Cell Reports
Article

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, San Francisco, 600 16th Street, GH-S572, UCSF Box 2280, San Francisco, CA 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
[4]Co-senior author
*Correspondence: madanm@mrc-lmb.cam.ac.uk (M.M.B.), raul.andino@ucsf.edu (R.A.)
http://dx.doi.org/10.1016/j.celrep.2014.04.052

SUMMARY

Viruses interact extensively with host proteins, but the mechanisms controlling these interactions are not well understood. We present a comprehensive analysis of eukaryotic linear 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 been suggested, based on a literature survey, that viruses may employ short, unstructured elements, which 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 posttranslational modifications (e.g., phosphorylation sites).

These small interaction modules are usually composed of two to eight 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 Figures 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 picornavirus 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).

Despite their potential importance in mediating host-virus interactions, the set of studied ELMs is limited and is mostly biased toward 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

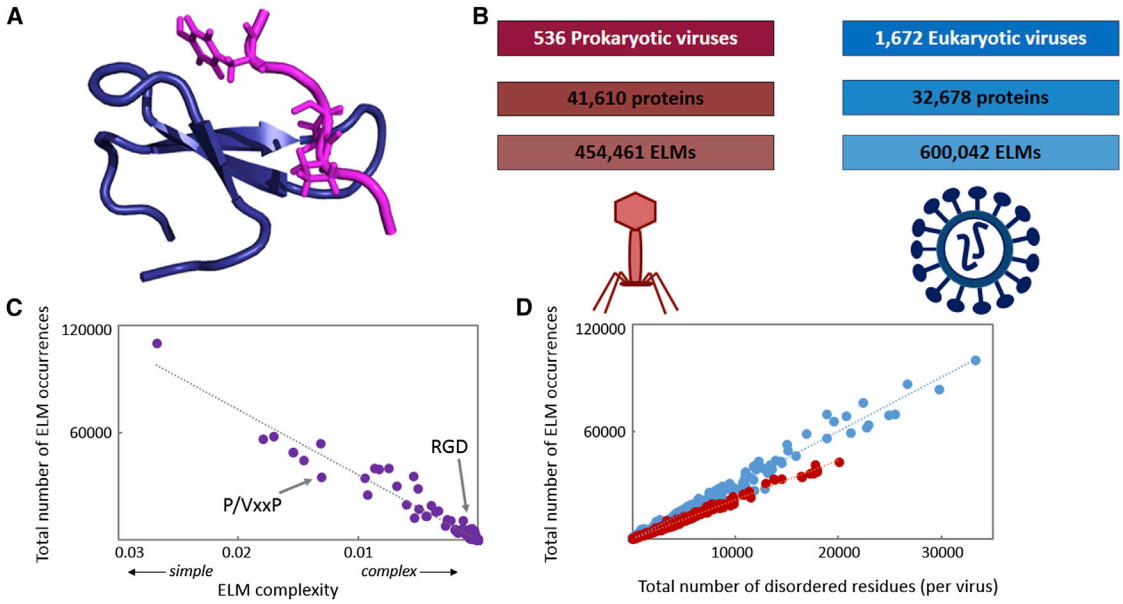

Figure 1. ELMs and Viral Proteins

(A) An example of a viral motif-host domain interaction. The PPxY motif of the EBV LMP2 protein (in magenta) interacts with the host E3 ligase WW domain (in purple) to promote degradation of Tyr kinases (PDB, 2JO9).

(B) A nonredundant set of 2,208 viruses, in which 1,672 eukaryotic viruses were compared to 536 prokaryotic viruses, is shown.

(C) A correlation between ELM complexity (according to its information content) and its observed occurrence in total in the entire viral proteome: $y = 4 \times 10^6 \times + 225$; $r^2 = 0.96$. 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; $y = 2 \times 10^6 \times +78$; $r^2 = 0.94$) and prokaryotic viruses (red; $y = 2 \times 10^6 \times +147$; $r^2 = 0.93$).

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 because 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 because 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 data set of ELMs to examine their occurrence in various virus families and to study ELM 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 data set of 2,208 nonredundant viruses, representing all orders and most known viral families (see Table S1 and Experimental Procedures). The data set 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 (Figure 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; Experimental Procedures).

