Title

Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions

Permalink
https://escholarship.org/uc/item/0h68c383

Journal
Cell Reports, 7(5)

ISSN
2639-1856

Authors
Hagai, Tzachi
Azia, Ariel
Babu, M Madan
et al.

Publication Date
2014-06-01

DOI
10.1016/j.celrep.2014.04.052

Published in final edited form as:
Cell Rep. 2014 June 12; 7(5): 1729–1739. doi:10.1016/j.celrep.2014.04.052.

Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions

Tzachi Hagai[1], **Ariel Azia**[2], **M. Madan Babu**[3,4,*], and **Raul Andino**[1,4,**]

[1]Department of Microbiology and Immunology, University of California, 600 16th Street, GH-S572, UCSF Box 2280, San Francisco, California 94143-2280, USA

[2]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel

[3]The Medical Research Council Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK

Abstract

Virus interact extensively with host proteins, but the mechanisms controlling these interactions are not well understood. We present a comprehensive analysis of eukaryotic linear-peptide motifs (ELMs) in 2,208 viral genomes and reveal that viruses exploit molecular mimicry of host-like ELMs to possibly assist in host-virus interactions. Using a statistical genomics approach, we identify a large number of potentially functional ELMs and observe that the occurrence of ELMs is often evolutionarily conserved but not uniform across virus families. Some viral proteins contain multiple types of ELMs, in striking similarity to complex regulatory modules in host proteins, suggesting that ELMs may act combinatorially to assist viral replication. Furthermore, a simple evolutionary model suggests that the inherent structural simplicity of ELMs often enables them to tolerate mutations and evolve quickly. Our findings suggest that ELMs may allow fast rewiring of host-virus interactions, which likely assists rapid viral evolution and adaptation to diverse environments.

Introduction

Viruses face a formidable challenge: they must invade their hosts, outwit their defense systems and successfully replicate to ensure their survival. Despite possessing small genomes and few proteins, viruses are equipped with high adaptive capacity to engage with their host to maximize successful viral replication. One mechanism often used by viruses is molecular mimicry, where a virus adopts a host's characteristics to successfully interact with host factors (Elde and Malik, 2009; Gorbalenya, 1992; Shackelton and Holmes, 2004). It has

[*,**]Corresponding authors: M. Madan Babu - madanm@mrc-lmb.cam.ac.uk, Raul Andino - Raul.Andino@ucsf.edu.
[4]Co-senior authors

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.

been suggested, based on a literature survey, that viruses may employ short, unstructured elements, that are called Eukaryotic Linear Motifs (ELMs), to mediate interactions with their host (Davey et al., 2011). ELMs appear to function in various regulatory interactions, by acting as docking sites for several protein domains (e.g. SH3 and WW domains), as subcellular targeting signals (e.g. Nuclear Localizing Signal) and as recognition sites for protease cleavage (e.g. Caspase) or for post-translational modifications (e.g. Phosphorylation sites).

These small interaction modules are usually composed of 2–8 residues and are often located within disordered regions of proteins (Davey et al., 2012b; Fuxreiter et al., 2007; Teyra et al., 2012). Disordered regions are polypeptide segments that do not adopt a defined tertiary structure but contribute to various regulatory functions (Babu et al., 2012; Dunker et al., 2008; Dyson and Wright, 2005; Tompa, 2002; Zhang et al., 2013). Unlike structured domains that are not easy to evolve or need to be acquired from the host's genome (Gorbalenya, 1992), ELMs can rapidly evolve in viral proteins, which might facilitate the formation of myriad networks of interactions with host proteins.

Literature-based analysis of a limited number of experimentally identified ELMs in viral proteins suggested that these modules participate in many stages of viral replication (see Figs 1A, and S1 for examples) (Davey et al., 2011). Indeed, recent evidence indicated that ELMs can modulate virulence, host-tropism, immune escape mechanisms, disease length and severity of infection (Boon and Banks, 2013; Das et al., 2010; Igarashi et al., 2008; Lu et al., 2012; Pantua et al., 2013; Sun et al., 2011). Evolutionary conservation of ELMs among orthologs of viral proteins might further support their importance in mediating specific interactions of many viruses in the same family. For instance, a host Ser/Thr kinase phosphorylates a conserved ELM within several Flaviviruses RNA-polymerases, thereby this motif presumably plays a conserved role in the flavivirus' life cycle (Reed et al., 1998). On the other hand, the simplicity of ELMs may allow them a greater evolutionary plasticity so that their rapid loss and gain can support a quick rewiring of virus interactions with the host. This is observed for example, in the binding of several different Picornaviruses capsid proteins to the integrin receptors using the RGD motif (where this motif was lost and gained several times in the course of picornavirus evolution) (Jackson et al., 2003)).

In spite of their potential importance in mediating host-virus interactions, the set of studied ELMs is limited and is mostly biased towards a few viruses. A major challenge of studying ELMs stems precisely from their low complexity. Indeed, ELM patterns can be often found in viral proteins; however, it has been difficult to discriminate between ELM-like sequences that appear by chance from those that truly represent functional ELMs (moreover, it is possible that viral proteins contain a higher fraction of nonfunctional ELMs, since cellular proteins are under tighter regulation and are selected to avoid nonfunctional ELMs (Landry et al., 2009)). Here, we overcome this obstacle by employing a simple metric that (1) assesses the probability of each ELM occurring serendipitously in a random disordered sequence and (2) compares this assessment in eukaryotic and prokaryotic viruses; the latter serving as a negative control, since ELMs are predominant in eukaryotes, and their occurrence in prokaryotic viruses is assumed to be due to chance. Our analysis allows us to identify potentially functional ELMs in a comprehensive set of viruses. We use this dataset of ELMs to examine their occurrence in various virus families and to study ELMs co-occurrence. Our observations suggest that viruses may use ELMs in a combinatorial manner to mediate their interactions with host cell networks. Importantly, ELMs might be simple means to promote robust and evolvable interactions with host pathways, and may explain how viruses achieve rapid adaptation to changing environments.

Results

Patterns that match ELMs are prevalent in viral sequences

To study ELMs in viral proteins, we composed a dataset of 2,208 non-redundant viruses, representing all orders and most known viral families (see Table S1 and **Methods**). The dataset contains 536 prokaryotic viruses and 1,672 eukaryotic viruses (of which 787 are animal viruses, 816 are plant viruses and 69 are other eukaryotic viruses) (Fig 1B). We then scanned the predicted disordered regions in the 74,288 viral proteins to identify regions that match 173 previously described ELM patterns (Dinkel et al., 2012)(Table S3 and **Methods**).

We found that the total number of occurrences of each ELM-matching region (hereafter referred as ELMs) in viral proteins significantly correlates with its sequence complexity (as calculated by the composition of its matching regular expression): ELMs with low information content tend to be common whereas complex ELMs are rare (Fig 1C). Furthermore, the total number of ELMs that occur in each virus is directly related to the total length of its disordered regions. Interestingly, we observed that the number of ELMs per disordered unit is higher in eukaryotic viruses than in prokaryotic viruses and that each of these sets has its own linear fit (Fig 1D). Thus, it appears that ELM-like patterns occur in a manner that correlates with the proportion of protein disorder and with ELM complexity. These characteristics confound the identification of additional, functional ELMs, given the potential for high proportions of ELM patterns that occur by chance. However, we hypothesized that the larger proportion of ELMs in eukaryotic viruses in comparison with prokaryotic viruses, as a negative control (Fig 1D) may serve as a basis to identify ELMs that are likely to be functional.

