Use of viral motif mimicry improves the proteome-wide discovery of human linear motifs

Bishoy Wadie*[1], Vitalii Kleshchevnikov*[1,2], Elissavet Sandaltzopoulou[1,3], Caroline Benz[4], Evangelia Petsalaki[+1]

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK

[2]Current address: Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

[3]Current address: Technical University of Dresden, Dresden, Germany

[4]Department of Chemistry – BMC, Uppsala University, Uppsala, Sweden

*equal contribution

[+]Correspondence should be addressed to: petsalaki@ebi.ac.uk

Abstract

Linear motifs have an integral role in dynamic cell functions including cell signalling, the cell cycle and others. However, due to their small size, low complexity, degenerate nature, and frequent mutations, identifying novel functional motifs is a challenging task. Viral proteins rely extensively on the molecular mimicry of cellular linear motifs for modifying cell signalling and other processes in ways that favour viral infection. This study aims to discover human linear motifs convergently evolved also in disordered regions of viral proteins, under the hypothesis that these will result in enrichment in functional motif instances. We systematically apply computational motif prediction, combined with implementation of several functional and structural filters to the most recent publicly available human-viral and human-human protein interaction network. By limiting the search space to the sequences of viral proteins, we observed an increase in the sensitivity of motif prediction, as well as improved enrichment in known instances compared to the same analysis using only human protein interactions. We identified > 8,400 motif instances at various confidence levels, 105 of which were supported by all functional and structural filters applied. Overall, we provide a pipeline to improve the identification of functional linear motifs from interactomics datasets and a comprehensive catalogue of putative human motifs that can contribute to our understanding of the human domain-linear motif code and the mechanisms of viral interference with this.

Keywords:

Host-viral, short linear motifs, motif mimicry, protein-protein interaction, protein domains, convergent evolution, sequence analysis, network analysis, structural bioinformatics, variants.

Introduction

Short Linear Motifs (SLiMs; also, often referred to as ELMs - Eukaryotic Linear Motifs) are short linear peptides approximately 3-10 amino acids long, that have a specific sequence pattern that is recognized by interacting domains (Van Roey et al., 2014). They typically lie in disordered regions of proteins, and they are mediators of critical, usually transient interactions involved in dynamic cell processes such as cell signaling, protein degradation, the cell cycle and others (Davey et al., 2011). Therefore, mutations in such motifs often lead to disease, and indeed they have been found to be enriched in cancer (Mészáros et al., 2017; Reimand et al., 2015; Uyar et al., 2014), and several other diseases (Ahmed et al., 2003; Furuhashi et al., 2005; Meyer et al., 2018; Pandit et al., 2007). Gene fusions in cancer often remove SLiM-containing protein regions and the capacity for gene regulation (Latysheva et al., 2016). Due to their short length, small number of specificity-determining residues and disordered nature, they are characterized by high evolutionary plasticity (Davey et al., 2015). They have therefore been used throughout evolution to rewire biological networks, increase their complexity and can often arise through convergent evolution (E. Davey et al., 2012).

A prime example of convergent evolution for the formation of linear motifs are those present in viruses and other pathogens (Davey et al., 2011; Via et al., 2015). Viruses often use molecular mimicry to target hubs or other strategic nodes in host networks and hijack cellular functions. For example the HIV-1 Nef protein interacts with the SH3 domain of the Src family kinases Hck, Lyn and c-Src via a PxxP motif, which leads to their strong activation (Trible et al., 2006). It has been shown that viruses often act in similar ways as oncogenic variations to cause cancer (Rozenblatt-

Rosen et al., 2012). A well-studied example of this is the LxCxE motif in the E7 protein of HPV and other viral proteins (Felsani et al., 2006) which acts to inactivate the retinoblastoma protein leading to deregulation of cell cycle control; this inactivation is also common in non-viral cancers (Horowitz et al., 1989; Weinberg, 1995). Viral motifs are generally similar to the host ones, however they tend to be of higher affinity so as to outcompete them (Sheng et al., 2006).

Due to their small size and often low complexity, linear motifs tend to be difficult to discover both by experimental and computational methodologies. In the past decade the availability of high-throughput protein interaction and genomic datasets (Huttlin et al., 2017; Luck et al., 2017; Orchard et al., 2014) have allowed significant advances to be made in developing computational tools for motif discovery. DiliMot (Neduva et al., 2005) and SliMFinder (Edwards et al., 2007) are the first methods able to predict the presence of human motifs at scale, by evaluating over-representation of specific motifs in the interactors of a specific protein and combining it with conservation measures. QSLiMFinder (Palopoli et al., 2015) provides additional functionality by requiring the hits to be present in additional query proteins. FIRE-Pro has also been used to discover yeast motifs in proteome scale datasets based on an information theoretic framework (Lieber et al., 2010). The ELM database and server (Kumar et al., 2020), and SLiMSearch (Krystkowiak and Davey, 2017) can search individual protein sequences or entire proteomes respectively for a given motif pattern. Edwards and colleagues (J. Edwards et al., 2012) demonstrated that using these tools to scan the entire human interactome allows the identification of not only new instances of known motifs, but also entirely novel ones. There are multiple other methods available for discovery of *de-novo*, known and user-defined motifs which can be used for simple protein search of motif patterns or at a much larger scale (Edwards and Palopoli, 2015; Hraber et al., 2020).

By scanning the disordered proteome of a large number of viruses for instances of ELMs, several instances of known SLiMs have been identified (Becerra et al., 2017; Hagai et al., 2014; Liu-Wei et al., 2021). And during the current pandemic, multiple resources have emerged to delineate SLiM based viral-host interactions and discover SLiM instances in SARS-CoV-2 (Li et al., 2021; Mészáros et al., 2021; Yang and Shi, 2021). However, this kind of approach doesn't consider the wealth of information included in the interacting partners of viral proteins. It also doesn't allow the discovery of motifs that are not already annotated in public databases. Zheng et al. (Zheng et al., 2014) used the available host-viral interactomes to identify domain-domain interactions between them but did not search for enrichment of linear motif-mediated interactions.

Despite advances in computational discovery of linear motifs the problem remains that only a small fraction of identified motifs is likely to be functional. Thus, additional layers of evidence for each hit can be important to help prioritize them. Here, we hypothesize that requiring human motifs to be present also in viral proteins with common interactors will improve their identification, while also pointing to nodes of viral interference with human host networks. We have thus combined proteome-wide motif search methods with domain enrichment, structural bioinformatics and integration with pathogenic variant information, to perform the first systematic search for human linear motifs also present in common viral interactors, based on the latest available human-viral interactome.

Results

A workflow for the discovery of functional linear motifs

Since viruses commonly interfere with human protein interaction networks through viral motif mimicry, we hypothesized that requiring putative new linear motifs to be present both in human and viral protein interactors of specific protein domains would result in enrichment of functional motif instances. Our workflow is based on this principle and in addition filters motifs according to other relevant parameters, such as domain enrichment, predicted binding on domain structure and others (**Figure 1**; Methods).

In brief, for each known human-viral protein interaction pair (H1-V), we first used QSLiMFinder (Palopoli et al., 2015) to identify sequence patterns that were both present in the respective viral protein and enriched in the protein interacting partners of the human protein H1 of the query pair (H1-H2, H1-H3 etc). For brevity, throughout the text, V denotes a viral protein that potentially carries a motif, H2, H3 etc human putative motif-carrying proteins and H1 the human domain-carrying protein that interacts with both V and the H2, H3 etc proteins. We thus identified 8,403 putative linear motif instances on 2,757 unique proteins (2457 human and 300 viral) **(Supplementary Table 1)**. An interaction-mediating motif is typically recognized by specific domains. To reduce the potential spurious motifs from our identified set, we thus imposed as an additional filter a requirement for a domain enrichment in the interacting partners of each motif-bearing protein **(**adjusted p-value <0.05 & present in at least 5 interacting partners **(Supplementary Table 1**)**.** This reduced our initial list to 1,970 motif instances presenting 1.8-fold higher enrichment in proteins including true SLiMs compared to raw predicted hits (**Figure 2**

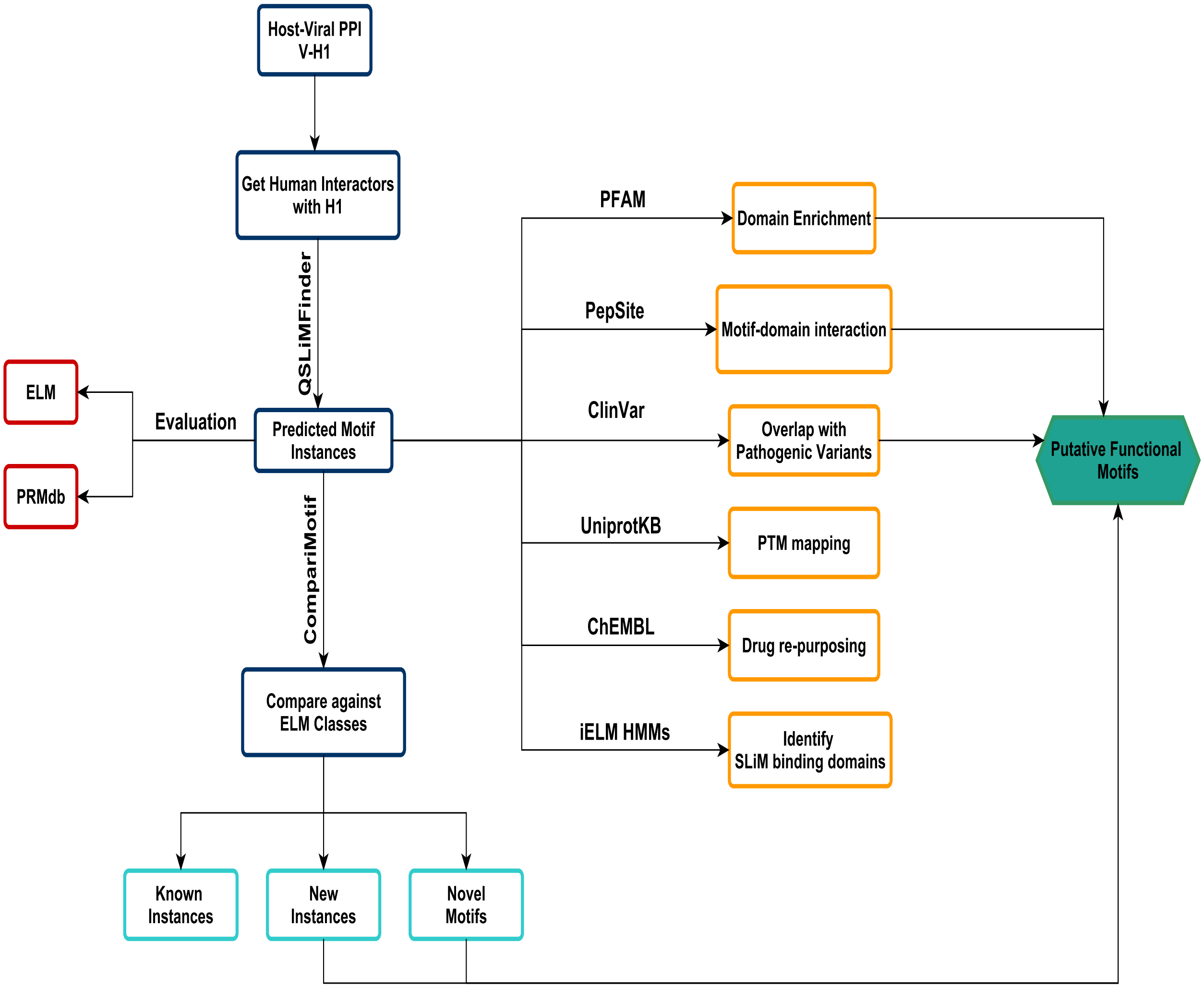


Figure 1: Workflow for discovery of short linear motifs. A schematic describing the overall pipeline for this study. For each from viral-human (V-H1) protein interaction downloaded from IntAct, the respective set of the human interacting partners of H1 were submitted to QSLiMFinder for motif prediction restricting hits to those present also in the respective viral interactor. Then, the predicted hits were evaluated against both ELM and PRMdb to measure how well they capture known hits. In addition, they were also compared against known ELM classes to further differentiate between known, new and novel motif instances. Finally, they were subjected to multiple downstream filters to reduce false positives and further enrich functional instances. Arrow labels show the method/ resource used for each step. V: Viral Protein; H: Human protein which directly interacts with viral protein.

C,D; Methods). As part of the workflow, we also use HMM profiles from the iELM web server (Weatheritt et al., 2012) to identify domains that have a match to domain profiles known to bind to known ELM motifs. While this leads to an improvement in the enrichment of known motifs (**Figure S1**), and is provided for information **(Supplementary Table 1**), we don't use it as a filter, as there are only available domain-specific HMMs for a small fraction of our H1 interacting proteins (125/517).

Using a structural bioinformatics strategy, we then sought to further reduce false positive hits. Specifically, we used PepSite2 (Petsalaki et al., 2009; Trabuco et al., 2012), which identifies binding pockets for peptides on protein structures, and tested whether the domains we identified above indeed carry a pocket that can accommodate the peptides carrying our putative new motifs. This step was dependent on the availability of the relevant protein domain structure, and thus was applicable to 7,941/8,403 of our identified motif instances. The result was 1,033 high quality putative linear motif instances for which we found evidence at the sequence, network and structural level and showed a further enrichment in known motif-carrying proteins **(Supplementary Table 1; Figure 2 C,D)**.

Finally, to provide functional support for our predicted motifs and further study the crosstalk between viral SLiMs and disease pathways, we mapped the predicted motifs to pathogenic variants from ClinVar (Landrum et al., 2018). We found that ~13% of our predicted motif instances were associated with a ClinVar variant, with 27% of these mapping to a pathogenic variant and ~85% having a significant fit on the structure, as predicted by PepSite2, to a pocket of a relevant domain (OR = 3.47, p-value = $5.3x10^{-18}$). We thus identified 105 'gold' functional linear motif instances that present a 2.15-fold enrichment in proteins including true SLiMs compared to the initial QSLiMFinder scan **(Supplementary Table 1; Figure 2 C,D).**

To evaluate the benefit of requiring motifs to be present in both human and viral interaction partners of a protein, this workflow was repeated as described also on the human interactome without this requirement (**Supplementary Table 2).** Starting from 3,994 initial predicted motifs from SLiMFinder (Edwards et al., 2007), our filtering process resulted in 29 'gold' human motifs. We found that the inclusion of the viral motif requirement as a filter resulted in a 2.1 fold higher number of motifs and in improvement in almost all metrics of performance (**Figure S2 A,B**; 8,403 vs 3,994). We additionally used SLiMEnrich (Idrees et al., 2018), to evaluate the level of enrichment of known Domain-motif interactions (DMIs) compared to random predictions for both analyses. Using ELM as background, 86 and 45 non-redundant DMIs were predicted compared