We found that the total number of occurrences of each ELM-matching region (hereafter referred to 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 (Figure 1C). Furthermore, the total number of ELMs that occurs 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 (Figure 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 (Figure 1D), may serve as a basis to identify ELMs that are likely to be functional.

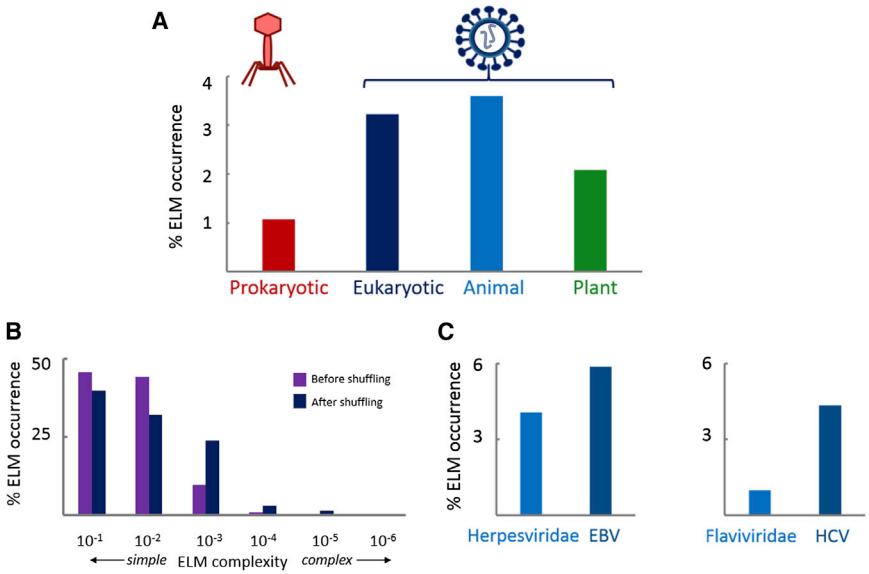

Figure 2. Occurrence of ELMs that Are Rare in Shuffled Sequences

(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) are shown.

(C) The percentage of ELMs that occur in less than 0.1% of the shuffled sequences in two viral families and two species (three strains of EBV and six strains of HCV).

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 Figure S2 and Experimental Procedures). Our premise is that regions matching known ELMs that are more rare 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 Experimental Procedures). For each of the 173 different ELM types that occur in each natural viral protein, 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://misc.hpse.ucsf.edu/~tzachi.hagai/viral_elms/.

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 (Figures 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) (Figure 2A). Only 1.1% of the ELMs observed in prokaryotic viruses occur in less than 0.1% of the shuffled sequences, in comparison with 3.2% in eukaryotic viruses and 3.6% in animal viruses. This represents a highly significant enrichment ($p < 10^{-15}$, Fisher's 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 (Figure S3; Experimental Procedures) (Davey et al., 2012a). Furthermore, the enrichment we observe is independent of the shuffling method (Figure 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 Figure 1A). We found a significant enrichment of ELMs that we identified to be potentially functional in a set of 42 experimentally validated functional viral ELMs (Dinkel et al., 2012) (a 6-fold enrichment with respect to their occurrence in the rest of the ELM data set; 19% overlap; $p = 6.5 \times 10^{-6}$, Fisher's exact test; Figure S4). This observation supports the notion that the set we identified is indeed enriched with functional ELMs.

Table 1. The Relative Occurrence of Inferred Functional ELMs and Disordered Regions in Groups of Viral Proteins with Specific Functions

