinase complex primarily required to solve the end replication problem of linear viral chromosomes.

Viral topoisomerases—Topoisomerases are primarily a feature of large DNA viruses (Fig. 1). Bacteriophage N4 (genome size: 70Kb) and the bovine papular stomatitis virus (genome size: 134Kb) are respectively the smallest bacterial and eukaryotic viruses containing a topoisomerase. Furthermore, over 90% of viruses with topoisomerases are linear genomes (supplementary material). This suggests that DNA manipulations mediated by these enzymes are predominantly adaptive in such genomes, where, in addition to end replication, alteration of linking number, supercoiling and decatenation of chromosomes might pose potential problems. Viruses possess either ATP-independent Type-I topoisomerases, which introduce single strand breaks, or ATP-dependent type-II topoisomerases, which introduce double-strand breaks and, in certain cases, both types (Figure 1) [13]. In evolutionary terms, type-IA and type II enzymes share a TOPRIM domain in their breakage/reunion module [104]. In contrast, the breakage/reunion module of type-IB enzymes belongs to the vast family of α -helical site-specific tyrosine recombinases which includes the integrase domain of caudate phages such as lambda and the bacterial XerC/D recombinase [116, 117]. Additionally, type II enzymes possess a bilobed ATPase module/subunit, which is related to ATPase modules of Hsp90, MutL and the MORCs [15, 118]. This module is in turn comprised of a histidine kinase-like ATP-binding module combined with a S5-like domain which provides a conserved “lysine finger” to the ATPase active site [15].

Type-IA enzymes, though universally found in cellular organisms, are very rare in viruses and are currently known only from the mimivirus (Top1A). Phylogenetic analysis suggests that it might have been recently acquired by this virus via gene transfer from an endosymbiotic bacterium in its *Acanthamoeba* host [13]. Type IB enzymes (Top1B) are known from all poxviruses, certain phycodnaviruses (including mimiviruses) and certain caudate phages (e.g. *Ralstonia* phage RSL1) (Fig. 1). In cellular organisms Top1B is found in all eukaryotes and sporadically across bacteria [13]. Studies on poxviral Top1B show that it is not required for replication; however, it is needed to organize viral chromatin for early transcription [119] probably via alteration of linking number and relaxing supercoiling. Unlike its cellular counterparts, it possesses sequence specificity for a 5'-(T/C)CCTT-3' pentamer [120]. This feature is reminiscent of α -helical site-specific recombinases, such as Cre, Lox and the lambda integrase [116], and suggests that Top1B could have emerged early

in viral evolution from the former class of enzymes. Type II topoisomerases with both ATPase and nuclease modules are seen across the entire NCLDV clade, except for most poxviruses [26]. The only poxvirus with a type II topoisomerase is the crocodilepox virus [121]. Interestingly, it displays a unique gene cluster with one complete Type-II topoisomerase gene (i.e. with both ATPase and TOPRIM containing modules) accompanied by six other genes encoding only the Hsp90-like ATPase modules [121]. However, the complete type-II topoisomerase homolog in this virus apparently lacks the conserved tyrosine suggesting that it only possesses nuclease but not religation activity. Among bacteriophages, Type-II topoisomerases are seen in large caudate phages such as T4 and have served as models for the biochemistry of type II enzymes [113]. Interestingly, as in the crocodile poxvirus, several bacteriophages like T4 have an additional gene rIIA, frequently adjacent to the topoisomerase gene (Fig. 4), which encodes a standalone homolog of the ATPase module [122]. Widespread presence of type-II topoisomerases in both the NCLDV clade and large bacteriophages suggests that decatenation and relaxation catalyzed by these proteins is critical for the successful propagation of these viral chromosomes, just as type-II enzymes are absolutely necessary for the propagation of cellular chromosomes [123].

At least in NCLDVs, the ancestral virus is inferred to have possessed a type-II topoisomerase, which in turn could be related to version found in bacteriophages [26]. However, its distribution in poxviruses is puzzling and might indicate a secondary reacquisition for a different role (especially given the lost the tyrosine). The exact role of the additional standalone ATPase modules like rIIA in phages and the multiple copies in crocodile poxvirus remains unclear. The multiplicity in the latter case points to a possible interaction with host defenses. Similarly Top1B might have been present in the ancestral NCLDV, although if this was the case it appears to have been lost more frequently.

Viral enzymes that covalent modify chromatin components

Viral Protein-modifying enzymes—Several viruses possess their own enzymes that modify either DNA or chromatin proteins. With a few exceptions most of these enzymes have been poorly characterized. The best studied of these is the SET domain histone methyltransferase from the *Paramecium bursaria Chlorella* virus (PBCV), a phycodnavirus. This protein has been shown to have profound regulatory effects on the host by catalyzing histone H3K27 methylation [34]. This modification has been proposed to recruit *Chlorella*'s polycomb complex for host chromosome heterochromatinization and accumulation of cells at the G2/M cell cycle transition which is favorable for viral replication. SET domain histone methylases are found across the phycodnavirus clade (including mimiviruses)[26] suggesting that such regulatory processes might have a widespread role in this clade. The phycodnaviruses also often possess a SWIB domain protein whose cellular homologs are protein-binding components of chromatin remodeling complexes [124, 125]. In several bacteria, such as *Chlamydia*, a SET domain protein is found along with SWIB domain protein and the two proteins along with Top1A are likely to constitute a minimal chromatin remodeling complex [124]. This, observation raises the possibility that even in the phycodnaviruses the SET and SWIB domain proteins might cooperate to form a minimal histone methylating complex. Iridoviruses also encode a SWIB domain protein but lack a SET domain methyltransferase [26]. It is plausible that in this case the viral protein recruits a host partner to direct specific chromatin modifications. Another well-studied viral chromatin modifying protein are the herpesviruses RING finger proteins typified by ICP0 protein. These proteins are ubiquitin E3 ligases [56] and its targets might include cellular chromatin proteins such as PML which it destabilizes [126]. It also interacts with several host histone deacetylases, HDAC5/6/7, and dissociates HDAC1 from the coREST/REST complex [127, 128]. Consistent with this ICP0 is critical for establishment of a euchromatic state for the viral chromosome [129]. Several poxviruses contain a protein

(variola D6R) which is expressed in virus factories and combines an N-terminal Kila-N domain (see below) a CCCH and RING finger domain [130]. It has been shown to bind DNA and inhibit host apoptosis [131] raising the possibility that it might define another host chromatin modifying viral ubiquitin E3 ligase. Herpesviruses also encode two distinct protein kinases that modify host chromatin proteins. The US3-type kinases have been implicated in phosphorylation of HDAC1/2 [132] whereas the UL97 kinases phosphorylate Rb [40]. While several NCLDVs also encode protein kinases it is yet unclear if they have a role in modifying chromatin proteins.