An approach to identify potentially functional ELMs

We next assessed the likelihood of each instance of ELM to occur by chance in prokaryotic and eukaryotic viruses. To this end, we shuffled the content of the disordered regions of each of the original viral proteins to create a large set of “shuffled” viral proteins. For each virus, we created two sets - each containing 100,000 randomly shuffled viral proteins – using two independent shuffling methods: (1) where the residues are shuffled within disordered regions of proteins belonging to the same virus, and (2) where the residues are shuffled between all viral proteins (see Fig S2 and **Methods**). Our premise is that regions matching known ELMs that are rarer in randomly shuffled sequences are more likely to represent truly functional ELMs (we note that with this method we cannot provide predictions regarding instances of ELMS with low complexity that only occur a few times in the natural sequences). We then ranked the observed ELMs based on their likelihood of occurrence in the shuffled set (obtaining an “expected value”; see Figure S2 and **Methods**). For each of the 173 different ELM types that occur in each natural viral proteins, we determined its frequency of occurrence in the shuffled sets. It should be noted that this frequency is affected by a complex combination of factors, including the number of occurrences in the natural sequence, ELM complexity and sequence composition (the complete list of putative ELMs in all the proteins identified in this study along with all their data and analysis can be found at http://andino.ucsf.edu/andino/viral_elms/viralELMs.html).

We next compared how many of the ELMs that appear in the natural viral sequences occur in shuffled sequences in prokaryotic and eukaryotic viruses. Interestingly, we observed that the fraction of ELMs that are less prevalent in shuffled sequences is significantly higher in eukaryotic viruses when compared with prokaryotic viruses. This trend is stronger in animal viruses and weaker in plant viruses (Figs 2A and S3, Table S4). As an example, we show the percentage of ELMs that occur in fewer than 100 of 100,000 shuffled sequences (0.1% of the total shuffled sequences)(Figs 2A). Only 1.1% of the ELMs observed in prokaryotic viruses occur in less than 0.1% of the shuffled sequences, in comparison to 3.2% in eukaryotic viruses and 3.6% in animal viruses. This represents a highly significant enrichment ($p<10^{-15}$, Fisher exact test). This trend remains consistent when we compare ELMs that occur in less than 10 of 100,000 shuffled sequences (0.01% of the total shuffled sequences), when we use only a subset of the viruses, or when we compare specific types of ELMs or a separate set of 117 “putative” ELMs (Fig S3 **and Methods**)(Davey et al., 2012a). Furthermore, the enrichment we observe is independent of the shuffling method (Fig S3).

Inferred functional ELMs are enriched in experimentally validated ELMs

The fact that eukaryotic viruses (especially animal viruses) contain higher fractions of ELMs that are less prevalent in shuffled sequences (in comparison with prokaryotic viruses), suggests that the set of ELMs identified by our approach as rare in shuffled sequences is likely to be enriched in functional ELMs. We therefore investigated the set of ELMs in eukaryotic viruses that occur in less than 0.1% of the shuffled sequences and further analyzed it in comparison with the rest of the ELMs. Indeed, many of the ELMs identified by our unbiased approach were previously reported as functional motifs in viral proteins. This includes the motif that mediates the binding of Epstein-Barr virus (EBV) LMP2 protein to host E3 ligase, to promote the degradation of several host kinases (shown in Fig 1A). We found a significant enrichment of ELMs we identified to be potentially functional in a set of 42 experimentally validated functional viral ELMs (Dinkel et al., 2012) (a six fold enrichment with respect to their occurrence in the rest of the ELM dataset; 19% overlap, $p=6.5\times10^{-6}$, Fisher exact test, Fig S4). This observation supports the notion that the set we identified is indeed enriched with functional ELMs.

Family and species level analysis reveals enormous heterogeneity of ELM usage among eukaryotic viruses

While some viruses contain proteins with many ELMs, others appear to have only few ELMs. For instance, many dsDNA virus families, such as Papillomaviridae, Adenoviridae and Herpesviridae, that are all known to use ELMs to mediate numerous interactions with their host, have been identified in our analysis to be relatively rich with ELMs (see Table S5 for a complete list of the 21 viral families enriched with ELMs). Surprisingly, other families such as the ssRNA viruses Picornaviridae have small fractions of disordered regions and their proteins seem to contain relatively few ELMs. A recent analysis, based on a smaller set of 267 viral proteins with known interactions with host proteins, suggested that viral proteins tend to contain higher numbers of ELMs in comparison with their cellular proteins (Garamszegi et al., 2013). Interestingly, most of the proteins in that study belong to viral families found to be enriched with ELMs in our analysis (e.g. the three dsDNA virus families mentioned above). Our analysis further suggests that different viruses greatly differ in the use of ELMs to mediate interactions with their host.

In addition, even within specific virus families, individual members contain different proportions of ELMs in their proteins. For example, Hepatitis C Virus (HCV) is enriched with ELMs not only in comparison to other flaviviruses (which tend to have few ELMs), but also relative to animal viruses in general (Fig 2B). The variation in ELMs content between viruses could be related to several factors, such as: (a) virus lifecycle – persistent viruses might require a more precise regulation of cellular pathways to ensure that the cell remains functional. Thus, they might need to carefully regulate the expression of disordered regions which might be harmful to the cell (Babu et al., 2011; Vavouri et al., 2009); and (b) - genome size and architecture; e.g. - overlapping genes often contain disordered regions (Rancurel et al., 2009) that might carry out their functions primarily through ELMs (Carter et al., 2013)).

Inferred functional ELMs occur in a broad spectrum of functional classes of proteins and are enriched in specific functional groups

Unlike structured domains, which are often specific to certain functional classes of proteins, a given ELM can be found in functionally diverse viral proteins. Conversely, ELMs can differ in their types and numbers among viral proteins that share similar functions. To examine whether specific groups of proteins are enriched or depleted of ELMs, we composed 30 sets of eukaryotic viral proteins with similar function, viral infection stage or subcellular location (Tables 1 and S6). Interestingly, the group that is most highly enriched with ELMs and disorder is the group of phosphoproteins (which is in agreement with our findings that many phosphosites tend to co-occur with other motifs – see sections on ELM co-occurrence below). The group that is significantly depleted of ELMs includes proteins that act as host domain mimics that were likely transferred from host genomes through horizontal gene transfer (other groups of viral proteins that modulate the host, such as virulence proteins, tend to be depleted of ELMs, but not in such a strong manner). In addition, viral proteins that function in different subcellular compartments tend to differ in their ELMs content - nuclear viral proteins tend to be relatively enriched with ELMs, whereas ER and mitochondrial proteins do not significantly differ from the entire viral set. We summarize the results in Tables 1 and S6, and note that several groups are heterogeneous in their ELM and disorder composition (for example, some proteins that are involved in DNA replication have very high fractions of disorder, whereas other proteins from the same functional group have very few disordered segments). This exemplifies how functionally similar viral proteins can use a diversity of molecular mechanisms to mediate their interactions.