Group	No. of Proteins	Average of Percentage of Functional ELMs	p Value	Median of Percentage of Disorder	p Value
Molecular mimicry and host modulation					
Inferred HGTs	795	0.34	9.91×10^{-7}[a]	4.27	6.03×10^{-11}[a]
Virulence	35	1.37	NS	2.35	0.021417[a]
Structural proteins					
Virion	2,941	1.79	3.33×10^{-5}	17.42	1.62×10^{-103}
Entry, exit, and movement within and between cells					
Viral budding via host ESCRT complexes	46	1.46	4.03×10^{-2}	39.15	$1.10 \times 10^{-1}1$
Viral movement protein	166	0.70	NS	19.00	1.88×10^{-14}
Subcellular location					
Cytoplasm	764	1.61	9.08×10^{-4}	18.38	1.41×10^{-31}
Endosome	80	1.90	1.05×10^{-2}	11.36	0.031649[a]
Mitochondrion	24	2.80	NS	17.92	NS
Nucleus	1,093	2.34	6.54×10^{-14}	23.95	6.32×10^{-81}
Early and late proteins					
Early	517	1.66	NS	13.72	2.35×10^{-7}
Late	468	2.43	1.54×10^{-2}	16.35	2.14×10^{-9}
Chemically modified proteins					
Lipoprotein	169	1.61	1.18×10^{-2}	17.20	5.54×10^{-8}
Phosphoprotein	545	2.83	6.20×10^{-15}	38.26	1.41×10^{-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 an "a" that denote significant depletion. NS, not significant (depletion or enrichment).

Family and Species Level Analysis Reveals Enormous Heterogeneity of ELM Usage among Eukaryotic Viruses

Although some viruses contain proteins with many ELMs, others appear to have only a few ELMs. For instance, many double-stranded DNA (dsDNA) virus families, such as *Papillomaviridae*, *Adenoviridae*, and *Herpesviridae*, which 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 single-stranded RNA (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 various 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 (Figure 2C). The variation in ELM content between viruses could be related to several factors, such as virus life cycle and length of infection. For example, some persistent viruses might require a more precise regulation of cellular pathways to ensure that the cell remains functional throughout their longer infection time. Thus, they might need to carefully regulate the expression of disordered regions that might be harmful to the cell (Babu et al., 2011; Vavouri et al., 2009). Other factors include 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

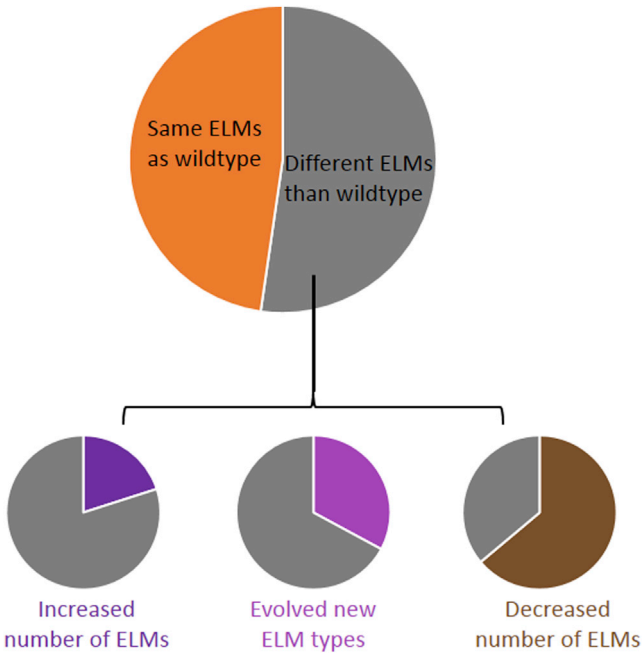

Figure 3. The Effects of Single Nonsynonymous Mutations on the Occurrence of ELMs in HIV-1 Genes

In the top view, 47.7% of the mutants remain with the same distribution of ELMs as occurs in the wild-type (in orange), 59% of them occur outside of ELM regions, whereas 41% occur within ELMs but still conform to the wild-type ELM. Of the remaining 52.3% mutants, which differ in their ELMs (gray fraction of the top circle), 33.3% have a reduced number of ELMs (brown, right circle), 27.7% have an increased number of ELMs (purple, left circle), and 32.9% have evolved a new type of ELM (pink, middle circle), with respect to the wild-type.