There are several other potential chromatin protein-modifying enzymes that are poorly understood. PBCV-like chlorellaviruses (PBCV: A654L) and the *Melanoplus sanguinipes* entomopoxvirus (MSV192) encode NH₂-acetyltransferases (GNATs) that could potentially acetylate histones. Likewise, the eukaryotic polyADP-ribose polymerase encoded by *Anticarsia gemmatilis* nucleopolyhedrovirus and another divergent ADP ribosyltransferase encoded by the Invertebrate iridescent virus 6 (an iridovirus; supplementary material) could potentially carry out ADP-ribosylation of histones or other nuclear proteins [8, 58]. However, a direct role for these in regulating host or viral chromatin is yet to be demonstrated. Interestingly, both iridoviruses as well as certain phycodnaviruses, such as *Ostreococcus* virus OsV5, encode homologs of the FCP1 phosphatase subunit of TFIIF. These phosphatases have a catalytic domain of the haloacid dehalogenase fold (HAD) and mediate the dephosphorylation of the RNA polymerase CTDs [133]. Given that viral RNA polymerases do not have the CTDs seen in their cellular counterparts, it is conceivable that the FCP1 acts on host CTDs, perhaps to release the associated mRNA capping enzymes [134]. The mimivirus encodes two proteins with the jumonji-related (JOR/JmjC) domain, which is the catalytic domain of protein demethylases [6, 135]. It is conceivable that these domains, like their cellular counterparts, catalyze demethylation of histones probably as a part of regulating host chromatin function. Much less is known of the protein-modifying functions in bacteriophages. Several bacteriophages encode GNATs, but it is unclear if they acetylate host or viral proteins in the context of chromatin dynamics or viral DNA packaging. The *Synechococcus* phage syn9 also encodes a JOR domain protein (gi: 113200554; gp49) related to the eukaryotic chromatin protein modifying asparagine hydroxylases (supplementary material), but again its role remains unclear. Several phages encode ADP ribosyltransferases that are distantly related to PARPs and the toxins which confer advantages to their host bacteria. Earlier gene context analysis of these enzymes suggests that ADP ribosylation of as yet unknown target proteins might have a role in packaging of viral DNA [28]. T5-like phages possess a Sir2 domain protein, the eukaryotic and archaeal counterparts of which are involved in chromatin protein deacetylase and ADP ribosylation activities [136]. However, the targets of these phage proteins remain unknown.

Viral DNA-modifying enzymes—DNA-modifying enzymes are best known from bacteriophages and catalyze a spectacular array of modifications that include N6-methyladenine, 5-methylcytosine, 5-hydroxymethylpyrimidines and their mono- or diglycosylated derivatives, α -putrescinylation or α -glutamylated thymines, sugar substituted 5-hydroxypentyluracil, and N6-carbamoylmethyl adenines (called Momylation after the phage Mu mom) [9, 50]. These viruses resort to one or both of two alternative means of modifying DNA: 1) One mechanism seen in T-even and certain Gram-positive phages is the prior synthesis of modified nucleotides followed by their incorporation into DNA during replication. Examples of this are the phage 5-hydroxymethylcytosine and 5-hydroxymethyluracil synthases that are related to classical thymidylate synthases [137]. 2) Direct *in situ* modification of bases in DNA. The most common of these are Dcm and Dam methylases that respectively methylate cytosine and adenine and are widely distributed in several caudate phages (this work). The Mom family of enzymes, typified by the phage mu Mom protein, directly modifies adenines in DNA by adding a carbamoylmethyl or a related

adduct, and has a catalytic domain of the GNAT fold [49]. Of these two means of DNA modification, incorporation of modified nucleotides in DNA results in near-complete substitution of the canonical base, whereas, *in situ* DNA modifications only change a fraction of the same. In T-even phages the pre-synthesized 5-hydroxymethylpyrimidines are subject to further modifications by two distinct families of glycosyltransferases. An initial glycosylation is catalyzed by either the alpha-glucosyltransferase or beta-glucosyltransferase of the glycogen synthase/glycogen phosphorylase fold [138]. In some viruses a second sugar moiety is added by beta-glucosyl-HMC-alpha-glucosyltransferase of the Fringe-like glucosyltransferase fold [139]. We recently showed that the gp2 protein from the actinophages Cooper and Nigel which infect mycobacteria and *Frankia* is an *in situ* 5-hydroxymethylcytosine-generating enzyme [10]. We predict that it is likely to synthesize this base by dioxygenase action on 5-methylcytosine. We also suggest that the gp42 protein with the GNAT fold domain from the *Pseudomonas* phage 201varphi2-1 [140] might be required for the synthesis of an unusual base with a polyamine adduct. These diverse DNA modifications of bacteriophages have in large part been implicated in conferring immunity against host restriction enzymes. This is supported by the evidence for an “arms-race” situation with multiple successive modifications as seen in the T-even phages [9, 50]. However, in some phages the methyltransferase genes are linked to the ParB gene encoding a protein involved in chromosome partitioning. A similar linkage is also seen for the 5-hydroxymethylcytosine generating enzyme of the actinophages [10]. Indeed, in phage P1 an epigenetic role for DNA methylation by its Dam methylase has been demonstrated in viral DNA partitioning (during replication in a plasmid form) and transcription regulation [141]. This suggests that an epigenetic role for DNA modifications might be far more widespread in viruses than currently known.

In *Chlorella* viruses the endogenous restriction-modification systems protect their own DNA from degradation while cutting the unmethylated host DNA with the corresponding restriction enzymes [63]. We have previously identified several distinct DNA methyltransferases in chlorophyte algae that are not found in other eukaryotic lineages [6]. It is possible that these methylases were acquired by these algae as a defense against the restriction enzymes of the phycodnaviruses. This might also explain the proliferation of such restriction-modification systems in the viral genome by virtue of the arms-race with the host. The iridovirus methyltransferases are related to bacterial Dcm methylases and are predicted to methylate cytosine [64, 142]. This is consistent with an early observation that a significant fraction of the cytosines in the iridoviral genome are methylated in a pattern distinct from the host genomes [64]. This methylation could be mediated by the virally encoded cytosine methylase and could have major consequences for epigenetic regulation of viral chromatin.

Viral transcription factors

Among the smallest dsDNA viruses specific transcription factors are usually “all-purpose” DNA binding proteins that also double as adaptor proteins in chromatin organization. Such proteins are exemplified by the papillomavirus E2 protein. As seen above, such multifunctional, virus-specific, DNA-binding adaptors that act as transcription regulators are also known from moderately sized and large viruses. However, in NCLDV there is at least one well-conserved dedicated basal transcription factor prototyped vaccinia A8L/variola A7L that is required for transcription of at least the early genes by the viral RNA polymerase [26, 143]. This protein appears to have distinctive set of conserved cysteines that suggest that it is a metal-chelating protein with an unusual fold [26]. In addition to this virus-specific basal transcription factor, some NCLDV, namely PBCV and the mimivirus also possess divergent homologs of the cellular basal transcription factors TBP and TFIIB [26]. This suggests that they might additionally assemble a minimal version of the RNA

PolII-type basal transcription factor complexes on some of their promoters. Most NCLDV also have a Zn-ribbon protein related to TFIIS (e.g. vaccinia E4L) that might act as a transcription elongation factor [26]. Virus-specific transcription factors of prokaryotic viruses remain largely uncharacterized. Larger dsDNA viruses additionally encode a variety of transcription factors that are ultimately related to host transcription factors. In eukaryotic viruses these include: 1) the basic-leucine zipper (bZIP) and IRF-like TFs in herpesviruses [144, 145]; 2) a homeodomain in mimivirus [26, 146]; 3) C2H2 zinc finger TFs in Lymphocystis disease iridovirus, and *Neodiprion lecontei* nuclear polyhedrosis virus. The bZIP protein of herpesviruses (e.g. ZEBRA of EBV) has been extensively studied [147] and is a key activator of genes related to lytic development and repressor of latency-associated genes. It also appears to target host genes and thereby modulates anti-viral responses of the host. This protein binds host chromatin proteins such as CBP and SNF5 to differentially regulate chromatin structure at different promoters [119]. Herpesviral IRF-like proteins are divergent versions of the host interferon regulatory factor and possess a winged helix-turn-helix (HTH) domain combined to a C-terminal MH2 domain. They are often present in several paralogous copies in each viral genome (e.g. four paralogs in KSHV). Each viral version targets distinct host promoters and helps in blocking the antiviral interferon pathway at different steps and also repressing p53 mediated apoptosis [148]. Viral IRFs have also been implicated in mediating a variety of chromatin modifications by recruiting host enzymes [149]. Thus, host-derived transcription factors of eukaryotic DNA viruses appear regulate both viral and host gene expression.