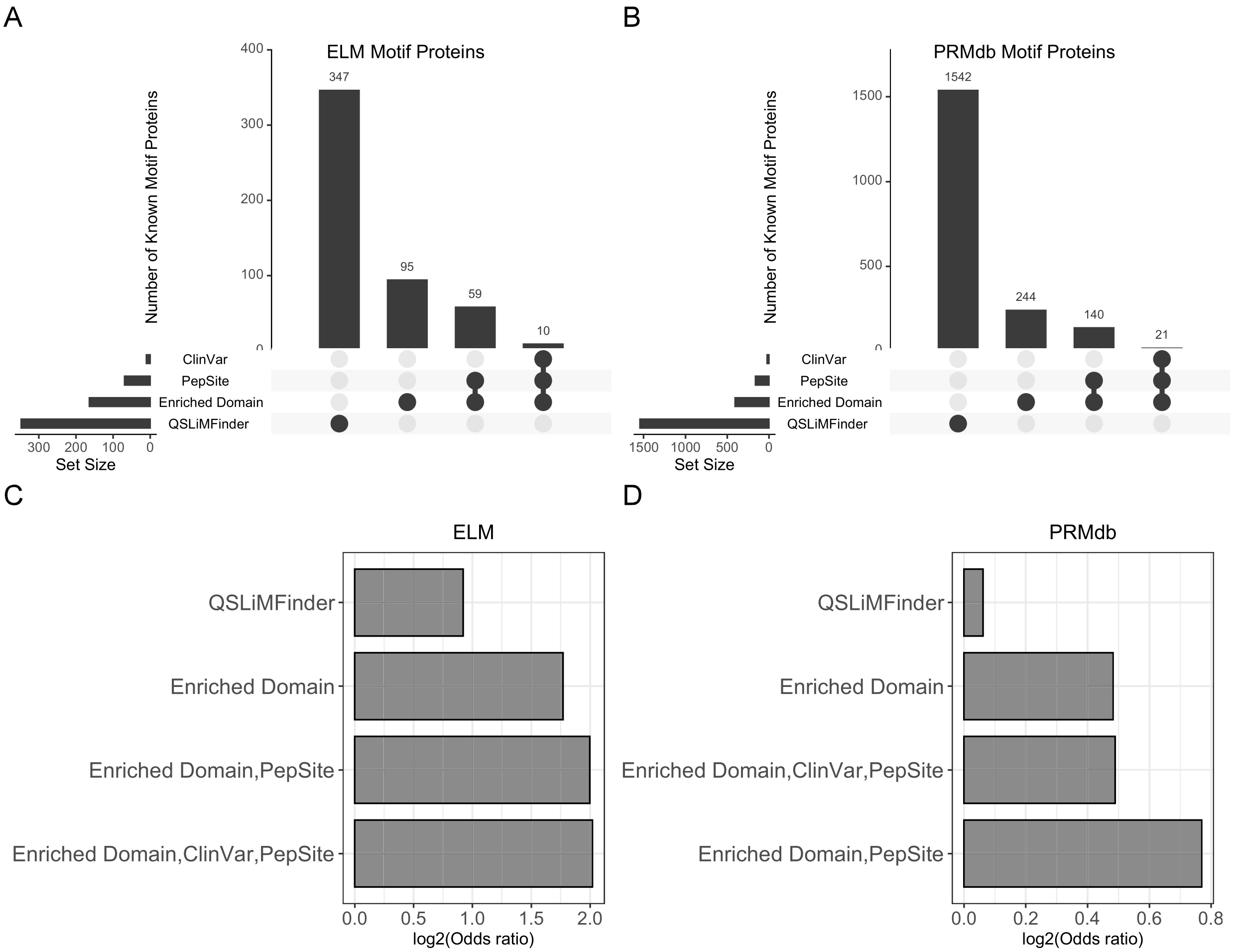


Figure 2: Evaluation of known motif-carrying proteins. (A & B) UpSet plots showing the overlapping number of known motif-carrying proteins that are found in either ELM (A) or PRMdb (B) and predicted datasets defined by a single or multiple filters. (C & D) Bar graph plotting the enrichment of known ELM (C) or PRMdb (D) motif-carrying proteins represented by log2(odds ratio) for each set of applied filters. Odds ratio was calculated based on a one-tailed fisher-exact test using query input to QSLiMFinder as background. All filters mentioned were applied on the raw output of the QSLiMFinder method (See also **Figure S1**)

to mean random DMI of 17.29 and 16.5 with an enrichment score of 5 and 2.7 (FDR 0.2 vs FDR 0.37) for host-viral and human-only hits, respectively (**Figure S2 C,D**). Therefore, the observed DMIs are significantly higher than expected for both datasets (P < 0.001) with the host-viral screen having ~2X fold higher enrichment than the human-only one.

Similarity between discovered motifs and known motif classes

To further classify the predicted motifs, we compared their regular expression patterns to those of known ELM identifiers using CompariMotif (Edwards et al., 2008), which scores each comparison based on the shared information content between two motifs. We found that the ELM classes were equally represented, with a slightly higher representation of Ligand binding sites (**Figure S3A**). When we only consider high-similarity comparisons (i.e having CompariMotif scores >= 3) and excluding CLV (cleavage) and TRG (targeting) classes (**Figure 3; Methods**), we found that 18 out of 77 motif patterns were part of previously known ELM instances, whereas the rest were novel ELM instances matching to 30 ELM identifiers. The top matched ELM identifiers are LIG_PDZ_Class_1, LIG_ULM_U2AF65_1, and DOC_MAPK_gen_1 showing high similarity with 30 out of 77 predicted motif patterns (**Figures 3, S3B**). The aforementioned ELM classes have been previously reported to be involved in multiple host-viral interactions throughout the viral life cycle (James and Roberts, 2016; Pabis et al., 2019; Evans et al., 2010). If we only consider the 'gold' instances, approximately 50% of the predicted motif patterns are associated with CFTR protein which bears multiple disease mutations, mapped through various host-viral interactions (**Figure 4A & B**).

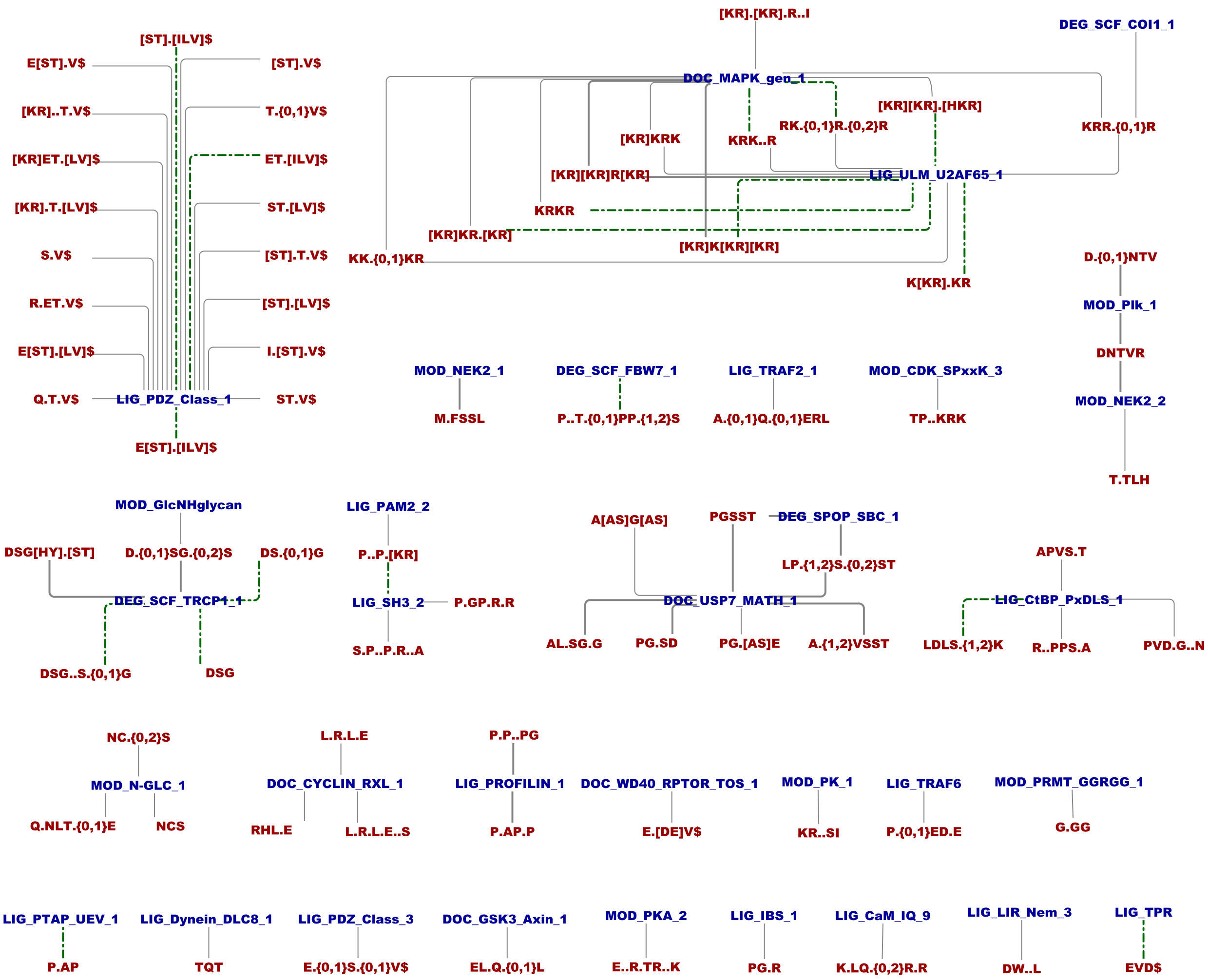


Figure 3: Comparison of predicted motif patterns to known ELM classes. Network showing the connection between predicted motif patterns with > 60% similarity to known ELM classes according to CompariMotif scores. For the sake of visualization, CLV and TRG classes were excluded as they tend to be easily discovered because of their shared location, and thus multiple patterns can be promiscuously mapped to few ELM classes. Predicted motif patterns are shown in red, while ELM classes are shown in blue. Green dash-dotted edges represent motif patterns that match to a known ELM instance, while solid edges represent new instances.

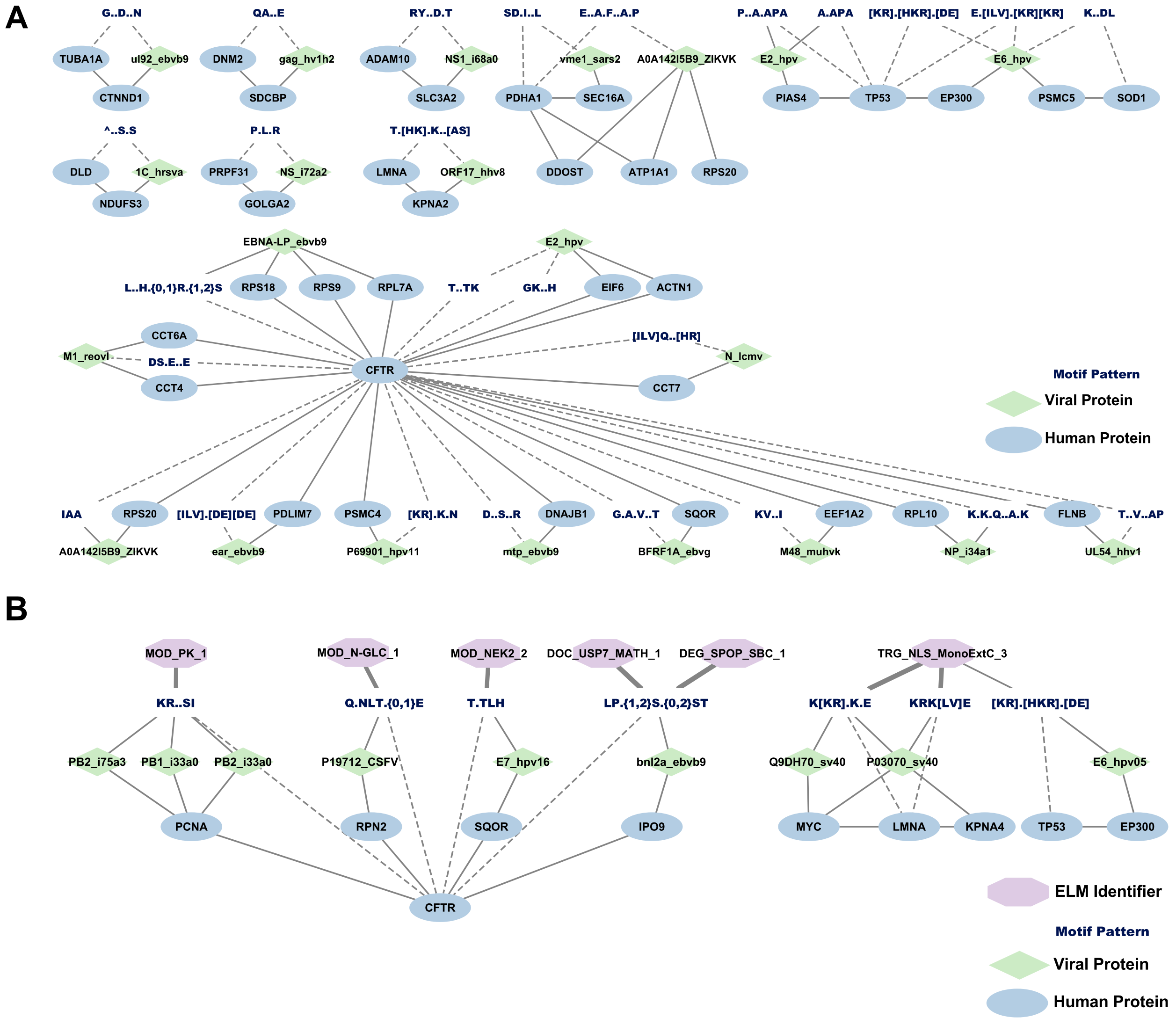


Figure 4: Putative functional motif instances. Network showing novel motif instances (A) and new ELM instances (B) associated with a ClinVar pathogenic variant in addition to confirmed structural binding with enriched domains according to PepSite. Stricter significance cutoffs were used for this figure compared to the 'gold set' (QSLiMFinder p-value < 0.05; PepSite p-value < 0.05) for simplicity of visualization. Nodes colored in green, cyan and dark blue represent viral, human and predicted motif patterns, respectively. Viral protein labels include protein name and viral strain separated by "_". Dashed edges are only drawn between motif patterns and human proteins, where a ClinVar variant is associated with that motif pattern that is predicted on the respective human protein. Solid edges between proteins represent PPI as retrieved from the IntAct database.

Functional enrichment, essentiality and pathway-specificity of linear motifs

Next, we sought to identify the functional pathways in which our predicted DMIs are involved. First, we performed Gene Ontology (biological process - GOBP) enrichment (Ashburner et al., 2000; Gene Ontology Consortium, 2021) for H1 proteins to find overrepresented pathways in each filter combination using all 517 H1 proteins as background (Methods). The top enriched GO terms can be summarized in 4 major categories: regulation of gene expression and nucleosome assembly, metabolic processes and protein-modification, interaction with host and symbiotic processes, and nuclear import/export and protein localization (**Figure 5**). Some filter combinations (e.g. Enriched Domains, ClinVar and PepSite) are almost exclusively enriched in protein transport and localization (**Figure 5**). Additionally, we performed GOBP enrichment analysis of either H1 domains or domains that interact with predicted human motifs, using the pfam2go (biological process - PFAM2GOBP) associations (see Methods) as background (Mitchell et al., 2015). Only 180 out of possible 679 H1 domains and 1,204 out of possible 4,674 enriched domains are annotated in PFAM2GOBP, thus this analysis was performed only on the subset for which annotations were available. As in H1 GOBP enrichment, we identified overrepresented pathways in each filter combination (Methods). Only 3 filter combinations had significantly enriched terms (adjusted p-value < 0.05): iELM, Enriched Domain & Pepsite, and ClinVar and PepSite. Domains in H1 proteins were mostly enriched in metabolic processes, specifically nitrogen-compound metabolism and RNA synthesis processes (**Figure 6 A,B**). Domains in interacting partners of predicted host motifs were mostly enriched in protein phosphorylation, RNA biosynthetic processes, and protein transport and localization (**Figure S4 A,B**). Moreover, we also measured the semantic similarity between the enriched GO terms for H1 domains and domains that pass either the PepSite filter, the Enriched Domain filter or both. In concordance with the aforementioned enriched terms, the most prominent PFAM domain-associated biological processes represented are transcription-related (Myc_N, Creb_binding, KIX), translation related (Ribosomal_S17, Ribosomal_S13), and phosphorylation-related (Pkinase, PK_Tyr_Ser-Thr, Pkinase_C) (**Figure 6C, S4C**).

