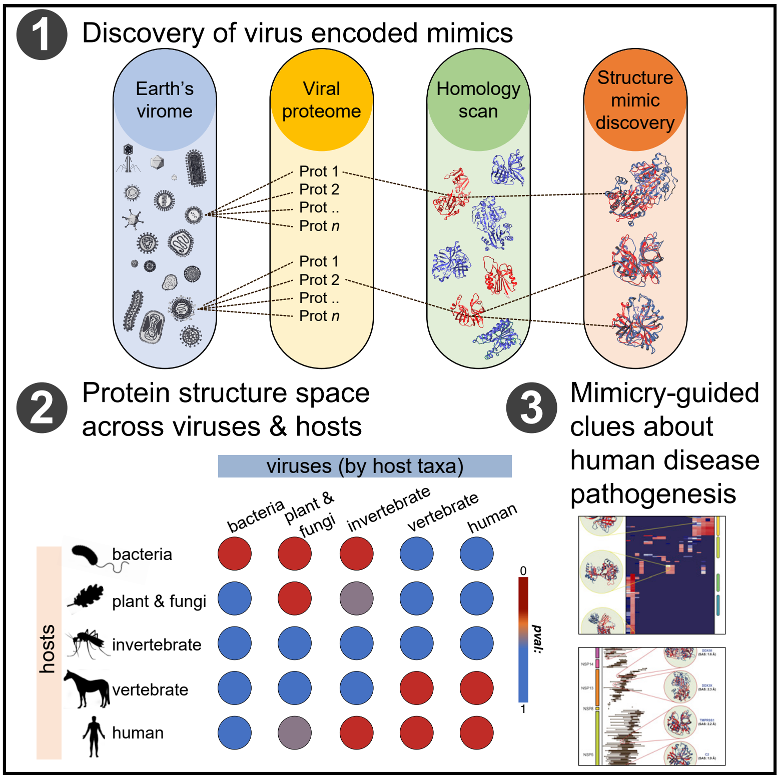
Cell Systems

Report

A Sweep of Earth's Virome Reveals Host-Guided Viral Protein Structural Mimicry and Points to Determinants of Human Disease

Graphical Abstract



Authors

Gorka Lasso, Barry Honig, Sagi D. Shapira

Correspondence

ss4197@columbia.edu

In Brief

Virally encoded factors resembling host factors hijack host molecular machines in a phenomenon called molecular mimicry. Lasso et al. use 3D protein structural information to identify over 6,000,000 instances of structural mimicry, >70% of which cannot be discerned through protein sequence alone. Their work provides the first systematic analysis of molecular mimicry across the earth's virome.

Highlights

- Structural mimicry of host proteins is a pervasive strategy across earth's virome
- Majority of the >6,000,000 cataloged mimics cannot be discerned through sequence
- Protein structure space utilized by viruses is dictated by host proteomes
- Knowledge of viral mimics provides clues about pathophysiology of COVID-19

Lasso et al., 2021, Cell Systems *11*, 82–91
January 20, 2021 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.09.006

CellPress

Cell Systems

Report

A Sweep of Earth's Virome Reveals Host-Guided Viral Protein Structural Mimicry and Points to Determinants of Human Disease

Gorka Lasso,[1,2,3] Barry Honig,[1,4,5,6] and Sagi D. Shapira[1,2,7,*]

[1]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA
[2]Department of Microbiology and Immunology, Columbia University Medical Center, New York, NY, USA
[3]Department of Microbiology and Immunology, Albert Einstein College of Medicine, New York, NY, USA
[4]Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY, USA
[5]Zuckerman Mind Brain Behavior Institute, Columbia University Medical Center, New York, NY, USA
[6]Department of Medicine, Columbia University, New York, NY, USA
[7]Lead Contact
*Correspondence: ss4197@columbia.edu

https://doi.org/10.1016/j.cels.2020.09.006

SUMMARY

Viruses deploy genetically encoded strategies to coopt host machinery and support viral replicative cycles. Here, we use protein structure similarity to scan for molecular mimicry, manifested by structural similarity between viral and endogenous host proteins, across thousands of cataloged viruses and hosts spanning broad ecological niches and taxonomic range, including bacteria, plants and fungi, invertebrates, and vertebrates. This survey identified over 6,000,000 instances of structural mimicry; more than 70% of viral mimics cannot be discerned through protein sequence alone. We demonstrate that the manner and degree to which viruses exploit molecular mimicry varies by genome size and nucleic acid type and identify 158 human proteins that are mimicked by coronaviruses, providing clues about cellular processes driving pathogenesis. Our observations point to molecular mimicry as a pervasive strategy employed by viruses and indicate that the protein structure space used by a given virus is dictated by the host proteome.

A record of this paper's transparent peer review process is included in the Supplemental Information.

INTRODUCTION

Viruses deploy an array of genetically encoded strategies to coopt host machinery and support viral replicative cycles. Among the strategies, protein-protein interactions (PPIs), mediated by promiscuous, multifunctional viral proteins are widely documented. Targeted discovery tools focused largely on viruses of public-health importance and have experimentally mapped thousands of virus-host protein complexes, and structure-informed prediction algorithms have allowed discovery of such interactions across all fully sequenced human-infecting viruses (Lasso et al., 2019). Molecular mimicry, manifested by structural similarity between viral and endogenous host proteins, allows viruses to harness or disrupt cellular functions including nucleic acid metabolism and modulation of immune responses. Yet, while examples of this latter strategy pepper the literature (Wimmer and Schreiner, 2015; Alcami, 2003; Guven-Maiorov et al., 2016), most have focused on human-infecting viruses (Elde and Malik, 2009; Felix and Savvides, 2017), and a systematic analysis of pathogen-encoded molecular mimics has not been performed.

The vast genomic landscape occupied by viruses hampers the discovery of evolutionary relationships between viral proteins and their hosts. As is well known, however, since three-dimensional (3D) protein structure is much better conserved than sequence, structural information can be used to interrogate evolutionary relationships (Lasso et al., 2019; Bamford et al., 2005) as well as uncover virus-encoded structural mimics that cannot be detected by sequence relationships (see STAR Methods). Here, we use protein structure similarity to identify virally encoded mimics of host proteomes. Briefly, we first employ sequence-based methods to identify proteins that have similar structures to queried viral proteins and then use structural alignment to find "structural neighbors" of viral proteins (Figure 1A). We refer to the corresponding viral proteins as mimics of their host-encoded neighbors. We applied the approach to a set of 337,493 viral proteins representing 7,486 viruses across a broad host taxonomic range, including bacteria, plants and fungi, invertebrates, and vertebrates. Our survey identified over 6,000,000 instances of structural mimicry, the vast majority of which (>70%) cannot be discerned through protein sequence alone (see below). Our results point to molecular mimicry as a pervasive strategy employed by viruses and indicate that the protein structure space used by a given virus is dictated, at least in part, by the host proteome.

