Structural insights into pathogen-induced perturbation of host protein interactome

Dissertation

by

Yangchun Frank Chen

Supervisor:
Yu (Brandon) Xia

Advisory committee members:
Amine Kamen and Jérôme Waldispühl

Biological and Biomedical Engineering
McGill University, Montreal

June 2021

A thesis submitted to McGill University
in partial fulfilment of the requirements of the degree of
Doctor of Philosophy in Biological and Biomedical Engineering

Table of Contents

List of Figures

List of Tables

ABSTRACT

La version française suit.

Pathogenic viruses and bacteria encode a multitude of virulence factors which, through extensive interactions with host proteins, perturb the host protein interactome and rewire host signalling pathways to the advantage of the pathogen. Despite insights gained from studies showing host-binding and modulatory properties of specific microbial virulence factors, there has yet to be a comprehensive, structural characterization of virulence factors as a general class of exogenous "perturbagens" of the host protein interactome. Here, I examine correlations between the structural composition of microbial virulence factors and their mechanistic involvement in host genetic diseases, as well as their propensity for disrupting or repurposing host-specific cellular processes, using domain-resolved host-pathogen protein interaction networks.

Based on principles of homology modelling, I first construct a domain-resolution, human-virus protein-protein interaction (PPI) network where human domains are annotated with causal mutations for diseases that have both genetic and viral etiologic factors. I show that point mutations and viral infections leading to similar diseases tend to perturb the same human domains and domain-domain interactions (DDIs). In addition, domains of human oncoproteins either physically targeted or structurally mimicked by oncoviruses are enriched for cancer driver rather than passenger mutations. These results demonstrate that similar perturbations of the human protein interactome at the domain level can have equivalent phenotypic consequences, regardless of whether the perturbation is initiated by endogenous genetic alterations or exogenous viral proteins.

Next, I construct a domain-resolution, eukaryote-bacteria PPI network and assess the potential of domains and short linear motifs within bacterial proteins to repurpose or disrupt eukaryote-specific PPIs. I show that compared to the rest of the pathogen proteome, effector proteins are enriched for domains that either structurally mimic or convergently target host domains involved in eukaryote-specific DDIs, as opposed to DDIs that are conserved between eukaryotes and bacteria. Moreover, in the absence of eukaryotic-like domains or among pathogen proteins without domain assignment, effector proteins harbour a higher variety and density of short linear motifs which are known to interact with eukaryotic domains.

Given the rapidly growing number of microbial genome sequences and the relative scarcity of host-pathogen PPI data, structure-function analysis based on homology modelling may help accelerate the discovery and mechanistic study of novel virulence factors, as well as the development of selective inhibitors of pathogen-subverted host signalling pathways.

RÉSUMÉ

Les virus et bactéries pathogènes encodent une multitude de facteurs de virulence qui, grâce à de nombreuses interactions avec les protéines de l'hôte, perturbent l'interactome protéique et re-câblent les voies de signalisation de l'hôte à l'avantage de l'agent pathogène. Malgré les connaissances obtenues lors d'études sur les propriétés de liaison à l'hôte, et les sur les propriétés modulatrices de certains facteurs de virulences microbiens spécifiques, il n'y a pas encore eu de caractérisation complète et structurelle des facteurs de virulence, en tant que classe générale de « perturbagènes » exogènes des l'interactome protéique d'un un hôte. Ici, j'examine les corrélations entre la composition structurelle des facteurs de virulence microbienne, et leur implication mécaniste dans les maladies génétiques de l'hôte, ainsi que leur propension à perturber ou à réutiliser des processus cellulaires spécifiques à l'hôte, en utilisant des réseaux d'interaction protéiques hôte-pathogène, à la résolution des domaines protéiques.

En me basant sur les principes de la modélisation homologique, je construis d'abord un réseau d'interactions protéine-protéine (IPP) entre humain et virus, à la résolution des domaines protéiques. Chaque domaine humain est, par la suite, annoté avec des mutations causant des maladies qui ont des facteurs étiologiques génétiques et viraux. Je montre que les mutations ponctuelles et les infections virales conduisant à des maladies similaires ont tendance à perturber les mêmes domaines humains, et les mêmes interactions domaine-domaine (IDDs). De plus, les domaines des oncoprotéines humaines physiquement ciblées ou structurellement imitées par les oncovirus sont enrichis en mutations pilotes pour le cancer, plutôt qu'en mutations passagères. Ces résultats démontrent que des perturbations similaires de l'interactome protéique humain au niveau des domaines peuvent avoir des conséquences phénotypiques équivalentes, indépendamment du

fait que la perturbation soit initiée par des altérations génétiques endogènes ou des protéines virales exogènes.

Ensuite, je construis un réseau d'IPP, à la résolution des domaines protéiques, entre humain et bactérie, et j'évalue le potentiel des domaines et des motifs linéaires courts dans les protéines bactériennes pour réutiliser ou perturber les IPP spécifiques aux eucaryotes. Je montre que, comparé au reste du protéome pathogène, les protéines effectrices sont enrichies en domaines qui imitent structurellement ou ciblent de manière convergente les domaines hôtes impliqués dans des IDDs spécifiques à l'eucaryote, par opposition aux IDDs qui sont conservés entre les eucaryotes et les bactéries. De plus, même en l'absence de domaines de type eucaryote, ou parmi les protéines pathogènes sans attribution de domaine, les protéines effectrices abritent une plus grande variété et densité de motifs qui sont connus pour interagir avec des domaines eucaryotes.

Compte tenu de l'augmentation rapide du nombre de séquences de génomes microbiens et de la rareté relative des données d'IPP hôte-pathogène, l'analyse structure-fonction basée sur la modélisation homologique peut aider à accélérer la découverte et l'étude des mécanismes de nouveaux facteurs de virulence, ainsi que le développement potentiel d'inhibiteurs sélectifs des voies de signalisation de l'hôte détournées par les pathogènes.

Acknowledgements

I am deeply grateful to my thesis supervisor, Brandon Xia, for urging me on as I fumble for a sense of urgency and purpose, for being patient as I struggle to find my way, and for sharing his philosophy of minimalism, which is at once theoretical and practical, towards scientific reasoning and problem solving. I thank members of my advisory committee, Amine Kamen and Jérôme Waldispühl, for providing constructive criticism and feedback during committee meetings. I thank the thesis examiners and manuscript reviewers, for their positive evaluation of my work and valuable intellectual insights. Special thanks to my colleague, Léah Pollet, for helping with the French translation of my thesis abstract. And finally, I would like to thank my parents for paving the way for me to study in Canada, and for their unceasing love and support throughout my tortuous and, at times, torturous PhD journey.

My work was funded by the McGill Engineering Doctoral Award, and by Natural Sciences and Engineering Research Council of Canada grants, Canada Foundation for Innovation grants, and Canada Research Chairs program awarded to my supervisor, Brandon Xia.

Contribution to Original Knowledge

Chapter 3 of this thesis produced the first comprehensive domain-resolution human-virus protein interaction network where human domains are annotated with disease variant information. Mapping of disease mutations with respect to virus-targeted human domains showed that similar perturbations of the human interactome at the domain level can have similar phenotypic consequences, regardless of the source of perturbation (endogenous genetic mutations *vs.* exogenous viral proteins). The study provides a framework for: (1) high-resolution, network-based comparison of the functional impacts of various genetic and environmental disease factors; and (2) identification of oncovirus-targeted or mimicked human domains and their interacting domains for focused screening of driver mutations across various types of cancer, which can reveal immune evasion strategies exploited in common by cancer cells and pathogens, and shed light on pathways dysregulated in other virally-implicated disorders.

Chapter 4 of this thesis compared the mechanism of binding site mimicry in host-endogenous *vs.* host-bacteria PPI network at the domain level, providing novel insight into the evolution of host-interacting domains in bacterial effectors. In particular, convergent evolution (or extreme divergent evolution) appears to be the more dominant mechanism behind binding site mimicry in host-bacteria interactions. To date, similar analysis has only been done for viral proteins. In addition, estimation of domain's relevance to eukaryote-specific domain-domain interactions (DDIs) provides quantitative, interaction-based criteria for identifying novel effectors, based on: (1) domains that exclusively or primarily mediate DDIs in eukaryotes; and (2) variety and density of short linear motifs targeting host domains that exclusively mediate DDIs in eukaryotes.

Contribution of Authors

Chen wrote Chapters 1, 2, 5, and 6. Thesis supervisor Brandon Xia, thesis examiners Reza Salavati and Jishnu Das, as well as reviewers of the manuscripts in Chapters 3 and 4 suggested relevant literature and points for discussion.

Chapter 3 contains materials from the published manuscript: Chen, Y.F. and Y. Xia, *Convergent perturbation of the human domain-resolved interactome by viruses and mutations inducing similar disease phenotypes.* PLoS Comput Biol, 2019. **15**(2): p. e1006762.

Chapter 4 contains materials from the initial submission of the recently published manuscript: Chen, Y.F. and Y. Xia, *Structural Profiling of Bacterial Effectors Reveals Enrichment of Host-Interacting Domains and Motifs.* Front Mol Biosci, 2021. **8**: p. 626600. Parts of Chapters 1, 2, 5, and 6 contain materials from the accepted version of this manuscript.

For Chapters 3 and 4, Chen was responsible for project conceptualization, methodology development, data curation and analysis, as well as writing and editing of the manuscripts. Xia contributed to project conceptualization and supervision, methodology development, and editing of the manuscripts.

Chapter 1: Introduction

Protein-protein interaction (PPI) networks provide key insights into the structural and functional organization of proteomes in physiological and disease conditions. In fact, systems biology often models disease as resulting from perturbation of PPI networks. Perturbations are mainly from two sources: endogenous genetic mutations, and exogenous factors such as pathogens and drug molecules. In much the same way genetic diseases can be viewed as resulting from mutation-induced perturbation of a patient's protein interactome, infectious diseases can be viewed as resulting from pathogen-induced perturbation of the host's protein interactome. There are clear advantages to using structurally-resolved, as opposed to whole-protein resolution PPI networks, because knowledge of the PPI interface is crucial to uncovering the mechanism of host-pathogen interactions, and the PPI interface is only preserved in structurally-resolved networks, but not in whole-protein resolution networks. For instance, previous studies found that human-virus PPI interfaces overlap extensively with human-endogenous PPI interfaces, suggesting that rather than creating new interfaces, viruses tend to repurpose existing interfaces in the human-endogenous network for human-virus interactions [1]. This insight would not have been gained using a whole-protein resolution network. It is conceivable that structurally-resolved host-pathogen PPI networks can help elucidate the location of disease mutations with respect to pathogen-targeted host domains, as well as the evolutionary origins of host-interacting structural modules within pathogen proteins. The overall objective of this thesis is to characterize pathogen proteins as exogenous perturbagens of the host protein interactome, using domain-resolved human-virus and eukaryote-bacteria PPI networks. PPIs without experimentally determined structures are resolved using domain-domain interaction templates derived from solved structures of protein complexes.

In Chapter 3, I used a domain-resolved human-virus PPI network to examine correlations between structural features of viral proteins and their mechanistic involvement in human disease. An important goal of systems medicine is to study disease in the context of genetic and environmental perturbations to the human interactome network. For diseases with both genetic and infectious etiologic factors, a key postulate is that similar perturbations of the human protein interactome by either disease mutations or pathogens can have similar disease consequences. This postulate has so far only been tested for a few viral species at the whole-protein level. Here, I expanded the scope of viral species and tested this postulate more rigorously at the higher resolution of protein domains, by constructing a domain-resolved human-virus PPI network where human domains are annotated with disease mutations. I show that missense mutations and viral infections leading to similar diseases tend to perturb the same human domains and domain-domain interactions. In addition, domains of human oncoproteins either physically targeted or structurally mimicked by oncoviruses are enriched for cancer driver rather than passenger mutations. These results suggest that similar perturbations of the human protein interactome at the domain level can have equivalent phenotypic consequences, regardless of the source of perturbation.

In Chapter 4, I used a domain-resolved eukaryote-bacteria PPI network to examine correlations between structural features of bacterial effector proteins and their potential for targeting eukaryote-specific cellular processes. Effector proteins are bacterial virulence factors secreted directly into host cells and, through extensive interactions with host proteins, rewire host signalling pathways to the advantage of the pathogen. Despite the crucial role of globular domains as mediators of PPIs, previous structural studies of bacterial effectors are primarily focused on individual domains, rather than domain-mediated PPIs, which limits their ability to uncover systems-level molecular recognition principles governing host-bacteria interactions. Studies of host-bacteria PPIs have so

far examined either individual interactions at the domain level [2], or interactome networks at the whole protein level [3], but never both at the domain level and on an interactome scale. Here, I ask the following new question: how do host-bacteria PPIs mimic and modulate host-endogenous PPIs at the protein domain level on an interactome scale? To answer this question, I carried out two analyses of host-interacting bacterial proteins: the first on mimicry of host-endogenous binding sites by bacterial effectors, and the second on enrichment of host-interacting domains and short linear motifs in bacterial effectors. In the first analysis, I examined the mechanism of host binding site mimicry by bacterial proteins where, rather than creating new binding sites, bacteria recruit existing binding sites involved in host-endogenous PPIs for host-bacteria PPIs [2]. Previous studies of host-virus interactions found that binding site sharing among human proteins is largely attributable to divergent evolution through gene duplication, whereas binding site mimicry by viral proteins tends to involve convergent evolution of unique host-interacting modules in viruses [1, 4]. Similar analyses have yet to be performed for host-bacteria interactions. In the second analysis, I tested the hypothesis that compared to non-effector proteins, bacterial effectors are enriched for domains that either mimic or target host domains involved in eukaryote-specific domain-domain interactions (DDIs). In addition to domains, I also tested whether effectors tend to contain a higher variety and density of short linear motifs that interact with host domains mediating DDIs exclusively in eukaryotes. I show that compared to the rest of the pathogen proteome, effectors are enriched for domains that either structurally mimic or convergently target host domains involved in eukaryote-specific DDIs, as opposed to DDIs that are conserved between eukaryotes and bacteria. Moreover, in the absence of eukaryotic-like domains or among pathogen proteins without domain assignment, effectors harbour a higher variety and density of motifs which are known to interact with eukaryotic domains.

Chapter 2: Literature Review

Cellular processes are driven and regulated by highly coordinated biomolecular interaction networks – a prime example being the protein-protein interaction (PPI) network, often referred to as the protein interactome [5, 6]. When a cell is subjected to stress conditions such as genetic mutations or infectious agents, PPIs are often perturbed and as a result, the interactome network is rewired. Pathogenic viruses and bacteria encode a multitude of virulence factors that mimic or target host proteins involved in actin remodelling, protein degradation and cell cycle regulation, which helps the pathogen propagate in the host while bypassing immune surveillance [7, 8]. Despite insights gained from studies showing host-binding and modulatory properties of specific microbial virulence factors, there has yet to be a comprehensive, structural characterization of virulence factors as a general class of exogenous "perturbagens" of the host protein interactome. This thesis aims to examine correlations between the structural composition of microbial virulence factors and their mechanistic involvement in host genetic diseases, as well as their propensity for repurposing or redirecting host-specific cellular processes, using a domain-resolved interaction network consisting of within-human, within-pathogen and human-pathogen PPIs. Extraction of these molecular insights requires careful consideration of the physicochemical and structural features of PPI interfaces, which are preserved in structurally-resolved PPI networks, but are obscured in low-resolution PPI networks that treat proteins and PPIs as indistinguishable nodes and edges. Below is a review of current literature on high-throughput screens for protein-protein interactions, methods for constructing structurally-resolved PPI networks, and applications of structural interaction networks in the systems biology of genetic and infectious diseases.

High-throughput screens for protein-protein interactions. A cell is often compared to an ensemble of molecular machines, sustained by networks of interacting biomolecules which include proteins, nucleic acids, and metabolites. There are, broadly speaking, two distinct types of interactions: physical and genetic. Physical associations among proteins dictate the formation of protein complexes and signalling cascades, which are crucial components of the molecular machinery. Genetic interactions, on the other hand, reflect functional relationships between genes and are responsible for non-additive effects of genetic variation on complex phenotypic traits [9]. Proteome-wide PPI networks of varying degrees of completeness have been mapped for several viral, bacterial, and eukaryotic species, as well as for interspecies interactions [10-13]. Two popular techniques for high-throughput interactome mapping are yeast two hybrid (Y2H) and affinity purification followed by mass spectrometry (AP-MS).

The original Y2H system operates on the principle that the DNA-binding domain (BD) and transcription-activating domain (AD) of the yeast transcription activator GAL4 function independently of each other and can be expressed separately as fusion constructs with bait and prey proteins. When BD and AD are brought together via bait-prey interaction, the reconstituted transcription activator recovers the ability to activate transcription of a reporter gene [14]. Since its introduction by Fields and Song, modified versions of Y2H have been developed to overcome some of the limitations of the original assay [15]. One such example is the repressed transactivator system for bait proteins that can transactivate reporter genes on their own, regardless of bait-prey interaction [16]. Here, the prey is fused to the repression domain of a transcription repressor, such that bait-prey interaction would instead lead to repression of reporter gene transcription. Another example is the RAS recruitment system for bait-prey interactions occurring outside the nucleus, where transcription happens in eukaryotes [17]. Here, cytosolic bait is fused to constitutively active

RAS, such that should the bait interact with a membrane-anchored prey, the bait-RAS fusion construct would be recruited to the plasma membrane where it would activate MAPK signalling, thereby replacing transcriptional activation as the readout. Other modified versions of Y2H use yeast ectopically overexpressing protein-modifying enzymes or bait fused to cognate modifying enzymes to screen for interactions that are dependent on post-translational modifications [18, 19]. Despite its ease of automation, scalability, and variants of Y2H having greatly improved coverage of the cellular proteome, some limitations of the Y2H system remain. A common cause of false positives is expression of bait and prey which, under normal circumstances, are not co-expressed in the same subcellular compartment at the same time [20]. Meanwhile, a common cause of false negatives is the lack of certain post-translational modifications in yeast, which may be required for protein folding and interaction in higher eukaryotes [21, 22]. For these reasons, mammalian two-hybrid systems have been developed to recapitulate the native cellular environment for mammalian proteins and verify interactions detected by Y2H [23].

AP-MS involves purification, digestion, and ionization of tagged bait proteins bound to interacting preys, followed by sequencing of peptides [24]. AP-MS offers several advantages over Y2H, such as: (1) interactions can be screened in native organisms and cell types; (2) post-translational modifications and cofactors which may be required for protein complex formation are preserved; and (3) quantitative MS allows identification of dynamic changes in the composition, stoichiometry, and post-translational modification of protein complexes in response to stimuli [25]. Limitations of AP-MS include: (1) the purification step may disrupt weak or transient interactions, or interactions involving membrane proteins, which can be stabilized to some extent with chemical crosslinkers [26]; and (2) distinct complexes with overlapping members may be arbitrarily clustered into a single complex, rather than resolved into biologically meaningful protein

assemblies [27]. Since AP-MS does not typically distinguish between direct and indirect bait-prey interactions, interactions detected by AP-MS should be verified by high-confidence biochemical assays such as coimmunoprecipitation, or biophysical assays based on fluorescence transfer or complementation [28]. An advantage of fluorescence-based methods is the ability to visualize the subcellular distribution of protein complexes. In addition, results of Y2H and AP-MS can also be scrutinized by bioinformatics analyses that integrate information from heterogeneous datasets such as co-expression [29], gene ontology annotation [30], and homologous interactions identified in other species [31]. Comparison of Y2H and AP-MS data has shown that while both are of equally high quality, the interaction networks produced are complementary in nature, with the binary network being enriched for transient and inter-complex interactions [32].

Structurally-resolved protein interaction networks. High-throughput Y2H data have played a crucial part in mapping the protein interactomes of several organisms, including yeast, worm, fly and human [33-36]. For instance, the Human Interactome Project used Y2H for unbiased screening of all pairwise combinations of 15,517 human open-reading frames (ORFeome), which span half of the interactome space, and identified 14,000 binary PPIs [37]. The authors found that in contrast to literature-curated PPIs, where frequently studied proteins tend to occupy a dense zone while poorly studied proteins are relegated to a sparse zone, PPIs identified via unbiased screening of the human ORFeome are distributed evenly over the interactome space, regardless of the "popularity" of a protein among investigators. Furthermore, the PPIs are significantly enriched for co-localized proteins and kinase-substrate pairs, and are also more susceptible to perturbation by missense disease mutations than by non-disease mutations. The latter finding is a motivation for edgetic perturbation, where genetic mutations are introduced in an unbiased fashion, and the loss, retention, or gain of interaction between the mutant protein and all other proteins, *i.e.* "edgotype",

is captured and serves as a molecular proxy for disease outcomes of genetic variation [38]. A subsequent proof-of-concept study systematically compared the edgotype profiles of disease mutations and common genetic variants, and found that disease mutations are more likely to destabilize protein structure and perturb protein-protein and protein-DNA interactions [39]. Moreover, different mutations on the same gene often lead to distinct perturbation profiles, which are broadly classified as: (1) quasi-null, where all PPIs involving the mutant protein are lost; (2) edgetic, where only a subset of PPIs are lost; and (3) quasi-wild-type, where all PPIs involving the mutant protein are retained. By comparing the edgotype profiles of mutant and wild-type proteins, the study provides an experimental framework for assessing the pathogenic potential of genetic variants identified in genome-wide association studies.

The phenomenon of edgetic perturbation, *i.e.* PPI perturbation is interface-dependent, highlights the need for incorporating protein structural information into interactome analysis pipelines [40, 41]. As experimental determination of protein structure can be time-consuming and resource-intensive, template-based modelling is an efficient alternative for structural annotation of protein complexes, based on alignment to previously solved 3D structures in the Protein Data Bank (PDB). Homology modelling is a popular template-based method that uses sequence alignment to map residues of a target protein with unknown structure onto residues of a template protein with known structure. It is based on the observation that tertiary structure is more conserved than amino acid sequence, such that even homologues having as low as 30% sequence identity can have similar structures [42]. In practice, homology modelling begins with a BLAST search to identify template structures that are homologous to the target. To improve robustness of the target-template alignment, target and template sequences can be converted to frequency profiles, in the form of multiple sequence alignments with their respective homologues, which can capture structural

features conserved among remote homologues within the same family [43, 44]. Target and template profiles are then aligned, and the resulting profile-profile alignment is used to build the homology model [45, 46]. In addition to atomic coordinates, domain assignment can also be transferred from template to target, which increases coverage of the proteome space, albeit at the cost of reduced resolution [47-49]. Overall, homology modelling is a reliable technique that can provide qualitative insight into the physical and evolutionary properties of residues, such as solvent accessibility and degree of conservation, which are useful in predicting functional sites [50]. A major limitation of homology modelling, however, is its dependency on homologous structures and sufficient sequence identity between target and template. Gaps in the alignment and template structure, such as in flexible loop regions, as well as low sequence identity between target and template, can seriously compromise the quality of homology models [51]. An alternative template-based modelling method is threading, which can be used for predicting the structure of proteins without close homologues at the sequence level. Threading is based on the premise that there are a limited number of unique protein folds, and that secondary structure is more conserved than amino acid sequence [52, 53]. Unlike homology modelling, which only uses sequence alignment for structure prediction, threading leverages structural information in solved structures (*e.g.* interactions among adjacent residues, local secondary structure, solvent accessibility, *etc.*) to predict how well a fold fits a particular sequence [54, 55].

Applications of structural interaction networks in systems biology. Atomic and domain-resolution homology models have been used to resolve both intra- and inter-species PPIs [1, 4, 56]. Structurally-resolved PPI networks have proved useful for exploring the mechanisms of genetic diseases as well as molecular recognition principles governing host-pathogen interactions. For instance, Wang *et al.* built a domain-resolution PPI network involving disease proteins and

found that mutations in different PPI-mediating domains of the same protein tend to be associated with clinically distinct disorders [57]. Using PDB structures and atomic-resolution homology models of within-human and human-virus PPIs, Franzosa and Xia discovered extensive overlap between the endogenous (within-human) and exogenous (human-virus) PPI interfaces on human proteins that have both viral and human binding partners [1]. Importantly, while two human proteins occupying the same interface tend to be structurally similar, a viral protein and a human protein occupying the same interface tend to be structurally distinct. This finding implies that sharing of endogenous interface among human proteins is largely a consequence of divergent evolution through gene duplication, whereas mimicry of endogenous interface by viral proteins tends to involve convergent evolution of unique host-interacting modules in viruses. Moreover, compared to exclusively endogenous interfaces, overlapping endogenous-exogenous interfaces tend to be used transiently by multiple human binding partners and evolve faster. A follow-up study expanded the coverage of the human-virus structural interaction network and confirmed, at the domain level, a lack of structural similarity between viral and human proteins binding to the same domain on a common human target [4]. Moreover, human-virus PPIs tend to involve viral proteins containing multiple, unique short linear motifs and human proteins containing linear motif-binding domains. The authors conclude that the economical and pleiotropic nature of domain-motif interactions helps virus overcome genome size constraints. In other words, by encoding highly degenerate motifs that are recognizable by multiple host domains, a few viral proteins can hijack many host pathways at once. These studies demonstrate that pathogens can induce massive rewiring of host interactome networks, similar to perturbation of the yeast genetic interaction network by DNA-damaging agents [58], and rewiring of the human protein interaction network by cancer mutations [59].