Inferred functional ELMs are evolutionarily more conserved among viral orthologs

We next examined the evolutionary conservation of viral ELMs. In general, we expect that functional ELMs should be more conserved than non-functional ELMs. We thus compared the conservation of the set of potentially functional ELMs we identified with the rest of the ELMs. We determined the conservation of each of the ELMs by calculating the fraction of occurrence in orthologs from the same genus or the same family (Fig S5 and **Methods**). We observed that the selected subset of ELMs (that are rare in shuffled sequences) is indeed significantly more conserved than the rest of the ELMs in both the genus-based and the family-based levels ($p=2.2\times10^{-55}$ and $p=1.2\times10^{-49}$; sign-test). Consistent with these results, we also observed the inferred functional ELMs in the 6 strains of HCV to be more conserved in an independent set of variant HCV sequences (**Methods**). These results provide additional evidence that the selected ELMs (that are rare in shuffled sequences) are likely to represent functional viral motifs and that the use of evolutionary conservation offers an orthogonal approach to identify truly functional ELMs.

The presence of ELMs in viral genomes might permit rapid adaption during evolution

We note that conservation of ELMs in a given instance does not necessarily mean an exact conservation of residues in the same region. In orthologs, ELMs can occur in different regions of the protein (Hagai et al., 2012; Nguyen Ba and Moses, 2010), can appear in different numbers and can change their primary sequence patterns, while still maintaining functionality (as observed for example in ELMs that mediate interactions with the ESCRT machinery in various retroviruses (Martin-Serrano and Neil, 2011)). Thus, in comparison to structured domains or to catalytic sites in enzymes, ELMs seem to tolerate changes in location and mutations better, in addition to their capacity to evolve rapidly. To investigate this, we modeled a population of all possible single-point mutations of the HIV-1 genome. It is believed that this large spectrum of mutants is created every 24 hours *in vivo* upon infection (Coffin, 1995). We then examined how non-synonymous mutations in disordered regions in this viral population affect the distribution of ELMs in comparison with their occurrence in the wildtype HIV-1 genome (Figs 3). Almost half of the mutants in the viral population had the same distribution of ELMs, despite the fact that ~40% of them occurred within ELM segments (Fig 2C, top). Of the other half of mutants – those that differ in their ELMs distributions – a significant part had either increased the number of existing ELMs or evolved new types of ELMs with respect to the wildtype (Fig 2C – bottom circles – purple and pink fractions, respectively). Many viruses have high mutations rates that are thought to be central to adaptation to dynamic environments and survival (Domingo et al., 2012; Lauring and Andino, 2010). In this scenario, as suggested by our simple simulation, ELMs can act as functional modules that are robust in the face of mutations yet allow fine-tuning of the host-virus interactions and viral adaptation to changing environments by their ability to rapidly evolve.

The evolutionary origins of viral ELMs: horizontal transfer from host genes and convergent evolution

Mimics in various pathogens can either be acquired from the host genome through horizontal gene transfer (HGT), or evolve independently in a convergent manner. In structured domain mimicry, it is generally assumed that domains that are found in pathogen's proteins and have high sequence similarity in a large portion of the domain, are likely to have been acquired by HGT, whereas mimics of short structural segments (such as repeats) or mimics with small sequence similarity or lack of structural similarity are likely to have arisen in a convergent manner (Doxey and McConkey, 2013; Elde and Malik, 2009) (Fig 4A). Since ELMs are short and easy to evolve, it was suggested that they belong to the latter category (Davey et al., 2011). Indeed, some ELM instances undoubtedly evolved convergently; for example, the RDG motif that mediates interactions with integrin receptors has evolved independently several times in capsid proteins of distantly related picornaviruses (Fig 4B). However, instances where ELMs were transferred from host genes and maintained during the pathogens' evolution are also known. For example, the occurrence of the actin-binding WH2 motif, which is a fairly complex and long motif – in the baculovirus p78/83 protein, is likely to be a result of HGT, since the motif and the regions surrounding it are relatively similar to the host WASP protein, from which these regions are thought to originate (Machesky et al., 2001)(Fig 4B). However, the latter example of ELM acquisition from host genome is likely to be rare and limited to mostly long and complex motifs.

ELMs have predominantly emerged in viral proteins by convergent evolution

To investigate the evolutionary origins of viral ELMs in a quantitative manner and their likelihood of originating from host genes, we chose to focus on a set of viral genes that were identified to be a result of HGT and to examine their disordered regions and ELMs composition in comparison with that of the host homologous proteins. For this, we extracted 135 non-redundant animal viral proteins and their inferred homologs in human and mouse from the PhEVER database (Palmeira et al., 2011) – a comprehensive database that clusters host and viral homologous genes, based on significant sequence similarity. Thus, we created 135 groups of proteins, where each group contains a viral protein with its best matching human and mouse homologs (see **Methods**). We compared the level of similarity in ordered and disordered regions between human and virus and between human and mouse protein pairs. As expected, human proteins are more similar to their mouse homologs than they are to their corresponding virus homologs. In addition, in both human-mouse and human-virus pairs, the similarity in ordered regions is higher than in disordered regions (Fig 4C). Importantly, while in both human-mouse and in human-virus pairs the disordered regions tend to evolve rapidly, the disordered regions in human-virus pairs have diverged significantly faster than what would be expected based on the divergence in human-virus ordered regions, and in comparison to human-mouse divergence ($p=8.1\times10^{-13}$, sign test, see **Methods**). Thus, we infer that after acquisition by pathogens, disordered regions tend to evolve fast – even faster than what would be expected – and, it is likely that ELMs that were transferred as part of these disordered regions, were later likely lost. Consequently, most ELMs that appear in extant viral proteins are the product of convergent evolution.

To further verify this possibility, we examined the mutual occurrence of ELMs in the 135 human-virus protein pairs. In each pair, we checked how many ELMs of the same type occurred in both the human and the virus homologs. Out of the 1,325 ELMs that occur in these viral proteins, there were 333 cases that the same type of ELM appeared in the human homolog as well. These co-occurrence events might be a result of HGT or they can present unrelated events of ELMs emergence in virus and host proteins. These scenarios can be discerned, by checking if there is a significant enrichment of ELM co-occurrence in these 135 pairs, above what would be expected by ELM propensity occurring by chance in disordered regions of the entire proteome (i.e. – the HGT scenario is more likely, if there is a significant enrichment of ELM co-occurrence). We thus tested for the likelihood of these co-occurrence events happening by chance, by comparing the observed co-occurrence frequencies with frequencies resulting from 10,000 random shufflings of ELMs occurrences in the entire proteome (see **Methods**). We observe that no ELM type had a co-occurrence level significantly higher than expected by chance – suggesting that most of the ELMs that appear in both human and virus could co-occur based on their propensity to occur in disordered regions (these ELMs are simple enough to rapidly evolve independently in each of the protein’s pair). Thus, we conclude that at least in this set, most ELMs that appear in viral proteins are likely to be a result of convergent evolution, and that cases of ELM acquisition by HGT that survive rapid viral evolution are likely to be relatively limited.

Specific types of ELM pairs tend to occur in unrelated viral proteins

We next searched for instances of two different types of ELMs in the same viral protein. For example, we wanted to determine if a WW-domain binding motif and a phosphorylation site are likely to be present in the same protein so that their functionality might be affected by their co-occurrence. In addition, the functionality of certain ELMs can be supported by the presence of other ELMs, even if they are separated in sequence, as they can be brought together in the three dimensional space or might cooperatively assist the binding of another ELM. While many viral proteins are depleted of ELMs (as discussed above and as shown in Fig S6), some viral proteins tend to contain numerous types of ELMs within their disordered regions (Fig S6). We searched for cases of ELMs that tend to co-occur in the same protein in a non-redundant set of viral proteins (see **Methods**). Since we have a total of 173 types of ELMs in our dataset, there are ~15,000 ELM pairs that could theoretically occur. We compared co-occurrence of ELMs in the viral protein set to 10,000 equivalent sets where we randomly shuffled the occurrences of the ELMs between the proteins (see **Methods**). This comparison yielded 242 pairs of ELMs that occur significantly ($p<0.05$, p-values were corrected using the Benjamini-Hochberg method) (Table S2C).