domain mimics that were likely transferred from host genomes through horizontal gene transfer (HGT). Other groups of viral proteins that modulate host pathways, 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 ELM content—nuclear viral proteins tend to be relatively enriched with ELMs, whereas endoplasmic reticulum 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 nonfunctional 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 (Figure S5; Experimental Procedures). 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 ($p = 2.2 \times 10^{-55}$ and $p = 1.2 \times 10^{-49}$; sign test). Consistent with these results, we also observed the inferred functional ELMs in the six strains of HCV to be more conserved in an independent set of variant HCV sequences (Experimental Procedures). 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 Adaptation 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), appear in different numbers, and change their primary sequence patterns, while still maintaining functionality, as observed, for example, in ELMs that mediate interactions with the endosomal sorting complexes required for transport (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 hr in vivo upon infection (Coffin, 1995). We then examined how nonsynonymous mutations in disordered regions in this viral population affect the distribution of ELMs in comparison with their occurrence in the wild-type HIV-1 genome (Figure 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 (Figure 3, top). Of the other half of mutants, i.e., those that differ in their ELM distributions, a significant part had either increased the number of existing ELMs or evolved new types of ELMs with respect to the wild-type (Figure 3, bottom circles, purple and pink fractions, respectively). Many viruses have high mutation 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 HGT or evolve independently in a convergent manner. In structured domain mimicry, it is generally assumed that domains that are found in pathogen proteins and have high sequence similarity in a large portion of the domain