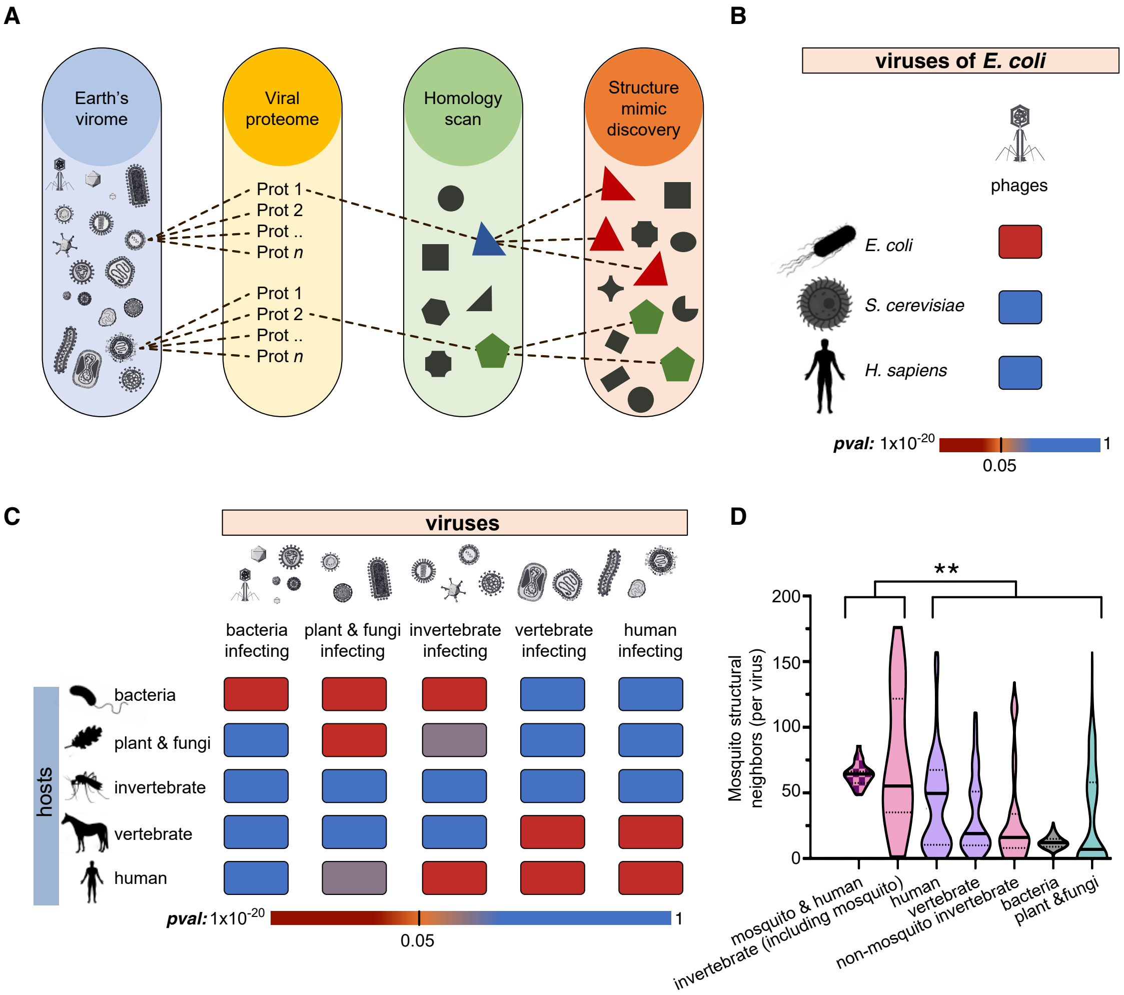


Figure 1. Identification and Characterization of Virus-Host Structural Relationships Reveals Utilization of Structure Space Guided by Host Taxonomic Division

(A) Graphical schematic of the strategy implemented to reveal virus-encoded protein mimics. Sequence-based searches are first used to identify proteins in the PDB that are structurally related to viral proteins. These intermediate, sequence-detected structural templates are then used, through structural alignment, to search the PDB for proteins that are geometrically similar to viral proteins—thus, revealing viral mimicry.

(B) Significance of overlap, measured in terms of numbers observed relationships (where SAS < 2.5 Å) between the structure space of *E. coli*-infecting phages and *E. coli*, *S. cerevisiae*, and human proteomes (see STAR Methods for details).

(C) Significance of overlap, measured in terms of numbers of observed relationships (where SAS < 2.5 Å) between the structure space of all cataloged viruses and hosts in different taxonomic divisions.

(D) Distribution of structural neighbors of the modeled mosquito proteome for different virus classes (brackets denote distributions used in Mann-Whitney, see methods; **p value < 0.0001).

We further observe that the manner and degree to which viruses exploit molecular mimicry varies by genome size and nucleic acid type. For example, while human-infecting and arthropod-infecting viruses occupy a structure space most similar to their host proteome, arboviruses, which are transmitted to humans by insect vectors, encode promiscuous proteins that mimic both human and insect proteins. In addition, we find that, relative to their proteome size, single-stranded RNA (ssRNA) viruses, including coronaviruses (CoVs), have circumvented the limitations of their small genomes by mimicking human proteins to a greater extent than their large dsDNA counterparts like Pox and Herpes viruses. Interrogation of proteins mimicked by human-infecting viruses points to broad diversification of cellular pathways targeted via structural mimicry, identifies biological processes that may underly autoimmune disorders, and reveals virally encoded mimics that may be leveraged to engineer synthetic metabolic circuits or may serve as targets for therapeutics. Finally, we identified over 150 cellular proteins (including members of the complement activation pathway and critical regulators of innate and adaptive immunity) that are mimicked by CoV, providing clues about the cellular processes underlying the pathogenesis driving the ongoing COVID-19 pandemic.

RESULTS

Mining the Virome-wide Structure Space to Identify Host Proteins Mimicked by Viruses

In order to identify mimicry relationships, we implemented the strategy summarized in Figure 1A. Sequence-based searches (see STAR Methods) of 337,493 viral proteins (269,077 unique sequences) against the PDB identified structural templates for 92,868 (34.5%), with large disparities in coverage: structural templates were identified for 64.7% of the proteins from human-infecting viruses, whereas the coverage for non-vertebrate-infecting viruses ranged from 31% to 38% (Figure S1A; Table S2). As shown in Figure S1B, structural templates for vertebrate and human-infecting viruses came primarily from other viruses while, for example, bacterial proteins provided most of the templates for bacteria-infecting viruses. Most of the templates for invertebrate-infecting viruses came from vertebrate and bacterial proteins, likely reflecting the limited structural coverage of non-vertebrate viruses and invertebrate hosts in the PDB. In order to measure structure similarity between protein structures we utilized Ska, an extensively utilized and validated tool for inference of structure-based functional relationships even in the absence of detectable sequence similarity (Garzón et al., 2016; Hwang et al., 2017; Lasso et al., 2019; Zhang et al., 2012; Petrey et al., 2003, 2009; Yang and Honig, 2000). In addition to Ska, we employed a conservative global structural similarity criteria (structural alignment score, SAS < 2.5 Å) (Budowski-Tal et al., 2010; Kolodny et al., 2005; Subbiah et al., 1993) to infer structural mimics and minimize biases imposed by local structural similarities (see STAR Methods). As shown in Figure S2, this approach enables rediscovery of known virus-encoded protein structure mimics. Structural alignment of the 92,868 templates with PDB proteins identified 6,083,167 mimicry relationships, involving 88,715 viral proteins and 26,542 host protein structures present in the PDB (data are available for download at https://github.com/gorkaLasso/VirusHost_mimicry). The vast majority of these mimicry relationships (72.2%) could not be retrieved through sequence similarity alone based on the requirement that no e value obtained from any one or three sequence-based methods was greater than 1×10^{-6} (see STAR Methods).

Host Range and Taxonomic Division Drives Viral Protein Structure Space and Mimicry

Taxonomic enrichment analysis (Figures 1B and 1C) reveals a clear correlation between the extent of virus structural mimicry and the identity of the host. Since their proteomes are well represented in the PDB, we first focused on the proteomes of *E. coli*-infecting phages, their natural host *E. coli*, the budding yeast *Saccharomyces cerevisiae*, and human. Significant structure similarity is only observed between *E. coli*-infecting phages and their natural host (Figure 1B). As shown in Figure 1C, with the exception of invertebrate-infecting viruses, which constitute a special case (see below), a broader analysis across all 7,486 cataloged viruses and 38,363 proteins from 4,045 putative hosts demonstrates that every virus group exhibits significant protein structural similarity to the taxonomic division of their hosts. In addition, viruses exhibit structural similarity to proteomes in other taxonomic divisions that are close in evolutionary distance to the host taxonomic division and reduced structural similarity to more distant taxonomic divisions (e.g., the structural space of human-infecting viruses are enriched for human, vertebrates, and invertebrate proteins but not bacteria or plant proteins, Figure 1C). Such similarities are reflected in evolutionary relationships between different taxonomic divisions. For instance, accumulating evidence describes some compartments of the phytosphere as environmental host spots for horizontal gene transfer (HGT) that facilitate genetic exchange between plants and bacteria (Pontiroli et al., 2009), underscoring the significant structural relationship between plant and fungi-infecting viruses and phages (Figure 1C). Moreover, in agreement with the results obtained for the broader virus categories, we found the protein structure space of human-infecting viruses to be enriched with structures found in human and non-human vertebrate proteomes (Figure 1C)—pointing to a possible path through which zoonotic infections may jump from animal reservoirs to cause human disease.