In addition to uncovering the molecular mechanisms of human-virus PPIs, structural characterization of pathogen proteins has also yielded insight into the function and evolution of bacterial effector proteins, which are virulence factors secreted directly into host cells, where they interact extensively with host proteins. Current literature contains many case studies of bacterial effectors targeting host domains involved in host-endogenous PPIs via eukaryotic-like domains or bacteria-exclusive domains. For instance, *Ralstonia solanacearum* has acquired a host-like F-box domain (PF00646) that competes with host-endogenous F-box protein for binding to SKP1, thus hijacking the ubiquitin-proteasome pathway in *Arabidopsis thaliana* [60], while *Shigella flexneri* has convergently evolved a GEF domain (PF03278) that competes with host Rho GEF, thus activating the Rho GTPase signalling pathway in humans [61]. In addition to mechanistic studies of individual host-targeting domains in bacteria, databases of eukaryotic-like domains and short linear motifs provide a snapshot of the extent to which bacterial pathogens mimic host structural modules. For instance, EffectiveDB, a database for predicting bacterial effectors based on several criteria including the presence of eukaryotic-like domains, currently reports 2,636 eukaryotic-like domains as being significantly enriched (Z-score ≥ 4) in the genomes of pathogenic *vs.* non-pathogenic bacteria [62]. Meanwhile, the Eukaryotic Linear Motif Resource currently contains ~100 instances of bacteria-mimicked eukaryotic short linear motifs from a small number of extensively studied pathogenic species [63]. Recent studies have demonstrated the usefulness of joint analysis of host-bacteria and within-bacteria PPI networks. For instance, Crua Asensio *et al.* found that bacterial virulence factors are fundamentally different from other bacterial proteins, in terms of their centrality in the host-bacteria *vs.* within-bacteria PPI network [64]. The authors show that bacterial proteins which are highly connected in the host-bacteria PPI network tend to be sparsely connected in the within-bacteria PPI network. Furthermore, deletion of virulence proteins

has a negative impact on pathogen fitness inside the host, but not outside the host. In other words, essentiality of pathogen proteins is context-dependent: proteins which increase pathogen infectivity in the host are dispensable for pathogen growth outside the host, and vice versa. A follow-up study compared within-eukaryote, within-bacteria and eukaryote-bacteria protein complexes at the level of residue interactions, and found that bacteria use imperfect mimicry of eukaryotic interfaces to maximize interactions with host proteins while minimizing interactions with other bacterial proteins, thereby maximizing infectivity and minimizing self-toxicity [65].

Owing to their ability to perturb host interactome networks, pathogens have found diverse applications in clinical research, ranging from etiologic studies to the development of oncolytic therapies. For instance, Gulbahce *et al.* found that proteins associated with virally-implicated diseases are either directly targeted by virus or are transcriptionally regulated by viral targets, as evidenced by their differential expression in virally-implicated disease tissues relative to healthy tissues [66]. A follow-up study confirmed that oncoviral infections and oncogenic mutations cause similar perturbations to the human interactome at the whole-protein level, suggesting that screening for oncoviral targets can complement functional genomics approaches and improve the specificity of cancer gene identification [67]. Oncolytic viro- and bacteriotherapy present alternative modes of cancer treatment, with some virotherapies having entered late phase clinical trials [68-71]. There are several advantages to virotherapy: (1) viruses naturally prefer to replicate in intra-tumoral environment, where overworked host machineries produce abundant nutrients; (2) viruses simultaneously impinge on multiple cellular processes, thereby overcoming drug-resistance by polymorphic subpopulations of cancer cells; and (3) viruses can redirect host immune response from eradicating viral infection to eliminating cancer, and alert distant immune cells to

the presence of cancer. Hence, virotherapy may have improved efficacy and safety profiles over conventional radio- and chemotherapy [72].

Chapter 3: Convergent perturbation of the human domain-resolved interactome by viruses and mutations inducing similar disease phenotypes

PLoS Computational Biology 15(2): e1006762 (2019)

Yangchun Frank Chen[1] and Yu Xia[1]

[1]Department of Bioengineering, McGill University, Montreal QC, Canada

3.1 Abstract

An important goal of systems medicine is to study disease in the context of genetic and environmental perturbations to the human interactome network. For diseases with both genetic and infectious contributors, a key postulate is that similar perturbations of the human interactome by either disease mutations or pathogens can have similar disease consequences. This postulate has so far only been tested for a few viral species at the level of whole proteins. Here, we expand the scope of viral species examined, and test this postulate more rigorously at the higher resolution of protein domains. Focusing on diseases with both genetic and viral contributors, we found significant convergent perturbation of the human domain-resolved interactome by endogenous genetic mutations and exogenous viral proteins inducing similar disease phenotypes. Pan-cancer, pan-oncovirus analysis further revealed that domains of human oncoproteins either physically targeted or structurally mimicked by oncoviruses are enriched for cancer driver rather than passenger mutations, suggesting convergent targeting of cancer driver pathways by diverse oncoviruses. Our study provides a framework for high-resolution, network-based comparison of various disease factors, both genetic and environmental, in terms of their impacts on the human interactome.

3.2 Author summary

Cellular function and behaviour are driven by highly coordinated biomolecular interaction networks. A prime example is the protein-protein interaction network, often simply referred to as the “interactome”. Recent advances in systems biology have spawned the view of human disease as a manifestation of genetic and environmental perturbations to the human interactome, a key postulate being that similar perturbation patterns lead to similar disease phenotypes. Here, we took a structural systems biology approach to compare mutation-induced and virus-induced perturbations of the human interactome in diseases with both genetic and viral contributors. Specifically, we constructed a domain-resolved human-virus protein interactome and characterized the distribution of genetic disease mutations with respect to human domains either physically targeted or structurally mimicked by virus. Overall, we found significant convergent perturbation of the human domain-resolved interactome by viruses and mutations inducing similar disease phenotypes. Structure-guided, integrated analysis of host genetic variation and host-pathogen protein interaction data may help elucidate the molecular mechanisms of infection and reveal its connections to genetic diseases such as cancer, autoimmunity, and neurodegeneration. On a broader note, our finding implies that similar perturbations of the human interactome at the domain level can have similar phenotypic consequences, regardless of the source of perturbation.

3.3 Introduction

Cellular function and behaviour are driven by highly coordinated biomolecular interaction networks. A prime example is the protein-protein interaction (PPI) network, also known as the protein "interactome" or interactome for short. A central focus of disease systems biology is to use interactome networks to study genotype-phenotype relationships in complex diseases [1]. The idea of using interactome networks to infer gene function and gene-disease association comes from the well-validated principle of "guilt by association", which states that physically interacting proteins tend to share similar functions and, by extension, tend to be involved in similar disease processes [1-4]. Recent advances in systems biology have spawned the view of human disease as a manifestation of genetic and environmental perturbations to the human interactome, a key postulate being that similar perturbation patterns lead to similar disease phenotypes [5-8]. A corollary is that, for diseases with both genetic and infectious contributors, similar perturbations of the human interactome by either disease mutations or pathogens can have similar disease consequences. This corollary has been tested for several viral species at the level of whole proteins [9, 10]. For example, Gulbahce *et al.* used yeast two-hybrid screens to map binary interactions between Epstein-Barr virus (EBV) and human papillomavirus (HPV) proteins and human proteins, and transcriptionally profiled human cell lines exogenously expressing HPV oncoproteins E6 and E7 [9]. They found that human genes associated with EBV- and HPV-implicated genetic diseases were often either directly targeted by the virus or transcriptionally regulated by viral targets. This finding led to the idea that oncoviral proteins may preferentially target host proto-oncogenes and tumour suppressors, which was experimentally validated in four families of DNA oncoviruses [10].

Despite insights from these studies on the etiology of virally-implicated genetic diseases, there has yet to be a systematic, structure-based comparison of mutation-induced and pathogen-induced perturbations of the human interactome. A high-resolution, structurally-resolved network biology approach is important for unravelling complex genotype-phenotype relationships, because mutations occurring in different PPI-mediating interfaces on the same protein often have distinct functional impacts and phenotypic consequences [5-8]. In this regard, structural systems biology has proved useful in uncovering evolutionary properties of single- and multi-interface PPI network hubs, systems-level principles governing human-virus interactions, and systems properties of disease variants [6, 11, 12]. For instance, by constructing atomic-resolution human-virus and within-human protein interactomes, Franzosa and Xia discovered that viral proteins tend to target existing endogenous PPI interfaces in the human interactome, rather than creating exogenous interfaces *de novo*, thereby efficiently perturbing multiple endogenous PPIs involved in cell regulation [12]. In a follow-up study, Garamszegi *et al.* expanded the coverage of the human-virus interactome using domain-resolved models of PPIs, and found that viral proteins tend to deploy short linear motifs to bind a variety of human protein domains [13]. The economical and pleiotropic nature of "host domain-viral motif" interactions reflects the efficiency with which viruses rewire the human interactome given limited genomic resources at their disposal. Meanwhile, Wang *et al.* constructed a domain-resolution within-human interactome where protein domains are annotated with disease variant information [6]. They found that mutations occurring in different PPI-mediating domains within the same protein tend to be associated with different disorders ("gene pleiotropy"). By contrast, mutations occurring in the domains of two different but interacting proteins, where the interaction is mediated by said domains, tend to be associated with

the same disorder ("locus heterogeneity"). These studies attest to the utility of structural systems biology in the study of infectious and genetic diseases.

Here, we apply structural systems biology to the study of virally-implicated genetic diseases (VIDs), and rigorously test the postulate that endogenous genetic mutations and exogenous viral proteins give rise to similar disease phenotypes by inducing similar perturbations of the human interactome at the level of protein domains. Specifically, we constructed a domain-resolved human-virus protein interactome and characterized the distribution of genetic disease mutations with respect to human domains targeted by virus. Overall, we found that viral proteins and VID mutations induce similar perturbations of the human domain-resolved interactome, for individual viruses with clearly defined VIDs and sufficient numbers of host-virus PPIs (including EBV, HPV and HIV), for oncoviruses, as well as for all viruses combined. We first analyzed the disease associations of host proteins targeted by viral proteins and confirmed that virus-targeted proteins tend to be causally associated with VIDs rather than non-VIDs. We then analyzed the domain-level distribution of disease mutations in virus-targeted proteins and found that virus-targeted domains are significantly enriched for mutations causing VIDs rather than non-VIDs. Using a pooled analysis of all oncoviruses and all oncomutations, we found oncovirus-targeted domains to be significantly enriched for mutations causing cancer rather than other diseases. Furthermore, domains of oncoproteins either physically targeted or structurally mimicked by oncoviruses are significantly enriched for cancer driver mutations rather than passenger mutations, which implies convergent perturbation of cancer driver pathways by diverse oncoviruses. Finally, we also assessed the extent to which viral proteins and VID mutations perturb the same domain-domain interactions (DDIs) in the human interactome. We found that viruses preferentially target DDI partners of domains harbouring VID mutations, regardless of whether the DDI partners

themselves are susceptible to known disease mutations. By correlating the equivalent pathogenicity of viral proteins and VID mutations with their convergent perturbation of the human domain-resolved interactome, we provide a framework for high-resolution, network-based comparison of the functional impacts of both genetic and environmental disease factors. On a broader note, our finding implies that similar perturbations of the human interactome at the domain level can have similar phenotypic consequences, regardless of the source of perturbation.

3.4 Results

3.4.1 Disease-annotated, domain-resolved human-virus protein interaction network

We first acquired human-endogenous and human-virus binary PPI data from IntAct, HPIDB 3.0, and the HIV-1 Human Interaction Database [14-18]. Only PPIs supported by at least one PubMed ID were included in the whole-protein resolution human-virus interactome, which consists of 173830 PPIs between 15995 human proteins, and 28531 PPIs between 7761 human proteins and 624 viral proteins. 7211 human proteins participate in both endogenous and exogenous PPIs. To build homology models of PPIs, we collected high-confidence domain-domain interaction (DDI) and domain-motif interaction (DMI) templates derived from 3D structures of protein complexes in the Protein Data Bank, and scanned protein sequences for the occurrence of Pfam domains and domain-binding linear motifs [19-23]. Structural models were assigned to each PPI by extracting all DDIs and DMIs possibly mediating the PPI. The resulting domain-resolved human-virus structural interaction network (hvSIN) consists of 61041 PPIs between 11596 human proteins, and 4654 PPIs between 1590 human proteins and 405 viral proteins. 1517 human proteins participate in both endogenous and exogenous portions of hvSIN.

We then obtained manually-curated disease variant data from UniProtKB and ClinVar [24, 25], selecting missense variants located inside Pfam domains for our analyses. Overall, 19047 mutations associated with 5383 diseases were mapped to 3585 domains of 2622 proteins. 14720 mutations associated with 4185 diseases were mapped to 2642 domains of 1864 human proteins in hvSIN. Table 1 lists the number of mutations by the type of domain in which they occur. Incidentally, 1272 domains of 957 human proteins in hvSIN are susceptible to disease mutations, but lack interacting domains or motifs. 850 of these 1272 domains harbour a total of 4154 mutations associated with 1381 diseases that are not accounted for by mutations occurring in PPI-mediating domains in hvSIN. Because the completeness of a domain's PPI profile depends largely

on the interactome search space and availability of 3D structures of protein complexes, and domains often have important biological functions besides mediating PPIs (*e.g.* enzymatic or nucleotide-binding activity), we included all domains of virus-targeted host proteins in a comprehensive analysis of the domain-level distribution of disease mutations.

Table 3.1 Number of disease mutations mapped to human protein domains in the human-virus structural interaction network (hvSIN).

	Proteins	Domains	Disease mutations	Diseases
All disease proteins in hvSIN	1864	2642	14720	4185
Disease proteins containing exclusively endogenously-interacting domains	924	1147	7073	2281
Disease proteins containing exclusively exogenously-interacting domains	9	9	19	15
Disease proteins containing overlapping endogenous-exogenous domains	200	214	1300	583
Disease proteins containing domains without annotated interacting domains or motifs	957	1272	6328	2224

3.4.2 Virus-targeted host domains are enriched for virally-implicated disease mutations

To relate the equivalent pathogenicity of viral proteins and VID mutations to their equivalent perturbation of the host interactome, we first characterized the mutational landscape of human proteins targeted by EBV, HPV and HIV, three viruses with clearly defined VIDs and sufficient numbers of host-virus PPIs. Since most oncoviruses are causally implicated in only a few site-specific malignancies (*e.g.* HBV/HCV in hepatocellular carcinoma, KSHV in Kaposi's sarcoma, and HTLV in adult T-cell lymphoma), and various types of cancer share common molecular hallmarks [26, 27], to increase the statistical power of our analysis and establish whether a general equivalence exists between endogenous and exogenous perturbagens of oncogenic

pathways, we also performed a pooled analysis of host proteins targeted by diverse oncoviruses, by considering all types of cancer as interchangeable diseases, all oncomutations as interchangeable endogenous perturbagens, and all oncoviral proteins as interchangeable exogenous perturbagens. We found that for EBV, HIV, HPV and a broad spectrum of oncoviruses, virus-targeted host proteins tend to be causally associated with VIDs (Fig 1), and virus-targeted host domains tend to harbour mutations causally associated with VIDs (Fig 2). We discuss our findings for each type of virus below. A full list of VIDs and disease-associated proteins for EBV, HPV and HIV can be found in S1 Table.

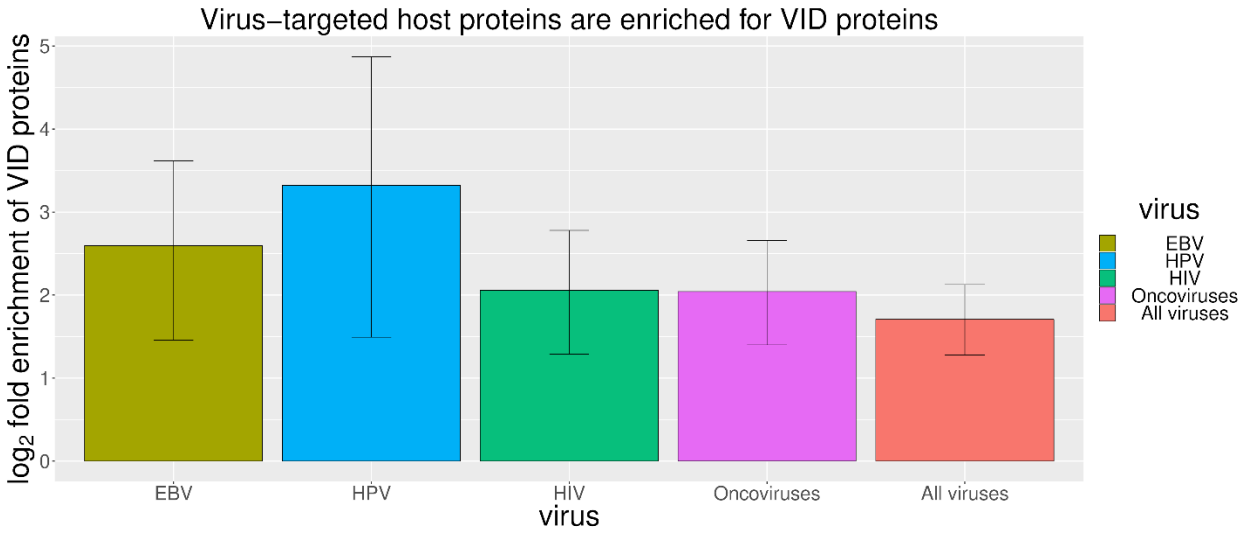


Figure 3.1 Virus-targeted host proteins tend to be causally associated with virally-implicated diseases (VIDs).

"VID proteins" have at least one missense variant that is causally associated with a VID, whereas all missense variants of "non-VID proteins" are exclusively associated with non-VIDs. Literature-curated, virus-specific diseases for EBV, HPV and HIV are listed in S1 Table. For pooled analysis of oncoviruses, VIDs include all types of cancer (Methods). For pooled analysis of all viruses, VIDs include all proliferative and immunological diseases (Methods). Error bars represent 95% confidence intervals.

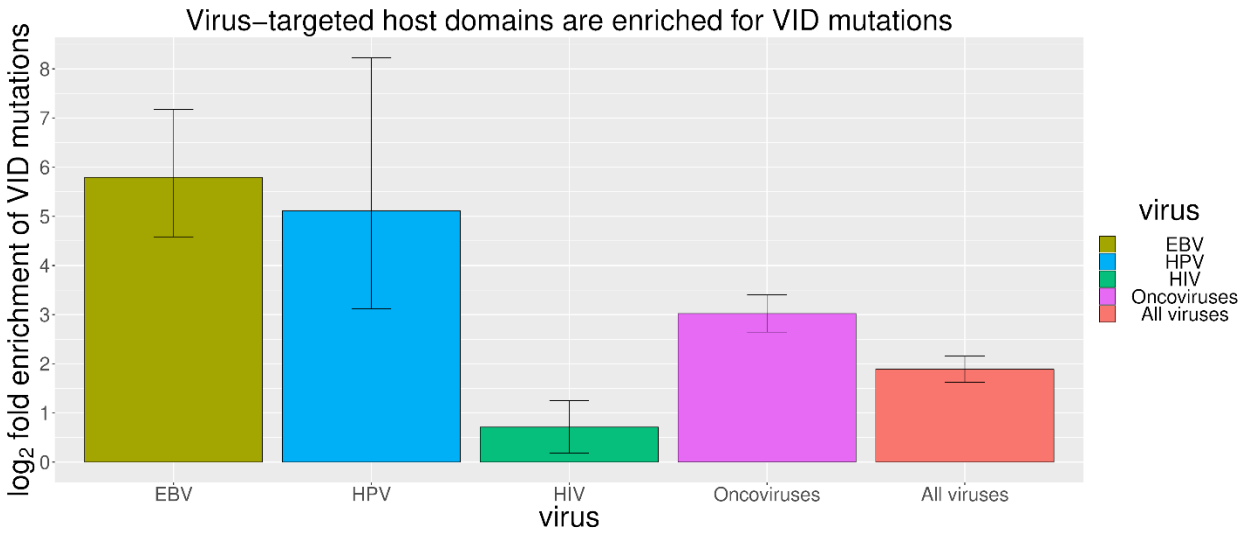


Figure 3.2 Virus-targeted host domains tend to harbour mutations causally associated with virally-implicated diseases (VIDs).

"VID mutations" are causally associated with at least one VID, whereas "non-VID mutations" are exclusively associated with non-VIDs. Error bars represent 95% confidence intervals.

EBV. EBV is involved in lymphomas of the B, T, and NK-cell lineages as well as in adenocarcinomas of epithelial cells [28-32]. EBV hijacks cellular signaling processes by encoding viral homologues of cellular proteins that play key roles in apoptosis and proliferation. Examples include *EBNA2* (mimics Notch signaling), *LMP1* (mimics CD40 receptor signaling), *LMP2* (mimics IgG receptor signaling), *BALF1* and *BHRF1* (homologues of cellular *Bcl-2*), and *BCRF1* (homologue of cellular *IL-10*) [27]. All EBV homologues share at least one PPI partner with their cellular counterparts. Overall, EBV targets 11/99 (11.1%) host proteins associated with EBV diseases, and 51/2523 (2%) host proteins associated with non-EBV diseases, *i.e.* EBV tends to directly target host proteins causally associated with EBV-implicated diseases (Fisher's exact test, two-tailed $P = 1 \times 10^{-5}$) (Fig 1). Analysis of the domain-level distribution of disease mutations found that 35/43 (81.4%) EBV-disease mutations and 62/856 (7.2%) non-EBV disease mutations occur in EBV-targeted domains, suggesting that EBV-targeted domains are significantly enriched for EBV-disease mutations (Fisher's exact test, two-tailed $P < 2.2 \times 10^{-16}$) (Fig 2). Fig 3A shows

the exclusive localization of mutations causing lung cancer, an EBV-implicated disease, in EBV-targeted tyrosine kinase domain (PF07714) of *EGFR* protein, while mutations causing other diseases such as brain cancer are evenly distributed among all domains of *EGFR*.

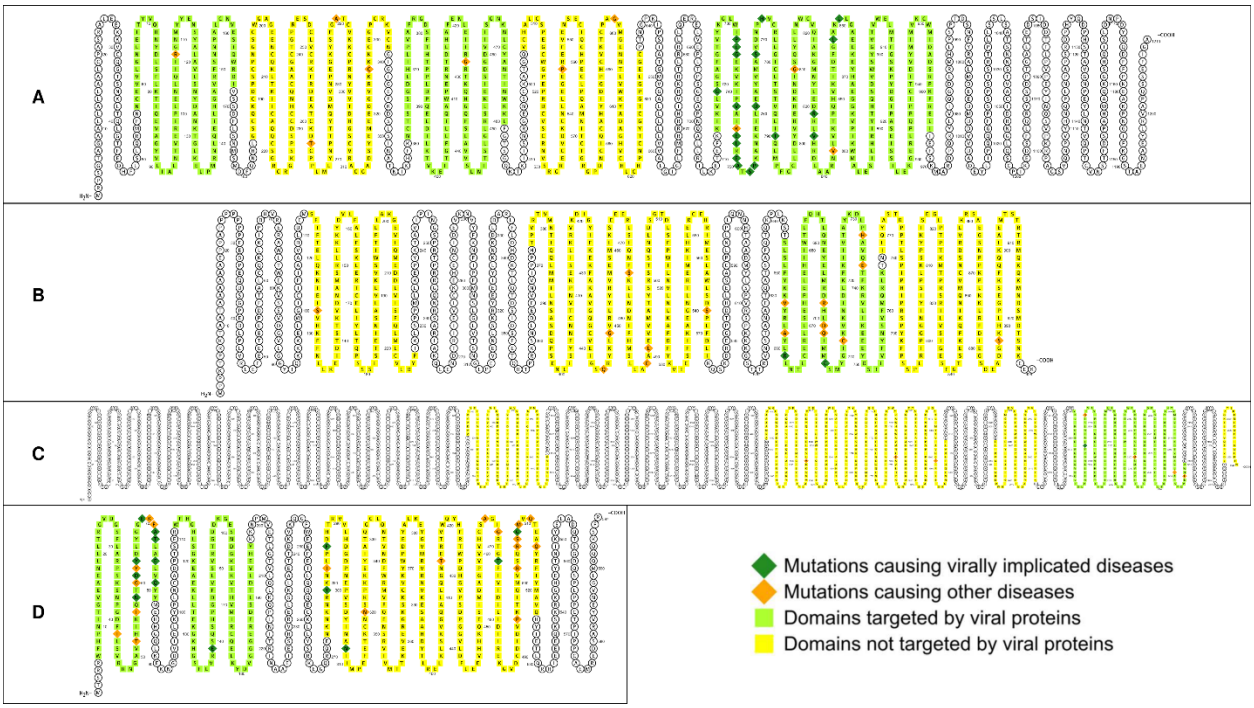


Figure 3.3 Exclusive localization or enrichment of VID mutations in virus-targeted domains.