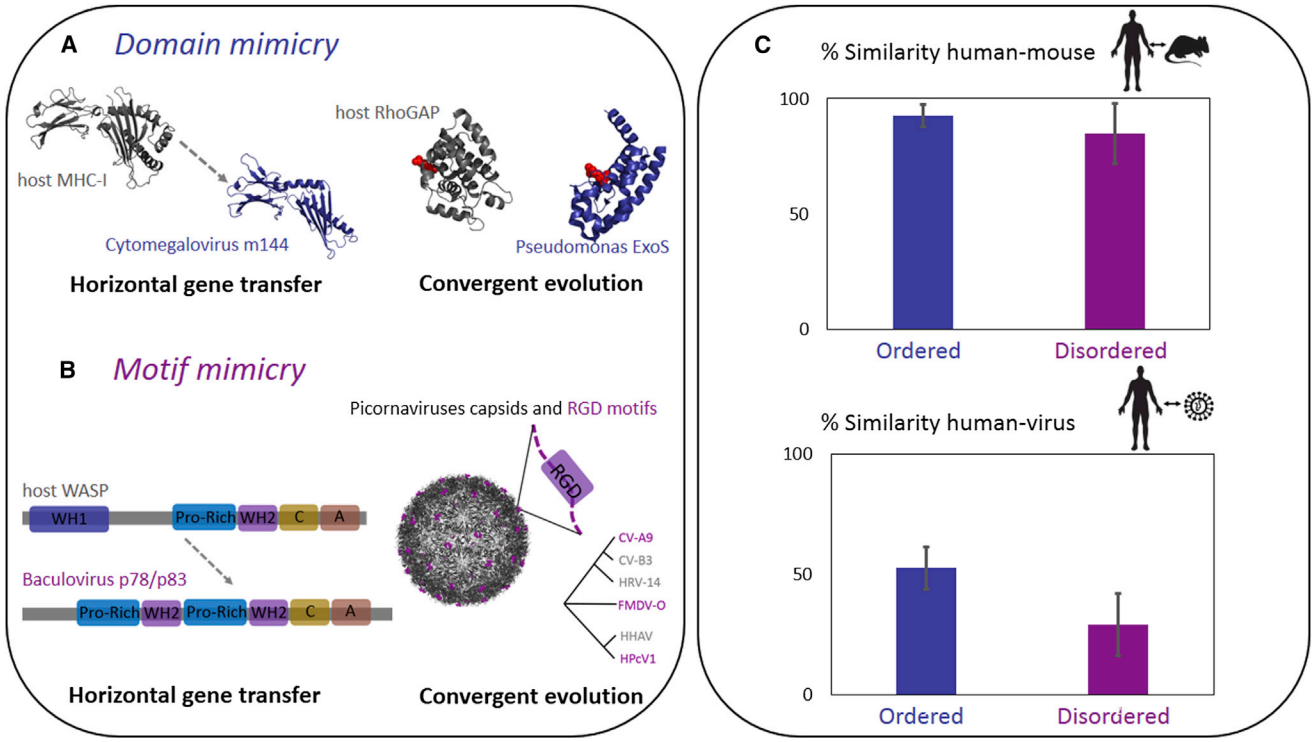

Figure 4. The Evolutionary Origins of Viral Mimics

(A) Structured domain mimics can be acquired from the host through HGT, such as in the case of the cytomegalovirus major histocompatibility complex-I (MHC-I) mimic m144 (PDB, 1U58, purple) that highly resembles in sequence and structure the murine homolog (PDB, 1VAC, gray), or evolve in a convergent manner, such as the pathogen RhoGAP mimic (PDB, 1HE1, purple) that has similar activity to that of the host RhoGAP (PDB, 1TX4, gray) despite of no sequence similarity (the two Arg fingers that are important for the GTPase reaction and are similarly positioned are shown in red).

(B) Motif mimics are less frequently acquired by HGT, such as in the case of 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 by Machesky et al. (2001). Many motifs emerge in pathogens in a convergent manner, such as the integrin receptor-binding 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 (error bars indicate SD).

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) (Figure 4A). Because 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 (Figure 4B). However, instances where ELMs were transferred from host genes and maintained during the pathogen's evolution are also known. For example, in the baculovirus p78/83 protein, the occurrence of the actin-binding WH2 motif, which is a fairly complex and long motif, is likely to be a result of HGT because the motif and the regions surrounding it are relatively similar to the host Wiskott-Aldrich syndrome protein (WASP), from which these regions are thought to originate (Machesky et al., 2001) (Figure 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 ELM composition in comparison with that of the host homologous proteins. For this, we extracted 135 nonredundant animal viral proteins and their inferred homologs in human and mouse from the PhEVER database (Palmeira et al., 2011), which is 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 Experimental Procedures). 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 (Figure 4C).

Importantly, whereas 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 ($p = 8.1 \times 10^{-13}$, sign test; see Experimental Procedures). 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 ELM 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 ELM occurrences in the entire proteome (see Experimental Procedures). We observed 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, because they can be brought together in the 3D space or might cooperatively assist the binding of another ELM. Although many viral proteins are depleted of ELMs (as discussed above and as shown in Figure S6), some viral proteins tend to contain numerous types of ELMs within their disordered regions. We searched for cases of ELMs that tend to co-occur in the same protein in a nonredundant set of viral proteins (see Experimental Procedures). Because we have a total of 173 types of ELMs in our data set, there are ~15,000 ELM pairs that could theoretically occur. We compared the 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 Experimental Procedures). 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 the 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 a 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 data set 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 ($p = 3.3 \times 10^{-16}$, Fisher's 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, suggests 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 that 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 switch to modulate host-virus-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 Figure 5A for the