Unlike viruses of bacteria and vertebrates, we do not observe significant structural similarity between proteomes of invertebrate-infecting viruses and their invertebrate hosts—perhaps owing to the fact that invertebrates account for the taxonomic group with the lowest structural coverage in the PDB (containing just 2,687 unique invertebrate proteins). To bridge this knowledge gap, we modeled the proteome of the mosquito *Ae. Aegypti* (see STAR Methods) and compared the protein structure space of mosquito with each group of viruses. As shown in Figure 1D, we find that mosquito-infecting viruses occupy a significantly larger mosquito structural space than viruses that infect hosts in other taxonomic divisions. Moreover, we observe that mosquito-borne arboviruses, human-infecting viruses that use mosquitoes as vectors, engage a larger space of mosquito protein mimics than human-infecting viruses that do not use mosquitoes as vectors (p value 2.3×10^{-3}; Figure 1D). Together, these data further highlight the evolutionary pressures imposed on viruses to utilize a protein structure space defined by their host's range.

Diversification of Mimicry Strategies among Human-Infecting Viruses

While we do not observe recursive structural mimicry of human proteins across the human virome, we find that viruses belonging to the same viral family (including astroviruses, coronaviruses, filoviruses, orthomyxoviruses, paramyxoviruses, rhabdoviruses, and picornaviruses; data available for download at https://github.com/gorkaLasso/VirusHost_mimicry) tend to share a large fraction of mimics (where >75% of viruses within the family share >75% structural mimics). Figure 2A displays the number of human structural mimics per virus for different viral families. A number of families, including *Hepadnaviridae*, *Bunyaviridae*, *Papillomaviridae*, *Parvoviridae*, *Retroviridae*, *Arenaviridae*, and *Polyomaviridae*, utilize structural mimicry far less than others. On the other hand, poxviruses and herpesviruses, large dsDNA viruses that are less constrained by genome size (Figure 2B) and more likely to retain horizontally transferred genes (Alcami, 2003; Elde and Malik, 2009), encode many human protein structure mimics (Figure 2A). Since viral genome sizes vary greatly both between and within viral families (Figure 2B), we computed the number of human mimics relative to viral proteome sizes in an effort to assess the relative utilizations of mimicry (Figure 2C). We find that relative to their proteome size, small RNA viruses, which are constrained

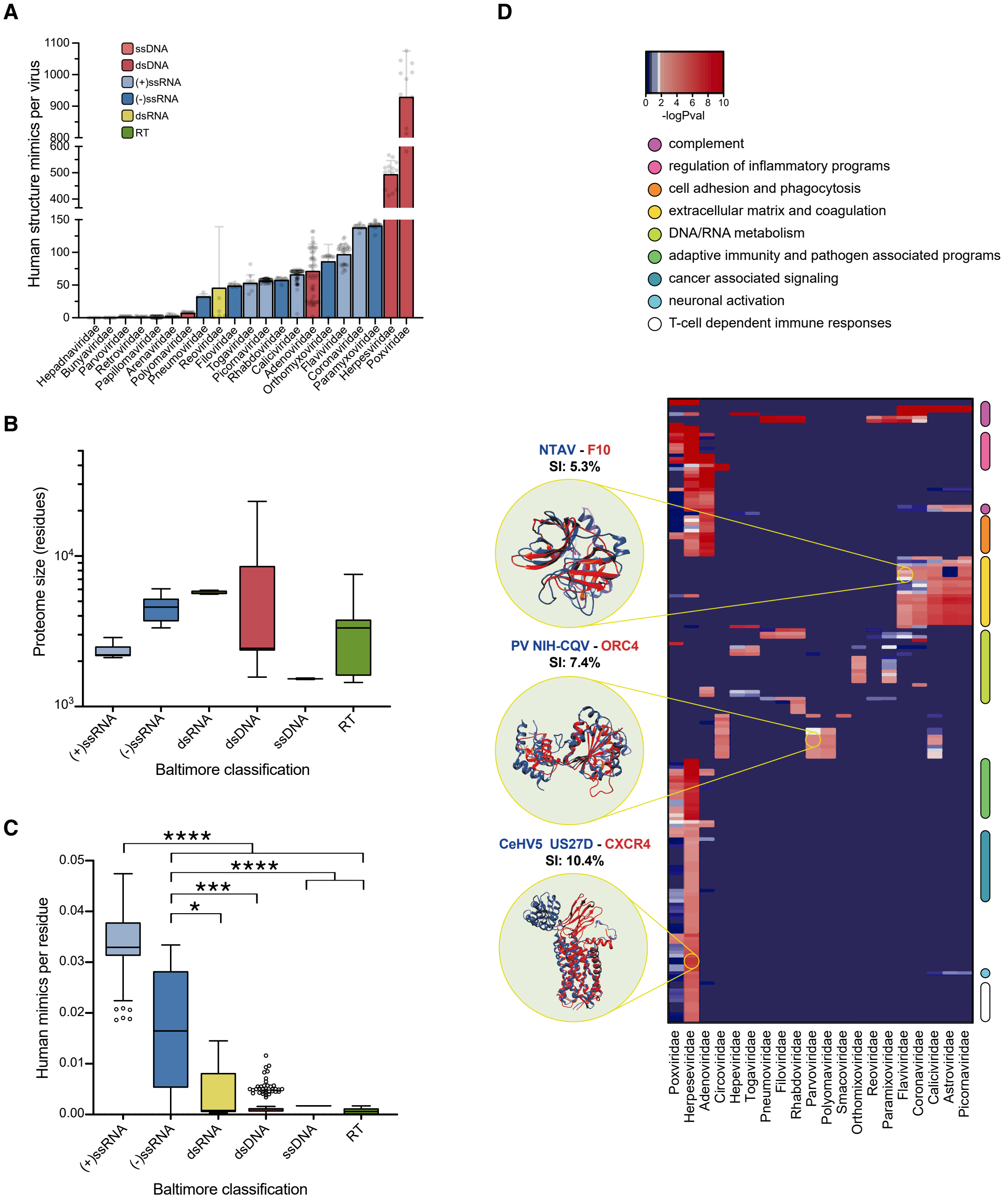


Figure 2. Mimicry Strategies Utilized by Human-Infecting Viruses Vary by Nucleic Acid Type and Proteome Size

(A) Number of virus-encoded human protein structure mimics for human-infecting viruses, grouped by viral family.

(B) Virus proteome sizes (amino acids) grouped by nucleic acid type.

(C) Number of structural mimics encoded per residue for human-infecting viruses, grouped by nucleic acid type. Virus family classification and nucleic acid type are as indicated by color key and labels (A–C); box and whisker plots (B) and (C) indicate lower and upper quartiles and span minimum and maximum values, brackets denote distributions used in Mann-Whitney (see STAR Methods; *p value < 0.05, ***p value < 0.001, ****p value < 0.0001).

(D) Hierarchical clustering of enriched biological pathways (rows) targeted through mimicry by human-infecting virus families (columns). Biological pathway categories are highlighted by color bars on the right of the heatmap. Examples of structural alignments between virus-encoded protein mimic (red) and human protein (blue); sequence identities (SIs) are indicated in black; p values indicated by color key. See also Burg et al. (2015), Fernandez et al. (2000), Kvansakul et al. (2007), and Li et al. (2007).

by genome size, display enhanced structural promiscuity, with a larger number of human protein mimics per residue. So, while viruses constrained by genome size may be restricted in their ability to expand their genomes through HGT, they have circumvented this limitation by encoding multifunctional proteins that mimic domains across multiple host proteins. These observations underscore the role of structural mimicry in evolving expanded molecular functionality in size constrained genomes.