Among bacteriophages the predominant specific TFs related to host TFs are HTH proteins. Examples of these proteins, such the phage lambda repressors cro and cI and activator cII, are the archetypal TFs of classical studies [150]. Like their herpesviral analogs these HTH TFs form a circuit to regulate the decision between lysogeny (latency) and the lytic cycle. Such repressors and activators are seen across a wide range of caudate phages and are encoded by a cluster of genes that are similarly organized in these diverse phages. This indicates that they were acquired very early in the evolution the caudoviruses, although on a few occasions they might have been displaced by homologous cellular TFs. Sequence analysis of these proteins suggests that they are rapidly diverging probably reflecting divergence of their target promoters to avoid cross-binding between different viruses infecting the same cell. On several occasions they have also been acquired by cellular genome and recruited as TFs. Another derived version of the HTH fold, the ribbon-helix-helix (RHH or MetJ/Arc) fold [98] is also prevalent among the transcription regulators of several viruses. These TFs with RHH domains might have on occasions displaced HTH TFs and vice versa in particular phage genomes. A few phages of actinobacteria also encode WhiB-like TFs [151] which are unique to this lineage of bacteria, suggesting that they are more recent acquisitions from their hosts. Archaea also possess an abundance of RHH fold TFs [98] and likewise some of their phages also possess RHH TFs (e.g. the Lipothrixviruses) (Supplementary material). Furthermore, like archaeal cells, some of their viruses, namely lipothrixviruses, fuselloviruses and rudiviruses also possess small C2H2 transcription factors [152]. While the role of these TFs in these archaeal viruses remains unstudied, it is conceivable that they might be regulatory switches in the viral life cycle as in the case of their bacteriophage counterparts. The repeated acquisition of host-derived TFs in the three superkingdoms of life is reflective of a similar selective pressure across these vast evolutionary distances. This most probably arises from the need for viruses to use a TF similar to that of the host to co-opt much of the host transcription and chromatin machinery for their regulatory purposes.

Other than these transcription factors several large DNA viruses such as NCLDVs, baculoviruses and certain bacteriophages encode an enigmatic group of proteins with either of two conserved domains known as the Bro-N (typified by baculovirus Bro N-terminal

domain) and Kila-N (typified by bacteriophage P1 Kila N-terminal domain) [130]. In certain viruses like baculoviruses and entomopoxviruses these domains show major lineage-specific expansions [130]. Both these domains might possess a similar structure and potentially bind DNA [131, 153, 154]. They are also combined with a variety of other domain include nucleases of the endonuclease fold and HTH domains, which are generally consistent with a DNA-binding role. In the case of baculoviruses they have been shown to localize to the host nucleus and associate with chromatin [154]. However, their exact significance in regulating viral or host transcription remains poorly understood. Hence, further investigation of their role as virus-specific TFs or chromatin associated adaptors would be of considerable interest.

EVOLUTIONARY CONSIDERATIONS ON VIRAL CHROMATIN PROTEINS

Lifestyle strategies shape viral proteomes

A number of interesting general evolutionary issues are raised by viral chromatin proteins: 1) General life style strategies adopted by DNA viruses; 2) Acquisition and adaptation of chromatin proteins of host origin by viruses; 3) Origin of virus-specific chromatin proteins; 4) host chromatin-level adaptations to viruses; 5) possible contributions of viruses to evolution of host chromatin proteins. Irrespective of their lifestyle, interactions with chromatin are vital for DNA viruses. However, the lifestyles influence the type of chromatin proteins typically encoded by viral genomes. Viral lifestyles can be understood at basic level by utilizing a model founded on the r/K selection theory from ecology (r and K are respectively constants in the logistic equation that indicate growth rates and carrying capacity) [155]. Smaller viruses typically follow a r-strategy; i.e. they aim at rapid replication and tend to depend mainly on producing a large number of copies for their survival. As a consequence they do not encode specialized chromatin-modifying function, instead possessing a few multi-functional proteins that perform several chromatin-related functions. In contrast, with increasing genome size, viruses tend to follow the K-strategy and become good competitors – they are adapted to survive against both strong host counter-measures as well resource competition from other viral and cellular parasites [155, 156]. From the view point of the virus, latency or lysogeny could be seen as such a survival strategy. As a consequence they have evolved a wide range of transcription factors and chromatin proteins that allow fine-tuned regulation of both their own genes and host immune responses in several independent ways. Thus, with increasing genome size viruses are characterized by the emergence of a series of regulatory mechanisms that back each other up and allow them to persist against various parallel defensive mechanisms of the host (e.g. the multiple IRF proteins of herpesviruses of the rhadinovirus clade [148]).

Provenance of viral chromatin proteins

Due to rapid divergence, the evolutionary affinities of several viral proteins are not easily established. However, structure determination and a combination of in-depth sequence and structure analysis have helped remedy this to a certain extent [26]. As a consequence we observe that several virus-specific proteins such as the E7 and E6 Zn-finger of papillomaviruses and the E1A Zn-finger of adenoviruses are ultimately derived from host proteins. Hence, right from some of the smallest viruses there has been acquisition of domains from host chromatin proteins followed by their rapid divergence and adaptation for virus-specific functions. Earlier studies have indicated that growing viral genomes have accreted genes from their host, other viruses and other parasites/symbionts which share hosts with them [26, 27, 38]. All these sources of gene accretion have contributed to the spread of various chromatin related adaptations. For example, the spread of the LXCXE motif, Kila-N and Bro-N appears to have occurred through genetic exchange between diverse DNA viruses [130]. The DNA-packaging proteins such as MC1 and IHF/HU and Top1A in

eukaryotic viruses could have potentially emerged through transfer from co-occurring bacterial parasites or endosymbionts [26, 38]. On the other hand histones and transcription factors like TBP, TFIIB, bZIP (e.g. ZEBRA) and the vIRFs appear to have been acquired from the host by various viruses. In many cases like the phage HTH transcription factors these transfers might have happened early in evolution followed by a long history of propagation within viruses [98]. In certain cases due to rapid divergence and domain architectural changes the exact point at which they were derived from host genomes is unclear. Thus, even though the viral SET domain and protein kinases were acquired from the host, the precise host representatives that gave rise to them remain unknown.