Regulation of host-virus interactions by a host-like ELM switch strategy

Recently, it was suggested that occurrence of ELMs in proximity to one another might act as a switch, whereby one ELM can act as a modulator of another ELM (by activating, blocking or modifying its functionality), as observed in a number of domain-interacting motifs that are localized next to phosphorylation sites (Akiva et al., 2012; Van Roey et al., 2012). Only few examples of ELM switches are known in viruses, including a complex module that supports cell transformation in the papillomavirus E6 (Boon and Banks, 2013; Pim et al., 2012). We used our dataset to examine the occurrence of ELM switches in viral proteins, by comparing the 242 co-occurring ELMs in viral proteins (which we found above) to an experimentally-validated set of ELM pairs that act as regulatory switches in eukaryotes (Van Roey et al., 2013) (Table S2C). Interestingly, out of the 68 switches that appear in this eukaryotic database, 17 overlap with ELM pairs in the viral set (a significant overlap when considering the possible ~15,000 pairs, $p=3.3\times10^{-16}$, Fisher exact test). Furthermore, both host and viral ELM pair sets are enriched with phosphorylation sites, much more than would be expected by their relative numbers in the ELM set. The significant overlap between ELM co-occurrence in viruses and their host, as well as the enrichment in phosphosites, which are known to modulate ELM's activity, suggest that viruses have extensively adopted mechanisms used by eukaryotes to tightly control important regulatory proteins. Viruses are likely to use these regulatory modules to coordinate complex and numerous interactions to achieve a successful and timely infection. In addition to ELM switches that are known to occur in their hosts, we identified a number of additional putative switches which have not yet been characterized in eukaryotes in our set of 242 ELM pairs. For example, we identified pairs of different subcellular localization signals that might target the same protein to different subcellular compartments in a controlled manner (this mechanism was shown to spatially regulate HIV-1 Rev (Henderson and Percipalle, 1997). We also identified cleavage sites in proximity to other ELMs, which might enable processing of viral proteins to further regulate their functions (see Table S2C for details). These observations suggest that the presence of multiple ELMs within a protein may act as a regulatory switches to modulate virus-host specific interactions.

Co-occurring ELMs evolved independently in different viral proteins

Finally, we were interested in seeing if instances of co-occurring ELMs tend to cluster in the same viral family, or if they tend to occur in various unrelated families. We found that in almost all cases, ELM pairs occur in different families, and almost always in at least one dsDNA viral family (see Figs 5A for the occurrence in different viral families of the 17 ELM switches, which are experimentally known to occur in hosts). These results suggest that ELM co-occurring pairs have evolved independently in various viral groups as a general mechanism which might support coordinated multiple interactions with their host.

Discussion

Recent large-scale analyses revealed an extensive and complex set of interactions between viruses and their host proteins(Ideker and Krogan, 2012). Given the fact that viral proteins are often shorter and contain less structured domains than host proteins (Fig S6A–B), it is intriguing how viruses establish such an extensive and fine-tuned network of interactions. Our analysis indicates that many viral proteins exploit the modular and simple architecture of ELMs to mediate these interactions. For example, we identified 5 different types of ELMs within the EBV EBNA-2 protein that can support its interactions with several known host factors (Fig 5B) (Calderwood et al., 2007). Our analysis is consistent with the idea that viruses have evolved ELM-mediated interactions because these motifs often enable transient interactions with cellular hubs, which are often targets of viral proteins (Dyer et al., 2008; Franzosa and Xia, 2011). In addition, the occurrence of ELMs within disordered regions allows for the rapid emergence of new interactions in response to different environmental challenges.

While the use of ELMs may be very common for certain viruses (Garamszegi et al., 2013), our analysis also indicates that a large fraction of viruses carry few recognizable ELMs (Fig

S6D). This is consistent with the fact that the fraction of disordered regions in virus proteins is not uniform (Goh et al., 2009; Ortiz et al., 2013; Pushker et al., 2013; Xue et al., 2012), which may restrict the number of ELMs that can be located in a given protein. However, our analysis likely underestimates the number of functional ELMs in viral proteins, given that the current list of annotated ELMs is probably incomplete and our conservative computational approach that may remove authentic functional ELMS. In addition to ELMs, viruses employ additional mechanisms, such as more complex forms of mimicry, to engage with their host during infection, as some interactions must be mediated by structured domains (Drayman et al., 2013; Handa et al., 2013). Recent studies have developed various approaches to identify functional domain mimics in various pathogens (such as by comparing host proteins with pathogenic and non-pathogenic bacteria (Doxey and McConkey, 2013), or by contrasting similarity scores of host proteins and specific viral families with scores of host proteins with other viral families (Odom et al., 2009)). In addition, structural similarities between viral proteins and host ligands have recently assisted in recognizing host receptors utilized by pathogens (Drayman et al., 2013). Our approach which tackles the difficulty of inferring the likelihood of ELM-matching sequences being functional mimicries thus complements these studies by focusing on motif mimicry.

One feature of ELMs in mediating host-virus interactions is their ability to tolerate mutations (Fig 2C). In addition, ELMs can evolve quickly to rewire the host-virus interaction network. Robustness and evolvability are observed for example in the PPxY motif that mediates interactions of EBV type-1 LMP2 protein with host E3-ligases (Fig 1A). This motif is conserved in LMP2 orthologs of the two additional strains of EBV in our dataset despite significant sequence divergence, but not in the distantly related Kaposi’s sarcoma-associated herpesvirus (KSHV) K15 protein. This might indicate that this interaction is not conserved across distant orthologs in the Gammaherpesvirinae subfamily. This observation is reminiscent of differences observed between EBV and KSHV in their use of other types of ELMs and the interactions they mediate (Tsai et al., 2009).

Notably, we observed that specific ELM pairs are significantly enriched in certain viral proteins. This observation suggests that viral ELMs, like host ELMs, might co-exist in the same protein to form regulatory modules that achieve tight regulation. The extensive occurrence of these modules in many viruses within the same family as well as in different families demonstrates the adeptness of these modules in host subversion. Many ELM modules uncovered here are targets for post-translational modification, such as phosphorylation or cleavage, suggesting that these modules might assist in temporal regulation of viral proteins. Similarly, the presence of subcellular localization signals in these modules argues that their activity is spatially regulated to assist in their multi-functionality.

Our analysis sheds light on ELM utilization in a large and unbiased set of viruses. As host-virus networks of additional viruses are elucidated, it will be possible to comprehensively assess the contribution of ELMs in shaping the interaction network with the host, and the rewiring of these networks in closely-related species. Future investigations of ELM involvement in host tropism, virus speciation and virulence might contribute to better understanding of biomedically important viruses and to assist in developing ways to overcome them.

Experimental procedures

We composed a set of 2,208 non-redundant viruses from 108 viral clades using available viral entries from the NCBI viral genomes resource (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/), and excluded viruses that had missing data or were too similar (Tables S1A–B). We predicted disorder values of protein sequences using the IUPred algorithm (Dosztanyi et al., 2005), and scanned disordered regions for sequences matching 173 known types of ELMs (Dinkel et al., 2012) and 117 "putative" motifs with as yet undiscovered functions (Davey et al., 2012a) (Table S2).