Pathways Targeted by Viral Mimicry

Human-Infecting DNA-Encoded Viruses

The critical nature of virus-host PPIs in engaging host molecular machinery to fulfill viral life cycles has led to convergence of evolutionarily unrelated viruses around key cellular pathways (e.g., human-infecting DNA, RNA and retro-transcribing viruses converge, via different sets of PPIs, on multiple cellular metabolic pathways) (Lasso et al., 2019). In contrast, as shown in Figure 2D; Table S3, structurally mimicked proteins revealed little such convergence. Moreover, when convergence was observed, it often involved 2 or 3 viral families belonging to the same Baltimore classification. For example, owing to their large genomes and in agreement with their propensity to acquire host genes through HGT, *Herpesviridae*, *Adenoviridae*, and *Poxviridae*, families of dsDNA viruses are enriched for the largest number of mimicked biological pathways, 113, 38, and 35, respectively (Figure 2D; Table S3). Yet, while these viruses harbor a large number of human protein mimics, they do not mimic cellular components of RNA and DNA metabolism—likely reflecting the fact that they encode polymerases (as is the case of *Poxviridae*) and/or integrate into the cellular genome and are therefore bystander participants in cellular DNA replication.

We observe an enrichment of cytokine-related pathways in poxvirus- and herpesvirus-mimicked proteins and chemokine-related pathways in proteins mimicked by herpesviruses (Table S3; Figure 2D). Indeed, poxviruses and herpesviruses have evolved a repertoire of immune evasion strategies by acquiring immunomodulatory host genes through genetic recombination (often involving cytokines and cytokine receptors). Undetectable sequence similarity with the host has complicated past attempts to identification virus-host relationships (Odom et al., 2009; Felix and Savvides, 2017). However, while many of the relationships between herpesvirus and chemokine or chemokine receptors can be discerned from sequence, we have expanded the list of known mimics to include 61 herpesviral structure mimics not discernible from sequence alone (Figure 2D; complete list of mimics is available at https://github.com/gorkaLasso/VirusHost_mimicry). For example, as shown in Figure 2D, we observed that US27D protein from Cercopithecine betaherpesvirus 5, a cytomegalovirus, mimics human CXCR4 (sequence identity: 10.4%), a chemokine receptor that regulates T cell inflammatory programs and can serve as a co-receptor for HIV (Busillo and Benovic, 2007). It is unclear if the structural mimics encoded by DNA viruses that we have identified are a result of sequence divergence of products acquired by HGT or through convergent evolution. Yet, they highlight the pervasive use of mimicry by DNA viruses and point to host factors that are targeted by structural mimicry.

Human-Infecting RNA-Encoded Viruses

As highlighted above, RNA viruses have circumvented limitations imposed on their small proteomes by encoding multifunctional and structurally promiscuous proteins that mimic domains across multiple host proteins (Figure 2C). In accordance with known phenotypes (like fibrinolysis and hemorrhagic fever) associated with Dengue virus (DENV; *Flaviridae*) infection (Chuang et al., 2014; Lin et al., 2011), we find that proteins mimicked by these RNA viruses are enriched for functions and pathways related to blood coagulation and the complement pathway—an observation that is mirrored by PPIs mediated by proteins encoded by these viruses (we previously found that PPI networks mediated by proteins of 51 of 56 flaviviruses were enriched for "complement and coagulation cascades") (Lasso et al., 2019). While examples of such mimicry have been identified through sequence homology between DENV C, prM, E, and NS1 proteins and human coagulation factors (Lin et al., 2011; Chuang et al., 2014), we find similar mimicry, which cannot be discerned through sequence, employed by four other positive-sense ssRNA ((+)ssRNA) virus families (168, 166, 1,376, and 1,293 structural relationships between coagulation factors and astroviridae, coronaviruses, caliciviruses, and picornaviruses, respectively; sequence identity <20%, Table S3, data available at https://github.com/gorkaLasso/VirusHost_mimicry). So, while DENV immunization or exposure elicits antibodies that cross-react with multiple coagulation factor-associated pathological states (Chuang et al., 2014; Lin et al., 2011), our observations suggest that infections with viruses belonging to these other families may also result in similar imbalances in immune responses. Indeed, viruses within these families (e.g., SARS-*Coronaviridae*, hepatitis A-*Picornaviridae*, and hepatitis C-*Flaviridae*) are known to be associated with coagulation disorders such as thrombocytopenia, thrombosis, and hemorrhage (Goeijenbier et al., 2012).

CoVs

Knowledge of the precise regulatory programs that control the viral life cycle and mediate immune pathology can provide valuable clues about disease determinants. A closer look at coronaviruses underscores the level of diversification and structural promiscuity of protein mimics encoded by these viruses. As highlighted in Figure 2C, relative to proteome size, (+)ssRNA viruses including CoVs utilize structural mimicry more so than other human-infecting viruses. We find that the majority (~96%) of the 158 structural mimics identified are shared by three or more coronaviruses (data available at https://github.com/gorkaLasso/VirusHost_mimicry). We also observe that seasonal CoVs tend to share greater global sequence identity with the human proteins they mimic than their more pathogenic CoV counterparts (including MERS, SARS-1, and SARS-CoV-2). However, amino acid identities at key functional sites ultimately shape the extent to which host proteins and pathways can be intervened through structural mimicry. To that end, further computational and experimental analysis and interrogation of CoV-encoded structure mimics will be necessary to determine the biological implications of sequence conservation at precise amino acid residues.

Among CoV-encoded structural mimics, we have identified a number that target key regulators of anti-viral immune programs. First, in the context of CoV infection, intracellular sensing of viral RNA through DDX58 (also known as RIG-I; retinoic acid-inducible gene I), a pathogen recognition receptor, triggers a signaling cascade that results in the production and secretion of type I interferons (IFNα/β) (Abe and Shapira, 2019; Jensen and Thomsen, 2012). IFNα/β in turn binds to cell surface receptors (IFNRs) on nearby cells and, following a signal-transduction cascade, leads to STAT1-mediated (signal transducer and activator of transcription 1) upregulation of hundreds of IFN-stimulated genes (ISGs) that govern cellular responses to infection and render cells refractory to viral infection. These responses are indispensable for control of viral replication and initiation of long-term immune responses (Abe and Shapira, 2019; Jensen and Thomsen, 2012). As shown in Figure 3, we find that NSP13 of CoV mimics DDX3X and RIG-I as well as other helicase proteins, critical cellular components that cooperate through both direct and indirect interactions to initiate immune responses to viral infection. RIG-I ligands include ssRNAs that contain a 5′-triphosphate moiety (Abe and Shapira, 2019; Jensen and Thomsen, 2012), and previous biochemical characterizations have attributed multiple enzymatic activities to NSP13, including hydrolysis of NTPs and dNTPs and RNA 5′-triphosphatase activity (Ivanov et al., 2004; Adedeji and Lazarus, 2016). So, mimicry of RIG-I may serve to reduce the amount of viral RNA that is available to be sensed in infected cells. In addition, the CoV replicase complex has recently been demonstrated to interact with host cell DDX proteins (V'Kovski et al., 2019; Chen et al., 2009), suggesting that coronaviruses may utilize both PPIs and structural mimicry to manipulate functions of these proteins.

In addition to mimicking critical components of the cellular nucleic acid sensing machinery, we have identified CoV-encoded structural mimics that target key regulators of downstream antiviral programs mediated by STAT1. As shown in Figure 3, we observe significant structural similarity between CoV NSP3 protein and human PARP9 and PARP14, ADP-ribosyltransferases with known cross-regulatory roles in controlling STAT1-mediated signaling (Iwata et al., 2016). Though the precise mode of action has remained unresolved, CoV ORF3, which encodes NSP3, has indeed been shown to be an antagonist of IFNα/β signaling and interferes with STAT1 nuclear translocation (Kopecky-Bromberg et al., 2007). Our results point to structural mimicry as a molecular explanation underlying these observations.