Yet, there remains a set of virus-specific proteins whose origins cannot be explained on the basis of genetic transfer from cellular sources. Examples of such proteins include the E2/LANA, VP16, RTA/BRLF and the adenoviral protein VII. Some of these proteins such as E/LANA contain an ancient protein fold (in this case the RRM) that is seen in several other viral and cellular proteins [29, 52]. Hence it is plausible that they emerged among the earlier replicons that were progenitors of both cellular and viral genomes [26]. Subsequently, the viral representatives appear to have evolved largely independently of their cellular counterparts. In contrast, virus-specific domains that are predominantly α -helical (e.g. RTA/BRLF and VP16) or compositionally biased (protein VII) possess potentially unique folds that are not found in cellular counterparts. Hence, such domains could have indeed emerged *de novo* in viruses by elaboration of simple helical structures such as coiled coils. A similar phenomenon has been earlier noted in the emergence of lineage-specific α -helical and low-complexity cellular proteins [157].

Chromatin level anti-viral defenses and viral contribution to origin of host chromatin proteins

Given the prominent interaction between viral and cellular chromatin proteins it is not surprising that cells use certain chromatin-based defenses against viruses. From the cellular view point, the simplest defensive strategy is targeting the viral DNA for degradation. This appears to be the main selective force that has favored the proliferation of a large number of restriction-modification systems in prokaryotes [158]. In eukaryotes, with the possible exception of counter-measures against the restriction-modification systems of chloroella viruses in algae, such systems appear to be largely absent. However, several eukaryotes possess conserved DNA methylases that were ultimately derived from bacterial restriction-modification systems [6]. There is evidence that cytosine methylation of invading DNA, and subsequent modulation of promoters contained in it or its heterochromatinization might be an important eukaryotic defensive strategy [64]. On the other hand several eukaryotic viruses have evolved means to harness host-mediated methylation to their own advantage, both in terms of gene regulation and host evasion [159]. Determination of the methylomes of several DNA viruses lends some support to these ideas [160]. The recent discovery of 5-hydroxymethylcytosine raises the possibility of further modifications of these methylcytosines being important in viral chromatin regulation [161]. Indeed, targeting viral DNA for heterochromatinization by means other than methylation might serve as a more general host strategy for attenuating an infection. Consistent with this, the chromatin associated PML or ND10 bodies in animals could form a potential defensive innovation against various DNA viruses by possibly sequestering viral genomes and modulating their expression [162]. Not unexpectedly, several DNA viruses have evolved strategies to target proteins such as PML that are constituents of these complexes. Eukaryotes show a proliferation of chromatin complexes, many of which show partially overlapping functions. Recent work on chemical genetics has revealed that this network of chromatin complexes has an important role in tailoring the cellular responses to chemical stresses, with different complexes possibly specializing in different stresses [163]. In this light it appears possible

that the proliferation of eukaryotic chromatin complexes is also a response to the selective pressure from DNA viruses –having many functionally overlapping complexes that might allow the cell to evade different chromatin targeting strategies of viruses. At the same time it might also allow cells to use a wide range of alternative mechanisms to target viral chromosomes for heterochromatinization.

The results of comparative genomics also point to more direct contribution of viral proteins to cellular proteomes. As noted earlier in bacteria, viral transcription factors might have on occasions given rise to cellular transcription factors. Such a possibility also exists in the case of the DNA-binding proteins of certain eukaryotic transcription factors such as the AP2 superfamily [164], which could have been derived either from the DNA-binding domains of viral integrases or transposases. Of particular interest is the possibility that certain eukaryotic chromatin proteins have an ultimately viral origin. It can be shown that various components of the eukaryotic chromatin have been derived from both the archaeal and bacterial precursors of the eukaryotic cell. However, the DNA-binding HEH fold domain that tethers chromosomes to the nuclear membrane (e.g. in Src1) is not represented in archaea [48]. The conserved bacterial representatives of the HEH domain are fused to the transcription terminator ATPase Rho [48] and do not appear to have a role in chromosome organization. In contrast, the evidence suggests that the HEH domains of caudate bacteriophages might have a role in chromosome organization, in particular tethering of DNA in viral capsids. This suggests that eukaryotes perhaps acquired their ancestral HEH fold domain from a bacteriophage that might have infected the endosymbiotic mitochondrial precursor. Thus, the very origin of a key feature of eukaryotic chromatin might have a viral provenance. Such acquisitions might have also played a role in later events in course of eukaryotic evolution. Recent studies indicate that the animal Tet protein which catalyze hydroxylation of 5-methylcytosine and the JBP proteins that catalyze hydroxylation of thymine in trypanosomes might have had their origin from phage proteins that catalyze similar DNA modifications [10]. Furthermore, the Topoisomerase Top1B is absent in archaea and only sporadically distributed in bacterial genomes [13]. However, it is the primary type-I topoisomerase in eukaryotes. As discussed above, its relationship to site-specific recombinases, which are widespread in viruses, raises the possibility that this enzyme arose in the viral world. Consistent with this, the viral versions represent a minimal version of this enzyme in contrast to their cellular counterparts. If this were so it could also be a potential contribution of large DNA viruses to cellular genomes.

CONCLUDING REMARKS

Investigation of virus-chromatin interactions has been at the heart of elucidating cellular chromatin functions. Over the years these studies have contributed an enormous body of literature that is only further burgeoning due to the multitude of high-throughput technologies. At the same time, there have been numerous sequencing efforts that have left us with genomic sequences, but few experimental studies on the corresponding viruses. In this review we have attempted to bridge the two areas by providing a comparative genomic scaffold for viral chromatin studies. We primarily try to summarize key results of case-by-case studies using a combination of genomic and protein structural insights. Our hope is that this might help in placing this large body of information under a common evolutionary framework to investigate the emergence and development of virus-host interactions. We also hope that this study might help in highlighting some of the more ignored viruses and proteins with possible chromatin-related functions. A sobering conclusion emerging from comparative genomic analysis is that, outside of the major tumor viral groups, a large number of chromatin proteins still remain poorly understood. A few proteins from such less-studied viral lineages, e.g. the SET domain of PBCV, have already offered successful alternative models for understanding the biochemistry of key chromatin-modifying enzymes

[34]. Hence, we hope that some of the discussion presented here might help inspire new experimental studies that better harness data from across the viral world.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Work by the authors is supported by the intramural funds of the National Library of Medicine, NIH. Due to space considerations we have been regrettably unable to cite a large number of original works in this field.