We then explored whether the domains involved in our predicted domain-motif interactions tend to be specific to certain GOBP or KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2021) pathways using a published network-based scoring approach (Shim et al., 2019). The top-scoring (z-score > 1) processes and pathways associated with at least 3 filter-combinations are: cell

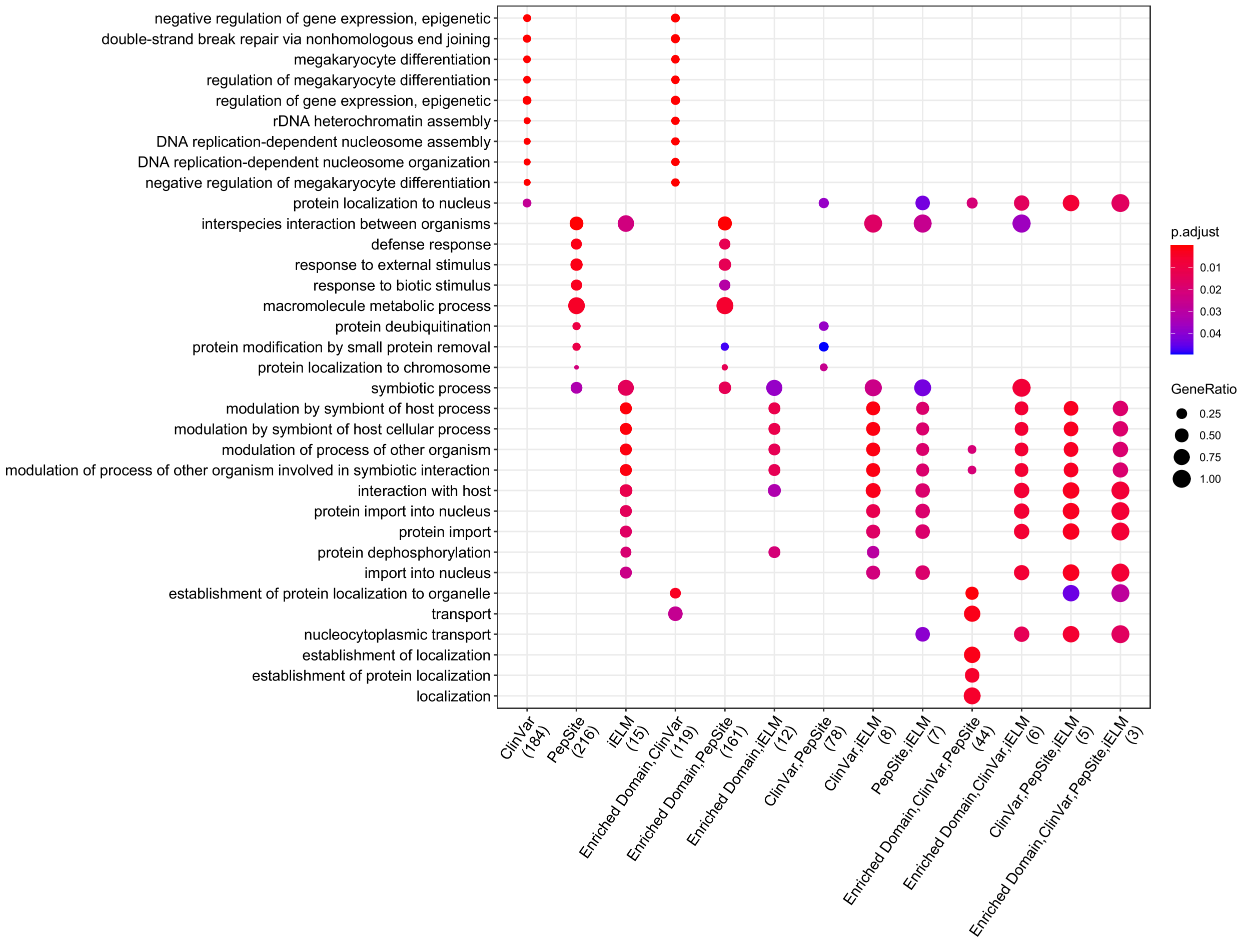


Figure 5: H1 protein GO enriched pathways. Dot plot showing the enriched GO terms (biological process) for each filter combination. This is based on an overrepresentation test using the 517 H1 proteins as background. The color scale corresponds to adjusted p-value and the size of dots corresponds to the proportion of overlapping proteins between input query and annotated proteins for each GO term. The number of enriched H1 proteins in each dataset is enclosed in parentheses for each filter combination annotated under x-axis labels.

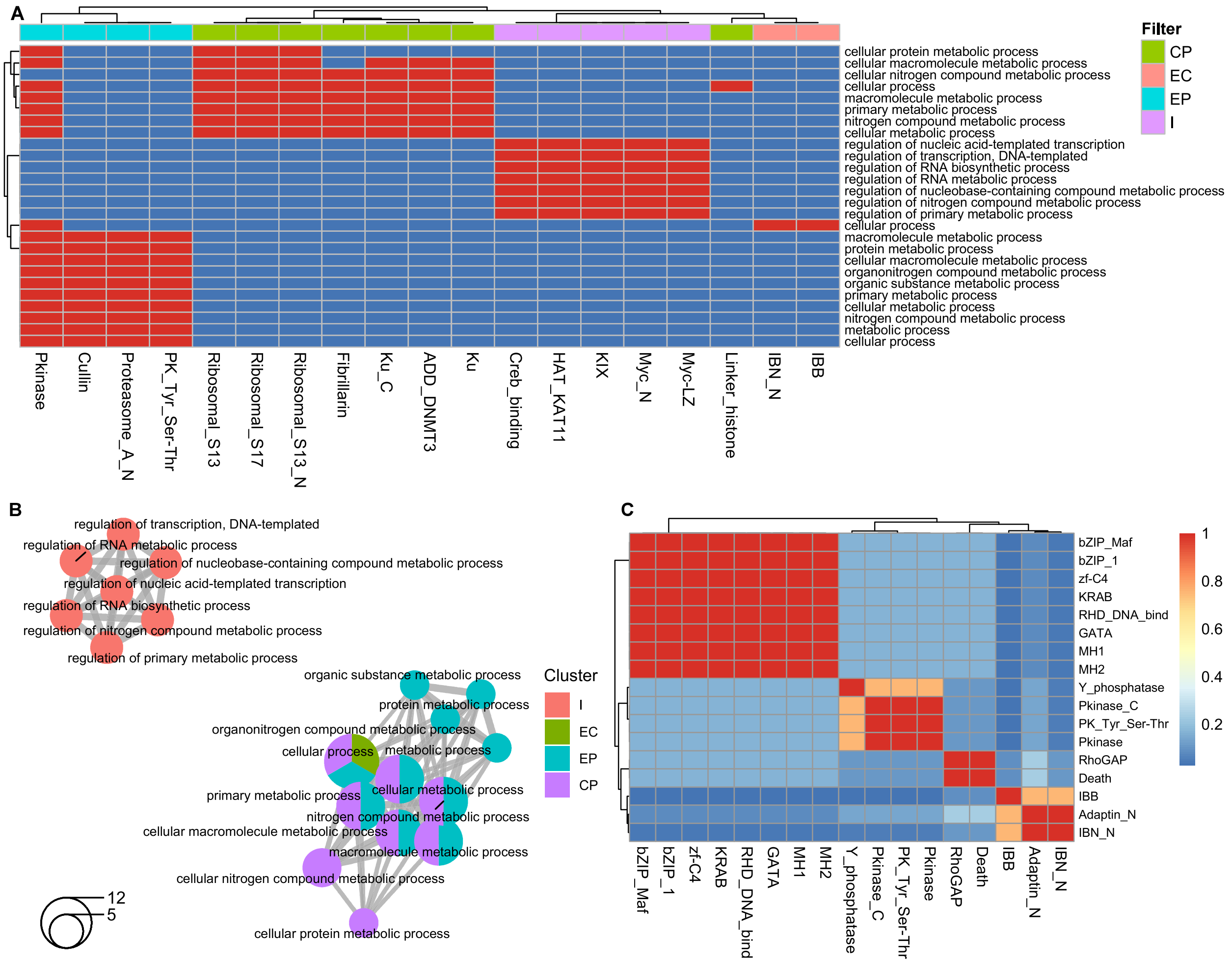


Figure 6: GO pathway enrichment of PFAM domains in H1 proteins. (A) Binary heatmap describing association between enriched GO terms (biological process, y-axis) and the corresponding overlapping domains between input dataset and PFAM2GOBP annotations (x-axis) for a given term, where red and blue boxes represent presence or absence of enriched terms for a given domain, respectively. Only Domain-pathway pairs with adjusted p-value < 0.05 are displayed. Domains are clustered by filter combinations shown as colored annotation bars above columns. E, Enriched Domains; P, PepSite; C,ClinVar; I, iELM. (B) Network showing the Jaccard similarity coefficient between enriched GO terms. Each node is a pie chart colored by filter combinations and the size of the nodes correspond to the number of domains. E, Enriched Domains; P, PepSite; C,ClinVar; I, iELM. (C) Heatmap showing GO semantic similarity between enriched domains calculated using the “Wang" method (Wang et al., 2007) between GO terms for each domain-pair. The color scale represents the similarity score where 1 denotes complete overlap, and 0 denotes no similarity..

motility, translation, and regulation of transcription (**Figure S5A**), and DNA repair and transcriptional misregulation in cancer (**Figure S5B**), respectively.

To gain more insight into the types of proteins that are involved in motif-domain interactions we also assessed the essentiality of both H1 proteins and predicted motif-carrying proteins. For our H1 proteins, all filter combinations including Enriched Domains, PepSite and ClinVar were significantly enriched in essential genes. Combinations including iELM datasets were mainly enriched in context-essential genes (**Figure S6**), although not significant due to the very low numbers of H1 and predicted motif-proteins covered.

Identification of candidates for Drug repurposing

Given that for all our human predicted motifs there was a corresponding viral motif targeting the same domain-carrying protein (H1), we sought therapeutic drugs that target the latter. Our hypothesis was that, if those drugs are indicated for the ClinVar diseases resulting from a pathogenic mutation in the relevant predicted motifs, they may be suitable for repurposing for the corresponding viral infection.

By mining the ChEMBL database (Gaulton et al., 2017; Mendez et al., 2019) we retrieved 13 molecules targeting 5 H1 proteins that also had drug indications matching the relevant ClinVar disease (**Figure 7; Supplementary Table 3**). Roscovitine (Seliciclib), which targets PSMC4 (26S proteasome regulatory subunit 6B), was one of the identified compounds. PSMC4 interacts with CFTR, where the pathogenic mutation (AlA412del) was found in the motif pattern ([KR].K.N). The same motif pattern was also observed in HPV11-E5B (Human Papillomavirus 11 – E5B) which is associated with PSMC4 through Tandem Affinity Purification (TAP) / MS (Rozenblatt-Rosen et al., 2012). Roscovitine has been previously reported to prevent the proteasomal degradation of F508del-CFTR through a CDK-independent mechanism, thus restoring the cell surface expression of CFTR in human CF (cystic fibrosis) airway epithelial cells (Norez et al., 2014). On the other hand, there have been several reports linking HPV E5 proteins with the host ubiquitin-proteasome system to modify and regulate host proteins (Wilson, 2014). Accordingly, Roscovitine can be potentially repurposed as pharmacological therapy for HPV 11 infections. Moreover, several other drugs that we identified (**Figure 7**) already have indications in ChEMBL or evidence for reducing the relevant viral infection, such as Sirolimus and Everolimus for Influenza A infection (Alsuwaidi et al., 2017; Murray et al., 2012), Vorinostat and Cytarabine for human herpesvirus

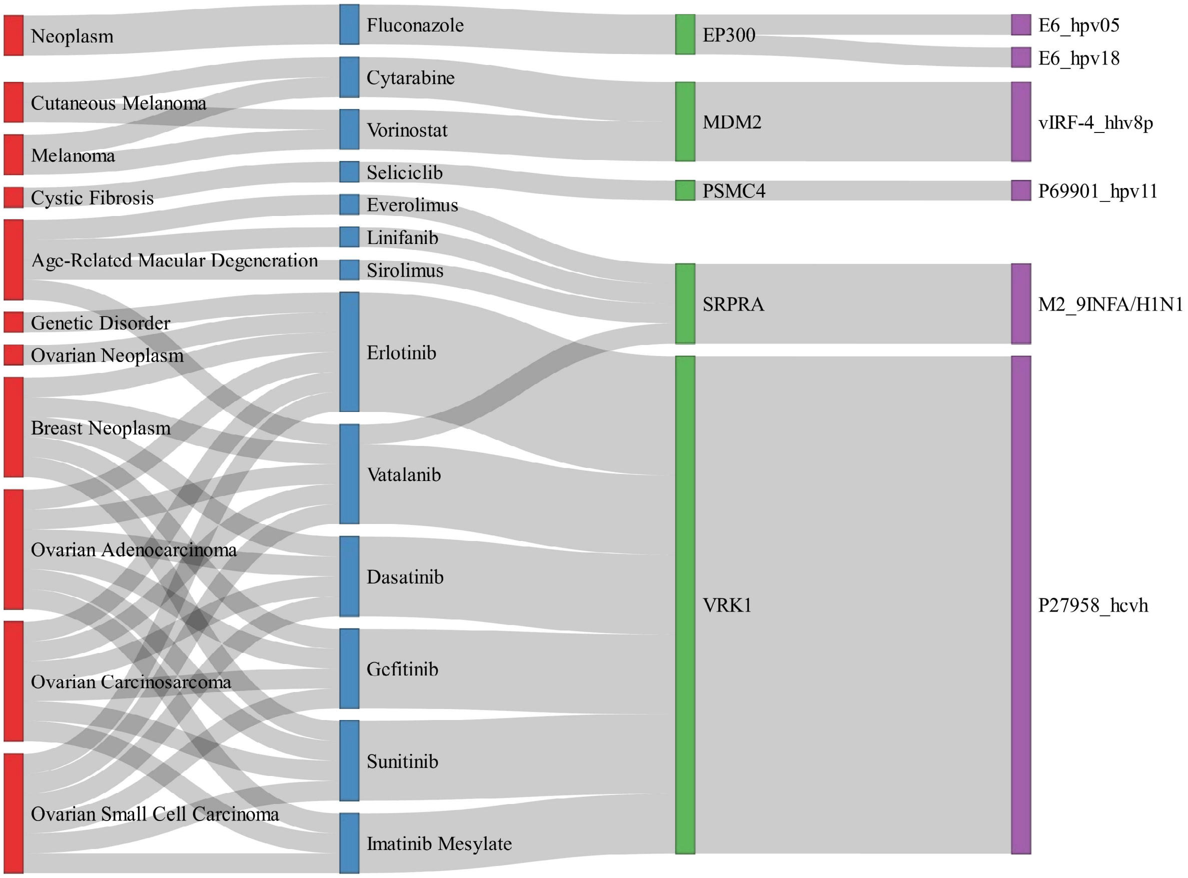


Figure 7: Potential Drug candidates for repurposing. Sankey network displaying the relationship between ClinVar diseases and their indicated ChEMBL drugs that can be potentially repurposed for the respective viral infection. Drugs target the host domain protein (H1) which directly interacts with the relevant viral protein (V1). Nodes colored in red, blue, green and violet represent ClinVar disease, ChEMBL drugs, host protein, and viral protein respectively.

type-8 (hhv8) (Hogan et al., 2018; Jc et al., 2020), and Erlotinib, Vatalanib, and Dasatinib for Hepatitis C (HCV) (Lupberger et al., 2011; Elsebai et al., 2016).

Discussion

Despite their genome size constraints, viruses manage to hijack a myriad of cellular processes and signalling pathways, often by means of SLiM-mediated interactions with the host proteome, thereby taking control of the host cellular machinery to their reproductive advantage (Davey et al., 2011). Moreover, because of their small size, degenerate nature, low complexity, and presence in intrinsically disordered regions, SLiMs are rapidly evolving *ex nihilo* (evolution from “nothing”) due to frequent mutations in these disordered regions, thus allowing a diverse repertoire of motifs to emerge in unrelated proteins through convergent evolution (Davey et al., 2015). Altogether, the nature of these short motifs explains the highly inflated false positive rates that can be computationally discovered throughout the proteome, due to their relatively poor signal to noise ratio (Gibson et al., 2015). While there has been a considerable effort to elucidate the molecular mechanisms by which viral proteins interact with the host proteome, there are relatively few resources (Hraber et al., 2020) highlighting SLiM-mediated interactions, the best of which is currently the ELM database (Kumar et al., 2020). In addition, resources and approaches to discover new linear motifs in general, rely on the largely manual identification of limited new motif instances or the re-discovery of known ELMs in different protein sequences, and in doing so, they fail to consider the wealth of information included in the interacting partners of viral proteins.

Here we describe a proteome-wide approach to identify novel functional human motifs by taking advantage of the latest available human-viral interactome (Orchard et al., 2014), and including evidence from domain enrichment, structural modelling of SLiM-domain interactions (Petsalaki et al., 2009; Trabuco et al., 2012), and mapping ClinVar (Landrum et al., 2018) pathogenic variants on our predicted motifs, to improve both the sensitivity and specificity of computational motif discovery. Overall, using the principle of viral motif convergent evolution, we improved both the sensitivity and the specificity of the motif identification (**Figure S2**) and discovered 74 and 3,673 known and new motif instances, respectively, in addition to 4,656 novel motifs.

Functional analysis of our putative motif-domain interactions showed enrichment in processes known to be mediated by motifs, and importantly also hijacked by viruses. These include protein translocation (Cook and Cristea, 2019), transcriptional regulation (Kropp et al., 2014; Tarakhovsky and Prinjha, 2018), cell signalling (Alto and Orth, 2012), and cellular metabolism (Mayer et al., 2019; Thaker et al., 2019) among others, highlighting that despite a potential high number of false positive motifs identified, our resource is enriched in functional hits that describe biologically relevant processes. These have also been identified in a similar study that performed a proteome-wide analysis of human motif-domain interactions in influenza A viral proteins (García-Pérez et al., 2018). Interestingly we found a significant enrichment of essential genes to be involved in motif-domain interactions. This highlights the importance of such interactions in cell functions and provides additional support for the functionality of our predicted motif-domain interaction set. Our observation can partially be explained by the fact that motif-binding domains and on occasion also motif-carrying proteins can interact with multiple diverse partners, i.e. be hubs, which are known to be more essential than other nodes in protein interaction networks (Puschnik et al., 2017; García-Pérez et al., 2018).