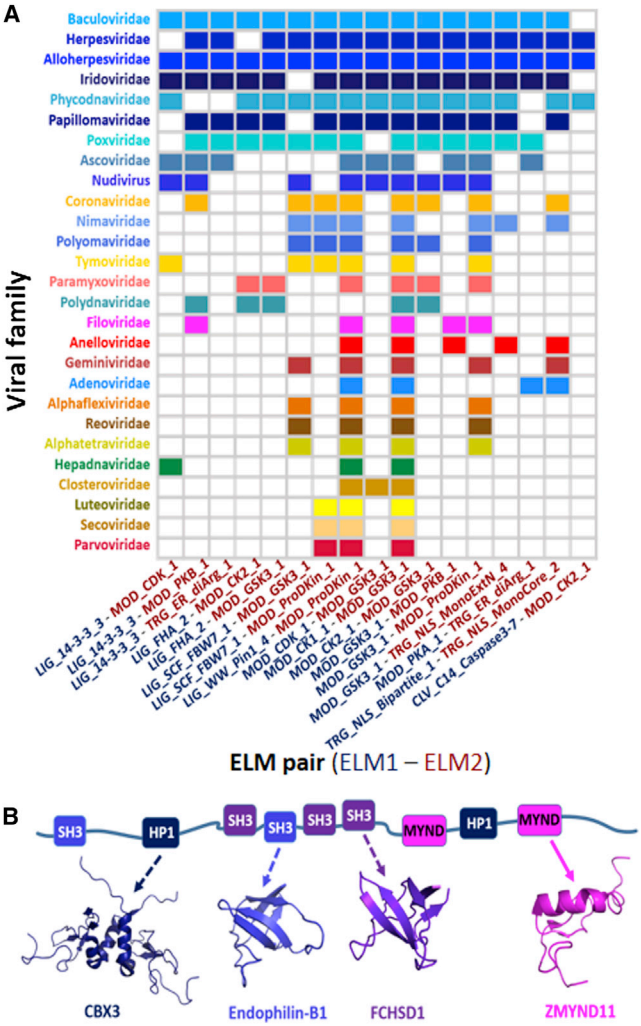

Figure 5. Occurrence of Multiple Types of ELMs

(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 indicates dsDNA, red is ssDNA, green is RT (reverse-transcribing viruses), yellow is +ssRNA, pink is −ssRNA, and dark brown is dsRNA. Each of these 17 pairs 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; the names of the two ELMs that compose the switch are marked in the x axis in blue and red in the bottom.

(B) EBNA-2 ELMs and their known or suggested interactions with host proteins are indicated with a 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 by 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 endophilin-B1, which contains an SH3 domain).

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 that 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 (Figures S6A and S6B), 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 five different types of ELMs within the Epstein-Barr nuclear antigen-2 (EBNA-2) protein that can support its interactions with several known host factors (Figure 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.

Although 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 (Figure 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 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 because 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 nonpathogenic 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 (Figure 3). 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 (Figure 1A). This motif is conserved in LMP2 orthologs of the two additional strains of EBV in our data set 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 coexist 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 posttranslational 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 multifunctionality.

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 a better understanding of biomedically important viruses and to assist in developing ways to overcome them.

EXPERIMENTAL PROCEDURES

We composed a set of 2,208 nonredundant viruses from 108 viral clades using available viral entries from the National Center for Biotechnology Information Viral Genomes Resource (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/) and excluded viruses that had missing data or were too similar (Tables S1A and S1B). We predicted disorder values of protein sequences using the IUPred algorithm (Dosztányi 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 (Figure 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 (Figures 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 in shuffled sequences) 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 (Figure S5). In addition, to examine whether our results hold outside our data set, we repeated the above analysis using extracted sequences of variants of HCV from the Los Alamos HCV database (http://www.hcv.lanl.gov).

For functional enrichment analysis, we composed sets of eukaryotic viral proteins based on keyword annotations in UniProt (http://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 a one-sided Kolmogorov-Smirnov test.

We examined the putative evolutionary origins of viral ELMs from host genes using a nonredundant 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 ELM occurrences were randomly shuffled.

We analyzed the biophysical characteristics and number of ELMs in a set of nonredundant animal viral proteins and compared them to a set of nonredundant human proteins (Figure S6). ELM co-occurrence analysis was done by comparing the occurrence of each pair of ELMs (2 different types of the 173 ELMs) in the nonredundant 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://misc.hpse.ucsf.edu/~tzachi.hagai/viral_elms/. See the Supplemental Experimental Procedures for additional details.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.04.052.

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, R.D. Hernandez, E.D. Levy, and H.S. Malik for helpful discussions, R.M. 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 (to R.A.) and by the Medical Research Council (MC_U105185859), HFSP (RGY0073/2010), and EMBO Young Investigator Program and ERASysBio+ (to M.M.B.). T.H. is supported by a Human Frontier Science Program Long-Term Fellowship.

Received: January 6, 2014
Revised: March 25, 2014
Accepted: April 24, 2014
Published: May 29, 2014

REFERENCES

Akiva, E., Friedlander, G., Itzhaki, Z., and Margalit, H. (2012). A dynamic view of domain-motif interactions. PLoS Comput. Biol. *8*, e1002341.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Babu, M.M., van der Lee, R., de Groot, N.S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. Curr. Opin. Struct. Biol. *21*, 432–440.

Babu, M.M., Kriwacki, R.W., and Pappu, R.V. (2012). Structural biology. Versatility from protein disorder. Science *337*, 1460–1461.

Boon, S.S., and Banks, L. (2013). High-risk human papillomavirus E6 oncoproteins interact with 14-3-3ζ in a PDZ binding motif-dependent manner. J. Virol. *87*, 1586–1595.

Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., Ewence, A.E., Li, N., Hirozane-Kishikawa, T., Hill, D.E., et al. (2007). Epstein-Barr virus and virus human protein interaction maps. Proc. Natl. Acad. Sci. USA *104*, 7606–7611.

Carter, J.J., Daugherty, M.D., Qi, X., Bheda-Malge, A., Wipf, G.C., Robinson, K., Roman, A., Malik, H.S., and Galloway, D.A. (2013). Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. Proc. Natl. Acad. Sci. USA *110*, 12744–12749.

Coffin, J.M. (1995). HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science *267*, 483–489.

Das, S.R., Puigbò, P., Hensley, S.E., Hurt, D.E., Bennink, J.R., and Yewdell, J.W. (2010). Glycosylation focuses sequence variation in the influenza A virus H1 hemagglutinin globular domain. PLoS Pathog. *6*, e1001211.

Davey, N.E., Travé, G., and Gibson, T.J. (2011). How viruses hijack cell regulation. Trends Biochem. Sci. *36*, 159–169.

Davey, N.E., Cowan, J.L., Shields, D.C., Gibson, T.J., Coldwell, M.J., and Edwards, R.J. (2012a). SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res. *40*, 10628–10641.

Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012b). Attributes of short linear motifs. Mol. Biosyst. *8*, 268–281.

Dinkel, H., Michael, S., Weatheritt, R.J., Davey, N.E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., et al. (2012). ELM—the database of eukaryotic linear motifs. Nucleic Acids Res. *40* (Database issue), D242–D251.

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. *76*, 159–216.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Doxey, A.C., and McConkey, B.J. (2013). Prediction of molecular mimicry candidates in human pathogenic bacteria. Virulence *4*, 453–466.

Drayman, N., Glick, Y., Ben-nun-shaul, O., Zer, H., Zlotnick, A., Gerber, D., Schueler-Furman, O., and Oppenheim, A. (2013). Pathogens use structural mimicry of native host ligands as a mechanism for host receptor engagement. Cell Host Microbe *14*, 63–73.

Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. *18*, 756–764.

Dyer, M.D., Murali, T.M., and Sobral, B.W. (2008). The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog. *4*, e32.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Elde, N.C., and Malik, H.S. (2009). The evolutionary conundrum of pathogen mimicry. Nat. Rev. Microbiol. *7*, 787–797.

Franzosa, E.A., and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. Proc. Natl. Acad. Sci. USA *108*, 10538–10543.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. Bioinformatics *23*, 950–956.

Garamszegi, S., Franzosa, E.A., and Xia, Y. (2013). Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. PLoS Pathog. *9*, e1003778.

Goh, G.K., Dunker, A.K., and Uversky, V.N. (2009). Protein intrinsic disorder and influenza virulence: the 1918 H1N1 and H5N1 viruses. Virol. J. *6*, 69.

Gorbalenya, A.E. (1992). Host-related sequences in RNA viral genomes. Semin. Virol. *3*, 359–371.

Hagai, T., Tóth-Petróczy, A., Azia, A., and Levy, Y. (2012). The origins and evolution of ubiquitination sites. Mol. Biosyst. *8*, 1865–1877.

Handa, Y., Durkin, C.H., Dodding, M.P., and Way, M. (2013). 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 *14*, 51–62.

Henderson, B.R., and Percipalle, P. (1997). Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-beta. J. Mol. Biol. *274*, 693–707.

Ideker, T., and Krogan, N.J. (2012). Differential network biology. Mol. Syst. Biol. *8*, 565.

Igarashi, M., Ito, K., Kida, H., and Takada, A. (2008). Genetically destined potentials for N-linked glycosylation of influenza virus hemagglutinin. Virology *376*, 323–329.

Jackson, T., King, A.M., Stuart, D.I., and Fry, E. (2003). Structure and receptor binding. Virus Res. *91*, 33–46.

Landry, C.R., Levy, E.D., and Michnick, S.W. (2009). Weak functional constraints on phosphoproteomes. Trends Genet. *25*, 193–197.

Lauring, A.S., and Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. *6*, e1001005.