References

- Swinger KK, Rice PA. *Curr Opin Struct Biol.* 2004; 14:28–35. [PubMed: 15102446]
- Sandman K, Reeve JN. *Curr Opin Microbiol.* 2005; 8:656–661. [PubMed: 16256418]
- Wong JT, New DC, Wong JC, Hung VK. *Eukaryot Cell.* 2003; 2:646–650. [PubMed: 12796310]
- White MF, Bell SD. *Trends Genet.* 2002; 18:621–626. [PubMed: 12446147]
- Schalch T, Duda S, Sargent DF, Richmond TJ. *Nature.* 2005; 436:138–141. [PubMed: 16001076]
- Iyer LM, Anantharaman V, Wolf MY, Aravind L. *Int J Parasitol.* 2008; 38:1–31. [PubMed: 17949725]
- Kouzarides T. *Cell.* 2007; 128:693–705. [PubMed: 17320507]
- Fuchs J, Demidov D, Houben A, Schubert I. *Trends Plant Sci.* 2006; 11:199–208. [PubMed: 16546438]
- Gommers-Ampt JH, Borst P, Faseb J. 1995; 9:1034–1042. [PubMed: 7649402]
- Iyer LM, Tahiliani M, Rao A, Aravind L. *Cell Cycle.* 2009; 8:1698–1710. [PubMed: 19411852]
- Berger SL, Kouzarides T, Shiekhhattar R, Shilatifard A. *Genes Dev.* 2009; 23:781–783. [PubMed: 19339683]
- Sasai N, Defossez PA. *Int J Dev Biol.* 2009; 53:323–334. [PubMed: 19412889]
- Forterre P, Gribaldo S, Gadelle D, Serre MC. *Biochimie.* 2007; 89:427–446. [PubMed: 17293019]
- Gangaraju VK, Bartholomew B. *Mutat Res.* 2007; 618:3–17. [PubMed: 17306844]
- Iyer LM, Abhiman S, Aravind L. *Biol Direct.* 2008; 3:8. [PubMed: 18346280]
- Losada A, Hirano T. *Genes Dev.* 2005; 19:1269–1287. [PubMed: 15937217]
- Allis, CD.; Jenuwein, T.; Reinberg, D. *Epigenetics.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2007.
- Berger SL. *Nature.* 2007; 447:407–412. [PubMed: 17522673]
- Wolffe AP, Guschin D. *J Struct Biol.* 2000; 129:102–122. [PubMed: 10806063]
- Casjens S, Sampson L, Randall S, Eppler K, Wu H, Petri JB, Schmieger H. *J Mol Biol.* 1992; 227:1086–1099. [PubMed: 1433288]
- Kim KS, Sharp DG. *J Immunol.* 1966; 97:197–202. [PubMed: 5921311]
- Yu TY, Schaefer J. *J Mol Biol.* 2008; 382:1031–1042. [PubMed: 18703073]
- Kuznetsov Y, Gershon PD, McPherson A. *J Virol.* 2008; 82:7551–7566. [PubMed: 18508898]
- Cyrklaff M, Risco C, Fernandez JJ, Jimenez MV, Esteban M, Baumeister W, Carrascosa JL. *Proc Natl Acad Sci U S A.* 2005; 102:2772–2777. [PubMed: 15699328]
- Chen J, Morral N, Engel DA. *Virology.* 2007; 369:411–422. [PubMed: 17888479]
- Iyer LM, Balaji S, Koonin EV, Aravind L. *Virus Res.* 2006; 117:156–184. [PubMed: 16494962]
- Filee J, Chandler M. *Res Microbiol.* 2008; 159:325–331. [PubMed: 18572389]
- Burroughs AM, Iyer LM, Aravind L. *Genome Dyn.* 2007; 3:48–65. [PubMed: 18753784]
- Hegde RS, Grossman SR, Laimins LA, Sigler PB. *Nature.* 1992; 359:505–512. [PubMed: 1328886]
- Iyer LM, Aravind L, Koonin EV. *J Virol.* 2001; 75:11720–11734. [PubMed: 11689653]

31. Iyer LM, Koonin EV, Aravind L. BMC Genomics. 2002; 3:8. [PubMed: 11914131]
32. Knipe DM, Cliffe A. Nat Rev Microbiol. 2008; 6:211–221. [PubMed: 18264117]
33. Berk AJ. Oncogene. 2005; 24:7673–7685. [PubMed: 16299528]
34. Muftaba S, Manzur KL, Gurnon JR, Kang M, Van Etten JL, Zhou MM. Nat Cell Biol. 2008
35. Brehm A, Miska EA, McCance DJ, Reid JL, Bannister AJ, Kouzarides T. Nature. 1998; 391:597–601. [PubMed: 9468139]
36. Izumiya Y, Ellison TJ, Yeh ET, Jung JU, Luciw PA, Kung HJ. J Virol. 2005; 79:9912–9925. [PubMed: 16014952]
37. Forterre P. Virus Res. 2006; 117:5–16. [PubMed: 16476498]
38. Raoult D, Forterre P. Nat Rev Microbiol. 2008; 6:315–319. [PubMed: 18311164]
39. Prangishvili D, Forterre P, Garrett RA. Nat Rev Microbiol. 2006; 4:837–848. [PubMed: 17041631]
40. Hume AJ, Kalejta RF. Cell Div. 2009; 4:1. [PubMed: 19146698]
41. Liu X, Clements A, Zhao K, Marmorstein R. J Biol Chem. 2006; 281:578–586. [PubMed: 16249186]
42. Nomine Y, Masson M, Charbonnier S, Zanier K, Ristriani T, Deryckere F, Sibler AP, Desplancq D, Atkinson RA, Weiss E, Orfanoudakis G, Kieffer B, Trave G. Mol Cell. 2006; 21:665–678. [PubMed: 16507364]
43. Ferreón JC, Martinez-Yamout MA, Dyson HJ, Wright PE. Proc Natl Acad Sci U S A. 2009; 106:13260–13265. [PubMed: 19651603]
44. Tucker PA, Tsernoglou D, Tucker AD, Coenjaerts FE, Leenders H, van der Vliet PC. Embo J. 1994; 13:2994–3002. [PubMed: 8039495]
45. Black BC, Center MS. Nucleic Acids Res. 1979; 6:2339–2353. [PubMed: 461192]
46. Chatterjee PK, Vayda ME, Flint SJ. Embo J. 1986; 5:1633–1644. [PubMed: 3743550]
47. Mohrmann L, Verrijzer CP. Biochim Biophys Acta. 2005; 1681:59–73. [PubMed: 15627498]
48. Mans BJ, Anantharaman V, Aravind L, Koonin EV. Cell Cycle. 2004; 3:1612–1637. [PubMed: 15611647]
49. Kaminska KH, Bujnicki JM. Cell Cycle. 2008; 7:120–121. [PubMed: 18204304]
50. Warren RA. Annu Rev Microbiol. 1980; 34:137–158. [PubMed: 7002022]
51. Viejo-Borbolla A, Ottinger M, Bruning E, Burger A, König R, Kati E, Sheldon JA, Schulz TF. J Virol. 2005; 79:13618–13629. [PubMed: 16227282]
52. Bochkarev A, Barwell JA, Pfuetzner RA, Furey W Jr, Edwards AM, Frappier L. Cell. 1995; 83:39–46. [PubMed: 7553871]
53. Tanguy Le Gac N, Villani G, Boehmer PE. J Biol Chem. 1998; 273:13801–13807. [PubMed: 9593724]
54. Mapelli M, Panjikar S, Tucker PA. J Biol Chem. 2005; 280:2990–2997. [PubMed: 15507432]
55. Hall KT, Stevenson AJ, Goodwin DJ, Gibson PC, Markham AF, Whitehouse A. J Virol. 1999; 73:9756–9763. [PubMed: 10559285]
56. Boutell C, Sadis S, Everett RD. J Virol. 2002; 76:841–850. [PubMed: 11752173]
57. Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE. Proc Natl Acad Sci U S A. 1996; 93:7481–7485. [PubMed: 8755499]
58. Koch-Nolte F, Reche P, Haag F, Bazan F. J Biotechnol. 2001; 92:81–87. [PubMed: 11640979]
59. Abhiman S, Iyer LM, Aravind L. Bioinformatics. 2008; 24:458–461. [PubMed: 18203771]
60. Wang HC, Wang HC, Ko TP, Lee YM, Leu JH, Ho CH, Huang WP, Lo CF, Wang AH. Proc Natl Acad Sci U S A. 2008; 105:20758–20763. [PubMed: 19095797]
61. Marmorstein R. Trends Biochem Sci. 2003; 28:59–62. [PubMed: 12575990]
62. Min J, Zhang X, Cheng X, Grewal SI, Xu RM. Nat Struct Biol. 2002; 9:828–832. [PubMed: 12389037]
63. Agarkova IV, Dunigan DD, Van Etten JL. J Virol. 2006; 80:8114–8123. [PubMed: 16873267]
64. Doerfler W. Epigenetics. 2008; 3:125–133. [PubMed: 18487948]
65. Dyson N, Howley PM, Munger K, Harlow E. Science. 1989; 243:934–937. [PubMed: 2537532]

66. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
67. Eddy SR. *PLoS Comput Biol.* 2008; 4 e1000069.
68. Li D, Zhao R, Lilyestrom W, Gai D, Zhang R, DeCaprio JA, Fanning E, Jochimiak A, Szakonyi G, Chen XS. *Nature.* 2003; 423:512–518. [PubMed: 12774115]
69. Luo X, Sanford DG, Bullock PA, Bachovchin WW. *Nat Struct Biol.* 1996; 3:1034–1039. [PubMed: 8946857]
70. Seif I. *Virology.* 1984; 138:347–352. [PubMed: 6093371]
71. Stubdal H, Zalvide J, Campbell KS, Schweitzer C, Roberts TM, DeCaprio JA. *Mol Cell Biol.* 1997; 17:4979–4990. [PubMed: 9271376]
72. Aronson MN, Meyer AD, Gyorgyey J, Katul L, Vetten HJ, Gronenborn B, Timchenko T. *J Virol.* 2000; 74:2967–2972. [PubMed: 10708410]
73. Rampalli S, Pavithra L, Bhatt A, Kundu TK, Chattopadhyay S. *Mol Cell Biol.* 2005; 25:8415–8429. [PubMed: 16166625]
74. Oh J, Broyles SS. *J Virol.* 2005; 79:12852–12860. [PubMed: 16188987]
75. Bernat A, Avvakumov N, Mymryk JS, Banks L. *Oncogene.* 2003; 22:7871–7881. [PubMed: 12970734]
76. Ohlenschlager O, Seiboth T, Zengerling H, Briesse L, Marchanka A, Ramachandran R, Baum M, Korbass M, Meyer Klaucke W, Durst M, Gorlach M. *Oncogene.* 2006; 25:5953–5959. [PubMed: 16636661]
77. Fiedler M, Sanchez-Barrena MJ, Nekrasov M, Mieszczynek J, Rybin V, Muller J, Evans P, Bienz M. *Mol Cell.* 2008; 30:507–518. [PubMed: 18498752]
78. Herbst LH, Lenz J, Van Doorslaer K, Chen Z, Stacy BA, Wellehan JF Jr, Manire CA, Burk RD. *Virology.* 2009; 383:131–135. [PubMed: 18973915]
79. Krishna SS, Majumdar I, Grishin NV. *Nucleic Acids Res.* 2003; 31:532–550. [PubMed: 12527760]
80. Wertz IE, O'Rourke KM, Zhou H, Eby M, Aravind L, Seshagiri S, Wu P, Wiesmann C, Baker R, Boone DL, Ma A, Koonin EV, Dixit VM. *Nature.* 2004; 430:694–699. [PubMed: 15258597]
81. Mammas IN, Sourvinos G, Giannoudis A, Spandidos DA. *Pathol Oncol Res.* 2008; 14:345–354. [PubMed: 18493868]
82. Pantry SN, Medveczky PG. *Semin Cancer Biol.* 2009; 19:153–157. [PubMed: 19429478]
83. Zheng G, Schweiger MR, Martinez-Noel G, Zheng L, Smith JA, Harper JW, Howley PM. *J Virol.* 2009; 83:8683–8692. [PubMed: 19553317]
84. Verma SC, Lan K, Robertson E. *Curr Top Microbiol Immunol.* 2007; 312:101–136. [PubMed: 17089795]
85. Stedman W, Deng Z, Lu F, Lieberman PM. *J Virol.* 2004; 78:12566–12575. [PubMed: 15507644]
86. Lim C, Gwack Y, Hwang S, Kim S, Choe J. *J Biol Chem.* 2001; 276:31016–31022. [PubMed: 11425857]
87. Shamay M, Krithivas A, Zhang J, Hayward SD. *Proc Natl Acad Sci U S A.* 2006; 103:14554–14559. [PubMed: 16983096]
88. Staudt MR, Dittmer DP. *Curr Top Microbiol Immunol.* 2007; 312:71–100. [PubMed: 17089794]
89. Aravind L, Landsman D. *Nucleic Acids Res.* 1998; 26:4413–4421. [PubMed: 9742243]
90. Gwack Y, Byun H, Hwang S, Lim C, Choe J. *J Virol.* 2001; 75:1909–1917. [PubMed: 11160690]
91. Sakakibara S, Ueda K, Chen J, Okuno T, Yamanishi K. *J Virol.* 2001; 75:6894–6900. [PubMed: 11435569]
92. Wysocka J, Herr W. *Trends Biochem Sci.* 2003; 28:294–304. [PubMed: 12826401]
93. Bochkarev A, Pfuetzner RA, Edwards AM, Frappier L. *Nature.* 1997; 385:176–181. [PubMed: 8990123]
94. Hollis T, Stattel JM, Walther DS, Richardson CC, Ellenberger T. *Proc Natl Acad Sci U S A.* 2001; 98:9557–9562. [PubMed: 11481454]
95. Taylor TJ, Knipe DM. *J Virol.* 2004; 78:5856–5866. [PubMed: 15140983]
96. Hackstadt T, Baehr W, Ying Y. *Proc Natl Acad Sci U S A.* 1991; 88:3937–3941. [PubMed: 2023942]