(A) Exclusive localization of mutations causing lung cancer, an EBV-implicated disease, in EBV-targeted tyrosine kinase domain of *EGFR* protein. (B) Exclusive localization of mutations causing vulvar cancer and lung cancer, both HPV-implicated diseases, in HPV-targeted B domain of *RB* protein. (C) Exclusive localization of mutations causing cervical cancer, an HIV-implicated disease, in HIV-targeted PI3-kinase domain of *MTOR* protein, while mutations causing other diseases such as focal cortical dysplasia and Smith-Kingsmore syndrome are evenly distributed among all domains of *MTOR*. (D) Moderate enrichment of oncomutations in KSHV-targeted SH2 domain of *PTPN11* protein, compared to mutations causing Noonan syndrome. Most of the oncomutations cause juvenile myelomonocytic leukemia, a disease although not caused by KSHV, is mimicked clinically and morphologically by other human herpesvirus infections, including EBV, CMV and HHV-6. VID mutations are shown as dark green diamonds. Non-VID mutations are shown as orange diamonds. Amino acid residues in virus-targeted domains are shown as light green squares. Residues in domains not targeted by virus are shown as yellow squares.

HPV. High-risk human papillomaviruses (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68), as defined by the Centers for Disease Control and Prevention (CDC) and International Agency for Research on Cancer (IARC), are established etiological agents for cervical, oropharyngeal and

anogenital cancers [33-35]. Several studies have also reported an association between HPV and cancers of the bladder [36], breast [37], lung [38], and prostate [39]. Overall, HPV targets 5/79 (6.3%) host proteins associated with HPV diseases, and 17/2543 (0.7%) host proteins associated with non-HPV diseases, *i.e.* HPV tends to directly target host proteins causally associated with HPV-implicated diseases (Fisher's exact test, two-tailed $P = 3 \times 10^{-4}$) (Fig 1). Analysis of the domain-level distribution of disease mutations found that 117/119 (98.3%) HPV-disease mutations and 94/150 (62.7%) non-HPV disease mutations occur in HPV-targeted domains, suggesting that HPV-targeted domains are significantly enriched for HPV-disease mutations (Fisher's exact test, two-tailed $P = 2 \times 10^{-14}$) (Fig 2). Fig 3B shows the exclusive localization of mutations causing vulvar cancer and lung cancer, both HPV-implicated diseases, in HPV-targeted B domain (PF01857) of *RB* protein, while mutations causing other diseases such as retinoblastoma are evenly distributed among all domains of *RB*.

HIV. HIV substantially raises the risk of Kaposi's sarcoma, non-Hodgkin's lymphoma and cervical cancer [40], as well as cancers of the anus, liver, lung, oropharynx and testes [41]. Although HIV-encoded accessory proteins such as *Tat* and *Nef* have demonstrated oncogenic properties on their own [42-44], HIV-associated cancers are mostly attributed to opportunistic infections with oncoviruses such as KSHV, EBV, HPV, and Hepatitis B/C virus. In addition, other HIV-associated complications such as cardiomyopathy and neurocognitive disorders have become increasingly common in the post-HAART era [45-50]. Overall, HIV targets 23/132 (17.4%) host proteins associated with HIV diseases, and 120/2490 (4.8%) host proteins associated with non-HIV diseases, *i.e.* HIV tends to directly target host proteins causally associated with HIV-implicated diseases (Fisher's exact test, two-tailed $P = 3 \times 10^{-7}$) (Fig 1). Analysis of the domain-level distribution of disease mutations found that 103/158 (65.2%) HIV-disease mutations and

479/898 (53.3%) non-HIV disease mutations occur in HIV-targeted domains, suggesting that HIV-targeted domains are significantly enriched for HIV-disease mutations (Fisher's exact test, two-tailed $P = 7 \times 10^{-3}$) (Fig 2). Fig 3C shows the exclusive localization of mutations causing cervical cancer, an HIV-implicated disease, in HIV-targeted PI3-kinase domain (PF00454) of *MTOR* protein, while mutations causing other diseases such as focal cortical dysplasia and Smith-Kingsmore syndrome are evenly distributed among all domains of *MTOR*. In addition to offering general insights on human-HIV interaction, our domain-resolved PPI models also provide useful information about specific HIV proteins. For instance, our model for the interaction between human *Akt1* and HIV *Nef* involves the protein kinase domain (PF00069) of *Akt1* and a region of *Nef* matching three overlapping motifs: MOD_NEK2_1 (residues 100-105), DOC_MAPK_gen_1 (residues 105-112) and DOC_MAPK_MEF2A_6 (residues 105-114). Notably, our predicted *Akt1*-binding region of *Nef* (residues 100-114) is consistent with the experimentally determined *Akt1*-binding region of *Nef* (residues 55-210) [51]. hvSIN also reveals a previously unreported similarity between the host interaction profiles of HIV *Nef* and the EBV oncoprotein *LMP2*, in that both can bind the SH2 domain (PF00017) of *Src* family kinases (*Lck*, *Lyn*, *Src*) and *Syk* family kinases (*Syk*, *ZAP70*), as well as the WW domain (PF00397) of the *Nedd4* family of E3 ubiquitin ligases (*Itch*, *Nedd4*), possibly revealing disease modules perturbed in common by HIV and EBV in AIDS-related lymphoma [52, 53].

Oncoviruses. Oncoviruses contribute to 12% of human cancers worldwide and can activate in a cancer cell the same molecular hallmarks shared among cancers of non-viral origin [27, 54]. In fact, some of the most potent oncogenes were first discovered in retroviruses [55]. Oncoviruses in hvSIN include human herpesviruses (HHV-4/EBV, HHV-5/CMV, HHV-8/KSHV), high-risk HPVs, human polyomaviruses (BKV, JCV, MCV), hepatitis B and C viruses, human T cell

lymphotropic virus (HTLV) and oncogenic retroviruses. Some oncoviruses, although not directly infectious to human, are tumorigenic in other species, can transform human cells *in vitro*, and serve as models for studying viral oncogenesis in human (*e.g.* murid herpesvirus 4) [56, 57]. Despite HIV being classified by IARC as a Group 1 carcinogen and the *in vitro* oncogenicity of HIV-encoded accessory proteins, we excluded it from the pooled analysis of oncoviruses, because there is insufficient data on HIV prevalence and cancer incidence among HIV-infected individuals to accurately assess the independent contribution of HIV to infection-attributable cancers [58]. Pooled analysis of all oncovirus-targeted host proteins found that oncoviruses target 34/194 (17.5%) oncoproteins and 119/2428 (4.9%) proteins associated with non-cancer diseases, *i.e.* oncoviruses tend to directly target oncoproteins (Fisher's exact test, two-tailed $P = 1 \times 10^{-9}$) (Fig 1). Analysis of the domain-level distribution of disease mutations found that 314/413 (76%) oncomutations and 371/1322 (28.1%) other disease mutations occur in oncovirus-targeted domains (OVTDs), *i.e.* the odds of finding cancer-causing over other disease-causing mutations in OVTDs is 8 times as high as that in non-OVTDs (Fisher's exact test, two-tailed $P < 2.2 \times 10^{-16}$) (Fig 2). Fig 3D shows a moderate enrichment of oncomutations in KSHV-targeted SH2 domain (PF00017) of *PTPN11* protein, compared to mutations causing Noonan syndrome. Most of the oncomutations cause juvenile myelomonocytic leukemia, a disease although not caused by KSHV, is mimicked clinically by other human herpesvirus infections, including EBV, CMV and HHV-6 [59, 60]. Finally, we also assessed the mutational landscape of 107 oncovirus-targeted pleiotropic proteins that are susceptible to both oncomutations and other disease mutations. Overall, 88/113 (77.9%) oncomutations and 110/179 (61.5%) other disease mutations were mapped to the OVTDs of these pleiotropic proteins, suggesting that enrichment of oncomutations in OVTDs holds even

at the level of individual proteins involved in both cancer and other diseases (Fisher's exact test, two-tailed $P = 4 \times 10^{-3}$).

Viruses in proliferative and immunological diseases. All viruses have evolved sophisticated mechanisms to subvert host transcriptional and signaling machineries for replication and persistence. Viruses are known to encode homologues of cellular proteins to mimic mutant oncoproteins (Fig 4A) or antagonize mutant cytokine receptors (Fig 4B). Viruses have also been shown to abuse peptide motifs to modulate host signaling pathways, potentially mimicking the effects of disease-causing mutations (Fig 4C). We suspect that viruses and mutations causing proliferative and immunological diseases (PIDs) target similar human domains involved in cell cycle progression, apoptosis, DNA repair and immune homeostasis. Proliferative diseases include various neoplasms, both benign and malignant. Examples include lung cancer (Fig 3A), vulvar and lung cancer (Fig 3B), cervical cancer (Fig 3C), juvenile myelomonocytic leukemia (Fig 3D), glioblastoma multiforme and non-small-cell lung cancer (Fig 4A), lung cancer, breast cancer and lymphoma (Fig 4C). Immunological diseases include autoimmune diseases, hypersensitivity, and immunodeficiency disorders. One example of an immunological disease, inflammatory bowel disease (IBD), is given in Fig 4B, where we show convergent perturbation of the *IL10*-binding domain of *IL-10R1* by both viral homologues of *IL-10* and IBD mutations.

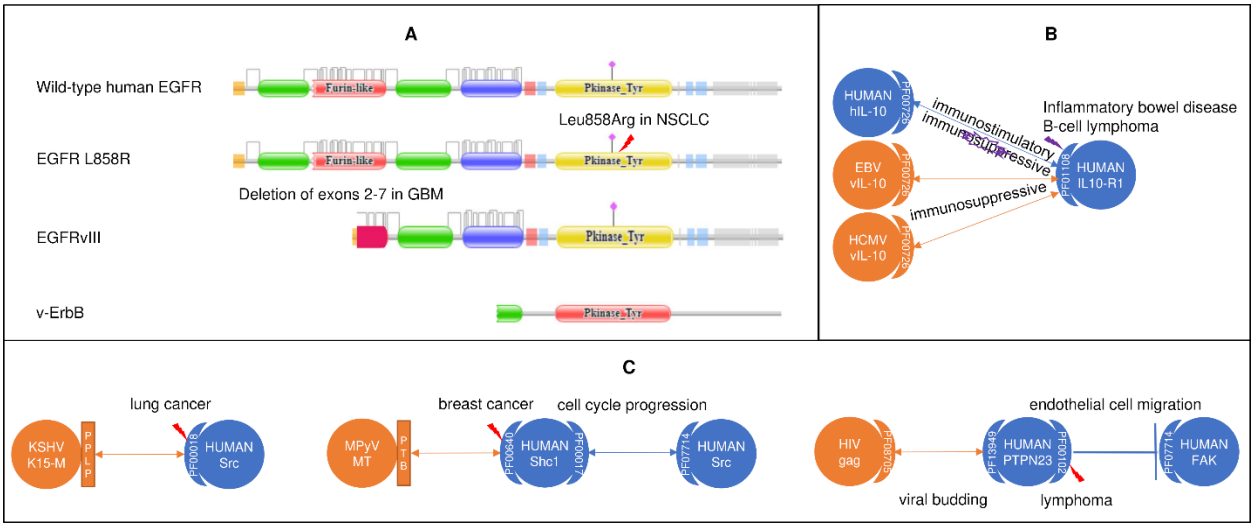


Figure 3.4 Viral and mutational perturbations of host domains are mechanistically similar.

(A) Viruses encode homologues of human proteins to mimic mutations in oncoproteins that cause uncontrolled cell proliferation. Top: EGFRvIII deletion mutation, frequently detected in glioblastoma multiforme (GBM) patients, and *v-ErbB*, encoded by avian leukosis virus, both lack the *EGFR* ligand-binding domain. Meanwhile, an L858R missense mutation in the *EGFR* kinase domain is frequently found in non-small-cell lung cancer (NSCLC). These alterations lead to conformational changes that result in ligand-independent, constitutive kinase activity [61, 62]. (B) Viruses encode homologues of human proteins to antagonize mutations in cytokine receptors that cause hypersensitivity. Human *IL-10* functions both as an immunosuppressant in the inhibition of proinflammatory cytokines, and as an immunostimulant in the induction of MHC II expression on B cells. Mutations in the *IL10*-binding domain of *IL-10R1* abrogate *hIL10*-induced phosphorylation, leading to loss of immunosuppression and inflammatory bowel disease [63]. In contrast, viral *IL-10* homologues encoded by Epstein-Barr virus (EBV) and human cytomegalovirus (HCMV) retain and amplify the immunosuppressive properties of *hIL-10*, thus facilitating viral persistence after lytic infection [64]. *ebvIL-10* selectively retains only the immunosuppressive properties of *hIL-10*. *cmvIL-10* binds with greater affinity to *IL-10R1* than *hIL-10*, while co-opting other *IL10*-associated pathways to amplify the immunosuppressive properties of *hIL-10*. Interestingly, transgenic expression of v*IL-10* has been tested in animal models as an immunosuppressant option for transplant recipients [65]. In addition, abnormal expression levels of *IL-10*, *IL10-R1* and *IL10-R2* has been suggested as a mechanism for diffuse large B-cell lymphoma, a disease with clear EBV involvement [66]. (C) Viruses abuse peptide motifs to modulate host signaling pathways, potentially mimicking the effects of disease-causing mutations. Left: Kaposi's sarcoma-associated herpesvirus (KSHV) protein *K15-M* uses a "PPLP" motif to bind the SH3 domain (PF00018) of *Src* [67], which possibly induces conformational opening of the *Src* kinase domain, thereby mimicking activating mutations such as Y527F [68]. Interestingly, a W121C mutation in the KSHV-targeted SH3 domain of *Src* has been identified in lung cancer [69]. Middle: Murine polyomavirus (MPyV) Middle T antigen (*MT*) uses a tyrosine-phosphorylated motif to recruit host *Shc1*, thereby promoting cell cycle progression [70]. Interestingly, a R175Q mutation in the MPyV-targeted PTB domain (PF00640) of *Shc1* has been found to regulate tumorigenesis in mouse models of breast cancer [71]. Right: HIV protein *gag* uses the late-budding domain to sequester host *PTPN23* and facilitate viral budding [72]. The

phosphatase domain (PF00102) of *PTPN23* regulates cell migration via dephosphorylation of *FAK* and is often mutated in cancer and developmental disorders [73, 74].

To establish whether a general equivalence exists between endogenous and exogenous perturbagens of pathways associated with PIDs, we performed a pooled analysis of all virus-targeted host proteins by considering all PIDs as a unique category of diseases with both genetic and viral contributors, all PID mutations as interchangeable endogenous perturbagens, and all viral proteins as interchangeable exogenous perturbagens. We found that overall, viruses tend to target host proteins associated with PIDs (85/338, 25.1%) rather than non-PIDs (213/2284, 9.3%) (Fisher's exact test odds ratio = 3.3, two-tailed $P = 1 \times 10^{-14}$) (Fig 1), and virus-targeted domains are enriched for mutations causing PIDs (525/737, 71.2%) rather than non-PIDs (803/2003, 40%) (Fisher's exact test odds ratio = 3.7, two-tailed $P < 2.2 \times 10^{-16}$) (Fig 2). Since the equivalence between oncoviruses and oncomutations has already been established in the previous section, we excluded proliferative diseases from consideration and further tested the equivalence between viral proteins and mutations in causing immunological diseases. Again, we found that viruses tend to target host proteins associated with immunological diseases (31/151, 20.5%) rather than other diseases (267/2471, 10.8%) (Fisher's exact test odds ratio = 2.1, two-tailed $P = 8 \times 10^{-4}$), and virus-targeted domains are enriched for mutations causing immunological diseases (101/179, 56.4%) rather than other diseases (1227/2561, 47.9%) (Fisher's exact test odds ratio = 1.4, two-tailed P = 0.03). Finally, we tested the equivalence between viral proteins and mutations in causing proliferative, but not immunological diseases. Overall, viruses tend to target host proteins associated with proliferative diseases (56/199, 28.1%) rather than other diseases (242/2423, 10%) (Fisher's exact test odds ratio = 3.5, two-tailed $P = 8 \times 10^{-12}$), and virus-targeted domains are enriched for mutations causing proliferative diseases (431/571, 75.5%) rather than other diseases (897/2169, 41.4%) (Fisher's exact test odds ratio = 4.4, two-tailed $P < 2.2 \times 10^{-16}$).

3.4.3 Oncovirus-targeted host domains are enriched for cancer driver mutations

A main challenge in cancer research is to distinguish mutations which confer clonal growth advantage (*i.e.* drivers), from mutations that do not cause clonal expansion (*i.e.* passengers) [75]. Large-scale cancer genome sequencing projects have enabled systematic identification of cancer driver proteins and mutations [76]. Rozenblatt-Rosen *et al.* previously constructed an oncovirus-human interactome and demonstrated, at the whole-protein level, comparability between oncoviral perturbation and conventional functional genomics approaches to cancer gene discovery [10]. However, by representing proteins and PPIs as generic nodes and edges, their approach is not sensitive enough to distinguish driver mutations from passenger mutations occurring in the same oncoprotein. As we demonstrated earlier in the case of pleiotropic oncoproteins, the oncogenicity or "driver-ness" of a mutation is often correlated with its occurrence in oncovirus-targeted domains (OVTDs).

To confirm that oncoviruses can help identify driver proteins, we first cross-classified human proteins in hvSIN by whether they are oncoviral targets, and whether they are curated by the Cancer Gene Census (CGC) as being causally implicated in cancer, *i.e.* driver proteins [76]. Out of 727 oncoviral targets, 93 (12.8%) are in CGC, whereas out of 10897 remaining human proteins in hvSIN, 514 (4.7%) are in CGC. In other words, there is a 3-fold enrichment of driver proteins among oncoviral targets (Fisher's exact test, two-tailed $P = 3 \times 10^{-16}$) (Fig 5A). Next, to find out if oncoviruses can also help identify driver mutations, we cross-classified mutations in oncoproteins by whether they are drivers or passengers, and by whether they map to OVTDs. Oncogenic and resistance mutations with a ClinVar clinical significance value of "pathogenic" or "likely pathogenic" are considered drivers, while passengers include all other missense mutations in oncoproteins that are catalogued by ClinVar and COSMIC. Out of 194 oncoproteins with annotated driver mutations, we identified 30 oncoproteins as having at least one OVTD. Pooled

analysis of all 30 oncoproteins mapped 340/398 (85.4%) driver mutations and 3673/7177 (51.2%) passenger mutations to OVTDs. In other words, the odds of finding a driver mutation in OVTDs is 5 times as high as that in non-OVTDs (Fisher's exact test, two-tailed $P < 2.2 \times 10^{-16}$) (Fig 5B). Closer inspection identified 19 candidates for focused investigations into the common basis of viral and mutational oncogenesis (Table 2): (I) 7 oncoproteins where all domains are OVTDs, and the driver:passenger ratio is higher than the average ratio across all oncoproteins; (II) 8 oncoproteins where some domains are OVTDs, and driver mutations are exclusively found in OVTDs; and (III) 4 oncoproteins where some domains are OVTDs, and driver mutations are significantly enriched in OVTDs (Fisher's exact test, two-tailed $P < 0.05$). An example of each type of candidate is given in Fig 6.

Table 3.2 Oncoproteins having at least one oncovirus-targeted domain (OVTD), where driver mutations are either exclusively found or enriched.

Type	Oncoprotein	OVTD
I	*BAX*	PF00452
I	*BCL10*	PF00619
I	*CDKN1B*	PF02234
I	*CHEK2*	PF00069; PF00498
I	*IRF1*	PF00605
I	*MAP2K2*	PF00069
I	*PPP2R1A*	PF02985; PF13646
II	*ABL1*	PF00017; PF00018; PF07714
II	*FBXW7*	PF00400
II	*FGFR4*	PF07714
II	*PDGFRA*	PF07714
II	*RAF1*	PF00130; PF07714
II	*RXRA*	PF00104
II	*SPOP*	PF00917
II	*TGFBR2*	PF07714
III	*AR*	PF00104
III	*ATM*	PF00454
III	*RB1*	PF01857
III	*TP53*	PF00870

Type I: All domains are OVTDs, and the driver:passenger ratio is higher than the average ratio across all oncoproteins. Type II: Some domains are OVTDs, and driver mutations are exclusively found in OVTDs. Type III: Some domains are OVTDs, and driver mutations are significantly enriched in OVTDs (Fisher's exact test, two-tailed $P < 0.05$).

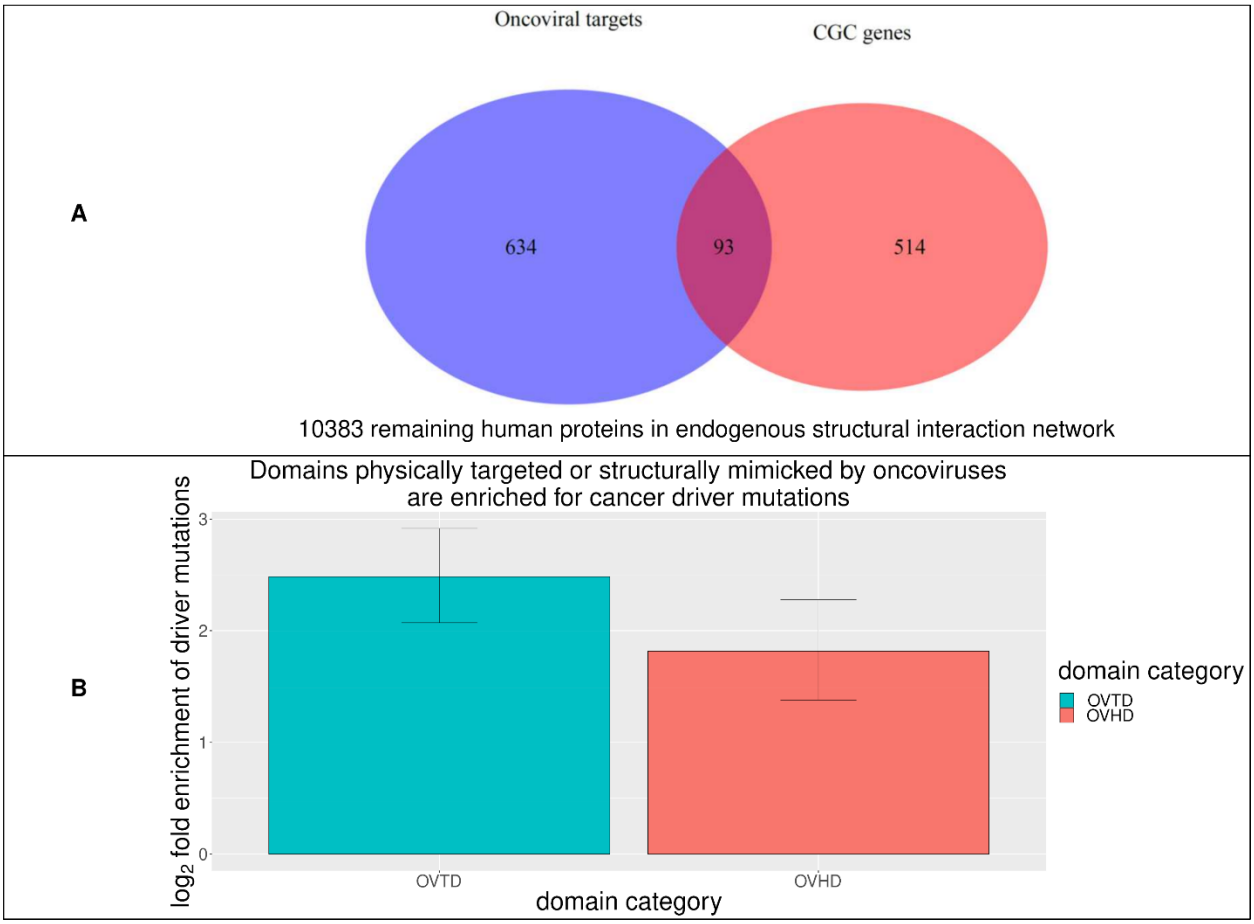


Figure 3.5 Oncovirus-targeted proteins are enriched for driver proteins, and oncovirus-targeted or mimicked domains are enriched for driver mutations.

(A) There is a 3-fold enrichment of Cancer Gene Census proteins in oncovirus-targeted proteins. (B) There are 5-fold and 3-fold enrichments of driver mutations in oncovirus-targeted domains (OVTDs) and oncoviral homology domains (OVHDs), respectively.