Lastly, as discussed above, we find that structural mimicry of complement components is a feature shared across all coronaviruses that were part of this study (including SARS-CoV-2). The complement system is a critical regulator of immune responses to microbial threats, but when dysregulated by age-related effects or excessive acute and chronic tissue damage, complement activation can contribute to pathologies mediated by inflammation. Our data implicate virally encoded structural mimics that may contribute to SARS (*Coronaviridae*)-associated coagulation disorders like thrombocytopenia, thrombosis, and hemorrhage (Goeijenbier et al., 2012)—disease outcomes shared with hepatitis A (*Picornaviridae*) and hepatitis C (*Flaviridae*) infections (which also encode coagulation and complement cascade mimics). In addition, while the age-related differences in susceptibility to CoV (including SARS-CoV-2) are likely a consequence of multiple underlying variables, one possibility is that prior exposure to coronavirus may potentiate complement-mediated responses to subsequent coronavirus infections. Moreover, a corollary of these observations is that complement-associated syndromes (including complement deficiencies or hyper-activation phenotypes like those associated with macular degeneration) may impact clinical outcome of SARS-CoV-2 infection. Indeed, as part of a separate retrospective clinical study, we have demonstrated that genetic and functional dysregulation of coagulation and complement functions (including history of macular degeneration), are associated with morbidity and mortality in SARS-CoV-2-infected patients—effects that could not be explained by age or sex (Ramlall et al., 2020).

DISCUSSION

The mimicry of host protein structures is a strategy used by viruses to harness or disrupt host cellular functions (Wimmer and Schreiner, 2015; Alcami, 2003; Guven-Maiorov et al., 2016; Elde and Malik, 2009; Felix and Savvides, 2017). Structural mimicry can occur at the level of entire protein domains or in the form of "interface mimicry," where the structure of host protein residues involved in PPIs is mimicked on the surface of a viral protein (Franzosa and Xia, 2011; Guven-Maiorov et al., 2020, 2019). High-throughput experimental methods, focused largely on viruses of public-health importance, have experimentally mapped thousands of virus-host protein complexes while structure-based computational tools based on interface mimicry have reported notable successes in the prediction of PPIs (Franzosa and Xia, 2011; Guven-Maiorov et al., 2020, 2019). Recently, we reported a structure-informed prediction algorithm, P-HIPSTer, that exploits domain-level similarity and interface properties to predict virus/host PPIs across all fully sequenced human-infecting viruses (Lasso et al., 2019). Studies such as these have revealed valuable information about the molecular mechanisms that underlie viral infection. Yet, while most of these experimental and computational studies have focused on PPIs, a systematic analysis of viral mimicry of host proteins has not been reported. Here, rather than considering PPIs, we explore the extent to which viral mimicry of host proteins is a common phenomenon and ask whether the "structure space" occupied by a particular virus is related to that of the host it infects. Moreover, by analyzing the host proteins that are mimicked, we are able to gain functional insights about infection mechanisms that are complementary to those gained by predicting PPIs.

Structure enables identification of mimics between viral and host proteins that cannot be observed from sequence alone. Mimicry relationships have been detected through sequence similarity and linear motif co-occurrence. The method presented here relies on sequence similarity (albeit distant) to proteins with experimentally solved structures used to build structural relationships. This limitation is reflected in the lower structural coverage of non-vertebrate-infecting viruses (Figure S1A). However, we bypass limitations of sequence-based approaches by leveraging 3D protein structure to identify virally encoded proteins with significant structural similarity to host proteins. Although examples of pathogen-encoded protein mimicry have been reported in the literature, to our knowledge, this is the first comprehensive and systematic analysis of molecular mimicry across the earth's virome—spanning 337,493 viral proteins from 7,486 viruses (bacteria-, plant-fungi-, invertebrate-, and vertebrate-infecting viruses) and 38,363 proteins from 4,045 putative hosts.

Our pipeline uses a conservative global structural similarity criteria (SAS < 2.5 Å) to infer structural mimics (Subbiah et al., 1993; Kolodny et al., 2005; Budowski-Tal et al., 2010).

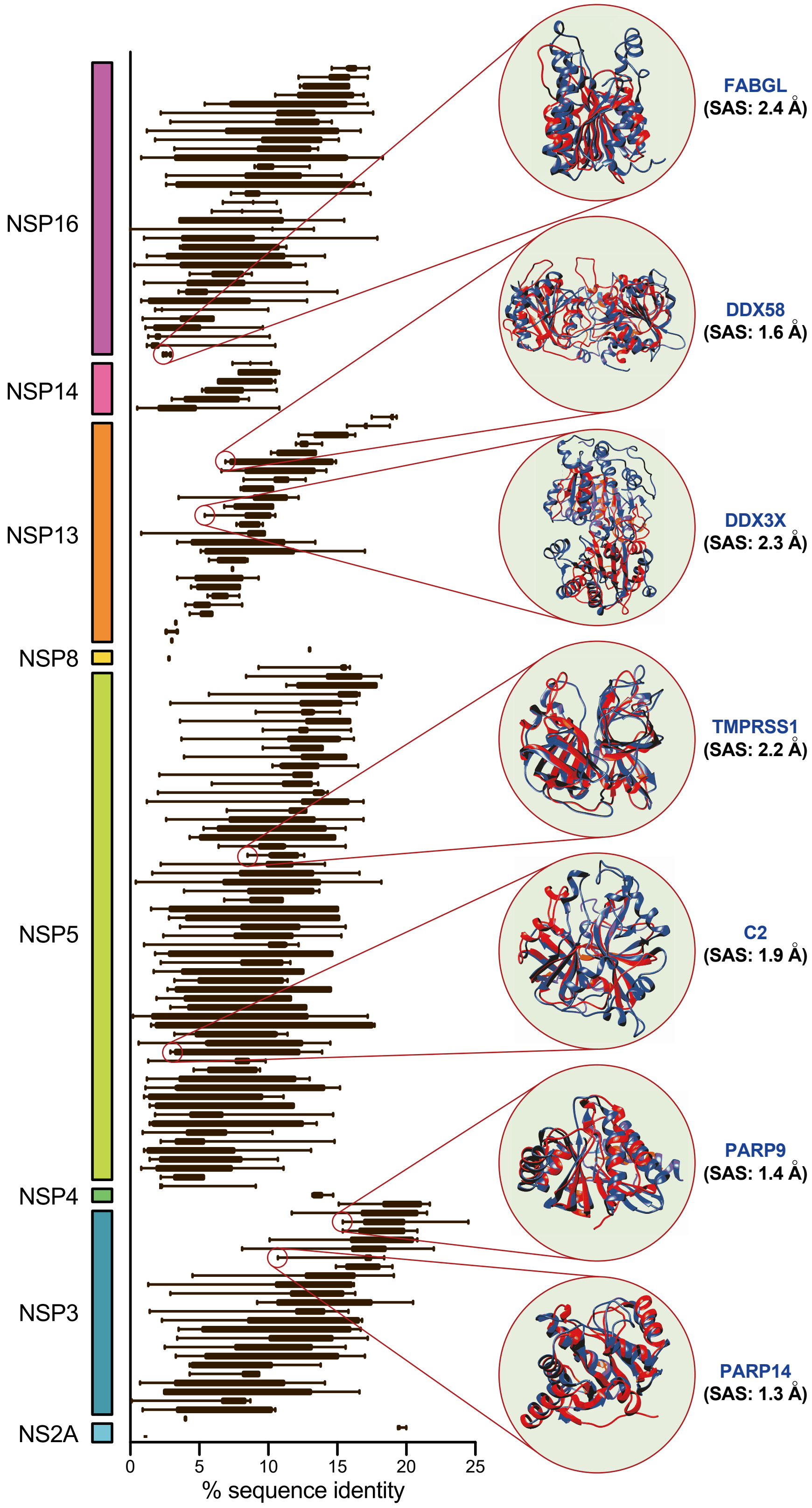


Figure 3. CoV-Encoded Proteins Mimic Known Regulators of Human Immune Response to Infection

A total of 158 human proteins are structurally mimicked by eight CoV proteins (grouped and indicated by color bars) for which structural information could be determined. Shown are box and whisker plots indicating lower and upper quartiles and spanning minimum and maximum pairwise % sequence identities between human and virus-encoded proteins (variability in sequence identity reflects CoV-specific sequence divergence; each human protein is mimicked by ≥3 viruses). Highlighted are seven examples of virus-encoded human protein mimics (virus protein structure in red; human protein structure and gene symbol in blue; SAS indicated in black).