We also highlight specific diseases associated with a disease-causing variant mapped to the shared motif pattern between host-viral and human PPIs, thus providing potential insights into the crosstalk between viral infection and the underlying disease. Interestingly, among the gold set of our hits about 50% putative motif patterns were associated with a pathogenic variant in *CFTR* gene (**Figure 4**). Viral infections have been heavily implicated in the pathogenesis of cystic fibrosis by increasing the susceptibility to other bacterial infections leading to further complications in the respiratory tract in cystic fibrosis patients (Frickmann et al., 2012). It has also been reported that multiple viral proteins with homology to CFTR interact with the same targets as CFTR, to commandeer the CFTR interactome. Several pathways are common between CFTR interactome and the viral life cycle including: endosomal pathways, lysosomal trafficking and protein processing (Carter, 2010). This could explain why we got multiple hits with CFTR and these putative motif-domain interactions shed light on the potential mechanisms by which viral proteins interfere with the CFTR interactome.

In a search for drug repurposing candidates, we also identified drugs that target domain-carrying properties that we predicted to be relevant both for the viral infection and associated ClinVar disease. While only 13 drugs targeting 5 of our proteins were retrieved, several of them already had indications for the relevant viral infection in addition to the disease used to extract the drugs from the database (Alsuwaidi et al., 2017; Elsebai et al., 2016; Hogan et al., 2018; Jc et al., 2020; Lupberger et al., 2011; Murray et al., 2012). This suggests that our collection of paired viral-human motifs could be a good starting point for identifying key nodes to target in the pathogenesis of both the disease associated with the motif mutations and the viral infections.

Although we observed evidence for the presence of functional novel motifs and the recovery of known instances, computational *de novo* SLiM prediction remains a difficult challenge despite the plethora of available SLiM prediction methods, and thus the number of false positive predictions is expected to be high (Prytuliak et al., 2017). Our approach is vulnerable to additional uncertainties, which can be attributed to multiple confounders throughout our workflow in addition to the inherent false positive rate of the motif prediction method. Starting off with the PPI data from IntAct (Orchard et al., 2014), some of the binary interactions reported in IntAct database have weaker interaction evidence and thus an overall low Molecular Interaction score (MIscore) compared to “true” interactions which has been confirmed by several methods and has more literature evidence (Villaveces et al., 2015). Therefore, some of the input host-viral interactions are merely association events as opposed to having strong structural binding evidence and could be introducing noise to the analysis. Furthermore, as with all proteome-wide motif discovery methods, the precision and recall were relatively low when we evaluated the predicted instances against that of ELM’s. Nevertheless, when we used SLiMFinder (Edwards et al., 2007) to apply the same pipeline on all human proteins without restricting the motif space to viral motifs, we obtained even lower precision and recall in most of the evaluations (**Figure S2)**. Therefore, the predicted hits conditioned on the presence of viral motifs recover more known instances with higher accuracy compared to their absence despite their apparent high false positive rate.

Our study provides a considerable number of potential motif instances with several levels of support for functionality that can be prioritized for further investigation and validation studies (e.g. using ITC or Fluorescent Polarization). We show that taking advantage of the principle of convergent evolution allows us to reduce the false positive rate inherent to computational screens for linear motifs, which we then further reduce using additional orthogonal filters. This not only provides an improved catalogue of putative human motifs but also provides indications for points of motif-mediated viral interference that can be easier to target than interactions mediated by larger interfaces. This resource will be valuable towards improved understanding of the biological functions of motif-mediated interactions in humans and viruses. As more human-viral protein interactions, domain structures and other information become available, our workflow will be able to provide evidence for an increasing number of human motifs.

Acknowledgments

The authors would like to thank EMBL-EBI for funding this project and Ylva Ivarsson for input on the manuscript and validation.

Author Contributions

Conceptualization: BW, VK, EP, Methodology: BW, VK, EP, Software: BW, VK with contribution from ES, Validation: BW with contribution from CB, Formal Analysis: BW, VK, Investigation: BW, VK, EP, Resources: EP, Data Curation: BW, VK, Writing - Original Draft: BW, EP, with some contribution from VK, Writing - Review & Editing: BW, EP, Visualization: BW, Supervision: EP, Project Administration: EP, Funding Acquisition: EP

Declaration of interests

The authors declare no competing interests.

Data and code availability

To ensure reproducibility, all data, code, intermediate and final analysis files, have been made available as follows:

- All, raw and analyzed data that are not already included in Supplementary Tables in this manuscript are publicly available at the Biostudies database with project ID: S-BSST668 (https://www.ebi.ac.uk/biostudies/studies/S-BSST668/)
- All original code is publicly available at https://gitlab.ebi.ac.uk/petsalakilab/HVSlimPred.

Methods

Interactome Datasets

We extracted all human-viral interactions from the IntAct database (Orchard et al., 2014) on 18/05/2020 by using the query:“(taxidA:9606 AND taxidB:10239) OR (taxidA:10239 AND taxidB:9606) AND ptypeA:protein AND ptypeB:protein” on 20/05/2020. This resulted in 22,839 binary interactions. We then removed the duplicate entries resulting in a Human-Viral dataset of 15,559 interactions including unique 5,575 proteins from 325 viruses (types and isolates) and humans (**Supplementary Table 4**). Distributions of interactions per virus type are shown in **Supplementary table 5.**

In addition, we extracted all human interactions that had more than 2 types of evidence and at least one of the two interacting proteins was present in the Human-Viral dataset we had already collected. The resulting dataset included 49,000 interactions amongst 10,707 proteins (**Supplementary Table 6**).

Validation datasets

Our main benchmarking dataset comprises 1,936 known human SLiMs in 1,153 human proteins, extracted from the Eukaryotic Linear Motif (ELM) database on 24.07.2020 (Kumar et al., 2020) (**Supplementary Table 7**). In addition, we considered PRMdb (Peptide Recognition Modules database) which is based on large-scale peptide phage-display methods (Teyra et al., 2020) and comprises 386 PRM modules covering 50 structural families in 12,771 human proteins spanning 47,657 SLiMs (**Supplementary Table 8**).

Identification of linear motifs in viral-human protein interaction networks

QSLiMFinder (Palopoli et al., 2015) was used for the identification of motifs in our networks with the default settings in terms of filtering except for setting the maximum number of sequences to evaluate to 800 (dismask=T, consmask=T, cloudfix=T, maxseq=800, gnspacc=F). Specifically, for each viral protein (V1) interacting with a specific human protein (H1) we searched for motifs that are enriched in the known human protein interaction partners of H1 and the viral protein. We also imposed the restriction that the motif must also be present in the viral protein.

Enrichment of human protein domains as potential motif interactors

For each viral-human V-H1 protein interaction we extracted all human interactors of H1 from the IntAct database (Orchard et al., 2014). To avoid bias towards specific domain types due to sequence similarity of these interactors, we then used cd-hit (Fu et al., 2012; Li and Godzik, 2006) to reduce redundancy at 70% sequence identity level. We then used a two-tailed fisher-test (Fisher, 1934) to calculate the enrichment of observing a specific domain for each interaction set compared to the background. These p-values were FDR corrected using the Benjamini-Hochberg approach (Benjamini and Hochberg, 1995). We only considered as enriched the domains that had a significant adjusted p-value in terms of enrichment compared to the background (adjusted p-value < 0.05). As an additional stringency filter, we required that the identified domain was present in at least 5 interacting partners of the motif carrying protein. This choice was made after evaluating multiple different cutoffs for the adjusted p-value and domain numbers and selecting the best performing combination (**Figure S7 – also see next section**).

Selection of domain enrichment filter

To select possible filters for functional motif enrichment, we first tested the effect of considering enriched domains on the enrichment of known hits. We tried different combinations of parameters concerning the minimum number of interaction partners for a specific domain (1-10), odds ratio (>1 or >4), and adjusted p-value (0.01, 0.05, 0.1) for a total of 60 datasets. Based on the scores for all 3 levels, a minimum of at least 5 interaction partners returned the best results, where the difference was more pronounced in PRMdb (Teyra et al., 2020) compared to ELM (Kumar et al., 2020) (**Figure S7**). Therefore, the predicted hits were filtered so that the identified domain was present in at least 5 interaction partners of the motif carrying protein with an adjusted p-value of < 0.05. The previous dataset was used in further downstream analyses as another functional filter to reduce false positives and will be referenced as “Enriched Domain”. Additionally, we applied the same evaluation pipeline using combinations of functional filters described below including: ClinVar, PepSite, Enriched Domain, and iELM HMMs to check which functional filters are better at capturing known instances.

Comparison of discovered motifs with known ELM classes

To further classify the predicted motifs, we compared their regular expression patterns to those of known ELM identifiers using CompariMotif (Edwards et al., 2008), which scores each comparison based on the shared information content between two motifs. We grouped the matched relationships identified by CompariMotif into 5 confidence levels ranked by their proposed heuristic score. This considers the number of matched positions and a normalized score corresponding to the information content between the respective motifs. For each of the six major ELM classes, (ligand (LIG), cleavage (CLV), docking (DOC), degradation (DEG), post-translational modification (MOD), and targeting (TRG)), we performed an all Vs all comparison against the relevant ELM identifiers (**Supplementary Table 1**). The predicted instances were also classified into known and new instances based on whether the instance overlaps with a true positive instance in the ELM database or not, respectively (**Figure S3A**). In addition, we also checked the concordance between CompariMotif similarity score and the known instances that we predicted, confirming that indeed high CompariMotif scores were able to identify true ELM class labels (AUC 0.791-0.829 depending on the set evaluated).

Evaluation of predicted motifs

Fair evaluation of linear motif identification methods is complicated partly due to: a) the difficulty to define terms like ‘true positive’, because motif instances often overlap and are found in duplicates and b) the very unbalanced nature of the validation dataset with very few known motifs and undefined, but likely large, numbers of true but not yet discovered motif instances. Thus to provide performance metrics from different points of view, we performed the evaluation at 3 levels (motif-carrying protein, motif instances, protein-domain interaction) using both ELM (Kumar et al., 2020) and PRMdb (Teyra et al., 2020) as separate validation datasets.

For the protein-level enrichment, we measured the enrichment of true-positives in our predicted dataset using a one-tailed Fisher’s exact test (Fisher, 1934), where the odds ratio represents the magnitude of the enrichment. True positives are the number of motif-carrying proteins present in both the predicted dataset and the validation dataset regardless of whether the predicted protein has the right motif or is found in the right location. For this level, we used the odds ratio to evaluate enrichment of true positives in each filtered dataset, as a measure of improved quality of the predictions after each filter.

For the motif and protein-domain levels the evaluation was restricted to the proteins present in the validation dataset as they are not applicable, for proteins that don’t carry a known motif. For the motif-level enrichment, due to the possibility for partially correct hits, we used a re-implemented version of the evaluation protocol proposed in (Prytuliak et al. 2017) to compute common performance metrics (recall, precision F1, etc) both residue-wise and site-wise. Given the skewed imbalance towards higher “false positives” compared to “false negatives”, we use the

F0.5 score instead of the F1 score, which puts more weight on the precision than the recall and is calculated as follows:

$$F0.5 = \frac{1.25 * Precision * Recall}{0.25 * Precision * Recall}$$

Finally, for evaluating protein-domain interactions we measured the enrichment of true-positive interactions between a given motif-carrying protein and its associated domains as reported in the validation dataset (Kumar et al. 2020). Here, true-positives represent the number of correctly associated domains for a given motif-carrying protein and then summed over all motif-carrying proteins in the predicted dataset.

Prediction of viral SLiM binding on identified human domains

We used PepSite2 (Petsalaki et al., 2009; Trabuco et al., 2012) to predict the binding interface of each SLiM on the respective domains. Note that the PepSite2 tool tends to have a high number of false negatives but a high accuracy. PDB structures for the respective domains were selected based on the protein interaction partners for a given motif-carrying protein harboring the specified domain according to the pdb_pfam data downloaded from (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/database_files/pdb_pfamA_reg.txt.gz). The lowest resolution structure was selected if there was more than one potential structure for each domain protein. Only hits with p-value < 0.1 were considered to indicate a significant motif-domain interaction. We also applied multiple testing correction for each peptide length (3-10) using the Benjamini-Hochberg approach (Benjamini and Hochberg, 1995) and the same hits were retained after selecting those with adjusted p-value < 0.1. From the total of 501,109 peptide-domain pairs we identified, we were able to find an available structure for 439,192, of which 284,707 were associated with a domain protein that directly interacts with the motif-carrying protein. Those pairs were sent to PepSite2, and 76,399 peptide-domain pairs were returned as significant hits (p-value < 0.1). Since PepSite2 takes as input PDB_ID, chain and peptide sequence, some of the hits' binding sites might not be within the specified domain region, so we filtered only those hits for a total of 46,325 peptide-domain pairs with the correct motif-domain binding interface.

Identification of SLiM binding domains in H1 proteins

We used the iELM method (Weatheritt et al. 2012) to identify SLiM binding domains using both iELM and Pfam HMMs which are trained to recognize SLiM binding interfaces (Weatheritt et al., 2012). The HMMs were downloaded from http://elmint.embl.de/program_file/ and we used *hmmsearch* from the HMMER3 toolkit (http://hmmer.org/) to search each HMM profile (both iELM and pfam HMMs) against all H1 proteins that are reported to bind to a predicated motif instance and a viral protein. In concordance with the iELM method, we also used an E-value cut-off of 0.01 and excluded all hits with a length of <80% of the annotated SLiM-binding domain's length. The returned domtblout output was parsed using the R package “rhmmer” (Zebulun, 2017) and ELM classes were updated to the current version of ELM database (Kumar et al., 2020) using renamed ELM classes table downloaded from ELM database (http://elm.eu.org/infos/browse_renamed.tsv).

Clinvar mapping

Data for mapping ClinVar (Landrum et al., 2018) variants to motif residues were downloaded from (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/archive/submission_summary_2021-03.txt.gz) on 03/04/2021. For each motif in a given protein, we check the amino acid change of each reported variant in the corresponding protein in ClinVar and we establish a mapping if there’s at least one overlapping residue. Some of the variants in ClinVar had only genomic information with no reported amino acid changes. In that case we used the EMBL-EBI Proteins API (https://www.ebi.ac.uk/proteins/api/doc/) to retrieve the genomic coordinates of each motif that can be directly compared to those ClinVar variants with no information about amino acid changes. The mapped variants were filtered to include only single nucleotide variants (SNVs) and deletions as opposed to insertions, duplications and CNVs, since we only consider variants potentially having deleterious effects on motif function and binding affinity.

Identification of potential Drug candidates

ChEMBL27 (Gaulton et al., 2017; Mendez et al., 2019) was downloaded from (ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/) and data tables corresponding to compound, target and assay information were concatenated according to the database schema. The following filters were applied to retrieve high-confidence Drug-target records: 1) Only human single-protein targets were considered as opposed to protein-complex, tissue, cell_type, etc.... 2)

Only compounds that are indicated to have a therapeutic application (Therapeutic_flag == 1) were considered as opposed to imaging agents, additives, etc … 3) Assays that are classified as Binding assays (Assay_type == “B”) and with high confidence score (confidence_score == 9) were considered to retrieve the assays where a precise molecular target is assigned. Using this curated dataset, we mapped the drugs that are reported to target H1 proteins (V1 → H1, where V is the viral protein). Additionally, we used the R packages: reactome.db, KEGG.db, and msigdbr to retrieve the pathways where H1 and H2 are involved, from Reactome, KEGG and MsigDB, respectively (Carlson, 2019; Dolgalev, 2020; Ligtenberg, 2019). Finally, we used the Experimental factor ontology (EFO) (Malone et al., 2010) to map the drugs and ClinVar diseases to disease ontology terms that can be systematically compared across the whole dataset. EFO disease ontologies corresponding to the identified drugs were directly retrieved from the Drug_indication table in ChEMBL database, while the ClinVar disease ontologies were cross-referenced with EFO using EMBL-EBI ontology xref service (OXO) (https://www.ebi.ac.uk/spot/oxo/index) which finds mapping between different ontology terms.