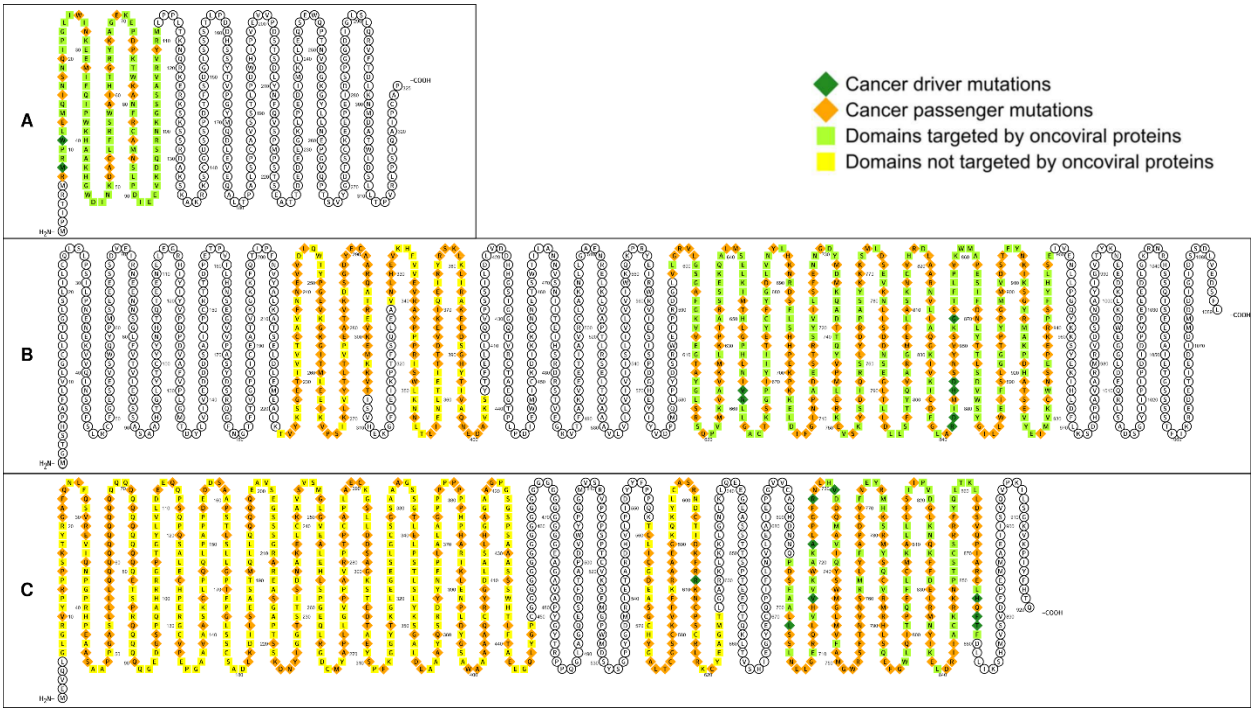


Figure 3.6 Oncoproteins having at least one oncovirus-targeted domain (OVTD), where driver mutations are either exclusively found or enriched.

(A) driver:passenger ratio in oncovirus-targeted PF00605 domain of *IRF1* is higher than the mean driver:passenger ratio for all oncoproteins; (B) driver mutations are exclusively found in oncovirus-targeted PF07714 domain of *PDGFRA*; (C) driver mutations are enriched in oncovirus-targeted PF00104 domain of *AR*.

3.4.4 Oncovirus-mimicked host domains are enriched for cancer driver mutations

Viruses are known to encode structural homologues that mimic host domains in order to modulate the biological activities of host targets. Such viral homology domains (VHDs) play key roles in mediating immune response (*e.g.* PF00048 in CMV and KSHV), apoptosis (*e.g.* PF00452 in EBV and KSHV), cell differentiation (*e.g.* PF07684 in feline leukemia virus), and protein phosphorylation (*e.g.* PF06734 in CMV), among other cellular processes involved in virally-implicated diseases. VHDs often compete with cellular counterparts for interaction partners, thereby rewiring host signaling networks to the virus's advantage. Table 3 lists instances of human proteins convergently targeted by human domains and oncoviral homology domains in hvSIN.

Table 3.3 Human proteins convergently targeted by human domains and oncoviral homology domains (OVHDs) in hvSIN.

Human domain/OVHD	Pfam description	Human proteins convergently targeted by human domain and OVHD
PF00001	7 transmembrane receptor (rhodopsin family)	*CX3CL1*
PF00017	SH2 domain	*CDC37;* ***HSP90AA1****;* ***HSP90AB1****; KHDRBS1;* ***NCKIPSD****;* ***PDGFRB****;* ***RAF1****; WASL*
PF00018	SH3 domain	*CDC37;* ***HSP90AA1****;* ***HSP90AB1****; KHDRBS1;* ***NCKIPSD****;* ***PDGFRB****; PPP2CA;* ***RAF1****; WASL*
PF00084	Sushi repeat (SCR repeat)	
PF00134	Cyclin, N-terminal domain	*CCT8; CDK2; CDK3;* ***CDK4****; CDK5;* ***CDK6****; CDK8;* ***CDKN1B****;* ***CDKN2A****; POLR2A*
PF00452	Apoptosis regulator proteins, Bcl-2 family	*BAK1;* ***BAX****;* ***BCL2****; BCL2L11; BIK; CCDC155; GPX8; PLD3; SPNS1; VRK2*
PF00489	Interleukin-6/G-CSF/MGF family	*IL6R;* ***IL6ST***
PF00605	Interferon regulatory factor transcription factor	
PF00726	Interleukin 10	*IL10RA; IL10RB*
PF01335	Death effector domain	***CASP8****; FADD; RIPK1*
PF07686	Immunoglobulin V-set domain	*NCR3LG1*
PF07714	Protein tyrosine kinase	*CDC37;* ***HSP90AA1****;* ***HSP90AB1****; KHDRBS1;* ***NCKIPSD****;* ***PDGFRB****;* ***RAF1****; WASL*
PF10401	Interferon-regulatory factor 3	***CREBBP****;* ***EP300****;* ***RB1***

OVHDs are structural homologues of human domains either exclusively occurring in oncoviruses or enriched in oncoviral proteomes (compared to generic viral proteomes). Cancer Gene Census proteins are in bold.

The preceding section established that oncovirus-targeted host domains are enriched for cancer driver mutations. Here, we test the hypothesis that oncovirus-mimicked host domains are also enriched for cancer driver mutations, independent of whether they are physically targeted by the virus. To this end, we identified 21 oncoproteins having at least one oncovirus-targeted domain (OVTD) and at least one viral homology domain (VHD). We further classified viral homology domains (VHDs) into those enriched in oncogenic viruses (oncoviral homology domains, or

OVHDs), versus those enriched in non-oncogenic, *i.e.* "generic" viruses (generic viral homology domains, or GVHDs) (Methods, S2 Table). We found that domains structurally mimicked by oncoviruses (OVHDs) are more likely to harbour driver mutations, compared to domains structurally mimicked by generic viruses (GVHDs), independent of whether the domain is physically targeted by oncoviruses (OVTD) (CMH test, common odds ratio = 2.2, $P = 5 \times 10^{-5}$).

We then analyzed the mutational landscape of 44 oncoproteins having at least one oncoviral homology domain (OVHD) but not physically targeted by the virus, *i.e.* having no OVTDs. Pooled analysis of all 44 oncoproteins mapped 245/298 (82.2%) driver mutations and 5422/9554 (56.8%) passenger mutations to OVHDs. In other words, the odds of finding a driver mutation in OVHDs is 3 times as high as that in non-OVHDs (Fisher's exact test, two-tailed $P < 2.2 \times 10^{-16}$) (Fig 5B). Closer inspection identified 23 candidates for focused investigations into the common basis of viral and mutational oncogenesis (Table 4): (I) 4 oncoproteins where all domains are OVHDs, and the driver:passenger ratio is higher than the average ratio across all oncoproteins; (II) 16 oncoproteins where some domains are OVHDs, and driver mutations are exclusively found in OVHDs; and (III) 3 oncoproteins where some domains are OVHDs, and driver mutations are significantly enriched in OVHDs (Fisher's exact test, two-tailed $P < 0.05$). An example of each type of candidate is given in Fig 7. In summary, oncovirus-mimicked host domains are enriched for cancer driver mutations, regardless of whether these domains are physically targeted by the virus.

Table 3.4 Oncoproteins having no oncovirus-targeted domain (OVTD) but at least one oncoviral homology domain (OVHD), where driver mutations are either exclusively found or enriched.

Type	Oncoprotein	OVHD
I	*ETV6*	PF00178; PF02198
I	*MAX*	PF00010
I	*MC1R*	PF00001
I	*MYC*	PF00010; PF01056; PF02344
II	*AKT3*	PF00169; PF00433
II	*ALK*	PF07714
II	*BTK*	PF00017; PF00018; PF00169; PF07714
II	*ESR1*	PF00104; PF00105
II	*FGFR1*	PF07714
II	*FLT3*	PF07714
II	*KIT*	PF07714
II	*MET*	PF07714
II	*NTRK1*	PF07714
II	*PLCG2*	PF00017; PF00018
II	*POLD1*	PF00136; PF03104
II	*POLE*	PF00136; PF03104
II	*RASA1*	PF00017; PF00018; PF00169
II	*RET*	PF07714
II	*REV3L*	PF00136; PF03104
II	*ROS1*	PF07714
III	*BRAF*	PF07714
III	*ERBB2*	PF07714; PF14843
III	*FGFR2*	PF07714

Type I: All domains are OVHDs, and the driver:passenger ratio is higher than the average ratio across all oncoproteins. Type II: Some domains are OVHDs, and driver mutations are exclusively found in OVHDs. Type III: Some domains are OVHDs, and driver mutations are significantly enriched in OVHDs (Fisher's exact test, two-tailed $P < 0.05$).

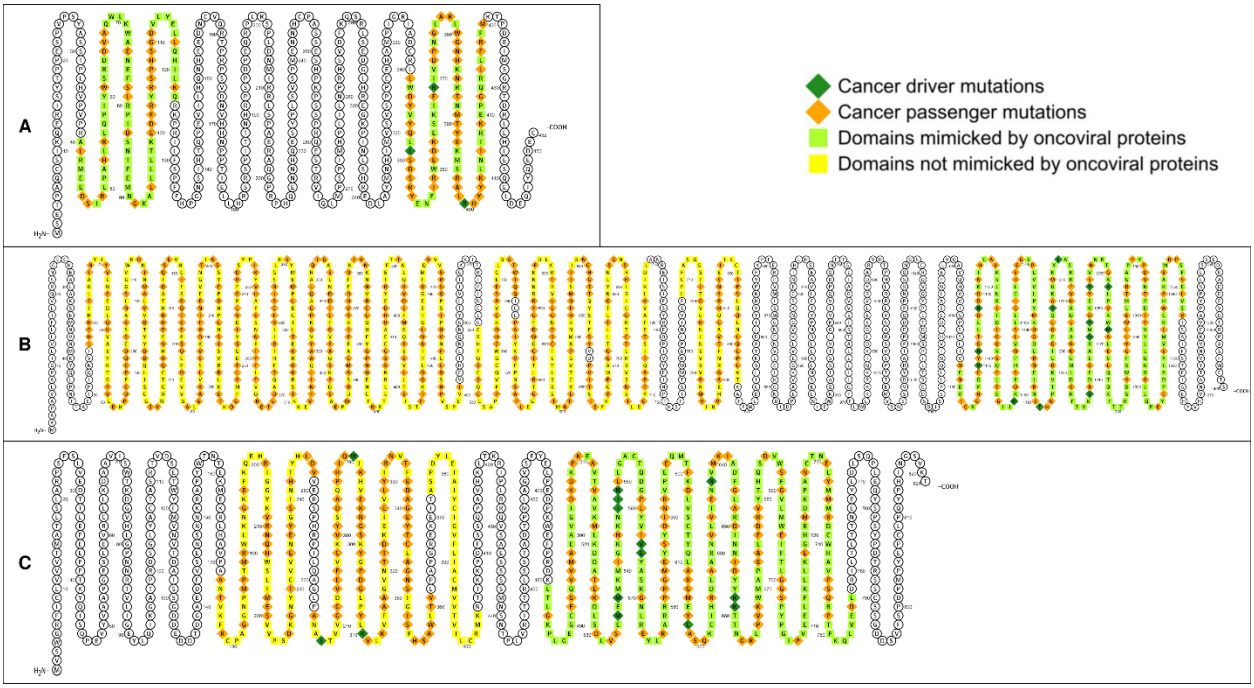


Figure 3.7 Oncoproteins having no oncovirus-targeted domain (OVTD) but at least one oncoviral homology domain (OVHD), where driver mutations are either exclusively found or enriched.

(A) driver:passenger ratio in oncovirus-mimicked PF00178 and PF02198 domains of *ETV6* is higher than the mean driver:passenger ratio for all oncoproteins; (B) driver mutations are exclusively found in oncovirus-mimicked PF07714 domain of *MET*; (C) driver mutations are enriched in oncovirus-mimicked PF07714 domain of *FGFR2*.

3.4.5 Viral proteins and virally-implicated disease mutations tend to perturb the same domain-domain interactions in the human interactome

Gulbahce *et al.* previously hypothesized, and established at the whole-protein level, that viruses and VID mutations induce similar perturbations of the human interactome [9]. Here, we test the same hypothesis at the higher resolution of protein domains, by examining whether viruses and VID mutations perturb the same domain-domain interactions (DDIs) in the human interactome. In other words, do viruses tend to target DDI partners of domains harbouring VID mutations (viral disease domain-interacting domains, or VDDiDs), rather than DDI partners of domains harbouring non-VID mutations (non-viral disease domain-interacting domains, or nVDDiDs) (Fig 8A)? As some domains can interact with both VID domains and non-VID

domains, we define VDDiDs as domains that interact with at least one VID domain, and nVDDiDs as domains that exclusively interact with non-VID domains. We found that EBV and HPV exhibit a slight preference for targeting VDDiDs, although the effect sizes are not statistically significant (42/62 VDDiDs *vs.* 58/104 nVDDiDs for EBV, and 20/29 VDDiDs *vs.* 41/69 nVDDiDs for HPV). HIV targets 218/309 (70.6%) VDDiDs and 193/346 (55.8%) nVDDiDs, representing a 1.9-fold enrichment of VDDiDs among HIV-targeted domains (Fisher's exact test, two-tailed $P = 1 \times 10^{-4}$). Similarly, oncoviruses target 204/285 (71.6%) VDDiDs and 164/291 (56.4%) nVDDiDs, *i.e.* a 1.9-fold enrichment of VDDiDs among oncovirus-targeted domains (Fisher's exact test, two-tailed $P = 1 \times 10^{-4}$). Finally, a meta-analysis on the common effect of all viral proteins and all mutations causing proliferative and immunological diseases found that viruses target 424/599 (70.8%) VDDiDs and 350/551 (63.5%) nVDDiDs, *i.e.* a 1.4-fold enrichment of VDDiDs among virus-targeted domains (Fisher's exact test, two-tailed P = 0.01) (Fig 8B).

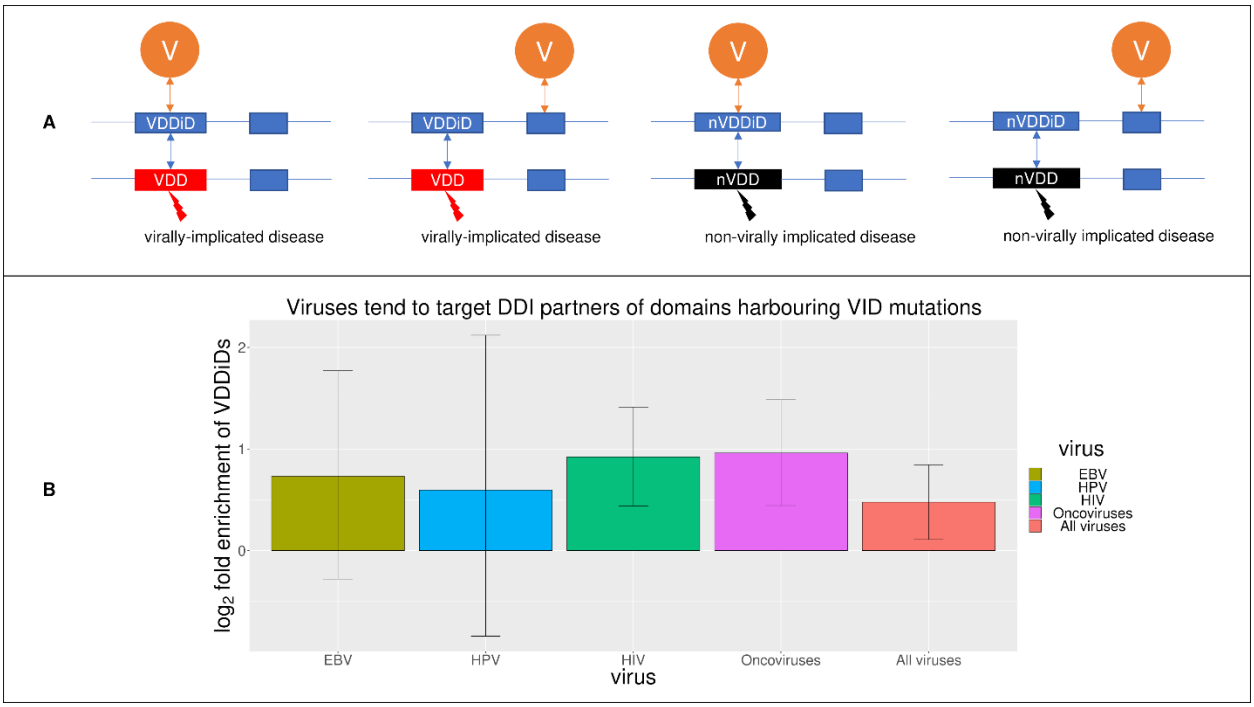


Figure 3.8 Viral proteins and VID mutations perturb the same domain-domain interactions in the human interactome.

(A) From left to right, domains are cross-classified as: interacting with a domain harbouring at least one VID mutation (VDDiD) and targeted by virus, VDDiD not targeted by virus, interacting with a domain harbouring only non-VID mutations (nVDDiD) and targeted by virus, and nVDDiD not targeted by virus. (B) Viruses tend to target VDDiDs rather than nVDDiDs, regardless of whether the VDDiDs and nVDDiDs are susceptible to known disease mutations. The results for EBV and HPV are not statistically significant, possibly due to small sample sizes.

Virus's preferential targeting of VDDiDs may be confounded by the tendency for viruses to target VID domains (Fig 2), and the tendency for VID domains to interact among themselves. We therefore excluded domains susceptible to known disease mutations and examined the extent to which virus targets "non-disease" domains that interact with VID domains. We found that HIV targets 179/250 (71.6%) VDDiDs and 164/285 (57.5%) nVDDiDs that do not harbour any known disease mutation (Fisher's exact test odds ratio = 1.9, two-tailed $P = 8 \times 10^{-4}$). Similarly, oncoviruses target 165/230 (71.7%) VDDiDs and 137/237 (57.8%) nVDDiDs that do not harbour any known disease mutation (Fisher's exact test odds ratio = 1.8, two-tailed $P = 2 \times 10^{-3}$). Pooled analysis of all viruses found that overall, viruses target 345/481 (71.7%) VDDiDs and 295/456 (64.7%) nVDDiDs that do not harbour any known disease mutation (Fisher's exact test odds ratio = 1.4, two-tailed P = 0.02). Virus's preferential targeting of VDDiDs supports our hypothesis that viruses and VID mutations inducing similar disease phenotypes convergently perturb the host domain interactome, possibly unveiling core disease modules underlying clinically heterogeneous virally-implicated diseases (Fig 9).

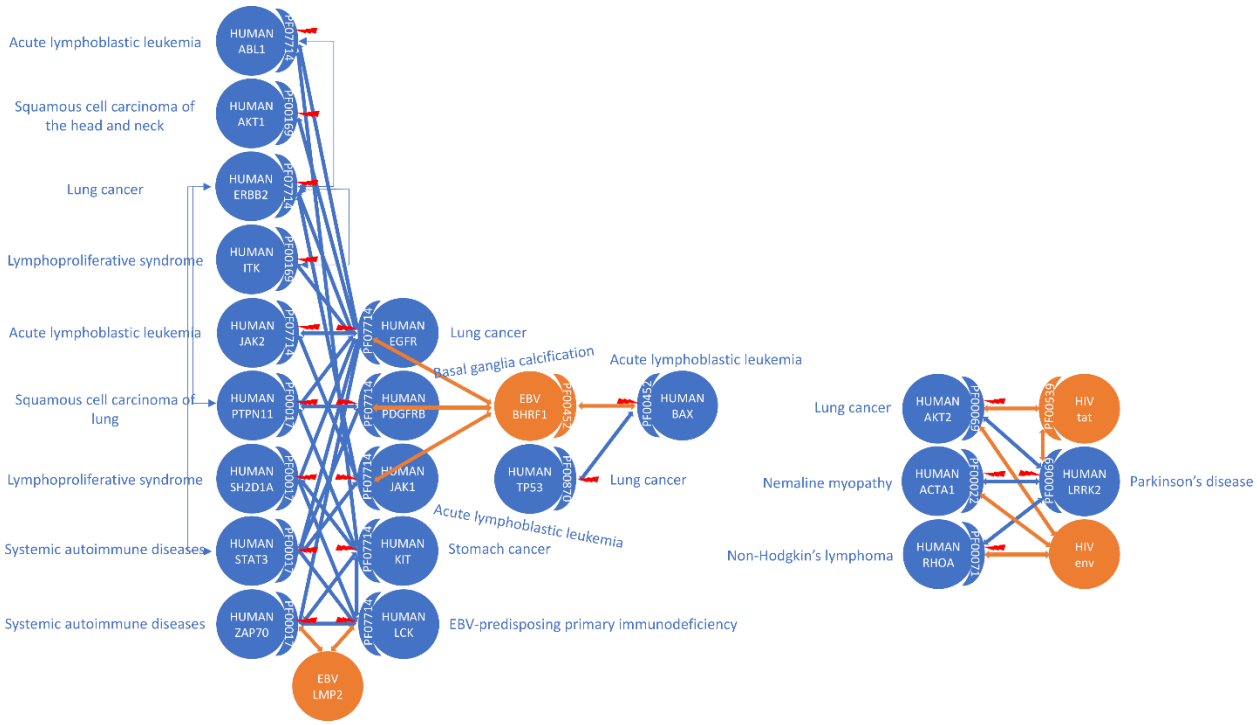


Figure 3.9 Viral proteins and VID mutations convergently perturb dense regions of the human domain interactome.

Examples are given for EBV (left) and HIV (right).

3.5 Discussion

Structural interaction networks serve as a valuable tool for understanding the molecular mechanisms of genetic diseases, as well as the fundamental differences between endogenous and exogenous PPI networks. As experimental determination of protein structure remains an arduous task, homology modelling offers an efficient alternative for the structural annotation of protein complexes. This is based on the observation that PPIs are often mediated by evolutionarily conserved structural modules, such as domains and short linear motifs [77]. Here, we reassess the role of viral proteins as surrogates for human disease variants in relating interactome network perturbation to disease phenotypes, using a domain-resolved human-virus protein interactome where human domains are annotated with disease variant information. Compared to previous work demonstrating general proximity between viral targets and VID proteins in the human interactome, our results provide a structural explanation for the equivalent pathogenicity of viral proteins and VID mutations. Whereas previous studies merely recognized the existence of viral homologues of cellular domains, we delve deeper into the functional implications of oncoviral domain homology. Our approach can readily identify domains convergently targeted or mimicked by diverse oncoviruses for focused screening of driver mutations across various types of cancer. Further characterization of cellular domains and motifs interacting with domains targeted or mimicked by viruses may uncover immune evasion strategies exploited in common by cancer cells and pathogens, and shed light on pathways dysregulated in other virally-implicated disorders.

Although most of our findings are statistically significant, there are notable differences in the enrichment of VID mutations in virus-targeted domains, both among individual viruses (EBV, HPV and HIV), as well as between single-virus analysis and pooled analysis on multiple viruses. For single-virus analysis, enrichment effect size and significance are impacted by the number of

virus-host protein-protein interactions and virus-specific diseases, which ultimately determine the statistical power. Pooled analysis on all oncoviruses detected trends in the same direction as analysis on single oncoviruses (EBV and HPV), but with higher statistical power. In addition to investigator bias resulting in some viruses having a higher number of mapped virus-host PPIs, it is also possible that certain viruses prefer to perturb host regulatory network, rather than host PPI network, which is beyond the scope of this work. Compared to direct targeting of VID domains (a "first-degree" effect), viral targeting of the interaction partners of VID domains is expected to have a weaker, "second-degree" effect on the VID domains. This partly explains why results of the "first-degree" analysis on EBV and HPV (Fig 2) are stronger than those of the "second-degree" analysis (Fig 8B).