We created two large sets, each with 100,000 shuffled viruses, by randomly shuffling the content of disordered regions either within the same virus or between all the disordered regions of all 2,208 viruses (Fig S2). The shuffled sets allow us to compare the occurrence of ELMs in the original virus to their numbers in the 100,000 shuffled sequences, thereby assessing the likelihood of each ELM observed in the original viruses to occur by chance. We hypothesize that an ELM that occurs in the original virus but occurs very rarely in the shuffled set is likely to be functional (whereas we cannot infer the functionality of ELMs that occur frequently in shuffled sequences).

We compared the fractions of ELMs that occur rarely in shuffled sequences in prokaryotic, eukaryotic, animal and plant viruses (Figs 2 and S3). We studied the enrichment of rarely shuffled ELMs in a small set of experimentally known ELMs in viruses (Dinkel et al., 2012). We analyzed the relative conservation of ELMs that occur rarely in shuffled sequences in comparison with other ELMs in the set (ELMs that occur more frequently), by comparing the fraction of occurrence of each ELM instance in orthologous proteins of viruses belonging to the same genus or to the same family (where a higher fraction of occurrence indicates a higher conservation) (Fig S5). In addition, to examine whether our results hold outside our dataset, we repeated the above analysis using extracted sequences of variants of HCV from the Los Alamos HCV database (hcv.lanl.gov).

For functional enrichment analysis, we composed sets of eukaryotic viral proteins based on keyword annotations in UniProt (www.uniprot.org/). The distributions of fractions of inferred functional ELMs and fractions of disordered regions of these sets were compared to the distributions of the entire eukaryotic viral set using one-sided Kolmogorov–Smirnov test.

We examined the putative evolutionary origins of viral ELMs from host genes, using a non-redundant set of 135 viral proteins that have significantly-similar homologs in human and mouse genomes from the PhEVER database (Palmeira et al., 2011). We compared the similarity levels of ordered and disordered regions in human-mouse and human-virus pairs by estimating the fraction of similar residues in each pair based on BLAST analysis (Altschul et al., 1997) and by comparing the similarity scores in ordered and disordered regions in human-mouse pair with the corresponding human-virus pair. The significance of

ELM co-occurrence in human and virus homologs was tested by comparing the frequency of each observed ELM co-occurrence with co-occurrence frequencies resulting from a set of 10,000 human-virus protein pairs in which the ELMs occurrences were randomly shuffled.

We analyzed the biophysical characteristics and number of ELMs in a set of non-redundant animal viral proteins, and compared them to a set of non-redundant human proteins (Fig S6). ELMs co-occurrence analysis was done by comparing the occurrence of each pair of ELMs (two different types of the 173 ELMs) in the non-redundant viral set to 10,000 equivalent sets in which the occurrences of the ELMs were randomly shuffled between the proteins. The resulting significantly-occurring 242 pairs were compared to a set of 68 experimentally-known functional ELM switches that occur in eukaryotes (Van Roey et al., 2013) (Table S2C).

All the details of the proteins and the viruses we used and the analyses performed are available publicly through our website: http://andino.ucsf.edu/andino/viral_elms/viralELMs.html

See the Supplemental Experimental Procedures for additional details.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank E. Akiva, H. Dawes, L. Gitlin, A. Marcovitz, S. Pechmann, A. Stern, A. Toth and C. Wright for helpful comments on the manuscript, RD Hernandez, ED Levy and HS Malik for helpful discussions, RM Ekman, O. Laufman**,** A. Retik, P. Wassam and Z. Whitfield for technical assistance, and A. Branch, N. Davey, and S. Fishman for providing data. This work was supported by grants from the NIAID (R01 AI36178 and AI40085) and DARPA "Prophecy" virus evolution program (RA), and by the Medical Research Council (MC_U105185859), HFSP (RGY0073/2010), EMBO Young Investigator Program and ERASysBio+ (MMB). TH is supported by a Human Frontier Science Program Long-Term Fellowship.

References

Akiva E, Friedlander G, Itzhaki Z, Margalit H. A dynamic view of domain-motif interactions. PLoS computational biology. 2012; 8:e1002341. [PubMed: 22253583]

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Babu MM, Kriwacki RW, Pappu RV. Structural biology. Versatility from protein disorder. Science. 2012; 337:1460–1461. [PubMed: 22997313]

Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol. 2011; 21:432–440. [PubMed: 21514144]

Boon SS, Banks L. High-risk human papillomavirus E6 oncoproteins interact with 14-3-3zeta in a PDZ binding motif-dependent manner. Journal of virology. 2013; 87:1586–1595. [PubMed: 23175360]

Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A, Holthaus AM, Ewence AE, Li N, Hirozane-Kishikawa T, Hill DE, et al. Epstein-Barr virus and virus human protein interaction maps. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:7606–7611. [PubMed: 17446270]

Carter, JJ.; Daugherty, MD.; Qi, X.; Bheda-Malge, A.; Wipf, GC.; Robinson, K.; Roman, A.; Malik, HS.; Galloway, DA. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. Proceedings of the National Academy of Sciences of the United States of America; 2013.

Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science. 1995; 267:483–489. [PubMed: 7824947]

Das SR, Puigbo P, Hensley SE, Hurt DE, Bennink JR, Yewdell JW. Glycosylation focuses sequence variation in the influenza A virus H1 hemagglutinin globular domain. PLoS pathogens. 2010; 6:e1001211. [PubMed: 21124818]

Davey NE, Cowan JL, Shields DC, Gibson TJ, Coldwell MJ, Edwards RJ. SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res. 2012a; 40:10628–10641. [PubMed: 22977176]

Davey NE, Trave G, Gibson TJ. How viruses hijack cell regulation. Trends Biochem Sci. 2011; 36:159–169. [PubMed: 21146412]

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. Attributes of short linear motifs. Mol Biosyst. 2012b; 8:268–281. [PubMed: 21909575]

Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, et al. ELM--the database of eukaryotic linear motifs. Nucleic Acids Res. 2012; 40:D242–251. [PubMed: 22110040]

Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiology and molecular biology reviews: MMBR. 2012; 76:159–216. [PubMed: 22688811]

Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005; 21:3433–3434. [PubMed: 15955779]

Doxey AC, McConkey BJ. Prediction of molecular mimicry candidates in human pathogenic bacteria. Virulence. 2013; 4:453–466. [PubMed: 23715053]

Drayman N, Glick Y, Ben-Nun-Shaul O, Zer H, Zlotnick A, Gerber D, Schueler-Furman O, Oppenheim A. Pathogens use structural mimicry of native host ligands as a mechanism for host receptor engagement. Cell host & microbe. 2013; 14:63–73. [PubMed: 23870314]

Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 2008; 18:756–764. [PubMed: 18952168]

Dyer MD, Murali TM, Sobral BW. The landscape of human proteins interacting with viruses and other pathogens. PLoS pathogens. 2008; 4:e32. [PubMed: 18282095]

Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nature reviews Molecular cell biology. 2005; 6:197–208.

Elde NC, Malik HS. The evolutionary conundrum of pathogen mimicry. Nature reviews Microbiology. 2009; 7:787–797.

Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:10538–10543. [PubMed: 21680884]

Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics. 2007; 23:950–956. [PubMed: 17387114]

Garamszegi S, Franzosa EA, Xia Y. Signatures of Pleiotropy, Economy and Convergent Evolution in a Domain-Resolved Map of Human–Virus Protein–Protein Interaction Networks. PLoS pathogens. 2013; 9

Goh GK, Dunker AK, Uversky VN. Protein intrinsic disorder and influenza virulence: the 1918 H1N1 and H5N1 viruses. Virology journal. 2009; 6:69. [PubMed: 19493338]

Gorbalenya AE. Host-related sequences in RNA viral genomes. Seminars in Virology. 1992; 3:359–371.