Lu, X., Shi, Y., Gao, F., Xiao, H., Wang, M., Qi, J., and Gao, G.F. (2012). 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. J. Virol. *86*, 12861–12870.

Machesky, L.M., Insall, R.H., and Volkman, L.E. (2001). WASP homology sequences in baculoviruses. Trends Cell Biol. *11*, 286–287.

Martin-Serrano, J., and Neil, S.J. (2011). Host factors involved in retroviral budding and release. Nat. Rev. Microbiol. *9*, 519–531.

Nguyen Ba, A.N., and Moses, A.M. (2010). Evolution of characterized phosphorylation sites in budding yeast. Mol. Biol. Evol. *27*, 2027–2037.

Odom, M.R., Hendrickson, R.C., and Lefkowitz, E.J. (2009). Poxvirus protein evolution: family wide assessment of possible horizontal gene transfer events. Virus Res. *144*, 233–249.

Ortiz, J.F., MacDonald, M.L., Masterson, P., Uversky, V.N., and Siltberg-Liberles, J. (2013). Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome Biol. Evol. *5*, 504–513.

Palmeira, L., Penel, S., Lotteau, V., Rabourdin-Combe, C., and Gautier, C. (2011). PhEVER: a database for the global exploration of virus-host evolutionary relationships. Nucleic Acids Res. *39* (Database issue), D569–D575.

Pantua, H., Diao, J., Ultsch, M., Hazen, M., Mathieu, M., McCutcheon, K., Takeda, K., Date, S., Cheung, T.K., Phung, Q., et al. (2013). Glycan shifting on hepatitis C virus (HCV) E2 glycoprotein is a mechanism for escape from broadly neutralizing antibodies. J. Mol. Biol. *425*, 1899–1914.

Pim, D., Bergant, M., Boon, S.S., Ganti, K., Kranjec, C., Massimi, P., Subbaiah, V.K., Thomas, M., Tomaić, V., and Banks, L. (2012). Human papillomaviruses and the specificity of PDZ domain targeting. FEBS J. *279*, 3530–3537.

Pushker, R., Mooney, C., Davey, N.E., Jacqué, J.M., and Shields, D.C. (2013). Marked variability in the extent of protein disorder within and between viral families. PLoS One *8*, e60724.

Rancurel, C., Khosravi, M., Dunker, A.K., Romero, P.R., and Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J. Virol. *83*, 10719–10736.

Reed, K.E., Gorbalenya, A.E., and Rice, C.M. (1998). The NS5A/NS5 proteins of viruses from three genera of the family *Flaviviridae* are phosphorylated by associated serine/threonine kinases. J. Virol. *72*, 6199–6206.

Shackelton, L.A., and Holmes, E.C. (2004). The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol. *12*, 458–465.

Sun, S., Wang, Q., Zhao, F., Chen, W., and Li, Z. (2011). Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. PLoS One *6*, e22844.

Teyra, J., Sidhu, S.S., and Kim, P.M. (2012). Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. FEBS Lett. *586*, 2631–2637.

Tompa, P. (2002). Intrinsically unstructured proteins. Trends Biochem. Sci. *27*, 527–533.

Tsai, Y.H., Wu, M.F., Wu, Y.H., Chang, S.J., Lin, S.F., Sharp, T.V., and Wang, H.W. (2009). 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. J. Virol. *83*, 622–632.

Van Roey, K., Gibson, T.J., and Davey, N.E. (2012). Motif switches: decision-making in cell regulation. Curr. Opin. Struct. Biol. *22*, 378–385.

Van Roey, K., Dinkel, H., Weatheritt, R.J., Gibson, T.J., and Davey, N.E. (2013). The switches.ELM resource: a compendium of conditional regulatory interaction interfaces. Sci. Signal. *6*, rs7.

Vavouri, T., Semple, J.I., Garcia-Verdugo, R., and Lehner, B. (2009). Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell *138*, 198–208.

Xue, B., Dunker, A.K., and Uversky, V.N. (2012). Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. *30*, 137–149.

Zhang, X., Perica, T., and Teichmann, S.A. (2013). Evolution of protein structures and interactions from the perspective of residue contact networks. Curr. Opin. Struct. Biol. *23*, 954–963.