Our pooled analysis of all oncoviral targets and all oncomutations is motivated by the assumption of convergent evolution and mimicry of endogenous oncogenic mechanisms by diverse oncoviruses. There is compelling evidence of different oncoviruses complementing each other's replication and persistence strategies, thus eliciting multiple cellular responses associated with the hallmarks of cancer. One example is primary effusion lymphoma, a disease causally linked to KSHV but also having an EBV component. While expression of KSHV lytic genes such as *vIL-6* and *K1* promote VEGF secretion and angiogenesis, concomitant expression of EBV latent genes confers additional anti-apoptotic properties to infected cells in the initial phase of lymphomagenesis [78, 79]. Given the paucity of context-dependent (*i.e.* tissue- and disease-specific) host-endogenous and host-pathogen PPI data, here we focus on establishing viral proteins and genetic mutations that induce similar disease phenotypes as generally equivalent perturbagens of the human interactome. Future work will also consider the diversity of host range and tissue

tropism among different viruses, and the potentially distinct functional impacts of the same mutation in different cell types and diseases.

One potential caveat of our interactome perturbation model is its incompleteness, due to the following reasons. Firstly, current mapping of the host-virus protein interactome is far from exhaustive. Secondly, some bona fide host-virus PPIs cannot be modelled by existing domain-based interaction templates. Thirdly, virus may not interact with a host protein via PPI, but rather regulate its expression via transcriptional or epigenetic mechanisms. Lastly, our study only considers missense mutations, because domain-based analysis of interactome perturbation requires precise positioning of mutations with respect to protein domains. Missense mutations can be unambiguously mapped to individual domains, whereas other types of mutations (*e.g.* nonsense or frameshift) may cause more drastic changes in the protein structure and are more difficult to map to individual domains. We are aware, however, of literature suggesting that nonsense and frameshift mutations tend to occur more frequently in tumour suppressor genes than in oncogenes [80]. Effects of these mutations on the integrity of the human interactome warrant further investigation. Still, despite the incompleteness of our model, we observed significant convergent perturbation of the human domain-resolved interactome by viruses and mutations inducing similar disease phenotypes.

The advent of high-throughput biotechnology has made it possible to comprehensively characterize genomic variations in and interspecies interactions between human and microbes, which play important roles in health and disease. As more data on pathogen-implicated diseases and host-pathogen interactions emerge, our approach may be extended to the study of bacterial diseases and co-infections involving multiple pathogenic species, such as the co-pathogenesis of HIV and Mycobacterium tuberculosis. By combining these data within the framework of structural

systems biology, our work sets the stage for multi-scale, integrative investigations into endogenous and exogenous perturbagens of the human interactome, thus helping to elucidate the molecular mechanisms of infection and its possible connections to genetic diseases such as cancer, autoimmunity, and neurodegeneration.

3.6 Methods

3.6.1 Construction of disease-annotated human-virus structural interaction network

Human-endogenous and human-virus binary PPI data were obtained from IntAct [14], HPIDB [15], and the HIV-1 Human Interaction Database [16-18]. Structural templates for domain-domain and domain-motif interactions were obtained from 3did [19], iPfam [21] and ELM [20]. Protein sequences were scanned for Pfam domains using InterProScan under default settings (version 5.30-69.0) [23, 81], and for the occurrence of domain-binding motifs as defined by 3did and ELM. Domain-based interaction models were assigned to each PPI by extracting all DDIs and DMIs possibly mediating the PPI. Disease association and clinical significance of variants were obtained from UniProtKB, ClinVar, and COSMIC [24, 25, 76]. Ensembl Variant Effect Predictor (VEP v93.0) was used for extracting variant genomic location, variation class, reference allele, HGVS notations, amino acid position, overlapping Pfam domains, among other features [82]. To facilitate counting of mutational events, variants are annotated with RefSNP IDs using VEP's check_existing flag. Variants not co-located with any known variant are merged based on identical genomic location, variation class, and shared alleles, as per NCBI guidelines for merging submitted SNPs into RefSNP clusters (https://www.ncbi.nlm.nih.gov/books/NBK44417/). Only missense mutations located inside Pfam domains were retained for analyses. Assignment of each virally-implicated disease (VID) to EBV, HPV and HIV was based on at least two literature sources (S1 Text). To minimize redundancy in disease annotation, UMLS and OMIM IDs given to subtypes of the same disease were merged into the more general Disease Ontology [83], Orphanet [84] and MeSH IDs.

3.6.2 Pooled analysis of viral proteins and disease mutations

Oncoviruses are as classified by CDC, IARC, and MeSH (https://www.ncbi.nlm.nih.gov/mesh/68009858). Cancer is defined as any disease whose parent terms include "DOID:162", "ORPHA:250908", or MeSH IDs beginning with "C04.557|C04.588|C04.619|C04.626|C04.651|C04.666|C04.682|C04.692|C04.697|C04.700|C04.730|C04.834|C04.850". Diseases without Disease Ontology, Orphanet or MeSH IDs are manually labelled as "cancer" if their names match the following regular expression: "blastoma|cancer|carcino*|glioma|leukemia|leukaemia|lymphoma|melanoma|neoplas*|sarcoma|tumour|tumor". Proliferative diseases have parent terms "DOID:14566", "ORPHA:250908", or MeSH IDs beginning with "C04". Immunological diseases have parent terms "DOID:2914", "ORPHA:98004", or MeSH IDs beginning with "C20". All statistical analyses were conducted in R [85]. Plots of domain-level distribution of disease mutations were created with Protter [86].

3.6.3 Classification of viral homology domains

Pfam domain annotation for all human and viral proteins in UniProt was retrieved from InterPro (Release 69.0) [87]. We define viral homology domains (VHDs) as Pfam domains conserved between human and viral proteins. For each VHD, the likelihood of it occurring in oncoviruses was calculated as the number of oncoviruses encoding the VHD, divided by the total number of unique oncoviral species in UniProt. Similarly, the likelihood of a VHD occurring in "generic" (*i.e.* non-oncogenic) viruses was calculated as the number of generic viruses encoding the VHD divided by the total number of unique generic viral species in UniProt. The observed likelihood ratio (LR) of an oncovirus *vs.* a generic virus encoding the VHD is then the ratio of the two likelihoods. We then permuted the label "oncovirus" and "generic virus" 10000 times among viruses encoding the VHD, thereby obtaining a null distribution for the LR. An empirical p-value

for the enrichment or depletion of a VHD in oncoviral proteomes was calculated according to [88]. VHDs whose observed LR > 1 and Benjamini-Hochberg adjusted p-values (q-values) < 0.1 are considered enriched in oncoviral proteomes. These VHDs and other VHDs exclusively occurring in oncoviruses are called oncoviral homology domains (OVHDs). Likewise, VHDs whose observed LR < 1 and q-values < 0.1 are considered enriched in generic viral proteomes. These VHDs and other VHDs exclusively occurring in generic viruses are called generic viral homology domains (GVHDs).

3.7 References

1. Vidal, M., M.E. Cusick, and A.L. Barabasi, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-98.
2. Oliver, S., *Guilt-by-association goes global.* Nature, 2000. **403**(6770): p. 601-3.
3. Barabasi, A.L., *Network medicine--from obesity to the "diseasome".* N Engl J Med, 2007. **357**(4): p. 404-7.
4. Goh, K.I., *et al.*, *The human disease network.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8685-90.
5. Zhong, Q., *et al.*, *Edgetic perturbation models of human inherited disorders.* Mol Syst Biol, 2009. **5**: p. 321.
6. Wang, X., *et al.*, *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nat Biotechnol, 2012. **30**(2): p. 159-64.
7. Sahni, N., *et al.*, *Edgotype: a fundamental link between genotype and phenotype.* Curr Opin Genet Dev, 2013. **23**(6): p. 649-57.
8. Sahni, N., *et al.*, *Widespread macromolecular interaction perturbations in human genetic disorders.* Cell, 2015. **161**(3): p. 647-60.
9. Gulbahce, N., *et al.*, *Viral perturbations of host networks reflect disease etiology.* PLoS Comput Biol, 2012. **8**(6): p. e1002531.
10. Rozenblatt-Rosen, O., *et al.*, *Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins.* Nature, 2012. **487**(7408): p. 491-5.
11. Kim, P.M., *et al.*, *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.

12. Franzosa, E.A. and Y. Xia, *Structural principles within the human-virus protein-protein interaction network.* Proc Natl Acad Sci U S A, 2011. **108**(26): p. 10538-43.

13. Garamszegi, S., E.A. Franzosa, and Y. Xia, *Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks.* PLoS Pathog, 2013. **9**(12): p. e1003778.

14. Orchard, S., *et al.*, *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D358-63.

15. Ammari, M.G., *et al.*, *HPIDB 2.0: a curated database for host-pathogen interactions.* Database (Oxford), 2016. **2016**.

16. Fu, W., *et al.*, *Human immunodeficiency virus type 1, human protein interaction database at NCBI.* Nucleic Acids Res, 2009. **37**(Database issue): p. D417-22.

17. Ptak, R.G., *et al.*, *Cataloguing the HIV type 1 human protein interaction network.* AIDS Res Hum Retroviruses, 2008. **24**(12): p. 1497-502.

18. Pinney, J.W., *et al.*, *HIV-host interactions: a map of viral perturbation of the host system.* AIDS, 2009. **23**(5): p. 549-54.

19. Mosca, R., *et al.*, *3did: a catalog of domain-based interactions of known three-dimensional structure.* Nucleic Acids Res, 2014. **42**(Database issue): p. D374-9.

20. Dinkel, H., *et al.*, *ELM 2016--data update and new functionality of the eukaryotic linear motif resource.* Nucleic Acids Res, 2016. **44**(D1): p. D294-300.

21. Finn, R.D., *et al.*, *iPfam: a database of protein family and domain interactions found in the Protein Data Bank.* Nucleic Acids Res, 2014. **42**(Database issue): p. D364-73.

22. Berman, H.M., *et al.*, *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

23. Finn, R.D., *et al., The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Res, 2016. **44**(D1): p. D279-85.

24. Famiglietti, M.L., *et al., Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation.* Hum Mutat, 2014. **35**(8): p. 927-35.

25. Landrum, M.J., *et al., ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Res, 2016. **44**(D1): p. D862-8.

26. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.

27. Mesri, E.A., M.A. Feitelson, and K. Munger, *Human viral oncogenesis: a cancer hallmarks analysis.* Cell Host Microbe, 2014. **15**(3): p. 266-82.

28. Epstein, M.A., B.G. Achong, and Y.M. Barr, *Virus Particles in Cultured Lymphoblasts from Burkitt's Lymphoma.* Lancet, 1964. **1**(7335): p. 702-3.

29. Wang, S., *et al., Identification and Characterization of Epstein-Barr Virus Genomes in Lung Carcinoma Biopsy Samples by Next-Generation Sequencing Technology.* Sci Rep, 2016. **6**: p. 26156.

30. Kimura, H., *et al., EBV-associated T/NK-cell lymphoproliferative diseases in nonimmunocompromised hosts: prospective analysis of 108 cases.* Blood, 2012. **119**(3): p. 673-86.

31. Lung, M.L., *et al., The interplay of host genetic factors and Epstein-Barr virus in the development of nasopharyngeal carcinoma.* Chin J Cancer, 2014. **33**(11): p. 556-68.

32. Fukayama, M., R. Hino, and H. Uozaki, *Epstein-Barr virus and gastric carcinoma: virus-host interactions leading to carcinoma.* Cancer Sci, 2008. **99**(9): p. 1726-33.

33. Schiffman, M., *et al., Human papillomavirus and cervical cancer.* Lancet, 2007. **370**(9590): p. 890-907.

34. Nulton, T.J., *et al., Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma.* Oncotarget, 2017. **8**(11): p. 17684-17699.

35. Moscicki, A.B., *et al., Updating the natural history of human papillomavirus and anogenital cancers.* Vaccine, 2012. **30 Suppl 5**: p. F24-33.

36. Li, N., *et al., Human papillomavirus infection and bladder cancer risk: a meta-analysis.* J Infect Dis, 2011. **204**(2): p. 217-23.

37. Li, N., *et al., Human papillomavirus infection and sporadic breast carcinoma risk: a meta-analysis.* Breast Cancer Res Treat, 2011. **126**(2): p. 515-20.

38. Klein, F., W.F. Amin Kotb, and I. Petersen, *Incidence of human papilloma virus in lung cancer.* Lung Cancer, 2009. **65**(1): p. 13-8.

39. Singh, N., *et al., Implication of high risk human papillomavirus HR-HPV infection in prostate cancer in Indian population--a pioneering case-control analysis.* Sci Rep, 2015. **5**: p. 7822.

40. Monforte, A., *et al., HIV-induced immunodeficiency and mortality from AIDS-defining and non-AIDS-defining malignancies.* AIDS, 2008. **22**(16): p. 2143-53.

41. Grulich, A.E., *et al., Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis.* Lancet, 2007. **370**(9581): p. 59-67.

42. De Falco, G., *et al., Interaction between HIV-1 Tat and pRb2/p130: a possible mechanism in the pathogenesis of AIDS-related neoplasms.* Oncogene, 2003. **22**(40): p. 6214-9.

43. Nunnari, G., J.A. Smith, and R. Daniel, *HIV-1 Tat and AIDS-associated cancer: targeting the cellular anti-cancer barrier?* J Exp Clin Cancer Res, 2008. **27**: p. 3.

44. Briggs, S.D., *et al., SH3-mediated Hck tyrosine kinase activation and fibroblast transformation by the Nef protein of HIV-1.* J Biol Chem, 1997. **272**(29): p. 17899-902.

45. Barbaro, G., *Cardiovascular manifestations of HIV infection.* Circulation, 2002. **106**(11): p. 1420-5.

46. Pugliese, A., *et al., Impact of highly active antiretroviral therapy in HIV-positive patients with cardiac involvement.* J Infect, 2000. **40**(3): p. 282-4.

47. Yunis, N.A. and V.E. Stone, *Cardiac manifestations of HIV/AIDS: a review of disease spectrum and clinical management.* J Acquir Immune Defic Syndr Hum Retrovirol, 1998. **18**(2): p. 145-54.

48. Hersh, B.P., P.R. Rajendran, and D. Battinelli, *Parkinsonism as the presenting manifestation of HIV infection: improvement on HAART.* Neurology, 2001. **56**(2): p. 278-9.

49. Koutsilieri, E., *et al., Parkinsonism in HIV dementia.* J Neural Transm (Vienna), 2002. **109**(5-6): p. 767-75.

50. Mirsattari, S.M., C. Power, and A. Nath, *Parkinsonism with HIV infection.* Mov Disord, 1998. **13**(4): p. 684-9.

51. Kumar, A., *et al., Tuning of AKT-pathway by Nef and its blockade by protease inhibitors results in limited recovery in latently HIV infected T-cell line.* Sci Rep, 2016. **6**: p. 24090.

52. Portis, T., P. Dyck, and R. Longnecker, *Epstein-Barr Virus (EBV) LMP2A induces alterations in gene transcription similar to those observed in Reed-Sternberg cells of Hodgkin lymphoma.* Blood, 2003. **102**(12): p. 4166-78.

53. Lamers, S.L., *et al., HIV-1 Nef in macrophage-mediated disease pathogenesis.* Int Rev Immunol, 2012. **31**(6): p. 432-50.

54. Plummer, M., *et al.*, *Global burden of cancers attributable to infections in 2012: a synthetic analysis.* Lancet Glob Health, 2016. **4**(9): p. e609-16.

55. Vogt, P.K., *Retroviral oncogenes: a historical primer.* Nat Rev Cancer, 2012. **12**(9): p. 639-48.

56. Stevenson, P.G., J.P. Simas, and S. Efstathiou, *Immune control of mammalian gamma-herpesviruses: lessons from murid herpesvirus-4.* J Gen Virol, 2009. **90**(Pt 10): p. 2317-30.

57. Parada, L.F., *et al.*, *Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene.* Nature, 1982. **297**(5866): p. 474-8.

58. de Martel, C., *et al.*, *Cancers attributable to infections among adults with HIV in the United States.* AIDS, 2015. **29**(16): p. 2173-81.

59. Manabe, A., *et al.*, *Viral Infections in Juvenile Myelomonocytic Leukemia: Prevalence and Clinical Implications.* J Pediatr Hematol Oncol, 2004. **26**(10): p. 636-641.

60. Koike, K. and K. Matsuda, *Recent advances in the pathogenesis and management of juvenile myelomonocytic leukaemia.* Br J Haematol, 2008. **141**(5): p. 567-75.

61. Kaplan, M., *et al.*, *EGFR Dynamics Change during Activation in Native Membranes as Revealed by NMR.* Cell, 2016. **167**(5): p. 1241-1251 e11.

62. Purba, E.R., E.I. Saita, and I.N. Maruyama, *Activation of the EGF Receptor by Ligand Binding and Oncogenic Mutations: The "Rotation Model".* Cells, 2017. **6**(2).

63. Glocker, E.O., *et al.*, *Inflammatory bowel disease and mutations affecting the interleukin-10 receptor.* N Engl J Med, 2009. **361**(21): p. 2033-45.

64. Slobedman, B., *et al.*, *Virus-encoded homologs of cellular interleukin-10 and their control of host immune function.* J Virol, 2009. **83**(19): p. 9618-29.

65. DeBruyne, L.A., *et al., Lipid-mediated gene transfer of viral IL-10 prolongs vascularized cardiac allograft survival by inhibiting donor-specific cellular and humoral immune responses.* Gene Ther, 1998. **5**(8): p. 1079-87.

66. Beguelin, W., *et al., IL10 receptor is a novel therapeutic target in DLBCLs.* Leukemia, 2015. **29**(8): p. 1684-94.

67. Pietrek, M., *et al., Role of the Kaposi's sarcoma-associated herpesvirus K15 SH3 binding site in inflammatory signaling and B-cell activation.* J Virol, 2010. **84**(16): p. 8231-40.

68. Myoui, A., *et al., C-SRC tyrosine kinase activity is associated with tumor colonization in bone and lung in an animal model of human breast cancer metastasis.* Cancer Res, 2003. **63**(16): p. 5028-33.

69. Imielinski, M., *et al., Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing.* Cell, 2012. **150**(6): p. 1107-20.

70. Campbell, K.S., *et al., Polyoma middle tumor antigen interacts with SHC protein via the NPTY (Asn-Pro-Thr-Tyr) motif in middle tumor antigen.* Proc Natl Acad Sci U S A, 1994. **91**(14): p. 6344-8.

71. Ahn, R., *et al., The ShcA PTB domain functions as a biological sensor of phosphotyrosine signaling during breast cancer progression.* Cancer Res, 2013. **73**(14): p. 4521-32.

72. Dussupt, V., *et al., The nucleocapsid region of HIV-1 Gag cooperates with the PTAP and LYPXnL late domains to recruit the cellular machinery necessary for viral budding.* PLoS Pathog, 2009. **5**(3): p. e1000339.

73. Castiglioni, S., J.A. Maier, and M. Mariotti, *The tyrosine phosphatase HD-PTP: A novel player in endothelial migration.* Biochem Biophys Res Commun, 2007. **364**(3): p. 534-9.

74. Manteghi, S., *et al.*, *Haploinsufficiency of the ESCRT Component HD-PTP Predisposes to Cancer.* Cell Rep, 2016. **15**(9): p. 1893-900.

75. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.

76. Forbes, S.A., *et al.*, *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.

77. Aloy, P. and R.B. Russell, *Structural systems biology: modelling protein interactions.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 188-97.

78. Haddad, L., *et al.*, *KSHV-transformed primary effusion lymphoma cells induce a VEGF-dependent angiogenesis and establish functional gap junctions with endothelial cells.* Leukemia, 2008. **22**(4): p. 826-34.

79. Keller, S.A., *et al.*, *NF-kappaB is essential for the progression of KSHV- and EBV-infected lymphomas in vivo.* Blood, 2006. **107**(8): p. 3295-302.

80. Mort, M., *et al.*, *A meta-analysis of nonsense mutations causing human genetic disease.* Hum Mutat, 2008. **29**(8): p. 1037-47.

81. Jones, P., *et al.*, *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-40.

82. McLaren, W., *et al.*, *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.

83. Kibbe, W.A., *et al.*, *Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1071-8.

84. Rath, A., *et al.*, *Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users.* Hum Mutat, 2012. **33**(5): p. 803-8.

85. Team, R.C., *R: A language and environment for statistical computing*. 2018, R Foundation for Statistical Computing: Vienna, Austria.

86. Omasits, U., *et al.*, *Protter: interactive protein feature visualization and integration with experimental proteomic data.* Bioinformatics, 2014. **30**(6): p. 884-6.

87. Finn, R.D., *et al.*, *InterPro in 2017-beyond protein family and domain annotations.* Nucleic Acids Res, 2017. **45**(D1): p. D190-D199.

88. Phipson, B. and G.K. Smyth, *Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn.* Stat Appl Genet Mol Biol, 2010. **9**: p. Article39.

Connecting Statement

In Chapter 3, I used a domain-resolved human-virus PPI network to examine correlations between the structural features of viral proteins and their mechanistic involvement in human disease. I demonstrated that by causing similar perturbations to the human protein interactome at the domain level, genetic mutations and viral proteins are mechanistically equivalent contributors to diseases having both genetic and viral etiologic factors, *i.e.* virally-implicated diseases (VIDs).

In Chapter 4, I use a domain-resolved eukaryote-bacteria PPI network to examine correlations between the structural features of bacterial proteins and their potential for targeting eukaryote-specific cellular processes. I demonstrate that bacterial effector proteins are significantly enriched for structural domains and short linear motifs that either mimic or convergently target host domains involved in eukaryote-specific cellular processes.

Chapter 4: Structural profiling of bacterial effectors reveals enrichment of host-interacting domains and motifs

Yangchun Frank Chen[1] and Yu Xia[1]

[1]Department of Bioengineering, McGill University, Montreal QC, Canada

4.1 Abstract

Effector proteins are bacterial virulence factors secreted directly into host cells and, through their extensive interactions with host proteins, rewire host signaling pathways to the advantage of the pathogen. Despite the crucial role of globular domains as mediators of protein-protein interactions (PPIs), previous structural studies of bacterial effectors were primarily focused on individual domains, rather than domain-mediated PPIs, which limits their ability to uncover systems-level principles underlying the host-pathogen PPI network. Here, we took an interaction-centric approach and systematically examined the potential of structural components within pathogen proteins to repurpose or disrupt host-endogenous PPIs. We demonstrate that compared to the rest of the pathogen proteome, effectors are significantly enriched for domains and motifs that either structurally mimic or convergently target host domains involved in eukaryote-specific PPIs. Our study lends novel structural insight into the virulence mechanism of bacterial effectors and may aid in the design of selective inhibitors of host-pathogen interactions.

4.2 Introduction

An important goal of systems microbiology is to understand how host-pathogen protein-protein interactions (PPIs) impact host-endogenous signaling networks. Effector proteins are virulence factors secreted by pathogenic bacteria and injected directly into the host cytoplasm via specialized secretion systems (Galan 2009). Effectors are key mediators of host-pathogen interactions throughout the infection cycle, from initial host attachment and pathogen internalization, to migration and proliferation in the host. Among the diverse biochemical activities of effectors discovered so far are guanine nucleotide exchange factors and dissociation inhibitors, GTPase-activating proteins, kinases and phosphatases, ubiquitin ligases, and so on (Fu and Galan 1998, Steele-Mortimer, Knodler *et al.* 2000, Janjusevic, Abramovitch *et al.* 2006). A common virulence mechanism of effectors is functional mimicry of host activities, whereby effectors compete with host proteins for control of host signaling pathways. This functional mimicry can be achieved in one of two ways: horizontal acquisition of eukaryotic globular domains, or convergent evolution of domains and short linear motifs in bacteria that bear little sequence or structural similarity to eukaryotic proteins (Stebbins and Galan 2001, Popa, Tabuchi *et al.* 2016, Scott and Hartland 2017). These structural modules allow effectors to interact seamlessly with host-endogenous factors involved in actin remodelling, protein degradation and cell cycle regulation, helping the pathogen to survive and thrive in the host while bypassing immune surveillance. Previous studies have uncovered a large repertoire of bacterial effectors that are structural homologs of eukaryotic proteins, giving rise to models for predicting effectors based on the premise that whereas most domains are uniformly distributed among all species of bacteria, eukaryotic domains are overrepresented in the genomes of pathogenic and symbiotic species (Jehl, Arnold *et al.* 2011, Marchesini, Herrmann *et al.* 2011). Although useful for identifying candidate effectors in

metagenomic analyses, these models were primarily focused on individual domains, rather than domain-mediated PPIs, which limits their ability to uncover systems-level principles underlying the host-pathogen PPI network. Eukaryotic domains and their domain-domain interaction (DDI) partners are of special interest to the study of host-pathogen interactions, as they are often mimicked or targeted by pathogens to subvert host signaling pathways (Arnold, Boonen *et al.* 2012). Here, we constructed a domain-resolved interaction network consisting of eukaryote-endogenous, bacteria-endogenous and host-bacteria PPIs, based on which we examined the evolutionary origins of host-interacting domains in bacteria, as well as the relative likelihood that a domain found in both eukaryotes and bacteria engage in DDIs that are unique to eukaryotes, as opposed to DDIs that are conserved between eukaryotes and bacteria. We then systematically probed the proteomes of bacterial pathogens for horizontally acquired, eukaryotic-like domains that mediate eukaryote-specific DDIs, and convergently evolved, bacteria-exclusive domains that disrupt eukaryote-specific DDIs. We found that compared to the rest of the pathogen proteome, effectors are enriched for domains that are either homologs of, or convergently target, host domains involved in eukaryote-specific DDIs. Moreover, in the absence of eukaryotic-like domains or among pathogen proteins without Pfam domain assignment, effectors harbor a higher variety and density of short linear motifs that target host domains involved in eukaryote-specific DDIs.