Functional protein and domain enrichment analysis

To identify the functional pathways associated with our predicted motifs and proteins that are directly targeted by viral proteins, we performed a hypergeometric overrepresentation test on H1 proteins and PFAM (Mistry et al., 2021) domains found in both H1 proteins and those that are found to bind with our putative motif instances. For H1 proteins, we performed the enrichment against GO (biological process) (Ashburner et al., 2000; Gene Ontology Consortium, 2021) using only the domain proteins (H1) as background. We used the “enrichGO” function in the R package ClusterProfiler (Yu et al., 2012) to perform GO enrichment using default parameters except the universe parameter where we selected all 517 H1 proteins as background. We applied the enrichment on H1 proteins for all the combinations of filtered datasets and considered enriched terms with an adjusted p-value cutoff of 0.05.

Regarding domain enrichment, we used the R package “dcGOR” (Fang, 2014) to perform PFAM domain enrichment against GOBP (biological process) and using pfam2go (Mitchell et al., 2015) as background annotation which was downloaded from (http://current.geneontology.org/ontology/external2go/pfam2go) on 20.5.2021. For each set of filter combinations (Enriched Domain, PepSite, ClinVar, iELM) we used as input either the domains found in H1 proteins of the corresponding motif instances or the domains that are found in the interaction partners of the predicted motif-carrying proteins. For filter sets including

"PepSite" or "Enriched Domain" we additionally required evidence of binding of the query domain to predicted motif instance or enrichment in at least 5 interacting partners of motif-carrying proteins with adjusted p-value < 0.05, respectively. Moreover, the background for each filter set was adjusted based on the initial query of the respective filter combination. For example, the background for sets including the PepSite filter would be the domains found in the original requests sent to PepSite for the respective motif instances involved. The function "dcEnrichment" was used to perform the enrichment of PFAM domains as input against their respective background. The default parameters were used except "ontology.algorithm" where we used "lea" algorithm to account for the hierarchy of the GO ontology. Specifically, once domains are already annotated to any children terms with more significance than itself, then all these domains are eliminated from the use for the recalculation of the significance at that term and the final p-values takes the maximum of the original p-value and the recalculated p-value. Finally, in order to compare the similarity of the enriched PFAM domains, we used the R package "GOSemSim" (Yu et al., 2010) to calculate the semantic similarity between the GOBP terms between each pair of PFAM domains using the "Wang" method (Wang et al., 2007). Visualization of enrichment output of "dcGOR" was implemented using the R package "ClusterProfiler" and heatmap visualization of semantic similarity was implemented using the "pheatmap" R package (Kolde, 2012).

Enrichment of Domain-Motif interactions

Given the high false positive rate in our predicted hits and the inherent challenges in de-novo SLiM discovery, we wanted to assess whether the predicted domain-motif interactions (DMIs) are merely due to chance or can be used to infer functional SLiMs. For this purpose, we used SLiM-Enrich (Idrees et al., 2018), which calculates the enrichment of DMIs in a given PPI (protein-protein interaction) network using a permutation approach to create a background distribution of expected DMIs. We used the SLiM-Enrich Shiny webserver (http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/) which takes as input PPI data in terms of domain and motif proteins. We used the ELMc-Domain strategy and either ELM as the default background for motifs or our own predicted hits from QSLiMFinder (Palopoli et al., 2015) as proxy for the motif compositions of the input motif proteins. We performed this analysis using both host-viral and human-only predicted hits in order to compare the enrichment of DMIs in both datasets.

Context Essentiality of H1 and Motif proteins

We used CEN-tools (Sharma et al., 2020) in order to assess the essentiality of the host-proteins that are directly targeted by viral proteins and the motif-carrying proteins in our predicted motif instances. We used the classifications already provided by CEN-tools which are integrated from both Project Score (Behan et al., 2019) and DepMap (GTEx Consortium et al., 2017; Tsherniak et al., 2017) large-scale CRISPR screens in addition to a gold standard gene set from BAGEL (Hart and Moffat, 2016). We grouped both context-essential and rare-context essential so that each protein can be classified as essential, context-essential, or non-essential. Finally, we used a one-tailed Fisher's exact test (Fisher, 1934) to perform overrepresentation of either H1 or motif-carrying proteins for each essentiality cluster in all filter combinations. Out of 517 and 2,457 H1 proteins and predicted motif proteins, 487 and 2,261 had essentiality information available, respectively and were considered for downstream analysis.

Identification of pathway-specific domains

We used a network-based scoring approach to quantify domain-pathway associations and their specificity (Shim et al., 2019) in human domains that are found in host-viral PPI. The approach is mainly based on first constructing a domain-based network using weighted mutual information between domain profiles of human proteins, then using a Bayesian framework to assign log-likelihood scores to links of co-pathway networks. Finally, the Gini index (GI) was used to account for the distribution of each domain across pathways which eventually yields a pathway specificity (PS) score for each domain-pathway association (Shim et al., 2019). We already used the provided InterPro (Blum et al., 2021) domain profile for a total of 17,013 human proteins and 8,362 InterPro domains on both GOBP (which was already provided) and KEGG pathways (Kanehisa et al., 2021; Kanehisa and Goto, 2000) which were retrieved using the KEGGREST R package (Tenenbaum, 2020). We mapped the corresponding PFAM (Mistry et al., 2021) domains according to InterPro-PFAM cross-reference which was downloaded from (https://www.ebi.ac.uk/interpro/entry/pfam/#table), resulting in a total of 4,283 PFAM domains mapped to 4,140 out of 8,362 InterPro domains. Then, for all filter combinations, we mapped all domain-pathway associations for both KEGG and GOBP against domains in H1 proteins and domain-proteins binding to predicted motif instances. Finally, we combined all the associations and scaled their pathway-specificity scores and selected only associations with z-scores > 1 which corresponds to pathway specificity scores > 0.2.

References

Ahmed, Z.M., Riazuddin, S., Riazuddin, S., Wilcox, E.R., 2003. The molecular genetics of Usher syndrome. Clinical Genetics 63, 431–444. https://doi.org/10.1034/j.1399-0004.2003.00109.x

Alsuwaidi, A.R., George, J.A., Almarzooqi, S., Hartwig, S.M., Varga, S.M., Souid, A.-K., 2017. Sirolimus alters lung pathology and viral load following influenza A virus infection. Respir Res 18. https://doi.org/10.1186/s12931-017-0618-6

Alto, N.M., Orth, K., 2012. Subversion of Cell Signaling by Pathogens. Cold Spring Harb Perspect Biol 4, a006114. https://doi.org/10.1101/cshperspect.a006114

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25–29. https://doi.org/10.1038/75556

Becerra, A., Bucheli, V.A., Moreno, P.A., 2017. Prediction of virus-host protein-protein interactions mediated by short linear motifs. BMC Bioinformatics 18, 163. https://doi.org/10.1186/s12859-017-1570-7

Behan, F.M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C.M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., Ansari, R., Harper, S., Jackson, D.A., McRae, R., Pooley, R., Wilkinson, P., van der Meer, D., Dow, D., Buser-Doepner, C., Bertotti, A., Trusolino, L., Stronach, E.A., Saez-Rodriguez, J., Yusa, K., Garnett, M.J., 2019. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature 568, 511–516. https://doi.org/10.1038/s41586-019-1103-9

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R.D., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Research 49, D344–D354. https://doi.org/10.1093/nar/gkaa977

Carlson, M., 2019. KEGG.db: A set of annotation maps for KEGG.

Carter, C., 2010. Bioinformatics analysis of homologies between pathogen antigens, autoantigens and the CFTR cystic fibrosis protein: A role for immunoadsorption therapy? Nature Precedings 1–1. https://doi.org/10.1038/npre.2010.5352.1

Cook, K.C., Cristea, I.M., 2019. Location is everything: protein translocations as a viral infection strategy. Curr Opin Chem Biol 48, 34–43. https://doi.org/10.1016/j.cbpa.2018.09.021

Davey, N.E., Cyert, M.S., Moses, A.M., 2015. Short linear motifs – ex nihilo evolution of protein regulation. Cell Communication and Signaling 13, 43. https://doi.org/10.1186/s12964-015-0120-z

Davey, N.E., Travé, G., Gibson, T.J., 2011. How viruses hijack cell regulation. Trends in Biochemical Sciences 36, 159–169. https://doi.org/10.1016/j.tibs.2010.10.002

Dolgalev, I., 2020. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format.

E. Davey, N., Roey, K.V., J. Weatheritt, R., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., J. Gibson, T., 2012. Attributes of short linear motifs. Molecular BioSystems 8, 268–281. https://doi.org/10.1039/C1MB05231D

Edwards, R.J., Davey, N.E., Shields, D.C., 2008. CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics 24, 1307–1309. https://doi.org/10.1093/bioinformatics/btn105

Edwards, R.J., Davey, N.E., Shields, D.C., 2007. SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins. PLOS ONE 2, e967. https://doi.org/10.1371/journal.pone.0000967

Edwards, R.J., Palopoli, N., 2015. Computational prediction of short linear motifs from protein sequences. Methods Mol Biol 1268, 89–141. https://doi.org/10.1007/978-1-4939-2285-7_6

Elsebai, M.F., Koutsoudakis, G., Saludes, V., Pérez-Vilaró, G., Turpeinen, A., Mattila, S., Pirttilä, A.M., Fontaine-Vive, F., Mehiri, M., Meyerhans, A., Diez, J., 2016. Pan-genotypic Hepatitis C

Virus Inhibition by Natural Products Derived from the Wild Egyptian Artichoke. J Virol 90, 1918–1930. https://doi.org/10.1128/JVI.02030-15

Evans, P., Sacan, A., Ungar, L., Tozeren, A., 2010. Sequence Alignment Reveals Possible MAPK Docking Motifs on HIV Proteins. PLOS ONE 5, e8942. https://doi.org/10.1371/journal.pone.0008942

Fang, H., 2014. dcGOR: An R Package for Analysing Ontologies and Protein Domain Annotations. PLOS Computational Biology 10, e1003929. https://doi.org/10.1371/journal.pcbi.1003929

Felsani, A., Mileo, A.M., Paggi, M.G., 2006. Retinoblastoma family proteins as key targets of the small DNA virus oncoproteins. Oncogene 25, 5277–5285. https://doi.org/10.1038/sj.onc.1209621

Fisher, R.A., 1934. Statistical methods for research workers. Statistical methods for research workers.

Frickmann, H., Jungblut, S., Hirche, T.O., Groß, U., Kuhns, M., Zautner, A.E., 2012. Spectrum of viral infections in patients with cystic fibrosis. Eur J Microbiol Immunol (Bp) 2, 161–175. https://doi.org/10.1556/EuJMI.2.2012.3.1

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Furuhashi, M., Kitamura, K., Adachi, M., Miyoshi, T., Wakida, N., Ura, N., Shikano, Y., Shinshi, Y., Sakamoto, K., Hayashi, M., Satoh, N., Nishitani, T., Tomita, K., Shimamoto, K., 2005. Liddle's Syndrome Caused by a Novel Mutation in the Proline-Rich PY Motif of the Epithelial Sodium Channel β-Subunit. The Journal of Clinical Endocrinology & Metabolism 90, 340–344. https://doi.org/10.1210/jc.2004-1027

García-Pérez, C.A., Guo, X., Navarro, J.G., Aguilar, D.A.G., Lara-Ramírez, E.E., 2018. Proteome-wide analysis of human motif-domain interactions mapped on influenza A virus. BMC Bioinformatics 19, 238. https://doi.org/10.1186/s12859-018-2237-8

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R., 2017. The ChEMBL database in 2017. Nucleic Acids Research 45, D945–D954. https://doi.org/10.1093/nar/gkw1074

Gene Ontology Consortium, 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49, D325–D334. https://doi.org/10.1093/nar/gkaa1113

Gibson, T.J., Dinkel, H., Van Roey, K., Diella, F., 2015. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. Cell Commun Signal 13. https://doi.org/10.1186/s12964-015-0121-y

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B., 2017. Genetic effects on gene expression across human tissues. Nature 550, 204–213. https://doi.org/10.1038/nature24277

Hagai, T., Azia, A., Babu, M.M., Andino, R., 2014. Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions. Cell Reports 7, 1729–1739. https://doi.org/10.1016/j.celrep.2014.04.052

Hart, T., Moffat, J., 2016. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics 17, 164. https://doi.org/10.1186/s12859-016-1015-8

Hogan, L.E., Hanhauser, E., Hobbs, K.S., Palmer, C.D., Robles, Y., Jost, S., LaCasce, A.S., Abramson, J., Hamdan, A., Marty, F.M., Kuritzkes, D.R., Henrich, T.J., 2018. Human Herpes Virus 8 in HIV-1 infected individuals receiving cancer chemotherapy and stem cell transplantation. PLoS One 13. https://doi.org/10.1371/journal.pone.0197298

Horowitz, J.M., Yandell, D.W., Park, S.H., Canning, S., Whyte, P., Buchkovich, K., Harlow, E., Weinberg, R.A., Dryja, T.P., 1989. Point mutational inactivation of the retinoblastoma antioncogene. Science 243, 937–940. https://doi.org/10.1126/science.2521957

Hraber, P., O'Maille, P.E., Silberfarb, A., Davis-Anderson, K., Generous, N., McMahon, B.H., Fair, J.M., 2020. Resources to Discover and Use Short Linear Motifs in Viral Proteins. Trends in Biotechnology 38, 113–127. https://doi.org/10.1016/j.tibtech.2019.07.004

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A.E., Schweppe, D.K., Rad, R., Erickson, B.K., Obar, R.A., Guruharsha, K.G., Li, K., Artavanis-Tsakonas, S., Gygi, S.P., Harper, J.W., 2017. Architecture of the human interactome defines protein communities and disease networks. Nature 545, 505–509. https://doi.org/10.1038/nature22366

Idrees, S., Pérez-Bercoff, Å., Edwards, R.J., 2018. SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. PeerJ 6, e5858. https://doi.org/10.7717/peerj.5858

James, C.D., Roberts, S., 2016. Viral Interactions with PDZ Domain-Containing Proteins—An Oncogenic Trait? Pathogens 5. https://doi.org/10.3390/pathogens5010008

Jc, R., Ja, S., A, C., Eg, R., Rf, A., Er, S., Pc, M., Pg, R., Cm, D., E, C., D, A., R, B., L, R., L, K., Aa, C., Jy, L., R, M., A, N., 2020. Impact of Myc in HIV-associated non-Hodgkin lymphomas treated with EPOCH and outcomes with vorinostat (AMC-075 trial). Blood 136. https://doi.org/10.1182/blood.2019003959

J. Edwards, R., E. Davey, N., O' Brien, K., C. Shields, D., 2012. Interactome-wide prediction of short, disordered protein interaction motifs in humans. Molecular BioSystems 8, 282–295. https://doi.org/10.1039/C1MB05212H

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M., 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 49, D545–D551. https://doi.org/10.1093/nar/gkaa970

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Kolde, R., 2012. Pheatmap: pretty heatmaps.

Kropp, K.A., Angulo, A., Ghazal, P., 2014. Viral Enhancer Mimicry of Host Innate-Immune

Promoters. PLOS Pathogens 10, e1003804. https://doi.org/10.1371/journal.ppat.1003804

Krystkowiak, I., Davey, N.E., 2017. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Research 45, W464–W469. https://doi.org/10.1093/nar/gkx238

Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J.A., Bukirova, D., Čalyševa, J., Palopoli, N., Davey, N.E., Chemes, L.B., Gibson, T.J., 2020. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res 48, D296–D306. https://doi.org/10.1093/nar/gkz1030

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J.B., Kattman, B.L., Maglott, D.R., 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46, D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Latysheva, N.S., Oates, M.E., Maddox, L., Flock, T., Gough, J., Buljan, M., Weatheritt, R.J., Babu, M.M., 2016. Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer. Molecular Cell 63, 579–592. https://doi.org/10.1016/j.molcel.2016.07.008

Li, J., Guo, M., Tian, X., Wang, X., Yang, X., Wu, P., Liu, C., Xiao, Z., Qu, Y., Yin, Y., Wang, C., Zhang, Y., Zhu, Z., Liu, Z., Peng, C., Zhu, T., Liang, Q., 2021. Virus-Host Interactome and Proteomic Survey Reveal Potential Virulence Factors Influencing SARS-CoV-2 Pathogenesis. Med (N Y) 2, 99-112.e7. https://doi.org/10.1016/j.medj.2020.07.002

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lieber, D.S., Elemento, O., Tavazoie, S., 2010. Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. PLoS One 5, e14444. https://doi.org/10.1371/journal.pone.0014444

Ligtenberg, W., 2019. reactome.db: A set of annotation maps for reactome.