Hagai T, Toth-Petroczy A, Azia A, Levy Y. The origins and evolution of ubiquitination sites. Mol Biosyst. 2012; 8:1865–1877. [PubMed: 22588506]

Handa Y, Durkin CH, Dodding MP, Way M. Vaccinia Virus F11 Promotes Viral Spread by Acting as a PDZ-Containing Scaffolding Protein to Bind Myosin-9A and Inhibit RhoA Signaling. Cell host & microbe. 2013; 14:51–62. [PubMed: 23870313]

Henderson BR, Percipalle P. Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-beta. Journal of molecular biology. 1997; 274:693–707. [PubMed: 9405152]

Ideker T, Krogan NJ. Differential network biology. Molecular systems biology. 2012; 8:565. [PubMed: 22252388]

Igarashi M, Ito K, Kida H, Takada A. Genetically destined potentials for N-linked glycosylation of influenza virus hemagglutinin. Virology. 2008; 376:323–329. [PubMed: 18456302]

Jackson T, King AM, Stuart DI, Fry E. Structure and receptor binding. Virus research. 2003; 91:33–46. [PubMed: 12527436]

Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. Trends in genetics: TIG. 2009; 25:193–197. [PubMed: 19349092]

Lauring AS, Andino R. Quasispecies theory and the behavior of RNA viruses. PLoS pathogens. 2010; 6:e1001005. [PubMed: 20661479]

Lu X, Shi Y, Gao F, Xiao H, Wang M, Qi J, Gao GF. Insights into avian influenza virus pathogenicity: the hemagglutinin precursor HA0 of subtype H16 has an alpha-helix structure in its cleavage site with inefficient HA1/HA2 cleavage. Journal of virology. 2012; 86:12861–12870. [PubMed: 22993148]

Machesky LM, Insall RH, Volkman LE. WASP homology sequences in baculoviruses. Trends in cell biology. 2001; 11:286–287. [PubMed: 11434350]

Martin-Serrano J, Neil SJ. Host factors involved in retroviral budding and release. Nature reviews Microbiology. 2011; 9:519–531.

Nguyen Ba AN, Moses AM. Evolution of characterized phosphorylation sites in budding yeast. Molecular biology and evolution. 2010; 27:2027–2037. [PubMed: 20368267]

Odom MR, Hendrickson RC, Lefkowitz EJ. Poxvirus protein evolution: family wide assessment of possible horizontal gene transfer events. Virus research. 2009; 144:233–249. [PubMed: 19464330]

Ortiz JF, MacDonald ML, Masterson P, Uversky VN, Siltberg-Liberles J. Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome biology and evolution. 2013; 5:504–513. [PubMed: 23418179]

Palmeira L, Penel S, Lotteau V, Rabourdin-Combe C, Gautier C. PhEVER: a database for the global exploration of virus-host evolutionary relationships. Nucleic Acids Res. 2011; 39:D569–575. [PubMed: 21081560]

Pantua H, Diao J, Ultsch M, Hazen M, Mathieu M, McCutcheon K, Takeda K, Date S, Cheung TK, Phung Q, et al. Glycan shifting on hepatitis C virus (HCV) E2 glycoprotein is a mechanism for escape from broadly neutralizing antibodies. Journal of molecular biology. 2013; 425:1899–1914. [PubMed: 23458406]

Pim D, Bergant M, Boon SS, Ganti K, Kranjec C, Massimi P, Subbaiah VK, Thomas M, Tomaic V, Banks L. Human papillomaviruses and the specificity of PDZ domain targeting. The FEBS journal. 2012; 279:3530–3537. [PubMed: 22805590]

Pushker R, Mooney C, Davey NE, Jacque JM, Shields DC. Marked variability in the extent of protein disorder within and between viral families. PloS one. 2013; 8:e60724. [PubMed: 23620725]

Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. Journal of virology. 2009; 83:10719–10736. [PubMed: 19640978]

Reed KE, Gorbalenya AE, Rice CM. The NS5A/NS5 proteins of viruses from three genera of the family flaviviridae are phosphorylated by associated serine/threonine kinases. Journal of virology. 1998; 72:6199–6206. [PubMed: 9621090]

Shackelton LA, Holmes EC. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends in microbiology. 2004; 12:458–465. [PubMed: 15381195]

Sun S, Wang Q, Zhao F, Chen W, Li Z. Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. PloS one. 2011; 6:e22844. [PubMed: 21829533]

Teyra J, Sidhu SS, Kim PM. Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. FEBS letters. 2012; 586:2631–2637. [PubMed: 22691579]

Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002; 27:527–533. [PubMed: 12368089]

Tsai YH, Wu MF, Wu YH, Chang SJ, Lin SF, Sharp TV, Wang HW. The M type K15 protein of Kaposi's sarcoma-associated herpesvirus regulates microRNA expression via its SH2-binding motif to induce cell migration and invasion. Journal of virology. 2009; 83:622–632. [PubMed: 18971265]

Van Roey K, Dinkel H, Weatheritt RJ, Gibson TJ, Davey NE. The switches. ELM resource: a compendium of conditional regulatory interaction interfaces. Sci Signal. 2013; 6:rs7. [PubMed: 23550212]

Van Roey K, Gibson TJ, Davey NE. Motif switches: decision-making in cell regulation. Curr Opin Struct Biol. 2012; 22:378–385. [PubMed: 22480932]

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell. 2009; 138:198–208. [PubMed: 19596244]

Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. Journal of biomolecular structure & dynamics. 2012; 30:137–149. [PubMed: 22702725]

Zhang X, Perica T, Teichmann SA. Evolution of protein structures and interactions from the perspective of residue contact networks. Curr Opin Struct Biol. 2013; 23:954–963. [PubMed: 23890840]

Highlights

1. Viruses exploit molecular mimicry through host-like peptide motifs for replication
2. Motifs are used by different viruses to different extent
3. Rapid evolution of viral peptide motifs may rewire host-virus interaction networks
4. Co-occurring motifs facilitate combinatorial regulation of host-virus interactions

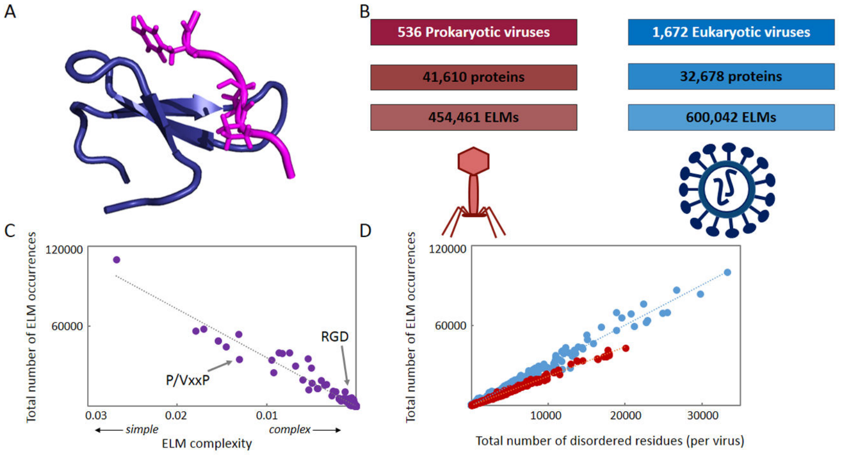

Figure 1. ELMs and viral proteins

(A) An example of a viral motif-host domain interaction: The PPxY motif of the Epstein–Barr virus LMP2 protein (in magenta) interacts with the host E3 ligase WW domain (in purple) to promote degradation of Tyr-kinases (PDB: 2JO9). (B) A non-redundant set of 2,208 viruses, in which 1,672 Eukaryotic viruses were compared to 536 prokaryotic viruses. (C) A correlation between ELM complexity (according to their information content) and their observed occurrence in total in the entire viral proteome ($y = 4*10^6x + 225$; r^2=0.96). P/VxxP (an SH3-domain binding motif) is an example of a simple ELM and RGD (an integrin binding motif) is an example of a complex ELM. (D) Correlations between disorder content (the total number of disordered residues in a virus) and the total number of ELM occurrences in that virus in eukaryotic (blue; $y=2*10^6x+78$; r^2=0.94) and prokaryotic viruses (red; $y=2*10^6x+147$; r^2=0.93).