4.3 Results

4.3.1 Horizontal acquisition *vs.* convergent evolution of host-interacting domains in bacteria

Previous studies have established binding site mimicry as a key feature of human-virus protein-protein interactions (PPIs) where, rather than securing new binding sites, viral proteins tend to compete with human proteins for the same binding sites. Moreover, while two human proteins sharing binding sites on a common target tend to be structurally similar, a viral protein and a human protein sharing binding sites on a common target tend to be structurally distinct (Franzosa and Xia 2011, Garamszegi, Franzosa *et al.* 2013). In other words, binding site sharing among human proteins is largely attributable to divergent evolution through gene duplication, whereas binding site mimicry by viral proteins tends to involve convergent evolution of unique host-interacting modules in viruses. As bacteria and viruses are both known to hijack host molecular machinery through interacting with host proteins, we performed similar analyses on a domain-resolved host-bacteria PPI network with regard to binding site mimicry and the evolutionary origins of host-interacting domains in bacteria. To this end, we acquired eukaryote-endogenous (within animals/plants/fungi), bacteria-endogenous and host-bacteria (between animals/plants and pathogenic bacteria) PPI data, and resolved each PPI into domain-domain interactions (DDIs) between interacting proteins, based on DDI templates previously derived from 3D structures of protein complexes (Materials and Methods). The resulting eukaryote-bacteria structural interaction network consists of: (1) 57,019 PPIs among 22,110 eukaryotic proteins, resolved into 4,953 DDIs among 2,859 eukaryotic domains; (2) 3,362 PPIs among 3,000 bacterial proteins, resolved into 1,434 DDIs among 1,120 bacterial domains; and (3) 173 PPIs between 107 host proteins and 103 bacterial proteins, resolved into 87 DDIs between 53 host domains and 63 bacterial domains.

We found that of the 103 host-targeting bacterial proteins, 95 (92%) bind to the same domains on their host target that are otherwise bound by host-endogenous proteins, suggesting that like viruses, bacteria also tend to recruit domains involved in host-endogenous PPIs for host-pathogen PPIs. We then determined whether bacterial and host proteins binding to the same domain on another host protein are structurally similar. We found that of 18,331 cases where two host proteins A and B bind to the same domain on a common target, 13,139 (72%) cases involve domains which are conserved between A and B, while in the remaining 5,192 (28%) cases there is no domain conserved between A and B. Contrarily, among 95 cases where host protein X and bacterial protein Y bind to the same domain on another host protein, only 8 cases (8%) involve domains which are conserved between X and Y, while in the remaining 87 (92%) cases there is no domain conserved between X and Y. In other words, compared to binding site sharing among host proteins, binding site mimicry by bacterial proteins is significantly more likely to arise via convergent evolution of bacteria-exclusive domains, rather than horizontal acquisition of host domains (Fisher's exact test, two-tailed $P < 2.2 * 10^{-16}$). Figure 1 shows examples of effectors targeting host domains involved in host-endogenous PPIs via: (A) a horizontally acquired, eukaryotic-like domain (Angot, Peeters *et al.* 2006); or (B) a convergently evolved, bacteria-exclusive domain (Huang, Sutton *et al.* 2009).

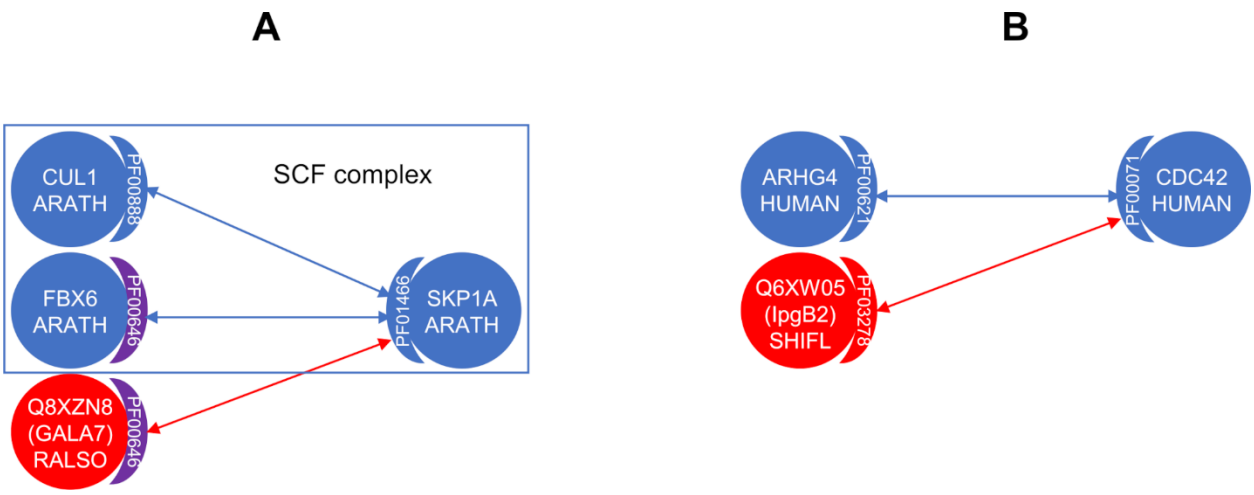


Figure 4.1 Horizontal acquisition vs. convergent evolution of host-interacting domains in bacteria.

(A) *Ralstonia solanacearum* acquired a host-like F-box domain (PF00646) that competes with host-endogenous F-box protein for binding to SKP1, thus hijacking the ubiquitin-proteasome pathway in *Arabidopsis thaliana*. (B) *Shigella flexneri* convergently evolved a GEF domain (PF03278) that competes with host Rho GEF, thus activating the Rho GTPase signaling pathway in humans. Host proteins and domains are colored in blue. Bacterial proteins and domains are colored in red. Bacteria-mimicked host domains are colored in purple. GEF: guanine nucleotide exchange factor.

4.3.2 Effectors structurally mimic host domains involved in eukaryote-specific PPIs

Having examined the evolutionary origins of host-interacting domains in bacteria, we then asked whether bacterial effectors tend to mimic or target domains that mediate DDIs predominantly in eukaryotes, as opposed to domains that mediate DDIs in eukaryotes and bacteria with similar likelihoods – the rationale being that the former are involved in eukaryote-specific processes such as protein ubiquitination (Grau-Bove, Sebe-Pedros *et al.* 2015), whereas the latter are involved in conserved, core cellular processes (Walhout, Sordella *et al.* 2000, Matthews, Vaglio *et al.* 2001), which are unlikely to be perturbed in host-pathogen interaction. We found that of the 63 host-binding bacterial domains in our dataset, 12 have homologs in eukaryotes, among which 7 mediate DDIs exclusively in eukaryotes (PF12796, PF00092, PF12799, PF02205, PF04564, PF00646, PF13676), 3 mediate DDIs primarily in eukaryotes (PF00069, PF00583,

PF00183), and 2 have similar numbers of DDI partners in eukaryotes and bacteria (PF13472, PF00085) (Supplementary Table 1). Meanwhile, of the 31 host domains convergently targeted by bacteria-exclusive domains, 28 otherwise mediate DDIs exclusively in eukaryotes, 2 mediate DDIs primarily in eukaryotes, and 1 has similar numbers of DDI partners in eukaryotes and bacteria (Supplementary Table 2). In summary, effectors tend to mimic or target domains that mediate DDIs predominantly in eukaryotes. Given that effectors comprise nearly half (43/103) of the host-targeting bacterial proteins in our PPI dataset, we hypothesized that compared to the rest of the pathogen proteome, effectors are generally enriched for: (1) eukaryotic-like domains that mediate DDIs predominantly in eukaryotes; and (2) bacteria-exclusive domains that target host domains which, when not involved in host-pathogen DDIs, mediate DDIs exclusively in eukaryotes. To test this hypothesis, we systematically compared the domain signatures of 238 effectors and 3,921 non-effectors encoded by 84 bacterial species of verified pathogenicity (Materials and Methods).

We first tested the hypothesis that effectors are enriched for domains that mediate DDIs exclusively in eukaryotes. We found that among 41 effectors and 1,478 non-effectors containing domains that mediate experimentally verified PPIs in eukaryotes, 8 effectors (20%) and 55 non-effectors (4%) contain domains that mediate PPIs exclusively in eukaryotes, suggesting that effectors are 6 times as likely as non-effectors to repurpose eukaryote-specific processes via structural mimicry (Fisher's exact test, two-tailed $P = 2 * 10^{-4}$) (Figure 2). Table 1 is a list of effectors containing domains involved in interprotein DDIs in eukaryotes, but neither inter- nor intra-protein DDIs in bacteria. Next, we tested the hypothesis that effectors are enriched for domains that mediate DDIs primarily in eukaryotes. For domains having DDI partners in both eukaryotes and bacteria, we estimated their propensity for mediating eukaryote-specific DDIs by computing the odds ratio of the domain's co-occurrence with DDI partners in eukaryotes *vs.* in

bacteria. If a bacterial protein contains multiple such domains, we computed a weighted average odds ratio (Table 2). We found that among 26 effectors and 635 non-effectors containing domains that have DDI partners in both eukaryotes and bacteria, the average domain in effectors is 7 times as likely as that in non-effectors to co-occur with DDI partners in eukaryotes than in bacteria (Wilcoxon test, two tailed $P = 4 * 10^{-7}$) (Figure 3). Table 3 is a list of effectors containing domains that are more likely to co-occur with DDI partners in eukaryotes than in bacteria.

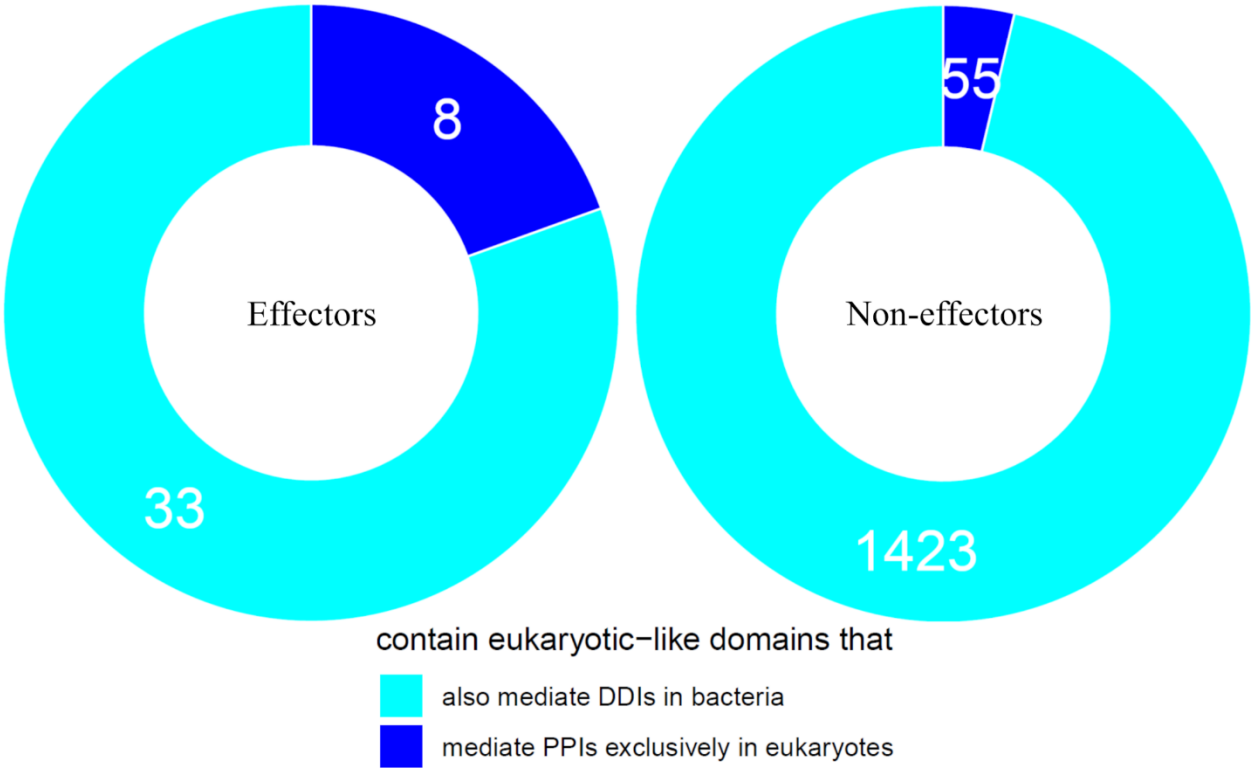


Figure 4.2 Effectors are enriched for domains that mediate PPIs exclusively in eukaryotes.

Among pathogen proteins containing domains that mediate experimentally verified PPIs in eukaryotes, 20% effectors and 4% non-effectors contain domains that mediate PPIs exclusively in eukaryotes, suggesting that effectors are 6 times as likely as non-effectors to repurpose eukaryote-specific processes via structural mimicry (Fisher's exact test, two-tailed $P = 2 * 10^{-4}$).

Table 4.1 Effectors containing domains that mediate PPIs exclusively in eukaryotes.

UniProt Accession	Species	Domains	Domains mediating PPIs exclusively in eukaryotes
B0RMF9	*Xanthomonas campestris*	PF00560	PF00560
A0A0A8VF40	*Yersinia ruckeri*	PF00560; PF13855	PF00560
A0A199P7E1	*Xanthomonas translucens*	PF00646	PF00646
A0A0S4VGA6	*Ralstonia solanacearum*	PF00646; PF13516	PF00646
F6G106	*Ralstonia solanacearum*	PF00646; PF13516; PF13855	PF00646
A0A1Y0FB05	*Ralstonia solanacearum*	PF00665; PF13276	PF00665
A0A286NT26	*Vibrio parahaemolyticus*	PF02205	PF02205
D8NFZ7	*Ralstonia solanacearum*	PF01535; PF12854; PF13812	PF13812

Table 4.2 Weighted average host-interacting potential of a multi-domain bacterial protein.

DDI	Eukaryotic species encoding both interacting domains	Eukaryotic species encoding either one or both interacting domains	Bacterial species encoding both interacting domains	Bacterial species encoding either one or both interacting domains	Odds ratio of domain mediating DDIs in eukaryotes *vs.* in bacteria	Weight
A_B	m	H_1	n	B_1	$OR_1 = \frac{m * (B_1 - n)}{n * (H_1 - m)}$	$w_1 = \frac{m * (B_1 - n)}{H_1 + B_1}$
A_C	p	H_2	q	B_2	$OR_2 = \frac{p * (B_2 - q)}{q * (H_2 - p)}$	$w_2 = \frac{p * (B_2 - q)}{H_2 + B_2}$
D_E	x	H_3	y	B_3	$OR_3 = \frac{x * (B_3 - y)}{y * (H_3 - x)}$	$w_3 = \frac{x * (B_3 - y)}{H_3 + B_3}$
Host-interacting potential = $\log\left(\frac{\sum_{i=1}^{3} OR_i * w_i}{\sum_{i=1}^{3} w_i}\right)$						

The host-interacting potential of a bacterial protein containing domains A and D, where A and D have DDI partners (domains B, C, E) in both eukaryotes and bacteria, is computed as the Mantel-Haenszel weighted average log odds ratio of domains A and D co-occurring with interacting domains in eukaryotes *vs.* in bacteria. The odds of domain co-occurring with DDI partners are the number of species encoding both interacting domains (*i.e.* DDI is possible) divided by the number of species encoding either one, but not both, of the interacting domains (*i.e.* DDI is not possible).

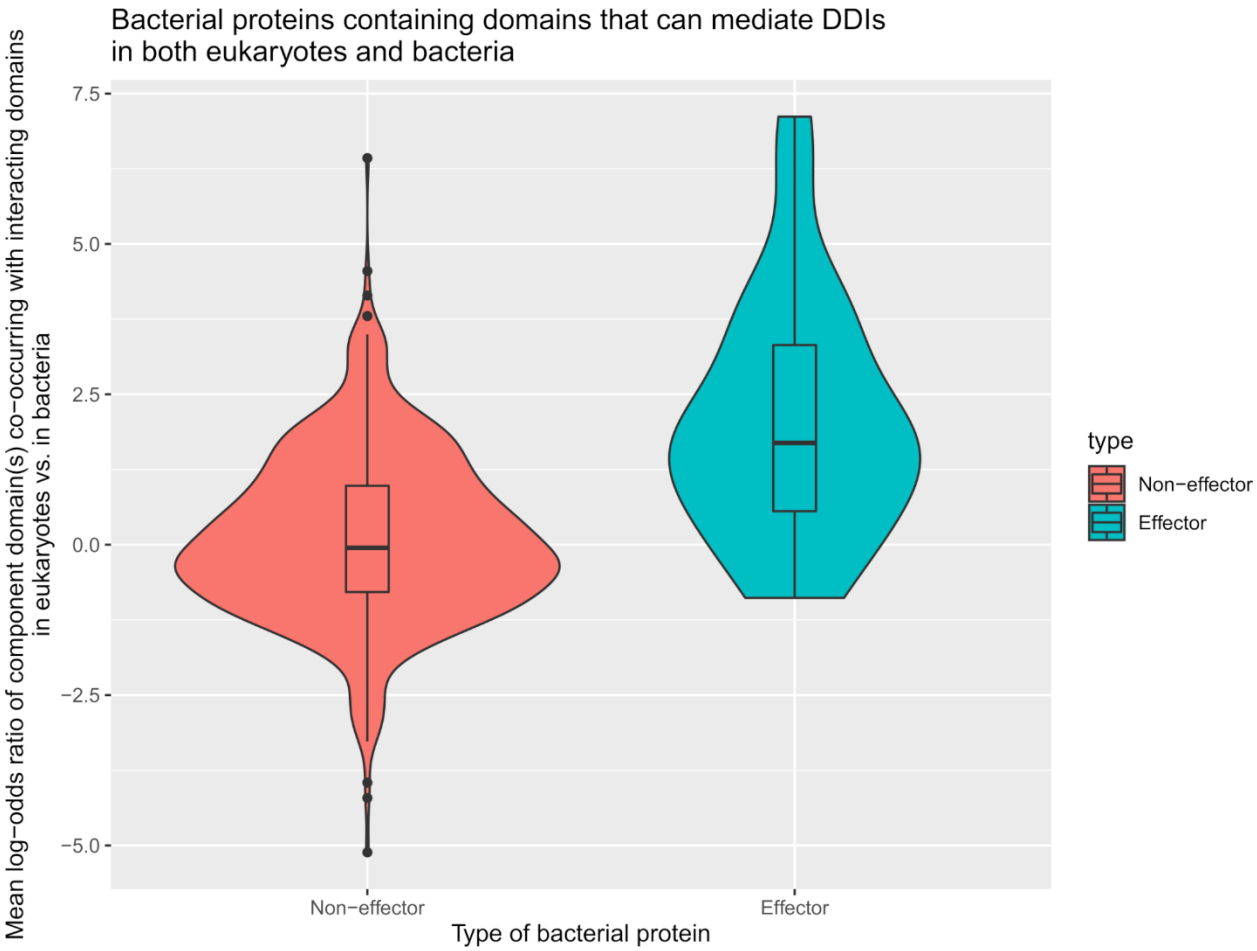


Figure 4.3 Effectors are enriched for domains that mediate PPIs primarily in eukaryotes.

Among pathogen proteins containing domains that have DDI partners in both eukaryotes and bacteria, the average domain in effectors is 7 times as likely as that in non-effectors to co-occur with DDI partners in eukaryotes than in bacteria (Wilcoxon test, two tailed P = 4 * 10-7).

Table 4.3 Effectors containing domains that mediate PPIs primarily in eukaryotes.

UniProt Accession	Species	Domains	Domains with DDI partners in both eukaryotes and bacteria	Log odds ratio of domains co-occurring with DDI partners in eukaryotes vs. in bacteria
Q5ZRQ0	*Legionella pneumophila*	PF04564	PF04564	7.1
O84875	*Chlamydia trachomatis*	PF02902	PF02902	6.2
P74873	*Salmonella enterica*	PF00102; PF03545; PF09119	PF00102	4.4
Q9KS43	*Vibrio cholerae*	PF01764	PF01764	3.9
Q3BQY9	*Xanthomonas euvesicatoria*	PF13202; PF13499	PF13202; PF13499	3.8
D8P6Z5	*Ralstonia solanacearum*	PF13516	PF13516	3.7
Q8XT98	*Ralstonia solanacearum*	PF00069	PF00069	3.5
D2TI55	*Citrobacter rodentium*	PF00557	PF00557	2.8
Q8XZN9	*Ralstonia solanacearum*	PF13516; PF13855	PF13516; PF13855	2.2
A0A6C9X110	*Escherichia coli*	PF00805; PF01391; PF13599	PF01391	2

4.3.3 Effectors convergently target host domains involved in eukaryote-specific PPIs

We then tested the hypothesis that effectors are enriched for bacteria-exclusive domains that target host domains which, when not involved in host-pathogen DDIs, mediate DDIs exclusively in eukaryotes. Given that experimental PPI data often suffer from limitations such as false negatives and investigator bias in pathogen selection, we supplemented host-interacting bacteria-exclusive domains supported by PPI data with host-interacting bacteria-exclusive domains supported by interprotein DDI templates (Mosca, Ceol *et al.* 2014). In this manner, we identified a total of 207 bacteria-exclusive domains with the potential to target host domains that mediate DDIs in eukaryotes, 52 of which target host domains that mediate DDIs exclusively in eukaryotes. We

found that among 30 effectors and 41 non-effectors with the potential to convergently target host domains that mediate DDIs in eukaryotes, 23 effectors (77%) and 11 non-effectors (27%) target host domains that mediate DDIs exclusively in eukaryotes, suggesting that effectors are 9 times as likely as non-effectors to disrupt eukaryote-specific processes via convergent evolution (Fisher's exact test, two-tailed $P = 4 * 10^{-5}$) (Figure 4). Supplementary Table 3 is a list of effectors containing bacteria-exclusive domains that convergently target host domains otherwise exclusively involved in host-endogenous DDIs.