Liu-Wei, W., Kafkas, Ş., Chen, J., Dimonaco, N.J., Tegnér, J., Hoehndorf, R., 2021. DeepViral:

prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab147

Luck, K., Sheynkman, G.M., Zhang, I., Vidal, M., 2017. Proteome-Scale Human Interactomics. Trends in Biochemical Sciences 42, 342–354. https://doi.org/10.1016/j.tibs.2017.02.006

Lupberger, J., Zeisel, M.B., Xiao, F., Thumann, C., Fofana, I., Zona, L., Davis, C., Mee, C.J., Turek, M., Gorke, S., Royer, C., Fischer, B., Zahid, M.N., Lavillette, D., Fresquet, J., Cosset, F.-L., Rothenberg, S.M., Pietschmann, T., Patel, A.H., Pessaux, P., Doffoël, M., Raffelsberger, W., Poch, O., Mckeating, J.A., Brino, L., Baumert, T.F., 2011. EGFR and EphA2 are host factors for hepatitis C virus entry and possible targets for antiviral therapy. Nat Med 17, 589–595. https://doi.org/10.1038/nm.2341

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H., 2010. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 26, 1112–1118. https://doi.org/10.1093/bioinformatics/btq099

Mayer, K.A., Stöckl, J., Zlabinger, G.J., Gualdoni, G.A., 2019. Hijacking the Supplies: Metabolism as a Novel Facet of Virus-Host Interaction. Front. Immunol. 10. https://doi.org/10.3389/fimmu.2019.01533

Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C.J., Segura-Cabrera, A., Hersey, A., Leach, A.R., 2019. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research 47, D930–D940. https://doi.org/10.1093/nar/gky1075

Mészáros, B., Kumar, M., Gibson, T.J., Uyar, B., Dosztányi, Z., 2017. Degrons in cancer. Sci. Signal. 10. https://doi.org/10.1126/scisignal.aak9982

Mészáros, B., Sámano-Sánchez, H., Alvarado-Valverde, J., Čalyševa, J., Martínez-Pérez, E., Alves, R., Shields, D.C., Kumar, M., Rippmann, F., Chemes, L.B., Gibson, T.J., 2021. Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. Sci. Signal. 14. https://doi.org/10.1126/scisignal.abd0334

Meyer, K., Kirchner, M., Uyar, B., Cheng, J.-Y., Russo, G., Hernandez-Miranda, L.R., Szymborska, A., Zauber, H., Rudolph, I.-M., Willnow, T.E., Akalin, A., Haucke, V., Gerhardt, H., Birchmeier, C., Kühn, R., Krauss, M., Diecke, S., Pascual, J.M., Selbach, M., 2018. Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs. Cell 175, 239-253.e17. https://doi.org/10.1016/j.cell.2018.08.019

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Research 49, D412–D419. https://doi.org/10.1093/nar/gkaa913

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J.A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P.D., Finn, R.D., 2015. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Research 43, D213–D221. https://doi.org/10.1093/nar/gku1243

Murray, J.L., McDonald, N.J., Sheng, J., Shaw, M.W., Hodge, T.W., Rubin, D.H., O'Brien, W.A., Smee, D.F., 2012. Inhibition of influenza A virus replication by antagonism of a PI3K-AKT-mTOR pathway member identified by gene-trap insertional mutagenesis. Antivir Chem Chemother 22, 205–215. https://doi.org/10.3851/IMP2080

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., Masi, F. de, Gibson, T.J., Lewis, J., Serrano, L., Russell, R.B., 2005. Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks. PLOS Biology 3, e405. https://doi.org/10.1371/journal.pbio.0030405

Norez, C., Vandebrouck, C., Bertrand, J., Noel, S., Durieu, E., Oumata, N., Galons, H., Antigny, F., Chatelier, A., Bois, P., Meijer, L., Becq, F., 2014. Roscovitine is a proteostasis regulator that corrects the trafficking defect of F508del-CFTR by a CDK-independent mechanism. Br J Pharmacol 171, 4831–4849. https://doi.org/10.1111/bph.12859

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., Hermjakob, H.,

2014. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research 42, D358–D363. https://doi.org/10.1093/nar/gkt1115

Pabis, M., Corsini, L., Vincendeau, M., Tripsianes, K., Gibson, T.J., Brack-Werner, R., Sattler, M., 2019. Modulation of HIV-1 gene expression by binding of a ULM motif in the Rev protein to UHM-containing splicing factors. Nucleic Acids Res 47, 4859–4871. https://doi.org/10.1093/nar/gkz185

Palopoli, N., Lythgow, K.T., Edwards, R.J., 2015. QSLiMFinder: improved short linear motif prediction using specific query protein data. Bioinformatics 31, 2284–2293. https://doi.org/10.1093/bioinformatics/btv155

Pandit, B., Sarkozy, A., Pennacchio, L.A., Carta, C., Oishi, K., Martinelli, S., Pogna, E.A., Schackwitz, W., Ustaszewska, A., Landstrom, A., Bos, J.M., Ommen, S.R., Esposito, G., Lepri, F., Faul, C., Mundel, P., López Siguero, J.P., Tenconi, R., Selicorni, A., Rossi, C., Mazzanti, L., Torrente, I., Marino, B., Digilio, M.C., Zampino, G., Ackerman, M.J., Dallapiccola, B., Tartaglia, M., Gelb, B.D., 2007. Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. Nat Genet 39, 1007–1012. https://doi.org/10.1038/ng2073

Petsalaki, E., Stark, A., García-Urdiales, E., Russell, R.B., 2009. Accurate prediction of peptide binding sites on protein surfaces. PLoS computational biology 5, e1000335. https://doi.org/10.1371/journal.pcbi.1000335

Prytuliak, R., Volkmer, M., Meier, M., Habermann, B.H., 2017. HH-MOTiF: de novo detection of short linear motifs in proteins by Hidden Markov Model comparisons. Nucleic Acids Res. 45, W470–W477. https://doi.org/10.1093/nar/gkx341

Puschnik, A.S., Majzoub, K., Ooi, Y.S., Carette, J.E., 2017. A CRISPR toolbox to study virus–host interactions. Nat Rev Microbiol 15, 351–364. https://doi.org/10.1038/nrmicro.2017.29

Reimand, J., Wagih, O., Bader, G.D., 2015. Evolutionary Constraint and Disease Associations of Post-Translational Modification Sites in Human Genomes. PLOS Genetics 11, e1004919. https://doi.org/10.1371/journal.pgen.1004919

Rozenblatt-Rosen, O., Deo, R.C., Padi, M., Adelmant, G., Calderwood, M.A., Rolland, T., Grace, M., Dricot, A., Askenazi, M., Tavares, M., Pevzner, S.J., Abderazzaq, F., Byrdsong, D., Carvunis, A.-R., Chen, A.A., Cheng, J., Correll, M., Duarte, M., Fan, C., Feltkamp, M.C., Ficarro, S.B., Franchi, R., Garg, B.K., Gulbahce, N., Hao, T., Holthaus, A.M., James, R., Korkhin, A., Litovchick, L., Mar, J.C., Pak, T.R., Rabello, S., Rubio, R., Shen, Y., Singh, S., Spangle, J.M., Tasan, M., Wanamaker, S., Webber, J.T., Roecklein-Canfield, J., Johannsen, E., Barabási, A.-L., Beroukhim, R., Kieff, E., Cusick, M.E., Hill, D.E., Münger, K., Marto, J.A., Quackenbush, J., Roth, F.P., DeCaprio, J.A., Vidal, M., 2012. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. Nature 487, 491–495. https://doi.org/10.1038/nature11288

Sharma, S., Dincer, C., Weidemüller, P., Wright, G.J., Petsalaki, E., 2020. CEN-tools: an integrative platform to identify the contexts of essential genes. Molecular Systems Biology 16, e9698. https://doi.org/10.15252/msb.20209698

Sheng, Y., Saridakis, V., Sarkari, F., Duan, S., Wu, T., Arrowsmith, C.H., Frappier, L., 2006. Molecular recognition of p53 and MDM2 by USP7/HAUSP. Nat Struct Mol Biol 13, 285–291. https://doi.org/10.1038/nsmb1067

Shim, J.E., Kim, J.H., Shin, J., Lee, J.E., Lee, I., 2019. Pathway-specific protein domains are predictive for human diseases. PLoS Comput Biol 15, e1007052. https://doi.org/10.1371/journal.pcbi.1007052

Tarakhovsky, A., Prinjha, R.K., 2018. Drawing on disorder: How viruses use histone mimicry to their advantage. J Exp Med 215, 1777–1787. https://doi.org/10.1084/jem.20180099

Tenenbaum, D., 2020. KEGGREST: Client-side REST access to KEGG.

Teyra, J., Kelil, A., Jain, S., Helmy, M., Jajodia, R., Hooda, Y., Gu, J., D'Cruz, A.A., Nicholson, S.E., Min, J., Sudol, M., Kim, P.M., Bader, G.D., Sidhu, S.S., 2020. Large-scale survey and database of high affinity ligands for peptide recognition modules. Molecular Systems Biology 16, e9310. https://doi.org/10.15252/msb.20199310

Thaker, S.K., Ch'ng, J., Christofk, H.R., 2019. Viral hijacking of cellular metabolism. BMC Biology 17, 59. https://doi.org/10.1186/s12915-019-0678-9

Trabuco, L.G., Lise, S., Petsalaki, E., Russell, R.B., 2012. PepSite: prediction of peptide-binding sites from protein surfaces. Nucleic Acids Res 40, W423–W427. https://doi.org/10.1093/nar/gks398

Trible, R.P., Emert-Sedlak, L., Smithgall, T.E., 2006. HIV-1 Nef Selectively Activates Src Family

Kinases Hck, Lyn, and c-Src through Direct SH3 Domain Interaction*. Journal of Biological Chemistry 281, 27029–27038. https://doi.org/10.1074/jbc.M601128200

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., Meyers, R.M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W.F.J., Howell, S., Merkel, E., Ghandi, M., Garraway, L.A., Root, D.E., Golub, T.R., Boehm, J.S., Hahn, W.C., 2017. Defining a Cancer Dependency Map. Cell 170, 564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010

Uyar, B., J. Weatheritt, R., Dinkel, H., E. Davey, N., J. Gibson, T., 2014. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? Molecular BioSystems 10, 2626–2642. https://doi.org/10.1039/C4MB00290C

Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J., Davey, N.E., 2014. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. Chem. Rev. 114, 6733–6778. https://doi.org/10.1021/cr400585q

Via, A., Uyar, B., Brun, C., Zanzoni, A., 2015. How pathogens use linear motifs to perturb host cell networks. Trends in Biochemical Sciences 40, 36–48. https://doi.org/10.1016/j.tibs.2014.11.001

Villaveces, J.M., Jiménez, R.C., Porras, P., del-Toro, N., Duesbury, M., Dumousseau, M., Orchard, S., Choi, H., Ping, P., Zong, N.C., Askenazi, M., Habermann, B.H., Hermjakob, H., 2015. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. Database (Oxford) 2015. https://doi.org/10.1093/database/bau131

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.-F., 2007. A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Weatheritt, R.J., Luck, K., Petsalaki, E., Davey, N.E., Gibson, T.J., 2012. The identification of short linear motif-mediated interfaces within the human interactome. Bioinformatics 28, 976–982. https://doi.org/10.1093/bioinformatics/bts072

Weinberg, R.A., 1995. The retinoblastoma protein and cell cycle control. Cell 81, 323–330. https://doi.org/10.1016/0092-8674(95)90385-2

Wilson, V.G., 2014. The Role of Ubiquitin and Ubiquitin-Like Modification Systems in Papillomavirus Biology. Viruses 6, 3584–3611. https://doi.org/10.3390/v6093584

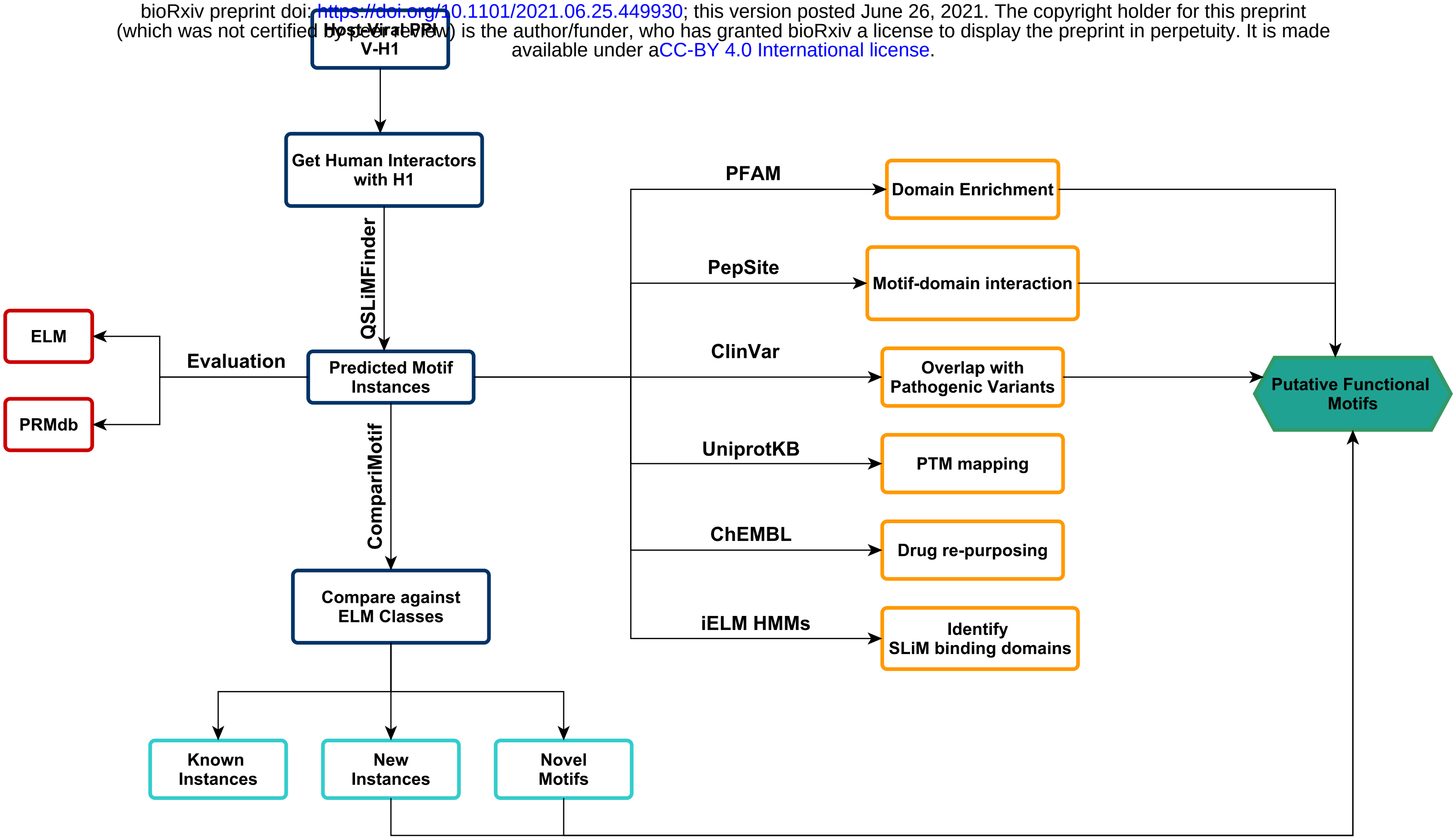
Yang, C.-W., Shi, Z.-L., 2021. Uncovering potential host proteins and pathways that may interact with eukaryotic short linear motifs in viral proteins of MERS, SARS and SARS2 coronaviruses that infect humans. PLOS ONE 16, e0246150. https://doi.org/10.1371/journal.pone.0246150