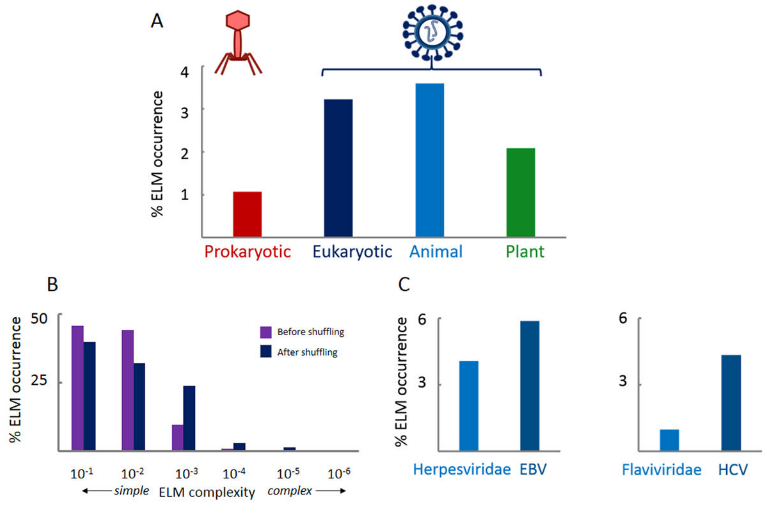

Figure 2. Occurrence of ELMs that are rare in shuffled sequences

(A) The percentage of ELMs that occur in less than 0.1% of the 100,000 shuffled sequences in prokaryotic (red), eukaryotic (dark blue), animal (cyan) and plant (green) viruses. Eukaryotic viruses have significantly higher fractions of ELMs that are hard to achieve by random shuffling. (B) The distribution of ELMs in eukaryotic viruses (as a function of complexity): the entire set of ELM-matching patterns (before shuffling, in purple) and the subset of ELMs that occur in less than 0.1% of the 100,000 shuffled sequences (in blue). (C) The percentage of ELMs that occur in less than 0.1% of the shuffled sequences in two viral families and two species (3 strains of EBV and 6 strains of HCV).

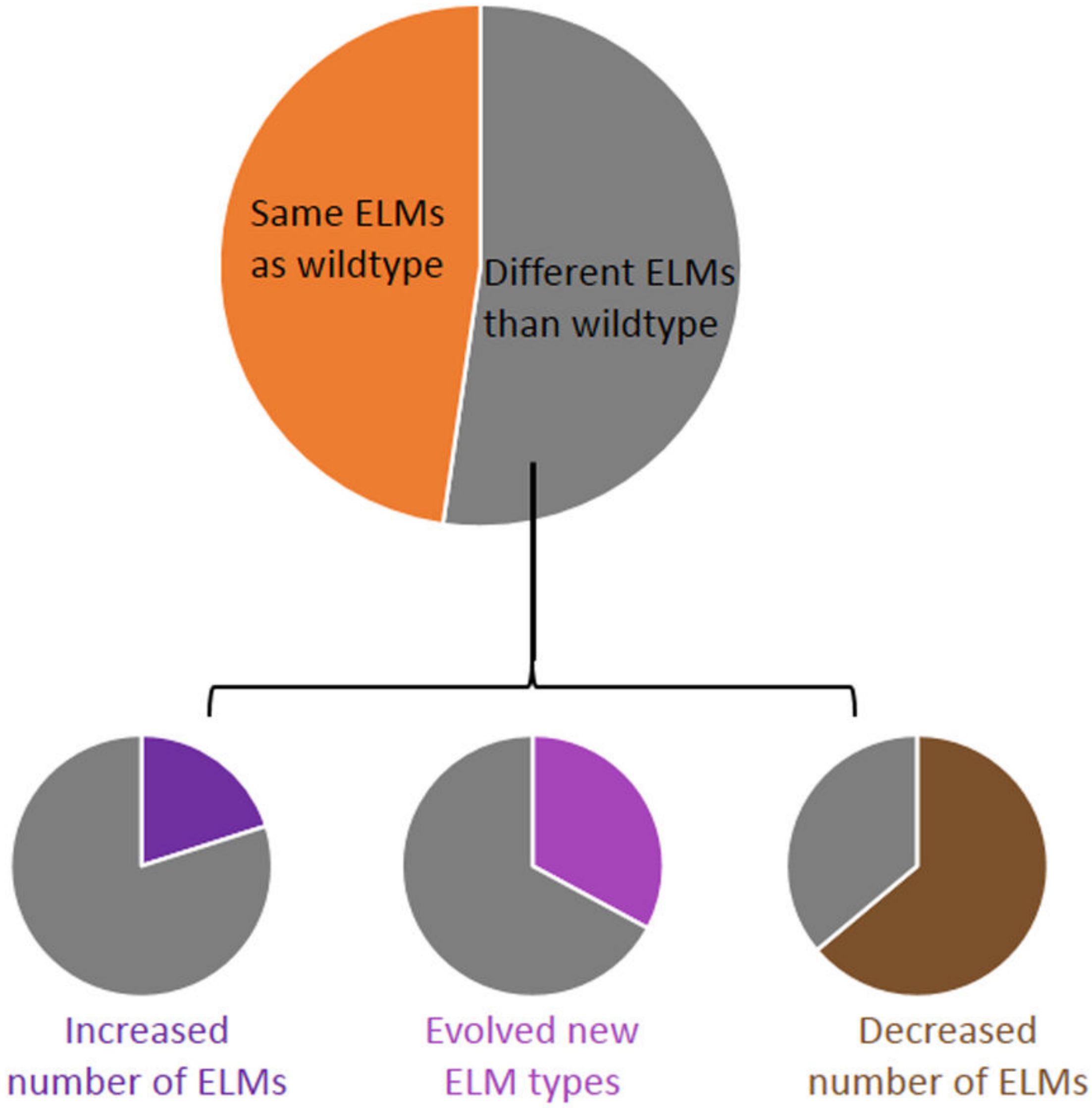

Figure 3. The effects of single non-synonymous mutations on the occurrence of ELMs in HIV-1 genes

Top - 47.7% of the mutants remain with the same distribution of ELMs as occurs in the wildtype (in orange) 59% of them occur outside of ELMs regions, while 41% occur within ELMs but still conform to the wildtype ELM(cyan)). Of the remaining 52.3% mutants (left circle in gray) – which differ in their ELMs – 33.3% have a reduced number of ELMs (red, top right circle), 27.7% have an increased number of ELMs (purple, middle circle), and 32.9% have evolved a new type of ELM (violet, bottom circle), with respect to the wildtype.