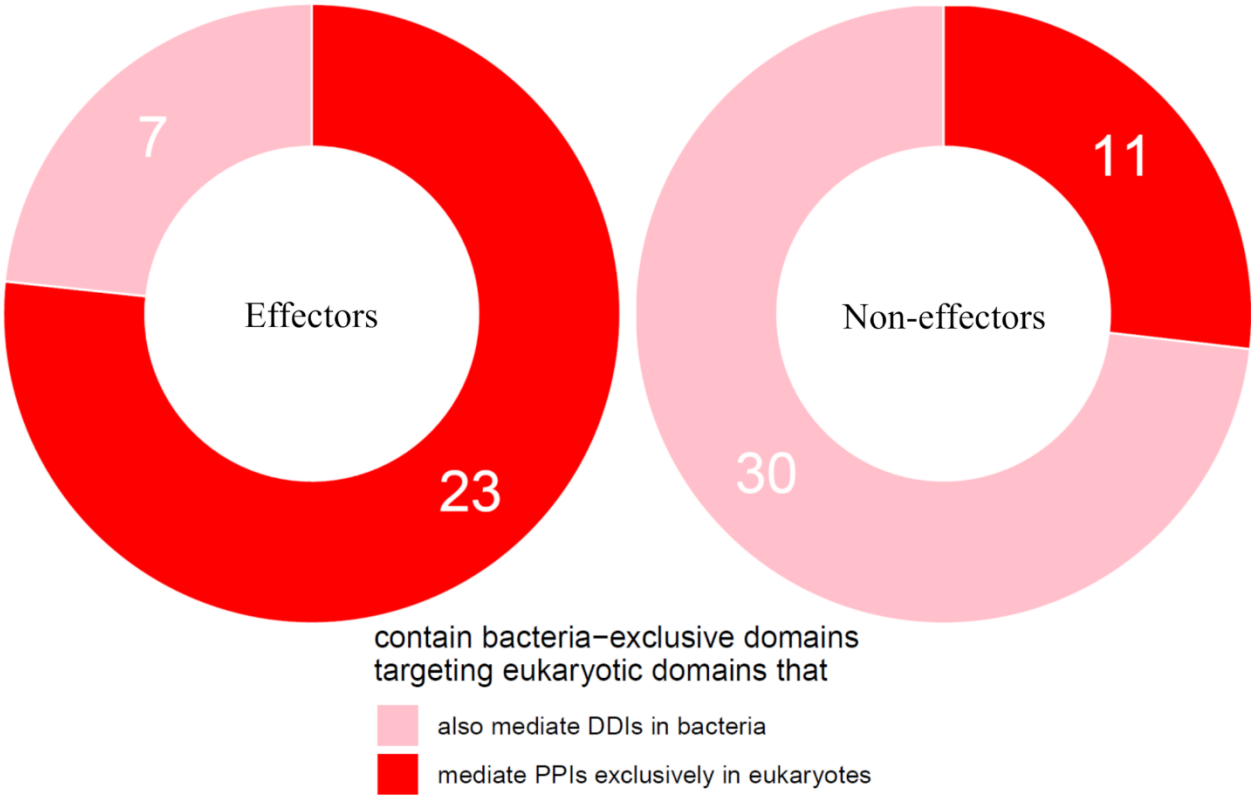
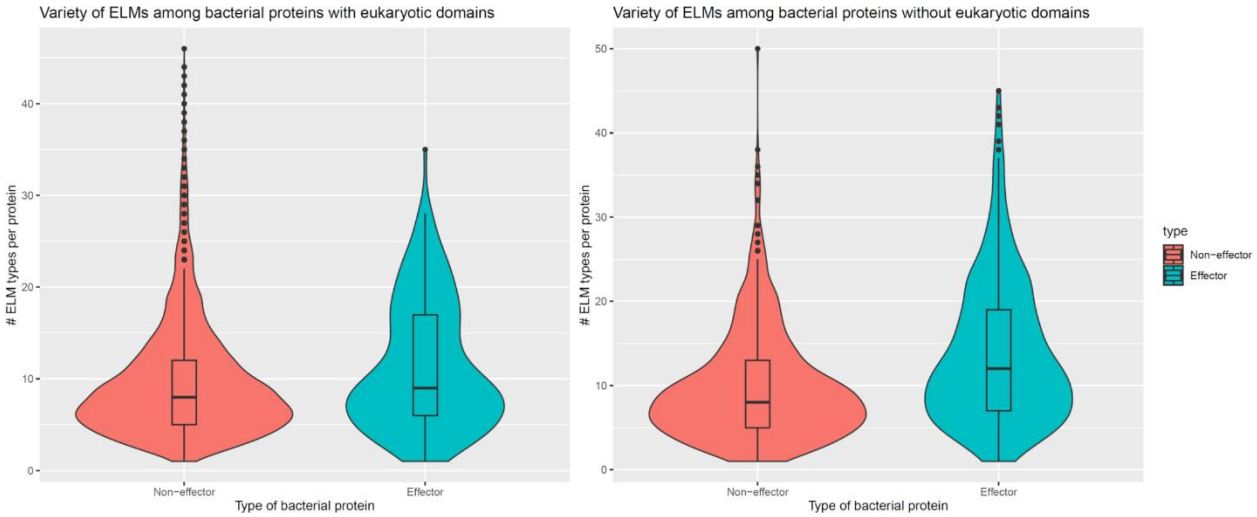


Figure 4.4 Effectors are enriched for bacteria-exclusive domains that target host domains otherwise exclusively involved in host-endogenous DDIs.

Among pathogen proteins with the potential to convergently target host domains that mediate DDIs in eukaryotes, 77% effectors and 27% non-effectors target host domains that mediate DDIs exclusively in eukaryotes, suggesting that effectors are 9 times as likely as non-effectors to disrupt eukaryote-specific processes via convergent evolution (Fisher's exact test, two-tailed $P = 4 * 10^{-5}$).

In addition to encoding globular domains that either mimic or convergently target host domains, effectors also encode short linear motifs that bind to host domains with similar specificities as host-endogenous proteins, while sharing little homology with the latter (Samano-Sanchez and Gibson 2020). These short linear motifs follow particular sequence patterns and are predominantly located in intrinsically disordered regions of proteins that are accessible to interacting partners (Davey, Van Roey *et al.* 2012). To determine whether effectors are enriched for host-interacting motifs, we counted the number of unique classes and instances of eukaryotic linear motifs (ELMs) (Kumar, Gouw *et al.* 2020) in long disordered regions of bacterial proteins (Piovesan, Tabaro *et al.* 2018). When comparing 162 effectors and 8,414 non-effectors with unique ELM compositions and containing eukaryotic-like domains, we found that effectors and non-effectors encode 9 and 8 ELM classes per protein (0.30 and 0.28 ELM instances per disordered residue), respectively. In other words, in the presence of eukaryotic-like domains, effectors encode a slightly higher variety (Wilcoxon test, two tailed $P = 3 * 10^{-3}$), but similar density of ELMs (Wilcoxon test, two tailed $P = 0.3$) compared to non-effectors. When comparing 521 effectors and 794 non-effectors with unique ELM compositions and not containing eukaryotic-like domains or any Pfam domains, however, we found that effectors and non-effectors encode 12 and 8 ELM classes per protein (0.31 and 0.27 ELM instances per disordered residue), respectively. In other words, in the absence of eukaryotic-like domains or among pathogen proteins without Pfam domains, effectors encode a higher variety (Wilcoxon test, two tailed $P < 2.2 * 10^{-16}$) as well as higher density of ELMs (Wilcoxon test, two tailed $P = 2 * 10^{-8}$) compared to non-effectors (Figure 5).

(A)



(B)

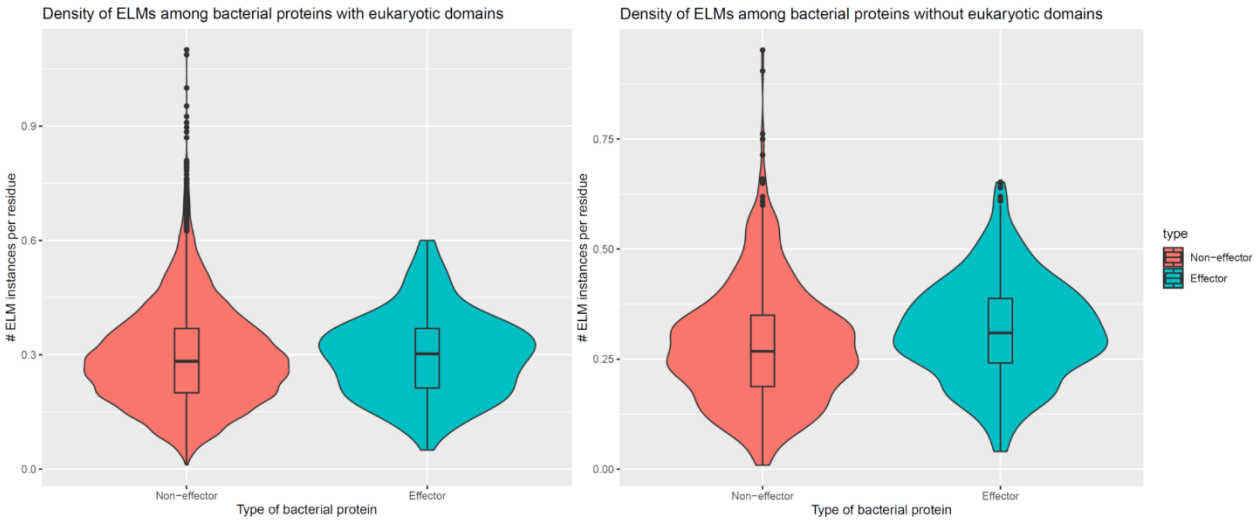


Figure 4.5 In the absence of eukaryotic-like domains or Pfam domains in general, effectors are enriched for eukaryotic linear motifs.

Compared to the rest of the pathogen proteome, effectors are enriched for eukaryotic linear motifs (ELMs) that target host domains. (A) Effectors encode more unique types of ELMs per protein; (B) Effectors encode more ELM instances per amino acid residue.

4.4 Discussion

Pathogenic bacteria have evolved a plethora of strategies to survive and thrive in eukaryotic hosts. A key strategy is functional mimicry of host activities, which is achieved through one of two orthogonal evolutionary mechanisms: horizontal acquisition of eukaryotic domains or convergent evolution of bacteria-exclusive domains. Based on a domain-resolved eukaryote-bacteria structural interaction network, we assessed domains' relative participation in eukaryote-specific DDIs, as opposed to DDIs conserved between eukaryotes and bacteria. We found that compared to the rest of the pathogen proteome, effector proteins are significantly enriched for domains and motifs that either mimic or convergently target host domains involved in eukaryote-specific DDIs. In the absence of eukaryotic-like domains or in the case of highly disordered proteins without Pfam domains, motif-based analysis can complement domain-based analysis in assessing the potential of a pathogen protein to disrupt host-endogenous PPIs.

To identify eukaryotic domains most likely acquired by bacteria for the express purpose of mimicking host-endogenous PPIs, we excluded domains which, while not mediating PPIs between different bacterial proteins, can nonetheless form crystal contacts with other domains within the same bacterial protein. Domains involved in such intraprotein DDIs in bacteria may serve structural functions, rather than engaging in host-pathogen PPIs. A case in point is the Fibronectin type III domain (PF00041), which in animals is involved in cell adhesion, migration and differentiation, and whose interaction with 44 domains leads to 1,156 interactions among 738 proteins in eukaryotes. While PF00041 does not mediate PPIs between bacterial proteins, it forms crystal contacts with the domain PF00704 within the *Bacillus thuringiensis* chitinase protein (PDB: 6BT9), and likely acts as a linker in the multi-domain chitinase (Juarez-Hernandez, Casados-Vazquez *et al.* 2019), rather than engaging in host-pathogen PPIs.

By relating the structural components of bacterial effectors to their propensity for repurposing or disrupting eukaryote-specific cellular processes, our study provides novel mechanistic and quantitative insight into the means by which pathogens hijack the host molecular machinery. Given the scarcity of host-bacteria PPI data and the rapidly increasing number of completely sequenced pathogen genomes, our framework for assessing the functional impact of structural modules within pathogen proteins, without needing direct experimental evidence of their interaction with host proteins, may help accelerate the discovery and mechanistic study of novel virulence factors, as well as the development of selective inhibitors of pathogen-subverted host signaling pathways.

4.5 Materials and Methods

4.5.1 Domain-resolved eukaryote-bacteria structural interaction network

Eukaryote-endogenous, bacteria-endogenous, and host-bacteria protein-protein interaction (PPI) data were obtained from IntAct and HPIDB 3.0 (Orchard, Ammari *et al.* 2014, Ammari, Gresham *et al.* 2016). Domain-domain interaction (DDI) templates were obtained from 3did and Pfam (Mosca, Ceol *et al.* 2014, El-Gebali, Mistry *et al.* 2019). Each PPI was resolved into DDIs between Pfam domains of the interacting proteins. When a PPI can be mediated by several possible DDIs, we gave the highest confidence to interchain DDIs (*i.e.* derived from PDB structures consisting of at least two distinct protein entities), followed by intrachain DDIs.

4.5.2 Inclusion criteria for effector and non-effector proteins

We included in our study proteins encoded by pathogenic bacterial species catalogued in PHI-base (Urban, Cuzick *et al.* 2020). Effector protein IDs were retrieved from UniProt (UniProt 2019) using two sets of keywords. **By gene name:** taxonomy:"Bacteria [2]" name:effector (name:"type 1" OR name:"type 2" OR name:"type 3" OR name:"type 4" OR name:"type 5" OR name:"type 6" OR name:"type 7" OR name:"type 8" OR name:"type 9" OR name:t*ss OR name:"secretion

system") **By cellular location:** taxonomy:"Bacteria [2]" (annotation:(type:function effector) OR locations:(note:"type 1") OR locations:(note:"type 2") OR locations:(note:"type 3") OR locations:(note:"type 4") OR locations:(note:"type 5") OR locations:(note:"type 6") OR locations:(note:"type 7") OR locations:(note:"type 8") OR locations:(note:"type 9") OR locations:(note:t*ss) OR locations:(note:"secretion system")) (locations:(location:"Secreted [SL-0243]") OR locations:(location:"Host [SL-0431]")) Non-effectors consist of cytoplasmic, membrane as well as other secreted proteins. **Cytoplasmic:** taxonomy:"Bacteria [2]" locations:(location:"Cytoplasm [SL-0086]") **Membrane:** taxonomy:"Bacteria [2]" (locations:(location:"Cell envelope [SL-0036]") OR locations:(location:"Membrane [SL-0162]")) **Secreted:** taxonomy:"Bacteria [2]" (locations:(location:"Secreted [SL-0243]") OR locations:(location:"Host [SL-0431]"))

4.5.3 Merging bacterial proteins with identical domain compositions

Taxonomy and Pfam domain annotations of proteins were obtained from UniProt and InterPro (Mitchell, Attwood *et al.* 2019). For each domain, we counted the number of eukaryotic and bacterial species encoding at least one protein containing that domain. To minimize the impact of spurious domains, such as arising from contaminated genomes or misannotated proteins, we required that each domain be found in at least three eukaryotic or bacterial proteomes – at least one of which must be a reference proteome or belong to a pan proteome. Bacterial proteins with identical domain compositions were merged into a single entry, as they are indistinguishable from one another at the domain resolution. To further reduce redundancy among highly related protein sequences (*e.g.* orthologs or fragments of the same protein) while also maintaining sufficient resolution, sequences belonging to the same UniRef50 cluster were ranked based on whether they are: (1) representative for the cluster; (2) manually reviewed; (3) assigned high annotation score

by UniProt; (4) from UniProt reference proteomes; and (5) longest. Only the top-ranking sequence was retained for each UniRef50 cluster. For domain compositions that are common to effectors and non-effectors, we assessed their relative frequency in effectors *vs.* non-effectors. Domain compositions that are significantly enriched (q-value < 0.1) in effectors were assigned to effectors, and domain compositions that are significantly depleted (q-value < 0.1) in effectors were assigned to non-effectors. Our final dataset thus contains 238 effectors and 3,921 non-effectors with unique domain signatures.

4.6 References

Ammari, M. G., C. R. Gresham, F. M. McCarthy and B. Nanduri (2016). "HPIDB 2.0: a curated database for host-pathogen interactions." Database (Oxford) **2016**.

Angot, A., N. Peeters, E. Lechner, F. Vailleau, C. Baud, L. Gentzbittel, E. Sartorel, P. Genschik, C. Boucher and S. Genin (2006). "Ralstonia solanacearum requires F-box-like domain-containing type III effectors to promote disease on several host plants." Proc Natl Acad Sci U S A **103**(39): 14620-14625.

Arnold, R., K. Boonen, M. G. Sun and P. M. Kim (2012). "Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space." Methods **57**(4): 508-518.

Davey, N. E., K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel and T. J. Gibson (2012). "Attributes of short linear motifs." Mol Biosyst **8**(1): 268-281.

El-Gebali, S., J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto and R. D. Finn (2019). "The Pfam protein families database in 2019." Nucleic Acids Res **47**(D1): D427-D432.

Franzosa, E. A. and Y. Xia (2011). "Structural principles within the human-virus protein-protein interaction network." Proc Natl Acad Sci U S A **108**(26): 10538-10543.

Fu, Y. and J. E. Galan (1998). "Identification of a specific chaperone for SptP, a substrate of the centisome 63 type III secretion system of Salmonella typhimurium." J Bacteriol **180**(13): 3393-3399.

Galan, J. E. (2009). "Common themes in the design and function of bacterial effectors." Cell Host Microbe **5**(6): 571-579.

Garamszegi, S., E. A. Franzosa and Y. Xia (2013). "Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks." PLoS Pathog **9**(12): e1003778.

Grau-Bove, X., A. Sebe-Pedros and I. Ruiz-Trillo (2015). "The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin." Mol Biol Evol **32**(3): 726-739.

Huang, Z., S. E. Sutton, A. J. Wallenfang, R. C. Orchard, X. Wu, Y. Feng, J. Chai and N. M. Alto (2009). "Structural insights into host GTPase isoform selection by a family of bacterial GEF mimics." Nat Struct Mol Biol **16**(8): 853-860.

Janjusevic, R., R. B. Abramovitch, G. B. Martin and C. E. Stebbins (2006). "A bacterial inhibitor of host programmed cell death defenses is an E3 ubiquitin ligase." Science **311**(5758): 222-226.

Jehl, M. A., R. Arnold and T. Rattei (2011). "Effective--a database of predicted secreted bacterial proteins." Nucleic Acids Res **39**(Database issue): D591-595.

Juarez-Hernandez, E. O., L. E. Casados-Vazquez, L. G. Brieba, A. Torres-Larios, P. Jimenez-Sandoval and J. E. Barboza-Corona (2019). "The crystal structure of the chitinase ChiA74 of Bacillus thuringiensis has a multidomain assembly." Sci Rep **9**(1): 2591.

Kumar, M., M. Gouw, S. Michael, H. Samano-Sanchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Calyseva, N. Palopoli, N. E. Davey, L. B. Chemes and T. J. Gibson (2020). "ELM-the eukaryotic linear motif resource in 2020." Nucleic Acids Res **48**(D1): D296-D306.

Marchesini, M. I., C. K. Herrmann, S. P. Salcedo, J. P. Gorvel and D. J. Comerci (2011). "In search of Brucella abortus type IV secretion substrates: screening and identification of four proteins translocated into host cells through VirB system." Cell Microbiol **13**(8): 1261-1274.

Matthews, L. R., P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent and M. Vidal (2001). "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"." Genome Res **11**(12): 2120-2126.

Mitchell, A. L., T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H. Y. Chang, S. El-Gebali, M. I. Fraser, J. Gough, D. R. Haft, H. Huang, I. Letunic, R. Lopez, A. Luciani, F. Madeira, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, G. Nuka, C. Orengo, A. P. Pandurangan, T. Paysan-Lafosse, S. Pesseat, S. C. Potter, M. A. Qureshi, N. D. Rawlings, N. Redaschi, L. J. Richardson, C. Rivoire, G. A. Salazar, A. Sangrador-Vegas, C. J. A. Sigrist, I. Sillitoe, G. G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, S. Y. Yong and R. D. Finn (2019). "InterPro in 2019: improving coverage, classification and access to protein sequence annotations." Nucleic Acids Res **47**(D1): D351-D360.

Mosca, R., A. Ceol, A. Stein, R. Olivella and P. Aloy (2014). "3did: a catalog of domain-based interactions of known three-dimensional structure." Nucleic Acids Res **42**(Database issue): D374-379.

Orchard, S., M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni and H. Hermjakob (2014). "The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases." Nucleic Acids Res **42**(Database issue): D358-363.

Piovesan, D., F. Tabaro, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztanyi, B. Meszaros, A. M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W. F.

Vranken and S. C. E. Tosatto (2018). "MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins." Nucleic Acids Res **46**(D1): D471-D476.

Popa, C. M., M. Tabuchi and M. Valls (2016). "Modification of Bacterial Effector Proteins Inside Eukaryotic Host Cells." Front Cell Infect Microbiol **6**: 73.

Samano-Sanchez, H. and T. J. Gibson (2020). "Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity." Trends Biochem Sci **45**(6): 526-544.

Scott, N. E. and E. L. Hartland (2017). "Post-translational Mechanisms of Host Subversion by Bacterial Effectors." Trends Mol Med **23**(12): 1088-1102.

Stebbins, C. E. and J. E. Galan (2001). "Structural mimicry in bacterial virulence." Nature **412**(6848): 701-705.

Steele-Mortimer, O., L. A. Knodler, S. L. Marcus, M. P. Scheid, B. Goh, C. G. Pfeifer, V. Duronio and B. B. Finlay (2000). "Activation of Akt/protein kinase B in epithelial cells by the Salmonella typhimurium effector sigD." J Biol Chem **275**(48): 37718-37724.

UniProt, C. (2019). "UniProt: a worldwide hub of protein knowledge." Nucleic Acids Res **47**(D1): D506-D515.

Urban, M., A. Cuzick, J. Seager, V. Wood, K. Rutherford, S. Y. Venkatesh, N. De Silva, M. C. Martinez, H. Pedro, A. D. Yates, K. Hassani-Pak and K. E. Hammond-Kosack (2020). "PHI-base: the pathogen-host interactions database." Nucleic Acids Res **48**(D1): D613-D620.

Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg and M. Vidal (2000). "Protein interaction mapping in C. elegans using proteins involved in vulval development." Science **287**(5450): 116-122.

Chapter 5: Discussion

An important goal of systems microbiology is to understand how pathogen-induced perturbation of the host protein interactome benefits the pathogen while contributing to disease in the host. Microbial virulence factors are ideal tools for probing the structural, functional, and evolutionary landscape of the host protein interactome. In this thesis, domain-resolved host-pathogen PPI networks are used to test the following hypotheses: (1) by causing similar perturbations to the human protein interactome at the domain level, genetic mutations and viral proteins are mechanistically equivalent contributors to diseases having both genetic and viral etiologic factors, *i.e.* virally-implicated diseases (VIDs); and (2) bacterial effector proteins are significantly enriched for structural modules that either mimic or convergently target host domains involved in eukaryote-specific cellular processes.

The main finding of Chapter 3 is that VID mutations are significantly enriched in human domains that are physically targeted (either directly or indirectly, by way of another human domain) or structurally mimicked by virus. In other words, VID mutations and viruses causing similar diseases tend to perturb the same domain-domain interactions in the human interactome and are therefore mechanistically equivalent. The general trend appears to be consistent across viruses; however, the extent of enrichment appears to differ slightly between viruses. There could be several reasons for this observation. One obvious explanation is investigator bias, which may lead to some genetic diseases having more established causal mutations, as well as some viruses having better domain annotation, more mapped human-virus PPIs, and epidemiologic studies supporting their causal role in more genetic diseases. Another plausible explanation is differences in pleiotropy among human proteins, where the same domain of a protein harbours mutations causing distinct types of

diseases. Whereas EBV and HPV cause various forms of cancer, many of which share similar disease genes and causal mutations, HIV contributes to diverse diseases with heterogeneous molecular mechanisms, including cancer, cardiomyopathy, hypertension, and Parkinson's disease. The observation that HIV-targeted domains appear less enriched for VID mutations (Figure 3.2) may be due to HIV-targeted domains being more pleiotropic. In other words, compared to domains targeted by EBV and HPV, HIV-targeted domains are more likely to be susceptible to both HIV-disease mutations as well as non-HIV-related disease mutations.

To establish whether a general equivalence exists between endogenous (cancer driver mutations) and exogenous (oncoviral proteins) perturbagens of oncogenic pathways, a pooled analysis was also conducted by treating all types of cancer as one class of disease, all oncomutations as one class of endogenous perturbagens, and all oncoviral proteins as one class of exogenous perturbagens of the human domain-resolved interactome. While this approach increases the statistical power of the analysis and reveals systems-level properties of oncomutations and oncoviral proteins, it may not be sensitive enough in scenarios where there is tissue specificity of oncogenic mutations or viral tropism for specific host tissue or cell type [73, 74]. While tissue-specific intra-species PPI data are available for model organisms and human [75], there is a lack of tissue-specific host-pathogen PPI data. There is evidence to suggest, however, that pathogens express different genes depending on the stage and site of infection. For instance, the opportunistic pathogen *Candida albicans* has been shown to activate different transcriptional programs in response to changes in host immunocompetence and tissue microenvironment, which allows it to transition from commensal to pathogen [76]. Similar dynamic regulation of pathogen and host gene expression by host-pathogen interplay has also been shown in *Staphylococcus aureus*, where differential expression of the virulence factor protein A determines whether the infection is

superficial or invasive [77], and in mice infected by the West Nile virus, where differential expression of genes involved in IFN and natural killer cell signalling determines tissue tropism for the spleen, but not the liver [78]. Therefore, pathogen perturbation of the host protein interactome likely hinges on the tissue-specific expression profiles of both pathogen virulence factors and host anti-virulence factors [79, 80]. It would be interesting to see whether domain-domain and domain-motif interactions can also be tissue-specific.

Domain-resolved interactomes offer distinct advantages over protein-level interactomes in explaining complex genotype-phenotype relationships, because mutations arising in different PPI-mediating domains of the same protein are likely to perturb its interaction with different PPI partners, often leading to distinct phenotypic outcomes [57]. Although there is statistically significant correlation between the distribution of VID mutations and whether VID mutation-carrying host domains are targeted or mimicked by virus, it is important to acknowledge that correlation does not imply causation. Causal inference of the etiologic role of viruses in complex diseases would require controlling for multiple confounding factors, which is beyond the scope of this thesis [81]. For instance, viruses may target only the non-membrane domains (*e.g.* found in cytoplasmic or extracellular compartments of the host cell) rather than the transmembrane domains of membrane-spanning host proteins. Enrichment of certain disease mutations in non-membrane domains may be due to factors unrelated to viral infection, such as physicochemical properties of amino acid residues, involvement of the domain in gene regulation, and so on [82]. Interestingly, these properties only emerge at the sub-protein level. As such, compared to previous work showing proximity between virus-targeted and VID-associated proteins in the human protein-level interactome, results of domain-level analyses are not only more informative of the molecular mechanism of disease, but also generate more precise hypotheses for further inquiries.