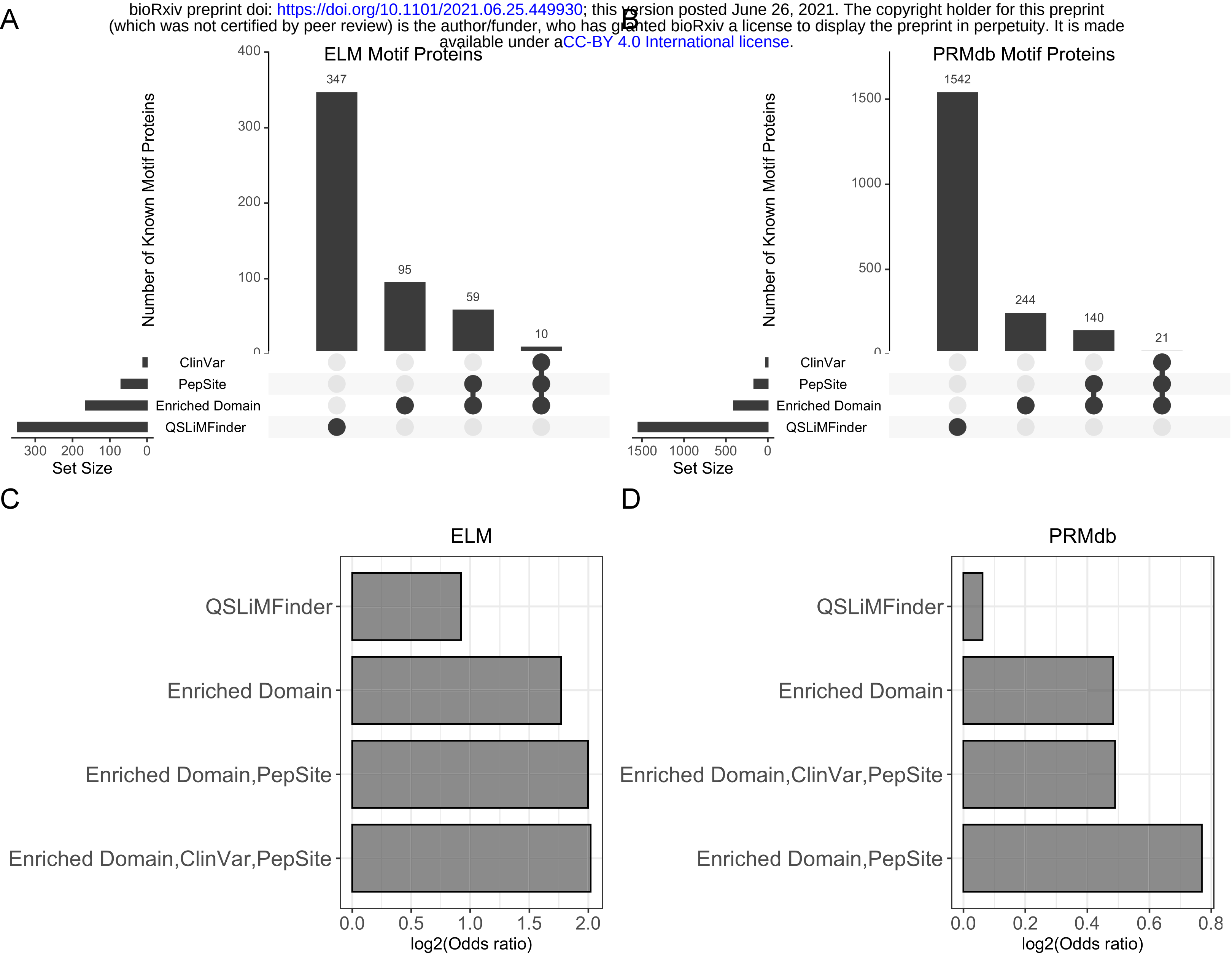
Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S., 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26, 976–978. https://doi.org/10.1093/bioinformatics/btq064

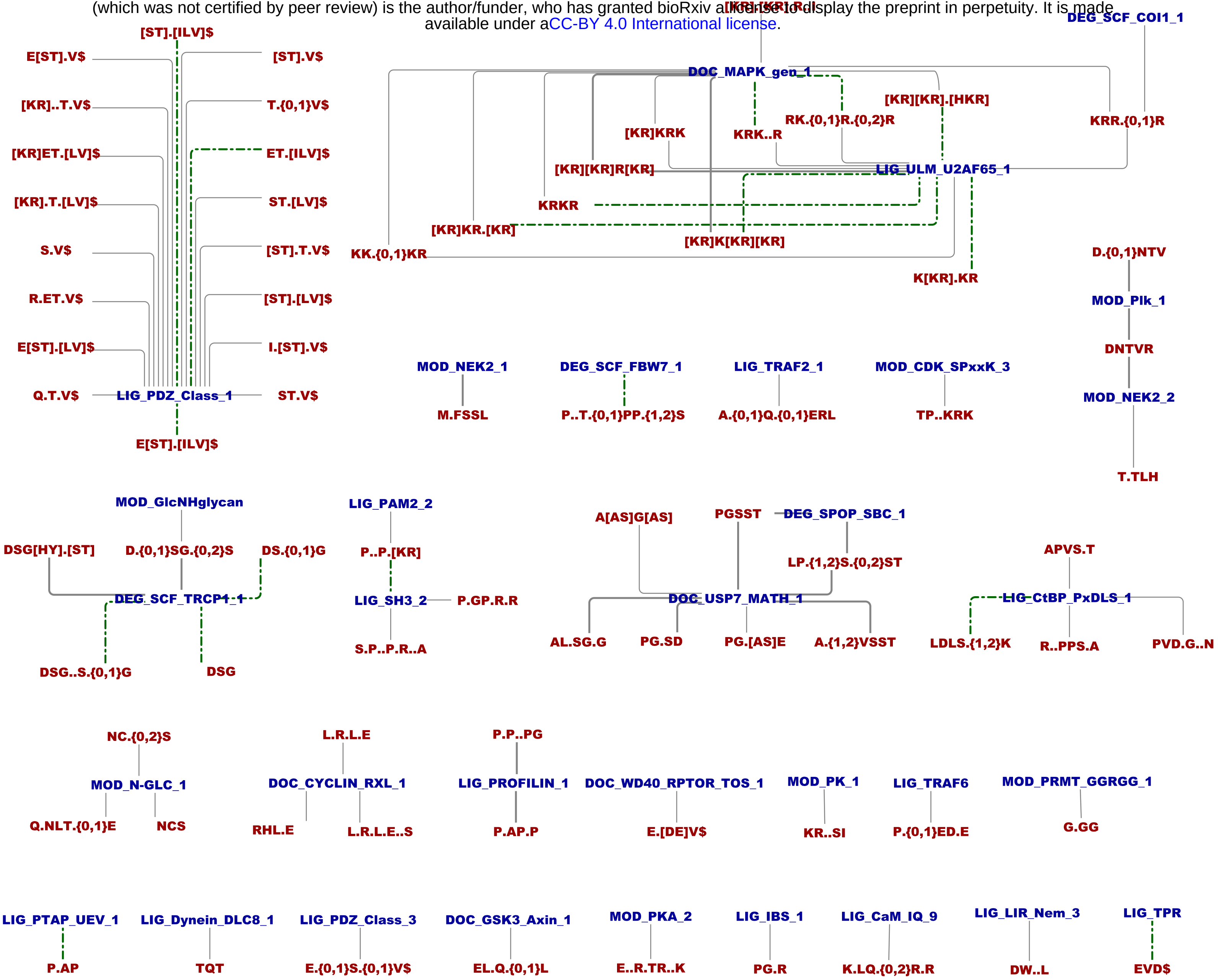
Yu, G., Wang, L.-G., Han, Y., He, Q.-Y., 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS 16, 284–287. https://doi.org/10.1089/omi.2011.0118

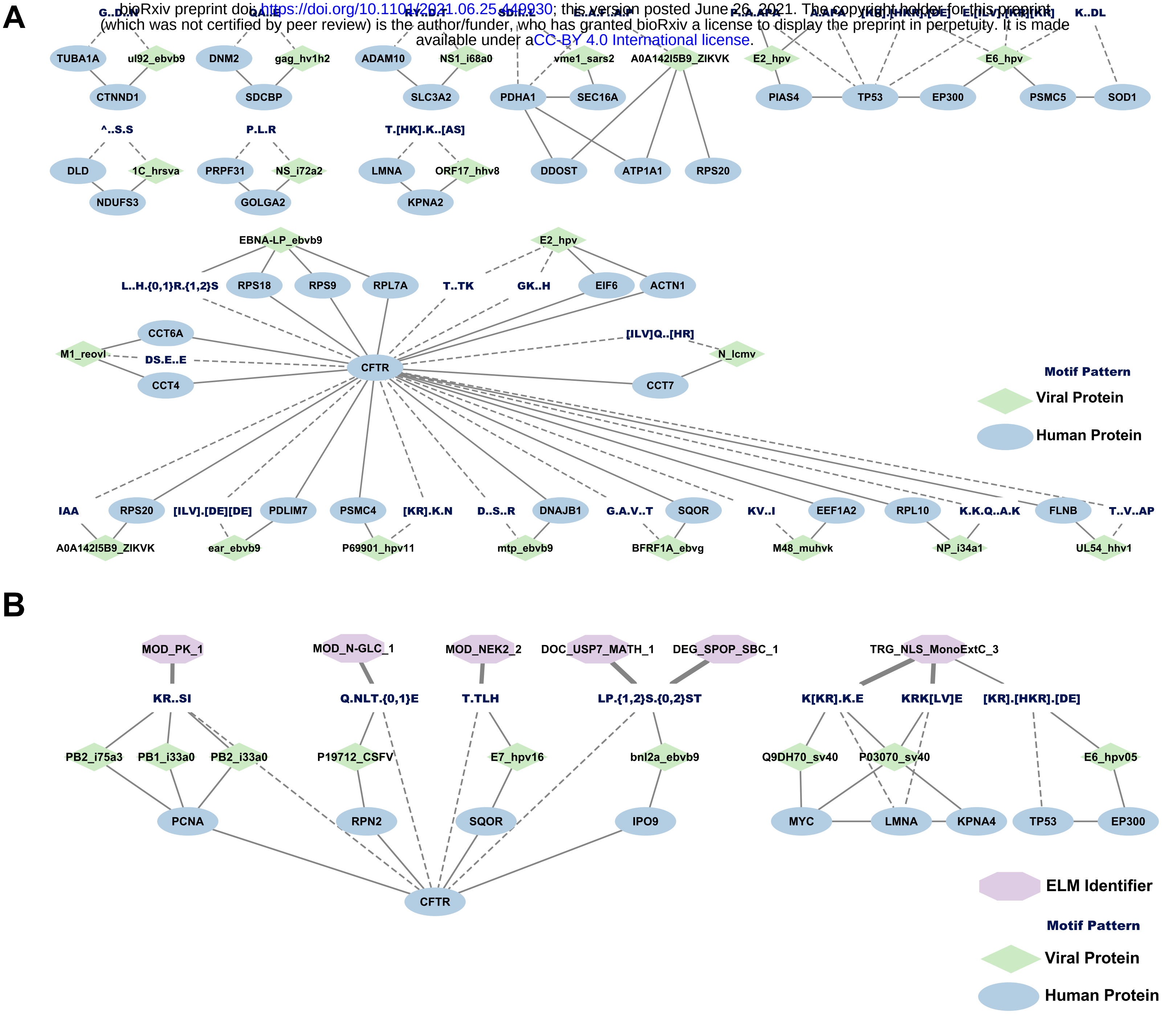
Zebulun, A., 2017. rhmmer: Utilities Parsing “HMMER” Results.

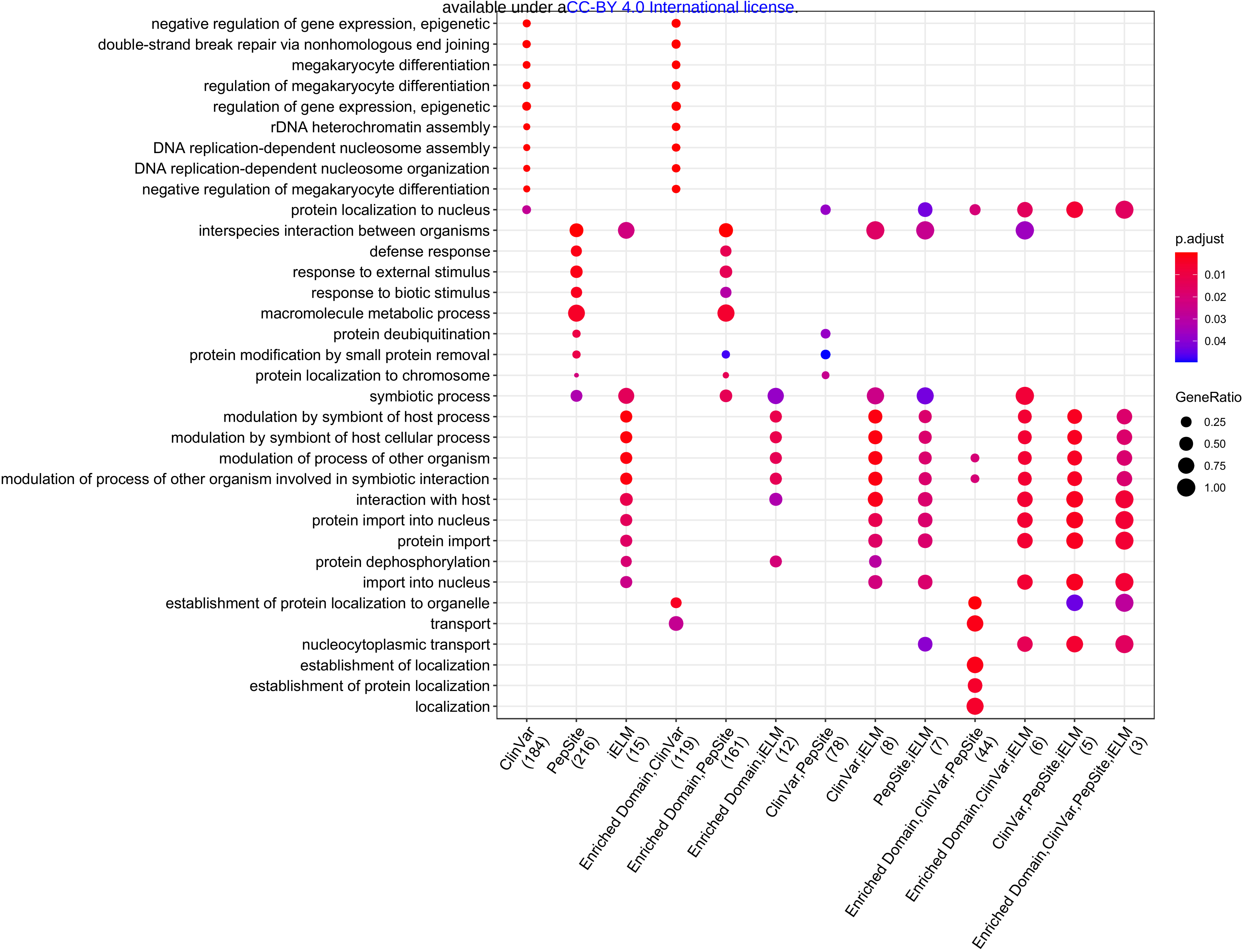
Zheng, L.-L., Li, C., Ping, J., Zhou, Y., Li, Y., Hao, P., 2014. The Domain Landscape of Virus-Host Interactomes. BioMed Research International 2014, e867235. https://doi.org/10.1155/2014/867235