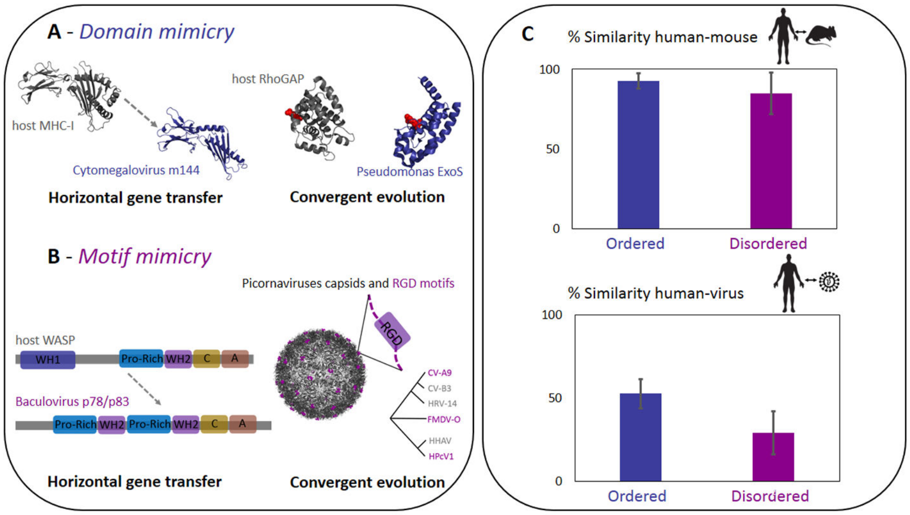

Figure 4. The evolutionary origins of viral mimics

(A) Structured domain mimics can be acquired from the host through HGT (such as in the case of the cytomegalovirus MHC-I mimic m144 (PDB: 1U58, purple) that highly resembles in sequence and structure the murine homolog (PDB: 1VAC, grey)); or evolve in a convergent manner (such as the pathogen RhoGAP mimic (PDB: 1HE1, purple) that has similar activity to that of the host (PDB: 1TX4, grey) (in red – the two Arg fingers which are important for the GTPase reaction and are similarly positioned) despite of no sequence similarity. (B) Motif mimics can less frequently be acquired by HGT, such as the WH2 motif occurrence in baculoviruses that is similar in sequence and in location of other regions to host WASP (several regions and motifs are shown, based on a previous annotation(Machesky et al., 2001) Many motifs emerge in pathogens in a convergent manner, such as the RGD motif which is found on the capsid surface of various unrelated picornaviruses to support their cell entry (a schematic clade with several picornaviruses is shown; RGD-containing species appear in purple). (C) The median of similarities of human-mouse and human-virus homolog pairs in ordered and disordered regions.

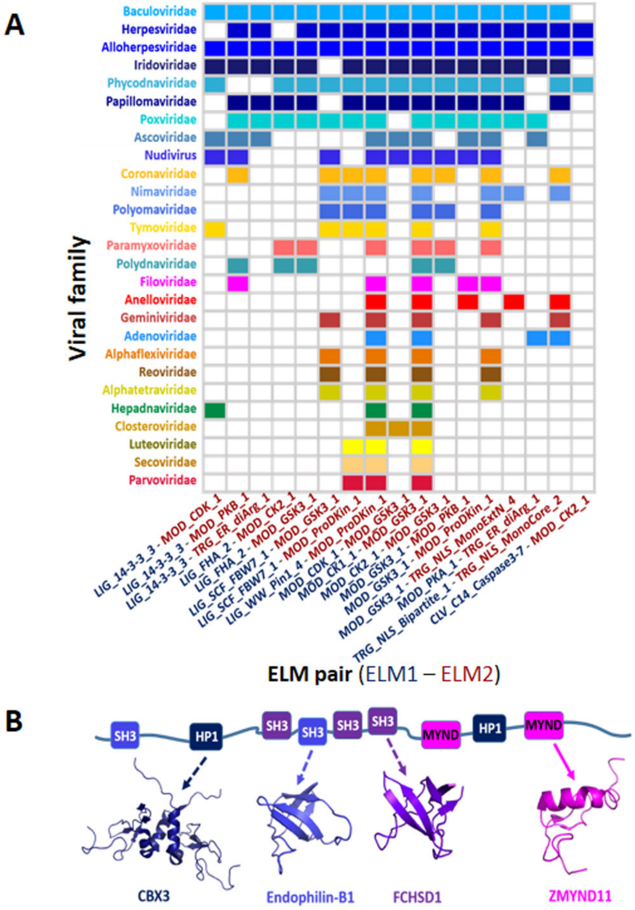

Figure 5. Occurrence of multiple types of ELMs

(A) Occurrence of 17 ELM switches in various viral families (In the Y-axis, families are colored with various shades according to their replication types: blue – dsDNA, red – ssDNA, green – RT, yellow - +ssRNA, pink - -ssRNA, dark brown – dsRNA). Each of these 17 pairs (the names of the two ELMs that compose the switch are marked in the X-axis in blue and red in the bottom) occurs in several viral families as well as in their host – suggesting convergent evolution in using ELM switches by unrelated viruses and to similarity to their host. (B) EBNA-2 ELMs and their known or suggested interactions with host proteins (solid or dashed arrow respectively) (structures of the host's interacting domains appear in matching colors). We associated identified ELMs in EBNA-2 with domains of host proteins that are known to interact with this viral protein according to a two-hybrid screening(Calderwood et al., 2007). In each case, the link was made based on the ELM type and the occurrence of a relevant domain in the host interacting proteins (e.g – an SH3-binding motif in EBNA-2 was linked to the host Endophillin-B1 which contains an SH3 domain).

Table 1

The relative occurrence of inferred functional ELMs and disordered regions in groups of viral proteins with specific functions

Group	number of proteins	average of % functional ELMs	p-value	median of % disorder	p-value
Molecular mimicry and host modulation					
inferred HGTs	795	0.34	9.91E-07*	4.27	6.03E-11*
virulence	35	1.37	NS	2.35	0.021417*
Structural proteins					
virion	2,941	1.79	3.33E-05	17.42	1.62E-103
Entry, exit and movement within and between cells					
viral budding via host ESCRT complexes	46	1.46	4.03E-02	39.15	1.10E-11
viral movement protein	166	0.70	NS	19.00	1.88E-14
Subcellular location					
cytoplasm	764	1.61	9.08E-04	18.38	1.41E-31
endosome	80	1.90	1.05E-02	11.36	0.031649*
mitochondrion	24	2.80	NS	17.92	NS
nucleus	1,093	2.34	6.54E-14	23.95	6.32E-81
Early and late proteins					
early	517	1.66	NS	13.72	2.35E-07
late	468	2.43	1.54E-02	16.35	2.14E-09
Chemically modified proteins					
lipoprotein	169	1.61	1.18E-02	17.20	5.54E-08
phosphoprotein	545	2.83	6.20E-15	38.26	1.41E-87
entire viral set	**32,672**	**1.62**		**7.34**	

The distribution of the fractions of functional ELMs and the fractions of disordered regions for each group was compared with that of the entire eukaryotic viral set. P-values imply enrichment in ELMs and disorder with respect to the entire viral set, except for cases marked with asterisks (*) that denote significant depletion. NS = not significant (depletion or enrichment).