The main finding of Chapter 4 is that, compared to non-effector proteins encoded by pathogenic bacteria, effector proteins are significantly enriched for globular domains and short linear motifs that either structurally mimic or convergently target host domains involved in eukaryote-specific cellular processes, thereby allowing host-bacteria PPIs to mimic host-endogenous PPIs on an interactome scale. Consistent with findings for host-virus interactions, there is a statistically significant difference in the dominant evolutionary mechanism behind binding site sharing in the host-endogenous *vs.* host-bacteria PPI network; namely, that binding site sharing among host proteins largely results from gene duplication followed by divergent evolution, whereas binding site mimicry by bacterial proteins seems to largely result from convergent evolution (or extreme divergent evolution) of structural modules in bacteria, which bear little resemblance to those in host. Horizontally acquired, eukaryotic-like domains allow pathogens to repurpose host-endogenous PPIs, whereas convergently evolved, bacteria-exclusive domains and short linear motifs redirect host-endogenous PPIs – both of which benefit the pathogen at the expense of the host. Compared to previous studies of bacterial domains outside the context of host-bacteria and within-bacteria PPIs, an interaction-centric approach allows for quantitative assessment of the potential for bacterial domains to target host-endogenous signalling pathways.

While the eukaryote-bacteria interactome does contain more within-host PPIs and domain-domain interactions (DDIs) compared to within-bacteria or host-bacteria PPIs and DDIs, this imbalance should not confound the analysis, as domain assignment and DDI templates are not taxonomy-specific, but rather are used to resolve all PPIs, regardless of the species involved. In fact, estimation of domain's relevance to eukaryote-specific DDIs anticipates and accounts for DDIs that are exclusive to host species, by giving more weight to domains engaging in such DDIs. In Tables 4.1 and 4.3 showing examples of effectors containing domains mediating DDIs either

exclusively or primarily in eukaryotes, the fact that many domains can be traced to a few species is a technical consequence of proteins containing the same domains being merged into UniRef50 clusters, and only the species of the representative member of each cluster being retained. It is also a testament to extensive domain sharing among diverse pathogenic species. Taxonomic information may be useful when comparing effectors that are indistinguishable at the domain level but exhibit more variations at the residue level.

Pooled analysis of proteins with identical domain compositions across different species can reveal general patterns in the host-pathogen PPI network that may not be obvious on a species-by-species or protein-by-protein basis. On the one hand, host domains targeted by multiple effector domains can reveal convergent evolution of common virulence mechanisms among different pathogenic species, which may prove useful in developing broad spectrum antibiotics. For instance, the human Ras domain (PF00071) is targeted by structurally distinct domains in *Legionella* (PF14860, PF18172, PF18641), *Pseudomonas* (PF03496), *Salmonella* (PF03545, PF05925, PF07487), *Shigella* (PF03278) and *Yersinia* (PF00069, PF09632) effectors. On the other hand, effector domains targeting multiple host domains and thus potentially perturbing multiple host pathways represent targets for multipronged therapeutic intervention. Of the 103 host-targeting bacterial proteins in the eukaryote-bacteria interactome, 71 interact with a single host protein, while 32 interact with multiple host proteins. For instance, the *Pseudomonas* effector ExoS contains the ADP ribosyltransferase domain (PF03496), which it uses to target host proteins containing either a 14-3-3 domain (PF00244) or Ras small GTPase domain (PF00071). These host domains participate in a wide array of signalling pathways [83, 84].

To identify domains in bacteria that are most likely involved in mimicking host-endogenous PPIs, eukaryotic-like domains were excluded if they engage in either interprotein or intraprotein DDIs

in bacteria. Previous studies suggest that intrachain DDIs often occur between adjacent domains of the same protein [85]; however, PPIs are much less likely attributable to DDIs derived solely from intrachain interactions, as opposed to DDIs derived from interchain interactions [86]. Although distinguishing biologically relevant interfaces from artifactual crystal contacts is beyond the scope of this work, several interface classification algorithms have been developed to address this specific issue, based on various criteria such as contact size and evolutionary conservation of interface residues [87, 88], thermodynamic prediction of interface stability [89], and interface conservation across multiple crystal forms of a protein [90]. Here, a conservative approach was taken, which excludes all eukaryotic-like domains engaging in intraprotein DDIs in bacteria, because while they may not mediate PPIs in bacteria, it is not clear whether they evolved to specifically interact with host proteins, or function in maintaining protein stability or metabolic processes in bacteria [91].

Since pathogenic viruses and bacteria both hijack host pathways involved in immune response and cell proliferation, it would be interesting to see whether results based on virally-implicated diseases and human-virus PPIs can be extrapolated to bacteria. Towards this end, the domain-resolved host-bacteria interactome was queried for human domains convergently targeted by bacterial proteins and missense mutations leading to bacteria-implicated genetic diseases. For immunological diseases, convergently perturbed domains were only found in 4 human proteins. Similarly, for proliferative diseases, convergently perturbed domains were only found in 7 human proteins. Given such small sample sizes, it is not possible at this time to assess the statistical significance of putative mechanistic equivalence between bacterial and mutational perturbations of human domains. As more data on bacteria-implicated genetic diseases and host-bacteria interactions

emerge, the domain-based convergent perturbation model may, in the future, be extended to the study of diseases and co-infections involving multiple pathogenic species.

Perhaps adding to the complexity of host-microbe interactome analysis is interaction among the microbial perturbagens themselves (*e.g.* pathogenic bacteria and viruses *vs.* commensal bacteria), which may be synergistic or antagonistic in nature, and cause non-additive perturbations to the human interactome network [92]. For instance, HIV-infected individuals are at a higher risk for co-infections by other pathogens, with *Mycobacterium tuberculosis* (TB) being a major co-infecting agent. There is literature implicating the TNFα signalling pathway in the co-pathogenesis of TB and HIV [93]. Query of the domain-resolved host-pathogen interactome suggests that the protein kinase domains of human MAPK8 (JNK1) and MAPK14 (p38α), both of which are involved in the TNFα signalling pathway, are targets for both TB-encoded protein tyrosine phosphatase ptpA as well as HIV-encoded Nef protein. Understanding the precise mechanism of TB and HIV-mediated cross-regulation of cytokine signalling pathways would require further investigation. Meanwhile, a study by Cohen *et al.* identified an effector gene family (Cbeg12) in commensal bacteria, which encodes for an enzyme whose metabolic product, an N-acylated small molecule, structurally mimics eukaryotic signalling molecules involved in NF-κB and GPCR activation [94]. The authors suggest that commensal bacteria may use structural mimicry as a mechanism for mutualistic interactions with the host. While the receptor-ligand interaction involves the metabolic product of a bacterial effector, as opposed to the effector itself, and no mention is made of the host domain bound by the bacterial ligand, it is conceivable that commensal bacteria may encode effector proteins that structurally mimic eukaryotic signalling proteins, but exert opposite immunomodulatory effects in comparison to pathogenic viruses and bacteria.

Chapter 6: Conclusion

In summary, domain-resolved host-pathogen PPI networks are useful for examining the molecular mechanisms of diseases with both genetic and viral components, as well as the evolutionary origin of host-interacting structural modules in bacteria. As recurrent and emerging infectious diseases pose a tremendous risk to public health, there is a strong need for unbiased, systematic investigation of the molecular mechanisms of host-pathogen interactions. While experimentally determined PPIs may be biased towards well-studied pathogen species, and domain-domain interaction templates may be biased towards well-defined protein structures, domain-level interactome analysis encompassing multiple pathogen and host species is nonetheless more powerful in uncovering general molecular recognition principles underlying the host-pathogen PPI network, compared to protein-level analysis focusing on one species at a time. To increase coverage of the host-pathogen structural interaction network, future efforts should focus on genome annotation of emerging pathogens, more systematic mapping of host-pathogen protein interactomes, as well as new molecular modelling methods to predict structures of proteins and PPIs which do not have homologs with known structure [95].

As discussed earlier, a major challenge in host-microbe interactome analysis is accounting for the presence of multiple microbial species, whose gene expression, and protein interactions with host, as well as amongst themselves, may be spatiotemporally regulated. Indeed, the emerging field of differential network biology aims to capture dynamic topological changes in the interactomes of single or multiple organisms as they adapt to genetic mutations, environmental stress and interspecies interactions [92, 96-98]. In this sense, pathogen proteins and host SNPs associated with susceptibility to infection can serve as candidate perturbagens for higher-order edgetic

perturbation experiments [99-102]. For instance, pathogen proteins and their host targets can be subjected to directed evolution experiments to identify gain- or loss-of-function variants that alter the binding affinity and specificity of host-pathogen PPIs. When creating mutant libraries, priority can be given to interacting domains and interface residues identified in structurally-resolved host-pathogen PPI networks, which may be enriched for mutations that alter binding properties. Multiple sequence alignments between pathogen proteins having similar functions may shed light on the molecular determinants and evolution of virulence.

While the cost of genome sequencing has become less prohibitive and community efforts such as 1000 Genomes [103] and the Human Interactome Project [37] continue to provide a clearer picture of the human genetic and physical interactome, the vast number of theoretically possible SNPs and PPIs, along with the difficulty of experimental determination of protein structure, rationalize the use of structural interaction networks built from template-based models in exploring the systems biology of genetic and infectious diseases, as well as dynamic changes in the human interactome upon perturbation. Although predicting new host-pathogen interactions is beyond the scope of this thesis, the work presented here is a first step towards resolving PPI interfaces in host-pathogen interactions: once domains involved in host-pathogen PPIs are identified, interface residues inside such domains can be predicted using machine learning methods [104]. It will then be possible to determine whether the interactome perturbation model can be generalized to the residue level, *i.e.* whether missense mutations and viruses inducing similar disease phenotypes convergently target the same interface residues. For host-pathogen PPIs without close homologues at the sequence level, structural models can be built using threading methods. Such models can "salvage" host-pathogen PPI interfaces and genetic disease mutations located outside Pfam domains, which are not considered in the current study. Furthermore, it may be possible to design drugs that precisely

inhibit host-pathogen PPIs, with minimal disruption to host-endogenous PPIs [105]. Given the scarcity of host-pathogen PPI data and the rapidly increasing number of completely sequenced pathogen genomes, the framework presented here for estimating the potential of pathogen domains to target eukaryote-specific cellular processes, without needing direct experimental evidence of their interaction with host domains, may help accelerate the discovery and mechanistic study of novel virulence factors, as well as the development of selective inhibitors of pathogen-subverted host signalling pathways.

Glossary

Term	Definition
Interactome	Usually refers to protein-protein interaction network, although generalizable to any interaction network
Domain	Conserved structural unit that can fold, function and evolve independently of the rest of a protein, often acting as protein-protein interaction module
Linear motif	Short stretch of adjacent amino acids in a protein sequence, often acting as protein-protein interaction module
Oncovirus	Virus causally linked to human cancer and activates common molecular hallmarks of cancer
Oncovirus-targeted domain (OVTD)	Human domain physically interacting with oncoviral proteins
Oncoviral homology domain (OVHD)	Human domain with viral homologue either exclusively occurring or enriched in proteomes of oncogenic viruses
Generic viral homology domain (GVHD)	Human domain with viral homologue either exclusively occurring or enriched in proteomes of non-oncogenic viruses
Effector	Virulence factor secreted by pathogenic bacteria and injected directly into host cytoplasm via specialized secretion systems

References

1. Franzosa, E.A. and Y. Xia, *Structural principles within the human-virus protein-protein interaction network.* Proc Natl Acad Sci U S A, 2011. **108**(26): p. 10538-43.

2. Cazalet, C., et al., *Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity.* Nat Genet, 2004. **36**(11): p. 1165-73.

3. Schweppe, D.K., et al., *Host-Microbe Protein Interactions during Bacterial Infection.* Chem Biol, 2015. **22**(11): p. 1521-1530.

4. Garamszegi, S., E.A. Franzosa, and Y. Xia, *Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks.* PLoS Pathog, 2013. **9**(12): p. e1003778.

5. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.

6. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.

7. Stebbins, C.E. and J.E. Galan, *Structural mimicry in bacterial virulence.* Nature, 2001. **412**(6848): p. 701-5.

8. Mesri, E.A., M.A. Feitelson, and K. Munger, *Human viral oncogenesis: a cancer hallmarks analysis.* Cell Host Microbe, 2014. **15**(3): p. 266-82.

9. Beyer, A., S. Bandyopadhyay, and T. Ideker, *Integrating physical and genetic maps: from genomes to interaction networks.* Nat Rev Genet, 2007. **8**(9): p. 699-710.

10. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.

11. Costanzo, M., et al., *The genetic landscape of a cell.* Science, 2010. **327**(5964): p. 425-31.

12. Jager, S., et al., *Global landscape of HIV-human protein complexes.* Nature, 2012. **481**(7381): p. 365-70.

13. Blasche, S., et al., *The EHEC-host interactome reveals novel targets for the translocated intimin receptor.* Sci Rep, 2014. **4**: p. 7531.

14. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.

15. Bruckner, A., et al., *Yeast two-hybrid, a powerful tool for systems biology.* Int J Mol Sci, 2009. **10**(6): p. 2763-88.

16. Hirst, M., et al., *A two-hybrid system for transactivator bait proteins.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8726-31.

17. Broder, Y.C., S. Katz, and A. Aronheim, *The ras recruitment system, a novel approach to the study of protein-protein interactions.* Curr Biol, 1998. **8**(20): p. 1121-4.

18. Osborne, M.A., S. Dalton, and J.P. Kochan, *The yeast tribrid system--genetic detection of trans-phosphorylated ITAM-SH2-interactions.* Biotechnology (N Y), 1995. **13**(13): p. 1474-8.

19. Guo, D., et al., *A tethered catalysis, two-hybrid system to identify protein-protein interactions requiring post-translational modifications.* Nat Biotechnol, 2004. **22**(7): p. 888-92.

20. Serebriiskii, I.G. and E.A. Golemis, *Two-hybrid system and false positives. Approaches to detection and elimination.* Methods Mol Biol, 2001. **177**: p. 123-34.

21. Huang, H., B.M. Jedynak, and J.S. Bader, *Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps.* PLoS Comput Biol, 2007. **3**(11): p. e214.

22. Koegl, M. and P. Uetz, *Improving yeast two-hybrid screening systems.* Brief Funct Genomic Proteomic, 2007. **6**(4): p. 302-12.

23. Luo, Y., et al., *Mammalian two-hybrid system: a complementary approach to the yeast two-hybrid system.* Biotechniques, 1997. **22**(2): p. 350-2.

24. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.

25. Gingras, A.C., et al., *Analysis of protein complexes using mass spectrometry.* Nat Rev Mol Cell Biol, 2007. **8**(8): p. 645-54.

26. Rashid, K.A., et al., *A proteomic approach identifies proteins in hepatocytes that bind nascent apolipoprotein B.* J Biol Chem, 2002. **277**(24): p. 22010-7.

27. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.

28. Morris, J.H., et al., *Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions.* Nat Protoc, 2014. **9**(11): p. 2539-54.

29. Kemmeren, P., et al., *Protein interaction verification and functional annotation by integrated analysis of genome-scale data.* Mol Cell, 2002. **9**(5): p. 1133-43.

30. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.

31. Matthews, L.R., et al., *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".* Genome Res, 2001. **11**(12): p. 2120-6.

32. Yu, H., et al., *High-quality binary protein interaction map of the yeast interactome network.* Science, 2008. **322**(5898): p. 104-10.

33. Fromont-Racine, M., J.C. Rain, and P. Legrain, *Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens.* Nat Genet, 1997. **16**(3): p. 277-82.

34. Walhout, A.J., et al., *Protein interaction mapping in C. elegans using proteins involved in vulval development.* Science, 2000. **287**(5450): p. 116-22.

35. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.

36. Lehner, B. and A.G. Fraser, *A first-draft human protein-interaction map.* Genome Biol, 2004. **5**(9): p. R63.

37. Rolland, T., et al., *A proteome-scale map of the human interactome network.* Cell, 2014. **159**(5): p. 1212-26.

38. Zhong, Q., et al., *Edgetic perturbation models of human inherited disorders.* Mol Syst Biol, 2009. **5**: p. 321.

39. Sahni, N., et al., *Widespread macromolecular interaction perturbations in human genetic disorders.* Cell, 2015. **161**(3): p. 647-60.

40. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.

41. Das, J., et al., *Exploring mechanisms of human disease through structurally resolved protein interactome networks.* Mol Biosyst, 2014. **10**(1): p. 9-17.

42. Launay, G. and T. Simonson, *Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations.* BMC Bioinformatics, 2008. **9**: p. 427.

43. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.

44. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

45. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.

46. John, B. and A. Sali, *Comparative protein structure modeling by iterative alignment, model building and model assessment.* Nucleic Acids Res, 2003. **31**(14): p. 3982-92.

47. Aloy, P. and R.B. Russell, *Structural systems biology: modelling protein interactions.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 188-97.

48. Mosca, R., et al., *3did: a catalog of domain-based interactions of known three-dimensional structure.* Nucleic Acids Res, 2014. **42**(Database issue): p. D374-9.

49. Finn, R.D., et al., *iPfam: a database of protein family and domain interactions found in the Protein Data Bank.* Nucleic Acids Res, 2014. **42**(Database issue): p. D364-73.

50. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nat Protoc, 2015. **10**(6): p. 845-58.

51. Dalton, J.A. and R.M. Jackson, *An evaluation of automated homology modelling methods at low target template sequence similarity.* Bioinformatics, 2007. **23**(15): p. 1901-8.

52. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.

53. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

54. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.

55. Peng, J. and J. Xu, *RaptorX: exploiting structure information for protein alignment by statistical inference.* Proteins, 2011. **79 Suppl 10**: p. 161-71.

56. Mosca, R., A. Ceol, and P. Aloy, *Interactome3D: adding structural details to protein networks.* Nat Methods, 2013. **10**(1): p. 47-53.

57. Wang, X., et al., *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nat Biotechnol, 2012. **30**(2): p. 159-64.

58. Bandyopadhyay, S., et al., *Rewiring of genetic networks in response to DNA damage.* Science, 2010. **330**(6009): p. 1385-9.

59. Taylor, I.W., et al., *Dynamic modularity in protein interaction networks predicts breast cancer outcome.* Nat Biotechnol, 2009. **27**(2): p. 199-204.

60. Angot, A., et al., *Ralstonia solanacearum requires F-box-like domain-containing type III effectors to promote disease on several host plants.* Proc Natl Acad Sci U S A, 2006. **103**(39): p. 14620-5.

61. Huang, Z., et al., *Structural insights into host GTPase isoform selection by a family of bacterial GEF mimics.* Nat Struct Mol Biol, 2009. **16**(8): p. 853-60.

62. Eichinger, V., et al., *EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems.* Nucleic Acids Res, 2016. **44**(D1): p. D669-74.

63. Samano-Sanchez, H. and T.J. Gibson, *Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity.* Trends Biochem Sci, 2020. **45**(6): p. 526-544.

64. Crua Asensio, N., et al., *Centrality in the host-pathogen interactome is associated with pathogen fitness during infection.* Nat Commun, 2017. **8**: p. 14092.

65. de Groot, N.S. and M. Torrent Burgas, *Bacteria use structural imperfect mimicry to hijack the host interactome.* PLoS Comput Biol, 2020. **16**(12): p. e1008395.

66. Gulbahce, N., et al., *Viral perturbations of host networks reflect disease etiology.* PLoS Comput Biol, 2012. **8**(6): p. e1002531.

67. Rozenblatt-Rosen, O., et al., *Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins.* Nature, 2012. **487**(7408): p. 491-5.

68. Cronin, M., et al., *Bacterial-mediated knockdown of tumor resistance to an oncolytic virus enhances therapy.* Mol Ther, 2014. **22**(6): p. 1188-97.

69. Roberts, N.J., et al., *Intratumoral injection of Clostridium novyi-NT spores induces antitumor responses.* Sci Transl Med, 2014. **6**(249): p. 249ra111.

70. Varghese, S. and S.D. Rabkin, *Oncolytic herpes simplex virus vectors for cancer virotherapy.* Cancer Gene Ther, 2002. **9**(12): p. 967-78.

71. Yu, Y.A., et al., *Visualization of tumors and metastases in live animals with bacteria and vaccinia virus encoding light-emitting proteins.* Nat Biotechnol, 2004. **22**(3): p. 313-20.

72. Mahoney, D.J., D.F. Stojdl, and G. Laird, *Virus therapy for cancer.* Sci Am, 2014. **311**(5): p. 54-9.

73. Haigis, K.M., K. Cichowski, and S.J. Elledge, *Tissue-specificity in cancer: The rule, not the exception.* Science, 2019. **363**(6432): p. 1150-1151.

74. McFadden, G., et al., *Cytokine determinants of viral tropism.* Nat Rev Immunol, 2009. **9**(9): p. 645-55.

75. Kotlyar, M., et al., *Integrated interactions database: tissue-specific view of the human and model organism interactomes.* Nucleic Acids Res, 2016. **44**(D1): p. D536-41.

76. Hube, B., *From commensal to pathogen: stage- and tissue-specific gene expression of Candida albicans.* Curr Opin Microbiol, 2004. **7**(4): p. 336-41.

77. Loughman, J.A., et al., *Virulence gene expression in human community-acquired Staphylococcus aureus infection.* J Infect Dis, 2009. **199**(3): p. 294-301.

78. Suthar, M.S., et al., *A systems biology approach reveals that tissue tropism to West Nile virus is regulated by antiviral genes and innate immune cellular processes.* PLoS Pathog, 2013. **9**(2): p. e1003168.

79. Lovegrove, F.E., et al., *Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria.* BMC Genomics, 2006. **7**: p. 295.

80. Teixeira, P.J., et al., *High-resolution transcript profiling of the atypical biotrophic interaction between Theobroma cacao and the fungal pathogen Moniliophthora perniciosa.* Plant Cell, 2014. **26**(11): p. 4245-69.

81. Foxman, E.F. and A. Iwasaki, *Genome-virome interactions: examining the role of common viral infections in complex disease.* Nat Rev Microbiol, 2011. **9**(4): p. 254-64.

82. Partridge, A.W., A.G. Therien, and C.M. Deber, *Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease.* Proteins, 2004. **54**(4): p. 648-56.

83. Xiao, B., et al., *Structure of a 14-3-3 protein and implications for coordination of multiple signalling pathways.* Nature, 1995. **376**(6536): p. 188-91.

84. Stenmark, H. and V.M. Olkkonen, *The Rab GTPase family.* Genome Biol, 2001. **2**(5): p. REVIEWS3007.

85. Littler, S.J. and S.J. Hubbard, *Conservation of orientation and sequence in protein domain--domain interactions.* J Mol Biol, 2005. **345**(5): p. 1265-79.

86. Itzhaki, Z., et al., *Evolutionary conservation of domain-domain interactions.* Genome Biol, 2006. **7**(12): p. R125.

87. Valdar, W.S. and J.M. Thornton, *Conservation helps to identify biologically relevant crystal contacts.* J Mol Biol, 2001. **313**(2): p. 399-416.

88. Duarte, J.M., et al., *Protein interface classification by evolutionary analysis.* BMC Bioinformatics, 2012. **13**: p. 334.

89. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J Mol Biol, 2007. **372**(3): p. 774-97.

90. Xu, Q., et al., *Statistical analysis of interface similarity in crystals of homologous proteins.* J Mol Biol, 2008. **381**(2): p. 487-507.

91. Tsoka, S. and C.A. Ouzounis, *Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion.* Nat Genet, 2000. **26**(2): p. 141-2.

92. Guven-Maiorov, E., C.J. Tsai, and R. Nussinov, *Structural host-microbiota interaction networks.* PLoS Comput Biol, 2017. **13**(10): p. e1005579.

93. Patel, N.R., et al., *HIV impairs TNF-alpha mediated macrophage apoptotic response to Mycobacterium tuberculosis.* J Immunol, 2007. **179**(10): p. 6973-80.

94. Cohen, L.J., et al., *Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist.* Proc Natl Acad Sci U S A, 2015. **112**(35): p. E4825-34.

95. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**(2): p. 547-56.

96. Ideker, T. and N.J. Krogan, *Differential network biology.* Mol Syst Biol, 2012. **8**: p. 565.

97. Fischbach, M.A. and N.J. Krogan, *The next frontier of systems biology: higher-order and interspecies interactions.* Genome Biol, 2010. **11**(5): p. 208.

98. Lambert, J.P., et al., *Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition.* Nat Methods, 2013. **10**(12): p. 1239-45.

99. Kumar, R. and B. Nanduri, *HPIDB--a unified resource for host-pathogen interactions.* BMC Bioinformatics, 2010. **11 Suppl 6**: p. S16.

100. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D358-63.

101. Stenson, P.D., et al., *The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution.* Curr Protoc Bioinformatics, 2012. **Chapter 1**: p. Unit1 13.

102. Amberger, J.S., et al., *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders.* Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98.

103. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

104. Meyer, M.J., et al., *Interactome INSIDER: a structural interactome browser for genomic studies.* Nat Methods, 2018. **15**(2): p. 107-114.

105. Voter, A.F. and J.L. Keck, *Development of Protein-Protein Interaction Inhibitors for the Treatment of Infectious Diseases.* Adv Protein Chem Struct Biol, 2018. **111**: p. 197-222.