Consequently, our approach does not consider local structural similarities defined by small protein regions and linear motifs (frequently found in disordered regions) (Franzosa and Xia, 2011; Guven-Maiorov et al., 2016). It is reasonable to consider that global structural mimicry relationships denote a common ancestral origin. However, we cannot rule out convergent evolution as a source of structural mimicry of smaller domains (e.g., <100 residues long) that would nevertheless meet the structural similarity criteria we have imposed. Importantly, structural mimicry does not necessarily imply functional mimicry. Furthermore, some of the mimics that we identify may represent spurious structural relationships that stem from underlying functional class of a given pair of proteins (e.g., virus and host-encoded proteases). Future experimental and functional interrogation will be necessary to demonstrate functional implications for the molecular mimics discovered as part of this work. Nevertheless, the results illustrate the ability of protein-structure-based analysis to infer structural, functional, and evolutionary relationships between viruses and their hosts.

Our results demonstrate that regardless of genome size, replicative cycle, or ecological niche, the evolutionary pressures imposed on viruses are reflected in the structure space they occupy and illustrate that mimicry may both constrain and enable host range. Of note, while structural similarity between viral and host proteins serves as a scaffold enabling viral proteins to coopt host pathways, amino acid identities at key functional sites will ultimately shape the extent to which host pathways can be intervened. Our observations offer a unique first step to investigating the role of amino acid variability at mimicked functional sites that might help explain differences in phenotypic outcome associated with viral infections.

Finally, the repertoire of structural mimics we discover opens new opportunities to identify potential mechanisms underlying autoimmune disorders of viral origin and new protein-based immune-modulatory therapeutics. For example, leveraging mimicry relationships between coronaviruses and human proteins informs about cellular targets and signaling cascades that are tuned during infection. Such knowledge can provide important clues about pathways that mediate pathology associated with infection and may point the way for designer therapeutics directed at these pathways. In addition, though beyond the scope of the current study, such knowledge may be useful in vaccine development as well as engineering of synthetic cellular operations where viral proteins are used as modules to build circuits with robust and predictable outputs. In the short term, and as highlighted by recent discovery of SARS-CoV-2 risk factors (Ramlall et al., 2020), information about virus-encoded structural mimics can help refine large-scale clinical studies and reveal determinants of immunity, susceptibility, and clinical outcome associated with human infection.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.09.006.

ACKNOWLEDGMENTS

We thank Donald Petrey for technical support. This work was funded by NIH grants 5R01GM109018-05 and 5U54CA209997-03 to S.D.S., 5R01GM030518-39 to B.H., and equipment grants S10OD012351 and S10OD021764 to the Columbia University Department of Systems Biology.

AUTHOR CONTRIBUTIONS

Conceptualization, S.D.S.; Methodology, G.L., B.H., and S.D.S.; Validation, G.L.; Formal Analysis, G.L.; Resources, B.H. and S.D.S.; Data Curation, G.L.; Writing, G.L., B.H., and S.D.S.; Visualization, G.L. and S.D.S.; Supervision, B.H. and S.D.S.; Project Administration, S.D.S.; Funding Acquisition, B.H. and S.D.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 24, 2020
Revised: September 3, 2020
Accepted: September 18, 2020
Published: October 13, 2020

REFERENCES

Abe, T., and Shapira, S.D. (2019). Negative regulation of cytosolic sensing of DNA. Int. Rev. Cell Mol. Biol. *344*, 91–115.

Adedeji, A.O., and Lazarus, H. (2016). Biochemical characterization of Middle East respiratory syndrome coronavirus helicase. mSphere *1*.

Alcami, A. (2003). Viral mimicry of cytokines, chemokines and their receptors. Nat. Rev. Immunol. *3*, 36–50.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *32*, D115–D119.

Bamford, D.H., Grimes, J.M., and Stuart, D.I. (2005). What does structure tell us about virus evolution? Curr. Opin. Struct. Biol. *15*, 655–663.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242.

Budowski-Tal, I., Nov, Y., and Kolodny, R. (2010). FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. Proc. Natl. Acad. Sci. USA *107*, 3481–3486.

Burg, J.S., Ingram, J.R., Venkatakrishnan, A.J., Jude, K.M., Dukkipati, A., Feinberg, E.N., Angelini, A., Waghray, D., Dror, R.O., Ploegh, H.L., and Garcia, C.K. (2015). Structural biology. Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor. Science *347*, 1113–1117.

Busillo, J.M., and Benovic, J.L. (2007). Regulation of CXCR4 signaling. Biochim. Biophys. Acta *1768*, 952–963.

Chen, J.Y., Chen, W.N., Poon, K.M., Zheng, B.J., Lin, X., Wang, Y.X., and Wen, Y.M. (2009). Interaction between SARS-CoV helicase and a multifunctional cellular protein (Ddx5) revealed by yeast and mammalian cell two-hybrid systems. Arch. Virol. *154*, 507–512.

Chuang, Y.C., Lin, Y.S., Liu, H.S., and Yeh, T.M. (2014). Molecular mimicry between dengue virus and coagulation factors induces antibodies to inhibit thrombin activity and enhance fibrinolysis. J. Virol. *88*, 13759–13768.

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

Federhen, S. (2012). The NCBI taxonomy database. Nucleic Acids Res. *40*, D136–D143.

Felix, J., and Savvides, S.N. (2017). Mechanisms of immunomodulation by mammalian and viral decoy receptors: insights from structures. Nat. Rev. Immunol. *17*, 112–129.

Fernandez, E.J., Wilken, J., Thompson, D.A., Peiper, S.C., and Lolis, E. (2000). Comparison of the structure of vMIP-II with eotaxin-1, RANTES, and MCP-3 suggests a unique mechanism for CCR3 activation. Biochemistry *39*, 12837–12844.

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

Garzón, J.I., Deng, L., Murray, D., Shapira, S., Petrey, D., and Honig, B. (2016). A computational interactome and functional annotation for the human proteome. eLife *5*, e18715.

Goeijenbier, M., Van Wissen, M., Van De Weg, C., Jong, E., Gerdes, V.E., Meijers, J.C., Brandjes, D.P., and Van Gorp, E.C. (2012). Review: viral infections and mechanisms of thrombosis and bleeding. J. Med. Virol. *84*, 1680–1696.

GraphPad Software (2019). GraphPad Software (California, USA: La Jolla). https://www.graphpad.com/.

Guven-Maiorov, E., Hakouz, A., Valjevac, S., Keskin, O., Tsai, C.J., Gursoy, A., and Nussinov, R. (2020). HMI-PRED: a web server for structural prediction of host-microbe interactions based on interface mimicry. J. Mol. Biol. *432*, 3395–3403.

Guven-Maiorov, E., Tsai, C.-J., Ma, B., and Nussinov, R. (2019). Interface-based structural prediction of novel host-pathogen interactions. Methods Mol. Biol. *1851*, 317–335.

Guven-Maiorov, E., Tsai, C.J., and Nussinov, R. (2016). Pathogen mimicry of host protein-protein interfaces modulates immunity. Semin. Cell Dev. Biol. *58*, 136–145.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Hwang, H., Dey, F., Petrey, D., and Honig, B. (2017). Structure-based prediction of ligand-protein interactions on a genome-wide scale. Proc. Natl. Acad. Sci. USA *114*, 13685–13690.

Ivanov, K.A., Thiel, V., Dobbe, J.C., Van Der Meer, Y., Snijder, E.J., and Ziebuhr, J. (2004). Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase. J. Virol. *78*, 5619–5632.

Iwata, H., Goettsch, C., Sharma, A., Ricchiuto, P., Goh, W.W., Halu, A., Yamada, I., Yoshida, H., Hara, T., Wei, M., et al. (2016). PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. Nat. Commun. *7*, 12849.

Jensen, S., and Thomsen, A.R. (2012). Sensing of RNA viruses: a review of innate immune receptors involved in recognizing RNA virus invasion. J. Virol. *86*, 2900–2910.

Johnson, N.L., Kemp, A.W., and Kotz, S. (1992). Univariate Discrete Distributions (Wiley).

Kolodny, R., Koehl, P., and Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J. Mol. Biol. *346*, 1173–1188.

Kopecky-Bromberg, S.A., Martínez-Sobrido, L., Frieman, M., Baric, R.A., and Palese, P. (2007). Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists. J. Virol. *81*, 548–557.

Kvansakul, M., Van Delft, M.F., Lee, E.F., Gulbis, J.M., Fairlie, W.D., Huang, D.C., and Colman, P.M. (2007). A structural viral mimic of prosurvival Bcl-2: a pivotal role for sequestering proapoptotic Bax and Bak. Mol. Cell *25*, 933–942.