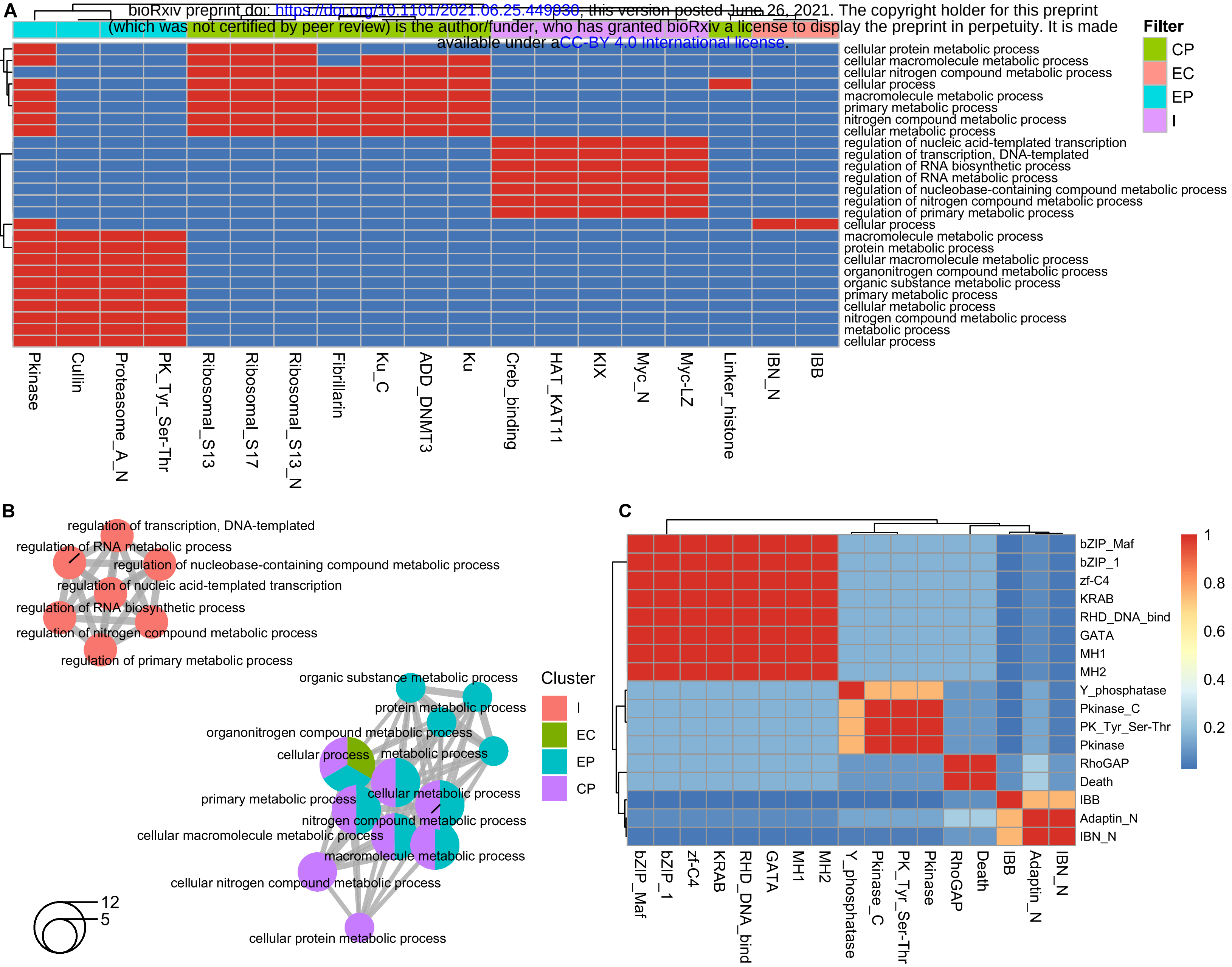

Host-Viral PPI
V-H1
Get Human Interactors
with H1
QSLiMFinder
Predicted Motif
Instances
Evaluation
ELM
PRMdb
CompariMotif
Compare against
ELM Classes
Known
Instances
New
Instances
Novel
Motifs
PFAM
Domain Enrichment
PepSite
Motif-domain interaction
ClinVar
Overlap with
Pathogenic Variants
UniprotKB
PTM mapping
ChEMBL
Drug re-purposing
iELM HMMs
Identify
SLiM binding domains
Putative Functional
Motifs

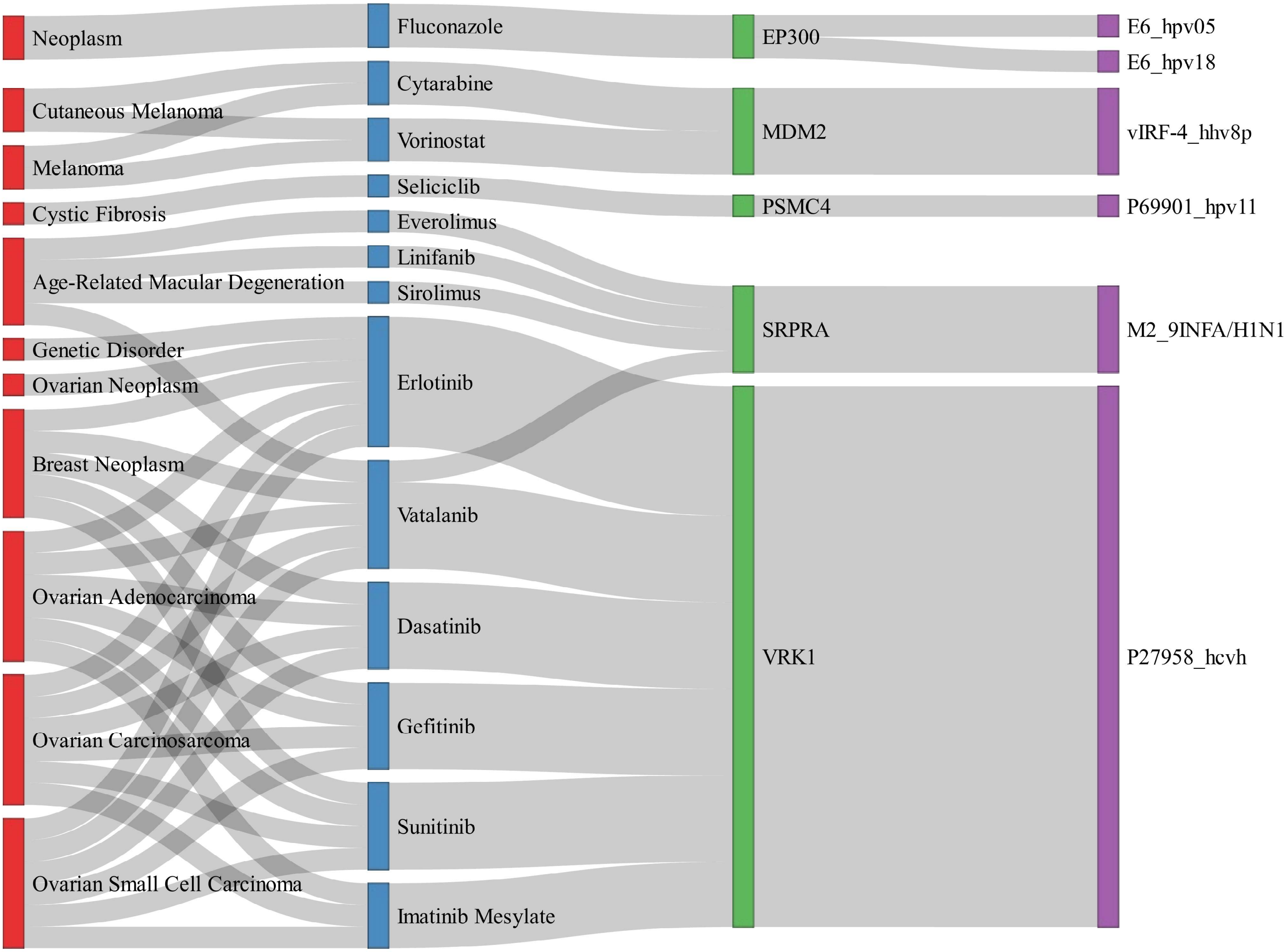

A
ELM Motif Proteins
Number of Known Motif Proteins
400
300
200
100
0
347
95
59
10
ClinVar
PepSite
Enriched Domain
QSLiMFinder
300
200
100
0
Set Size
B
PRMdb Motif Proteins
Number of Known Motif Proteins
1500
1000
500
0
1542
244
140
21
ClinVar
PepSite
Enriched Domain
QSLiMFinder
1500
1000
500
0
Set Size
C
ELM
QSLiMFinder
Enriched Domain
Enriched Domain,PepSite
Enriched Domain,ClinVar,PepSite
0.0
0.5
1.0
1.5
2.0
log2(Odds ratio)
D
PRMdb
QSLiMFinder
Enriched Domain
Enriched Domain,ClinVar,PepSite
Enriched Domain,PepSite
0.0
0.2
0.4
0.6
0.8
log2(Odds ratio)


[ST].[ILV]$
E[ST].V$
[ST].V$
[KR]..T.V$
T.{0,1}V$
[KR]ET.[LV]$
ET.[ILV]$
[KR].T.[LV]$
ST.[LV]$
S.V$
[ST].T.V$
R.ET.V$
[ST].[LV]$
E[ST].[LV]$
I.[ST].V$
Q.T.V$
LIG_PDZ_Class_1
ST.V$
E[ST].[ILV]$
[KR].[KR].R.I
DEG_SCF_COI1_1
DOC_MAPK_gen_1
[KR][KR].[HKR]
RK.{0,1}R.{0,2}R
KRR.{0,1}R
[KR]KRK
KRK..R
[KR][KR]R[KR]
LIG_ULM_U2AF65_1
KRKR
[KR]KR.[KR]
[KR]K[KR][KR]
KK.{0,1}KR
K[KR].KR
D.{0,1}NTV
MOD_Plk_1
DNTVR
MOD_NEK2_1
DEG_SCF_FBW7_1
LIG_TRAF2_1
MOD_CDK_SPxxK_3
MOD_NEK2_2
M.FSSL
P..T.{0,1}PP.{1,2}S
A.{0,1}Q.{0,1}ERL
TP..KRK
T.TLH
MOD_GlcNHglycan
LIG_PAM2_2
A[AS]G[AS]
PGSST
DEG_SPOP_SBC_1
DSG[HY].[ST]
D.{0,1}SG.{0,2}S
DS.{0,1}G
P..P.[KR]
LP.{1,2}S.{0,2}ST
APVS.T
DEG_SCF_TRCP1_1
LIG_SH3_2
P.GP.R.R
DOC_USP7_MATH_1
LIG_CtBP_PxDLS_1
AL.SG.G
PG.SD
PG.[AS]E
A.{1,2}VSST
LDLS.{1,2}K
R..PPS.A
PVD.G..N
S.P..P.R..A
DSG..S.{0,1}G
DSG
NC.{0,2}S
L.R.L.E
P.P..PG
MOD_N-GLC_1
DOC_CYCLIN_RXL_1
LIG_PROFILIN_1
DOC_WD40_RPTOR_TOS_1
MOD_PK_1
LIG_TRAF6
MOD_PRMT_GGRGG_1
Q.NLT.{0,1}E
NCS
RHL.E
L.R.L.E..S
P.AP.P
E.[DE]V$
KR..SI
P.{0,1}ED.E
G.GG
LIG_PTAP_UEV_1
LIG_Dynein_DLC8_1
LIG_PDZ_Class_3
DOC_GSK3_Axin_1
MOD_PKA_2
LIG_IBS_1
LIG_CaM_IQ_9
LIG_LIR_Nem_3
LIG_TPR
P.AP
TQT
E.{0,1}S.{0,1}V$
EL.Q.{0,1}L
E..R.TR..K
PG.R
K.LQ.{0,2}R.R
DW..L
EVD$


A
G..D..N
QA..E
RY..D.T
SD.I..L
E..A.F..A.P
P..A.APA
A.APA
[KR].[HKR].[DE]
E.[ILV].[KR][KR]
K..DL
TUBA1A
ul92_ebvb9
DNM2
gag_hv1h2
ADAM10
NS1_i68a0
vme1_sars2
A0A142I5B9_ZIKVK
E2_hpv
E6_hpv
CTNND1
SDCBP
SLC3A2
PDHA1
SEC16A
PIAS4
TP53
EP300
PSMC5
SOD1
^..S.S
P.L.R
T.[HK].K..[AS]
DLD
1C_hrsva
PRPF31
NS_i72a2
LMNA
ORF17_hhv8
DDOST
ATP1A1
RPS20
NDUFS3
GOLGA2
KPNA2
EBNA-LP_ebvb9
E2_hpv
L..H.{0,1}R.{1,2}S
RPS18
RPS9
RPL7A
T..TK
GK..H
EIF6
ACTN1
CCT6A
[ILV]Q..[HR]
M1_reovl
DS.E..E
N_lcmv
CFTR
CCT4
CCT7
Motif Pattern
Viral Protein
Human Protein
IAA
RPS20
[ILV].[DE][DE]
PDLIM7
PSMC4
[KR].K.N
D..S..R
DNAJB1
G.A.V..T
SQOR
KV..I
EEF1A2
RPL10
K.K.Q..A.K
FLNB
T..V..AP
A0A142I5B9_ZIKVK
ear_ebvb9
P69901_hpv11
mtp_ebvb9
BFRF1A_ebvg
M48_muhvk
NP_i34a1
UL54_hhv1
B
MOD_PK_1
MOD_N-GLC_1
MOD_NEK2_2
DOC_USP7_MATH_1
DEG_SPOP_SBC_1
TRG_NLS_MonoExtC_3
KR..SI
Q.NLT.{0,1}E
T.TLH
LP.{1,2}S.{0,2}ST
K[KR].K.E
KRK[LV]E
[KR].[HKR].[DE]
PB2_i75a3
PB1_i33a0
PB2_i33a0
P19712_CSFV
E7_hpv16
bnl2a_ebvb9
Q9DH70_sv40
P03070_sv40
E6_hpv05
PCNA
RPN2
SQOR
IPO9
MYC
LMNA
KPNA4
TP53
EP300
CFTR
ELM Identifier
Motif Pattern
Viral Protein
Human Protein


negative regulation of gene expression, epigenetic
double-strand break repair via nonhomologous end joining
megakaryocyte differentiation
regulation of megakaryocyte differentiation
regulation of gene expression, epigenetic
rDNA heterochromatin assembly
DNA replication-dependent nucleosome assembly
DNA replication-dependent nucleosome organization
negative regulation of megakaryocyte differentiation
protein localization to nucleus
interspecies interaction between organisms
defense response
response to external stimulus
response to biotic stimulus
macromolecule metabolic process
protein deubiquitination
protein modification by small protein removal
protein localization to chromosome
symbiotic process
modulation by symbiont of host process
modulation by symbiont of host cellular process
modulation of process of other organism
modulation of process of other organism involved in symbiotic interaction
interaction with host
protein import into nucleus
protein import
protein dephosphorylation
import into nucleus
establishment of protein localization to organelle
transport
nucleocytoplasmic transport
establishment of localization
establishment of protein localization
localization
ClinVar (184)
PepSite (216)
iELM (15)
Enriched Domain,ClinVar (119)
Enriched Domain,PepSite (161)
Enriched Domain,iELM (12)
ClinVar,PepSite (78)
ClinVar,iELM (8)
PepSite,iELM (7)
Enriched Domain,ClinVar,PepSite (44)
Enriched Domain,ClinVar,iELM (6)
ClinVar,PepSite,iELM (5)
Enriched Domain,ClinVar,PepSite,iELM (3)
p.adjust
0.01
0.02
0.03
0.04
GeneRatio
0.25
0.50
0.75
1.00


A
Filter
CP
EC
EP
I
cellular protein metabolic process
cellular macromolecule metabolic process
cellular nitrogen compound metabolic process
cellular process
macromolecule metabolic process
primary metabolic process
nitrogen compound metabolic process
cellular metabolic process
regulation of nucleic acid-templated transcription
regulation of transcription, DNA-templated
regulation of RNA biosynthetic process
regulation of RNA metabolic process
regulation of nucleobase-containing compound metabolic process
regulation of nitrogen compound metabolic process
regulation of primary metabolic process
cellular process
macromolecule metabolic process
protein metabolic process
cellular macromolecule metabolic process
organonitrogen compound metabolic process
organic substance metabolic process
primary metabolic process
cellular metabolic process
nitrogen compound metabolic process
metabolic process
cellular process
Pkinase
Cullin
Proteasome_A_N
PK_Tyr_Ser-Thr
Ribosomal_S13
Ribosomal_S17
Ribosomal_S13_N
Fibrillarin
Ku_C
ADD_DNMT3
Ku
Creb_binding
HAT_KAT11
KIX
Myc_N
Myc-LZ
Linker_histone
IBN_N
IBB
B
regulation of transcription, DNA-templated
regulation of RNA metabolic process
regulation of nucleobase-containing compound metabolic process
regulation of nucleic acid-templated transcription
regulation of RNA biosynthetic process
regulation of nitrogen compound metabolic process
regulation of primary metabolic process
organic substance metabolic process
protein metabolic process
organonitrogen compound metabolic process
cellular process
metabolic process
primary metabolic process
cellular metabolic process
nitrogen compound metabolic process
cellular macromolecule metabolic process
macromolecule metabolic process
cellular nitrogen compound metabolic process
cellular protein metabolic process
12
5
Cluster
I
EC
EP
CP
C
bZIP_Maf
bZIP_1
zf-C4
KRAB
RHD_DNA_bind
GATA
MH1
MH2
Y_phosphatase
Pkinase_C
PK_Tyr_Ser-Thr
Pkinase
RhoGAP
Death
IBB
Adaptin_N
IBN_N
1
0.8
0.6
0.4
0.2


Neoplasm
Cutaneous Melanoma
Melanoma
Cystic Fibrosis
Age-Related Macular Degeneration
Genetic Disorder
Ovarian Neoplasm
Breast Neoplasm
Ovarian Adenocarcinoma
Ovarian Carcinosarcoma
Ovarian Small Cell Carcinoma
Fluconazole
Cytarabine
Vorinostat
Seliciclib
Everolimus
Linifanib
Sirolimus
Erlotinib
Vatalanib
Dasatinib
Gefitinib
Sunitinib
Imatinib Mesylate
EP300
MDM2
PSMC4
SRPRA
VRK1
E6_hpv05
E6_hpv18
vIRF-4_hhv8p
P69901_hpv11
M2_9INFA/H1N1
P27958_hcvh