97. Thomas JO. *Biochem Soc Trans.* 2001; 29:395–401. [PubMed: 11497996]
98. Aravind L, Koonin EV. *Nucleic Acids Res.* 1999; 27:4658–4670. [PubMed: 10556324]
99. Kreuzer KN. *Trends Biochem Sci.* 2000; 25:165–173. [PubMed: 10754548]
100. Williams RS, Williams JS, Tainer JA. *Biochem Cell Biol.* 2007; 85:509–520. [PubMed: 17713585]
101. Graumann PL, Knust T. *Chromosome Res.* 2009; 17:265–275. [PubMed: 19308706]
102. Long DT, Kreuzer KN. *Proc Natl Acad Sci U S A.* 2008; 105:6852–6857. [PubMed: 18456838]
103. Myung H, Calendar R. *J Bacteriol.* 1995; 177:497–501. [PubMed: 7836278]
104. Aravind L, Leipe DD, Koonin EV. *Nucleic Acids Res.* 1998; 26:4205–4213. [PubMed: 9722641]
105. Yusufzai T, Kadonaga JT. *Science.* 2008; 322:748–750. [PubMed: 18974355]
106. Shaw G, Gan J, Zhou YN, Zhi H, Subburaman P, Zhang R, Joachimiak A, Jin DJ, Ji X. *Structure.* 2008; 16:1417–1427. [PubMed: 18786404]
107. Christen LM, Sanders M, Wiler C, Niles EG. *Virology.* 1998; 245:360–371. [PubMed: 9636376]
108. Hu X, Carroll LJ, Wolffe EJ, Moss B. *J Virol.* 1996; 70:7669–7677. [PubMed: 8892887]
109. Hu X, Wolffe EJ, Weisberg AS, Carroll LJ, Moss B. *J Virol.* 1998; 72:104–112. [PubMed: 9420205]
110. Iyer LM, Babu MM, Aravind L. *Cell Cycle.* 2006; 5:775–782. [PubMed: 16627993]
111. Lackner CA, Condit RC. *J Biol Chem.* 2000; 275:1485–1494. [PubMed: 10625702]
112. Singleton MR, Wentzell LM, Liu Y, West SC, Wigley DB. *Proc Natl Acad Sci U S A.* 2002; 99:13492–13497. [PubMed: 12370410]
113. Kreuzer KN. *Biochim Biophys Acta.* 1998; 1400:339–347. [PubMed: 9748648]
114. Bouchard JD, Moineau S. *J Bacteriol.* 2004; 186:3649–3652. [PubMed: 15150253]
115. Leipe DD, Aravind L, Grishin NV, Koonin EV. *Genome Res.* 2000; 10:5–16. [PubMed: 10645945]
116. Cheng C, Kussie P, Pavletich N, Shuman S. *Cell.* 1998; 92:841–850. [PubMed: 9529259]
117. Shuman S. *Biochim Biophys Acta.* 1998; 1400:321–337. [PubMed: 9748643]
118. Inoue N, Hess KD, Moreadith RW, Richardson LL, Handel MA, Watson ML, Zinn AR. *Hum Mol Genet.* 1999; 8:1201–1207. [PubMed: 10369865]
119. Da Fonseca F, Moss B. *Proc Natl Acad Sci U S A.* 2003; 100:11291–11296. [PubMed: 12972642]
120. Perry K, Hwang Y, Bushman FD, Van Duyne GD. *Mol Cell.* 2006; 23:343–354. [PubMed: 16885024]
121. Afonso CL, Tulman ER, Delhon G, Lu Z, Viljoen GJ, Wallace DB, Kutish GF, Rock DL. *J Virol.* 2006; 80:4978–4991. [PubMed: 16641289]
122. Kawabata T, Arisaka F, Nishikawa K. *Gene.* 2000; 259:223–233. [PubMed: 11163980]
123. Nitiss JL. *Nat Rev Cancer.* 2009; 9:338–350. [PubMed: 19377506]
124. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW. *Science.* 1998; 282:754–759. [PubMed: 9784136]
125. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP. *Science.* 1996; 274:948–953. [PubMed: 8875929]
126. Everett RD. *Bioessays.* 2000; 22:761–770. [PubMed: 10918307]
127. Gu H, Roizman B. *Proc Natl Acad Sci U S A.* 2007; 104:17134–17139. [PubMed: 17939992]
128. Lomonte P, Thomas J, Texier P, Caron C, Khochbin S, Epstein AL. *J Virol.* 2004; 78:6744–6757. [PubMed: 15194749]
129. Ferenczy MW, DeLuca NA. *J Virol.* 2009; 83:8514–8524. [PubMed: 19535445]
130. Iyer LM, Koonin EV, Aravind L. *Genome Biol.* 2002; 3 RESEARCH0012.
131. Brick DJ, Burke RD, Schiff L, Upton C. *Virology.* 1998; 249:42–51. [PubMed: 9740775]
132. Poon AP, Gu H, Roizman B. *Proc Natl Acad Sci U S A.* 2006; 103:9993–9998. [PubMed: 16785443]
133. Oelgeschlager T. *J Cell Physiol.* 2002; 190:160–169. [PubMed: 11807820]

134. Mandal SS, Chu C, Wada T, Handa H, Shatkin AJ, Reinberg D. *Proc Natl Acad Sci U S A*. 2004; 101:7572–7577. [PubMed: 15136722]
135. Klose RJ, Kallin EM, Zhang Y. *Nat Rev Genet*. 2006; 7:715–727. [PubMed: 16983801]
136. Smith JS, Avalos J, Celic I, Muhammad S, Wolberger C, Boeke JD. *Methods Enzymol*. 2002; 353:282–300. [PubMed: 12078503]
137. Song HK, Sohn SH, Suh SW. *Embo J*. 1999; 18:1104–1113. [PubMed: 10064578]
138. Holm L, Sander C. *Embo J*. 1995; 14:1287–1293. [PubMed: 7729407]
139. Morera S, Lariviere L, Kurzeck J, Aschke-Sonnenborn U, Freemont PS, Janin J, Ruger W. *J Mol Biol*. 2001; 311:569–577. [PubMed: 11493010]
140. Thomas JA, Rolando MR, Carroll CA, Shen PS, Belnap DM, Weintraub ST, Serwer P, Hardies SC. *Virology*. 2008; 376:330–338. [PubMed: 18474389]
141. Lobočka MB, Rose DJ, Plunkett G 3rd, Rusin M, Samojedny A, Lehnerr H, Yarmolinsky MB, Blattner FR. *J Bacteriol*. 2004; 186:7032–7068. [PubMed: 15489417]
142. Tidona CA, Schnitzler P, Kehm R, Darai G. *Virus Genes*. 1996; 12:219–229. [PubMed: 8883359]
143. Gershon PD, Moss B. *Proc Natl Acad Sci U S A*. 1990; 87:4401–4405. [PubMed: 2190222]
144. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F. *Trends Plant Sci*. 2002; 7:106–111. [PubMed: 11906833]
145. Sinclair AJ. *J Gen Virol*. 2003; 84:1941–1949. [PubMed: 12867624]
146. Gehring WJ. *Trends Biochem Sci*. 1992; 17:277–280. [PubMed: 1357790]
147. Petosa C, Morand P, Baudin F, Moulin M, Artero JB, Muller CW. *Mol Cell*. 2006; 21:565–572. [PubMed: 16483937]
148. Lee HR, Kim MH, Lee JS, Liang C, Jung JU. *J Interferon Cytokine Res*. 2009
149. Aravind L, Koonin EV. *Trends Biochem Sci*. 2000; 25:112–114. [PubMed: 10694879]
150. Ptashne, M. *A genetic switch : phage lambda revisited*. 3rd ed.. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2004.
151. Davis NK, Chater KF. *Mol Gen Genet*. 1992; 232:351–358. [PubMed: 1316997]
152. Bohm S, Frishman D, Mewes HW. *Nucleic Acids Res*. 1997; 25:2464–2469. [PubMed: 9171100]
153. Li Z, Pan L, Yu H, Li L, Gong Y, Yang K, Pang Y. *Virus Res*. 2006; 115:185–191. [PubMed: 16202468]
154. Zemskov EA, Kang W, Maeda S. *J Virol*. 2000; 74:6784–6789. [PubMed: 10888617]
155. Maynard, J. Smith, *Evolutionary genetics*. 2nd ed.. Oxford ; New York: Oxford University Press; 1998.
156. Hilleman MR. *Proc Natl Acad Sci U S A*. 2004; 101 Suppl 2:14560–14566. [PubMed: 15297608]
157. Aravind L, Iyer LM, Koonin EV. *Curr Opin Struct Biol*. 2006; 16:409–419. [PubMed: 16679012]
158. Tock MR, Dryden DT. *Curr Opin Microbiol*. 2005; 8:466–472. [PubMed: 15979932]
159. Tao Q, Robertson KD. *Clin Immunol*. 2003; 109:53–63. [PubMed: 14585276]
160. Fernandez AF, Rosales C, Lopez-Nieva P, Grana O, Ballestar E, Ropero S, Espada J, Melo SA, Lujambio A, Fraga MF, Pino I, Javierre B, Carmona FJ, Acquadro F, Steenbergen D, Snijders PJ, Meijer CJ, Pineau P, Dejean A, Lloveras B, Capella G, Quer J, Buti M, Esteban JI, Allende H, Rodriguez-Frias F, Castellsague X, Minarovits J, Ponce J, Capello D, Gaidano G, Cigudosa JC, Gomez-Lopez G, Pisano DG, Valencia A, Piris MA, Bosch FX, Cahir-McFarland E, Kieff E, Esteller M. *Genome Res*. 2009; 19:438–451. [PubMed: 19208682]
161. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. *Science*. 2009; 324:930–935. [PubMed: 19372391]
162. Everett RD. *Cell Microbiol*. 2006; 8:365–374. [PubMed: 16469050]
163. Venancio TM, Balaji S, Aravind L. *Molecular Biosystems*. 2009 Manuscript accepted for publication.
164. Balaji S, Babu MM, Iyer LM, Aravind L. *Nucleic Acids Res*. 2005; 33:3994–4006. [PubMed: 16040597]
165. S.I.o. *Bioinformatics*. 2009
166. Federici BA, Bigot Y. *J Insect Physiol*. 2003; 49:419–432. [PubMed: 12770621]