Lasso, G., Mayer, S.V., Winkelmann, E.R., Chu, T., Elliot, O., Patino-Galindo, J.A., Park, K., Rabadan, R., Honig, B., and Shapira, S.D. (2019). A structure-informed atlas of human-virus interactions. Cell *178*, 1526–1541.e16.

Li, M., Lee, H., Yoon, D.W., Albrecht, J.C., Fleckenstein, B., Neipel, F., and Jung, J.U. (1997). Kaposi's sarcoma-associated herpesvirus encodes a functional cyclin. J. Virol. *71*, 1984–1991.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Lin, Y.S., Yeh, T.M., Lin, C.F., Wan, S.W., Chuang, Y.C., Hsu, T.K., Liu, H.S., Liu, C.C., Anderson, R., and Lei, H.Y. (2011). Molecular mimicry between virus and host and its implications for dengue disease pathogenesis. Exp. Biol. Med. (Maywood) *236*, 515–523.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. *45*, D200–D203.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. (2016). Linking virus genomes with host taxonomy. Viruses *8*, 66.

Ramlall, V., Thangaraj, P.M., Meydan, C., Foox, J., Butler, D., Kim, J., May, B., de Freitas, J.K., Glicksberg, B.S., Mason, C.E., Tatonetti, N.P., and Shapira, S.D. (2020). Immune complement and coagulation dysfunction in adverse outcomes of SARS-CoV-2 infection. Nat. Med. https://doi.org/10.1038/s41591-020-1021-2.

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

Petrey, D., Fischer, M., and Honig, B. (2009). Structural relationships among proteins with different global topologies and their implications for function annotation strategies. Proc. Natl. Acad. Sci. USA *106*, 17377–17382.

Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., et al. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins *53*, 430–435.

Pontiroli, A., Rizzi, A., Simonet, P., Daffonchio, D., Vogel, T.M., and Monier, J.M. (2009). Visual evidence of horizontal gene transfer between plants and bacteria in the phytosphere of transplastomic tobacco. Appl. Environ. Microbiol. *75*, 3314–3322.

R Core Team (2016). R: a language and environment for statistical computing (R Foundation for Statistical Computing). https://www.R-project.org.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. Trends Genet. *16*, 276–277.

Subbiah, S., Laurents, D.V., and Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. Curr. Biol. *3*, 141–148.

Velankar, S., Dana, J.M., Jacobsen, J., Van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., and Kleywegt, G.J. (2013). SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res. *41*, D483–D489.

V'Kovski, P., Gerber, M., Kelly, J., Pfaender, S., Ebert, N., Braga Lagache, S., Simillion, C., Portmann, J., Stalder, H., and Gaschen, V. (2019). Determination of host proteins composing the microenvironment of coronavirus replicase complexes by proximity-labeling. eLife *8*, e42037.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2016). gplots: various R programming tools for plotting data. https://github.com/talgalili/gplots.

Wimmer, P., and Schreiner, S. (2015). Viral mimicry to usurp ubiquitin and SUMO host pathways. Viruses *7*, 4854–4872.

Xiang, Z., and Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. J. Mol. Biol. *311*, 421–430.

Yang, A.S., and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J. Mol. Biol. *301*, 691–711.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556–560.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
All generated data	GitHub	https://github.com/gorkaLasso/VirusHost_mimicry
Software and Algorithms		
CD-HIT	(Li and Godzik, 2006)	http://weizhongli-lab.org/cd-hit/
EMBOSS	(Rice et al., 2000)	http://emboss.sourceforge.net/
BLAST	(Altschul et al., 1990)	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Hhblits	(Remmert et al., 2011)	https://github.com/soedinglab/hh-suite
Ska	(Petrey et al., 2003)	http://honig.c2b2.columbia.edu/ska
DAVID	(Huang et al., 2009a)	https://david.ncifcrf.gov/
R	(R Core Team, 2016)	https://www.r-project.org/
Gplots package	(Warnes et al., 2016)	https://cran.r-project.org/web/packages/gplots/index.html
Prism	(GraphPad, 2019)	https://www.graphpad.com/
Honig modeling pipeline	(Garzón et al., 2016)	http://honig.c2b2.columbia.edu/preppi
NEST	(Petrey et al., 2003)	http://honig.c2b2.columbia.edu/nest
Other		
Database: virus-hostDB	(Mihara et al., 2016)	https://www.genome.jp/virushostdb
Database: NCBI taxonomy	(Federhen, 2012)	https://www.ncbi.nlm.nih.gov/taxonomy
Database: Uniprot	(Apweiler et al., 2004)	https://www.uniprot.org/
Database: Conserved Domain Database	(Marchler-Bauer et al., 2017)	https://www.ncbi.nlm.nih.gov/cdd/
Database: PDB	(Berman et al., 2000)	http://www.rcsb.org/
Database: SIFTS	(Velankar et al., 2013)	https://www.ebi.ac.uk/pdbe/docs/sifts/

RESOURCE AVAILABILITY

Lead Contact

Further information for resources should be directed to and will be fulfilled by the Lead Contact, Sagi Shapira (ss4197@cumc.columbia.edu).

Materials Availability

This study did not generate new materials.

Data and Code Availability

All data and code generated as part of this study is available at GitHub (https://github.com/gorkaLasso/VirusHost_mimicry)

- This paper analyzes existing, publicly available data. Appropriate links to these datasets is provided in the Key Resources Table
- This paper does not report original code.
- The scripts used to generate the figures presented in this paper are available at https://github.com/gorkaLasso/VirusHost_mimicry
- Any additional information required to reproduce this work is available from the Lead contact

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This section does not apply to our computational study

METHOD DETAILS

Viral Dataset Assembly

We compiled a dataset of 7,486 viruses, represented by a total of 337,493 viral protein amino acid sequences together with annotations of their corresponding hosts, from virus-hostDB as of October, 2016 (Mihara et al., 2016). Viruses were classified according to the taxonomic divisions of their corresponding host based on NCBI taxonomic divisions (Federhen, 2012) (Table S1).

Structural Neighbor Search

The 337,493 viral protein amino acid sequences were reduced to 269,077 non-redundant sequences filtered at 100% sequence identity with Cd-hit (Li and Godzik, 2006). Sequence-based methods were then used to identify experimentally determined protein structures that, based on their sequence relationship to the viral protein, are expected to have similar structures (referred to here as ''structural templates'' - Figure 1A). First, viral proteins were parsed into domains using CD-search (Marchler-Bauer et al., 2017). Next, to identify structural templates, full sequences and parsed domains were queried against the PDB (Berman et al., 2000). Sequence homology search was performed in three steps, where each step is run only if the preceding step fails to report a structural template with a conservative E-value $< 1x10^{-12}$: i) Blast-based search (Altschul et al., 1990); ii) HHblits-based search (Remmert et al., 2011) and; iii) the third step runs five iterations of PSI-Blast. The pipeline reports the best structural template mapping to non-overlapping sequence segments for each query protein. When multiple structural templates map to the same segment in a query protein, only the structural template(s) with the lowest E-value are reported. The set of structural templates, describing the structural space of viral proteins, is further refined in order to minimize the number of redundant templates. To this end, structural templates derived from the same pdb chain and with their start and end positions within 15 residues are clustered together, using the longest structural template as the cluster representative. Finally, the set of structural templates is searched against the PDB database with Ska (Petrey et al., 2003; Yang and Honig, 2000) using a structural alignment score (SAS) < 2.5Å as a cut-off to identify structurally similar proteins (Kolodny et al., 2005). The ska algorithm focuses on alignment of secondary structure elements and uses only C-alpha coordinates of residues. Hence any differences in loop or side-chain conformations that might be expected because of differences in resolution would have little to no effect on a ska alignment.

Sequence Homology Assessment between Viral Proteins and Their Corresponding Structural Neighbors

Sequence homology among a pair of structurally similar proteins was considered significant when neither a Blast, HHBlits or PSI-Blast alignment yielded an E-value $< 1x10^{-6}$. In addition, for Figures 2 and 3, sequence identities for pairs of structurally similar proteins, were computed using the Needleman-Wunsch algorithm implemented in the EMBOSS package (considering only the corresponding protein segments sharing structural similarity) (Rice et al., 2000).