167. O'Reilly, D.; Miller, L.; Luckow, V. *Baculovirus Expression Vectors: A Laboratory Manual*. USA: Oxford University Press; 1993.
168. Osmond C. *Computing in Science and Engineering*. 2003; 5:16–25.
169. Van Etten J. *Curr Top Microbiol Immunol*. 2009; 328:v–vii. [PubMed: 19216433]

	Tiny		Small		Medium		Large	
Chromatin proteins	E2 (RRM fold) E6 (ZNF) E7 (PHD-like) LxCxE		Protein VII E2A (ADDBP)		Histone (His-fold) BRCT SWIM ABC+TOPRIM ERF gp2.5 (OB fold) SAK4 (RecA-like) RAD52 SSB (OB fold) HU_IHF SAP (HEH) SbcC (ABC) A18 (SFII) SWI2/SNF2 (SFII)		H4 (His-fold) H4C1 N-hydroxylase (DSBH) RecA gp32 (OB fold) rHA (GHKL) GyrA/GyrB (TOP2A) ERF gp2.5 (OB fold) SAK4 (RecA-like) RAD52 SSB (OB fold) HU_IHF SAP (HEH) SbcC (ABC) TOP2B A18 (SFII) SWI2/SNF2 (SFII)	
	SAP (HEH)		LxCxE		LxCxE		H4 (His-fold) H4C1 N-hydroxylase (DSBH) RecA gp32 (OB fold) rHA (GHKL) GyrA/GyrB (TOP2A) ERF gp2.5 (OB fold) SAK4 (RecA-like) RAD52 SSB (OB fold) HU_IHF SAP (HEH) SbcC (ABC) TOP2B A18 (SFII) SWI2/SNF2 (SFII)	
	LxCxE		LxCxE		LxCxE		H4 (His-fold) H4C1 N-hydroxylase (DSBH) RecA gp32 (OB fold) rHA (GHKL) GyrA/GyrB (TOP2A) ERF gp2.5 (OB fold) SAK4 (RecA-like) RAD52 SSB (OB fold) HU_IHF SAP (HEH) SbcC (ABC) TOP2B A18 (SFII) SWI2/SNF2 (SFII)	
	LxCxE		LxCxE		LxCxE		H4 (His-fold) H4C1 N-hydroxylase (DSBH) RecA gp32 (OB fold) rHA (GHKL) GyrA/GyrB (TOP2A) ERF gp2.5 (OB fold) SAK4 (RecA-like) RAD52 SSB (OB fold) HU_IHF SAP (HEH) SbcC (ABC) TOP2B A18 (SFII) SWI2/SNF2 (SFII)	
group	Papovae	Phage	Adenoviridae	Phage	Herpesvirales	Baculoviridae	Polydnaviridae	WSSV
clade(s)	Papillomaviridae	Plasmaviridae	Adenoviridae	Lipothirixviridae	Herpesvirales	Baculoviridae	Polydnaviridae	WSSV
host	Eukaryotes	Bacteria	Eukaryotes	Archaea	Eukaryotes	Eukaryotes	Eukaryotes	Eukaryotes
number of proteins	5-15	16	26-39	57-73	70-223	136-156	70-461	531
genome size (bp)	5256-8607	11965	30570-36519	36895-41172	108409-241087	118584-228089	50930-316674	305107
capsid diameter (nm)	45-55	80	28-90	24-38 x 410-1950	100-110	40-50 x 200-400	90-110	120-150
Virion structure								
Color codes	sporadic		few homologs in one or two lineages		group specific		several homologs in one or two related lineages	
							diffuse	
							homologs in two or three independent lineages	
							widespread	
							homologs in at least four independent lineages	

Figure 1. Diversity of chromatin protein families among different lineages of dsDNA viruses
Labels at the top identify different genome size classes. Family names are the same as used in the main text and structural folds are given between parentheses. For the purpose of dividing the protein families according to spread and abundance among viral lineages (indicated by color codes), the Mimivirus was considered a lineage related to phycodnaviruses. Detailed description of each protein family is provided in the main text. Capsid/virion diameter data from [165–169].

Figure 2. Domain architectures of the (A) LXCXE motif and the (B) BEN domain in DNA viruses

Standard nomenclature is used for domain names, poorly characterized globular domains are denoted as 'X'. The architectures are labeled with the gene and species names of the virus in which they are found, separated by an underscore.

Figure 3. Cartoon representations and topology diagrams of various treble clef fold domains found in dsDNA viruses

Secondary structure elements of the shared core of the treble-clef fold are shown in color; helices are colored red and strands green. The characteristic N-terminal region of the treble-clef is colored blue. The unique helices and strands present only in certain representatives of the fold are shown in grey. Topology diagrams in the center depict the common and derived regions in the structures. The papillomavirus E6 Zn-finger is related to the LIM domain and the papillomavirus E7 Zn-finger is closest to the pygopus-like PHD fingers.

Figure 4. Gene neighborhoods and domain architectures of various chromatin domains found in double stranded DNA viruses

Genes are depicted as boxed arrows, with the arrowhead pointing from the 5' to the 3' end of the open reading frame. Domain architectures are shown as cartoon representations. Standard nomenclatures are used for both gene and domain names. Gene neighborhoods are labeled with the gene and species names of the primary domain discussed in the panel, separated by an underscore. The bar-graph in the top panel shows the genome position frequency of the SAP domain in completely sequenced viral genomes with linear chromosomes. There is a significant preference for positioning of this domain at one terminus of the viral chromosomes.