Calculating Overlap between Structural Space of a Virus and Proteins in a Given Taxonomic Division

Taxonomic division enrichment (Figures 1B and 1C) was computed with a hypergeometric test that describes the significance of having *k* structural neighbors belonging to a particular taxonomic division (out of *n* total structural neighbors for a group of viruses) given the entire set of structurally solved proteins in the PDB of size *N* that contains *K* proteins from the same taxonomic division (Johnson et al., 1992). To minimize experimental bias of multiple PDB entries for the same protein, structurally solved proteins were mapped to their Uniprot accession codes (Apweiler et al., 2004) using SIFTS mapping (Velankar et al., 2013).

Protein Modeling and Structural Neighbor Search for *Aedes aegypti* Proteome

Our validated, inhouse modeling pipeline (Garzón et al., 2016; Lasso et al., 2019) was used to model the 16,652 protein sequences (obtained from the Uniprot database (Apweiler et al., 2004)) in the *Aedes aegypti* proteome. Three-dimensional models for full-length proteins and protein domains, as defined by the Conserved Domain Database (Marchler-Bauer et al., 2017), were either taken directly from the PDB (Berman et al., 2000) or built by homology modeling as described previously (Garzón et al., 2016; Zhang et al., 2012). Protein structure models were built using NEST (Petrey et al., 2003; Xiang and Honig, 2001). Structural neighbor search was run on the set of modeled mosquito proteins using Ska (Petrey et al., 2003; Yang and Honig, 2000) and a structural alignment score (SAS) < 2.5Å was used as a cut-off to identify structurally similar proteins (Kolodny et al., 2005). The modeling and structural neighbor search pipeline reported structural neighbors for 10,895 (65%) mosquito proteins.

Identifying the Mosquito Structural Neighbor Space for Each Virus in the Dataset

The unique set of mosquito structural neighbors was computed for across all viruses. Structural relationships between viral and mosquito proteins was inferred by identifying common structural neighbors shared between a given viral and mosquito protein pair. Poorly annotated viruses with less than 4 annotated viral proteins were discarded. Invertebrate-infecting viruses were further subclassified into i) invertebrate-infecting viruses (including mosquito infecting viruses); ii) viruses infecting both human and mosquitoes and; iii) viruses infecting mosquitoes but not humans. We applied the non-parametric Mann-Whitney test to compare the size of the mosquito structural neighbor space per virus (normalized by the total number of annotated viral proteins in each virus) between different groups of viruses (Figure 1C), and outliers were removed with the Rout method (Q = 1%) implemented in Prism (GraphPad, 2019).

Human Protein Structural Neighbors of Human-Infecting Viral Families

To estimate the number of human protein structural neighbors per virus, viruses with 4 or more annotated proteins with structural template for at least 30% of the viral proteins were considered (Figures 2A and 2C). The total number of human structural neighbors per virus was corrected by the corresponding structural coverage (e.g. for a virus where 70% of its proteins have a structural template and shows structural similarity to 30 human proteins, the corrected number of human structural neighbors will be 30 / 0.7 = 42.9, Figure 1A). Viruses were grouped into viral families and families with less than 5 viruses were not plotted. In order to plot the number of human structural neighbors per residue for a given family of human-infecting viruses (Figure 2C), we normalized the total number of human mimics per virus (Figure 2A) by the corresponding proteome size (Figure 2B).

Functional Enrichment Analysis

Enrichment of biological ontologies (molecular function, biological pathways and diseases) was determined using David (Huang et al., 2009a, 2009b), background corrected using the 5,841 human protein structures extracted from PDB. A Bonferroni corrected p-value < 0.01 was used to identify enriched biological ontologies. Viral families and biological ontologies enriched in at least one viral family were clustered using R and the heatmap.2 function within gplots package (values in original matrix: -$\log_{10}$ P-value$_{\text{Bonferroni}}$; distance metric: Euclidean, method: complete) (R Core Team, 2016; Warnes et al., 2016).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameters, including the definition of center, dispersion and associated significance, are reported in the main text, Figures and Figure legends. We have applied hypergeometric test and Mann-Whitney test to calculate significance. P values for pathway enrichment analysis were adjusted for multiple comparison. The section entitled “Method Details” describes the statistical analyses performed. Data are judged to be statistically significant when $p < 0.05$ in applied statistical analyses. In Figures 1 and 2 asterisks denote statistical significance (Figure 1D: **, Pvalue < 0.0001; Figure 2C: *, Pvalue < 0.05, ***, Pvalue < 0.001, ****, Pvalue < 0.0001

ADDITIONAL RESOURCES

All data generated as part of this study is available at GitHub (https://github.com/gorkaLasso/VirusHost_mimicry)

Cell Systems, Volume *11*

Supplemental Information

A Sweep of Earth's Virome Reveals Host-Guided Viral Protein Structural Mimicry and Points to Determinants of Human Disease

Gorka Lasso, Barry Honig, and Sagi D. Shapira

Figure S1

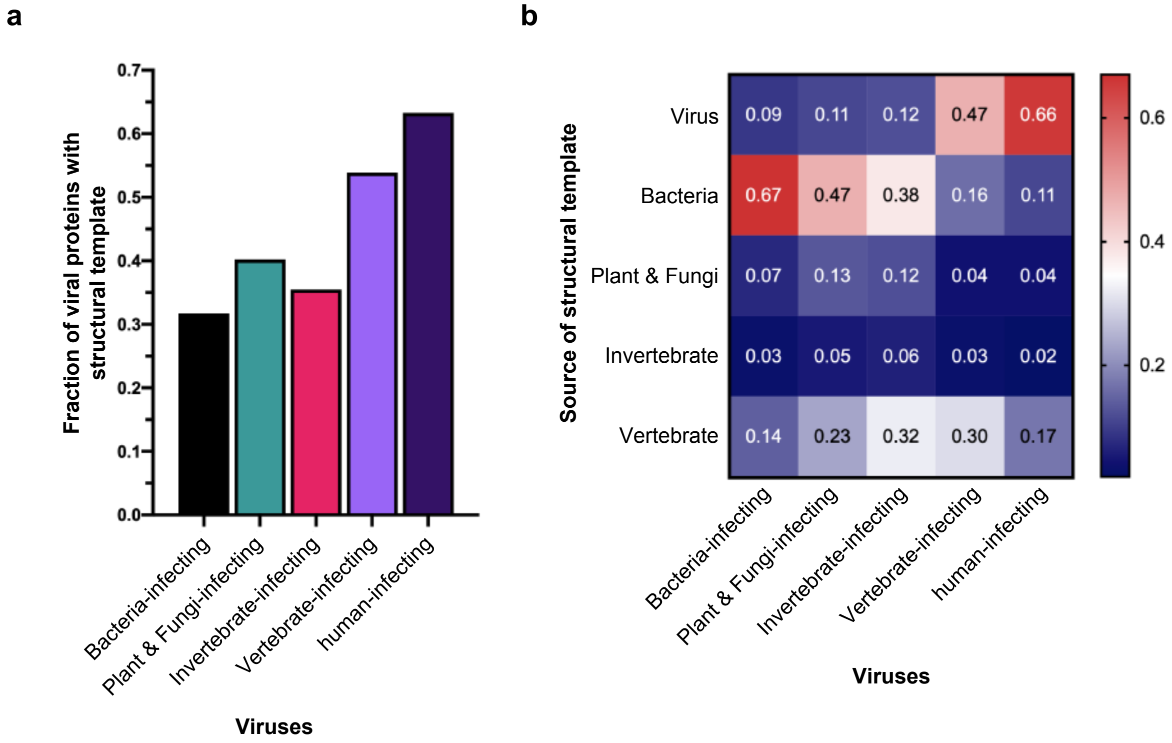


Figure S1l Structural coverage of viral proteins. a, Fraction of viral proteins, grouped based on the taxonomic division of the corresponding host, with a sequence homolog whose structure is experimentally known (structural template). b, Fraction of the structural templates belonging to a particular taxonomic division identified for different groups of viruses.

Figure S2

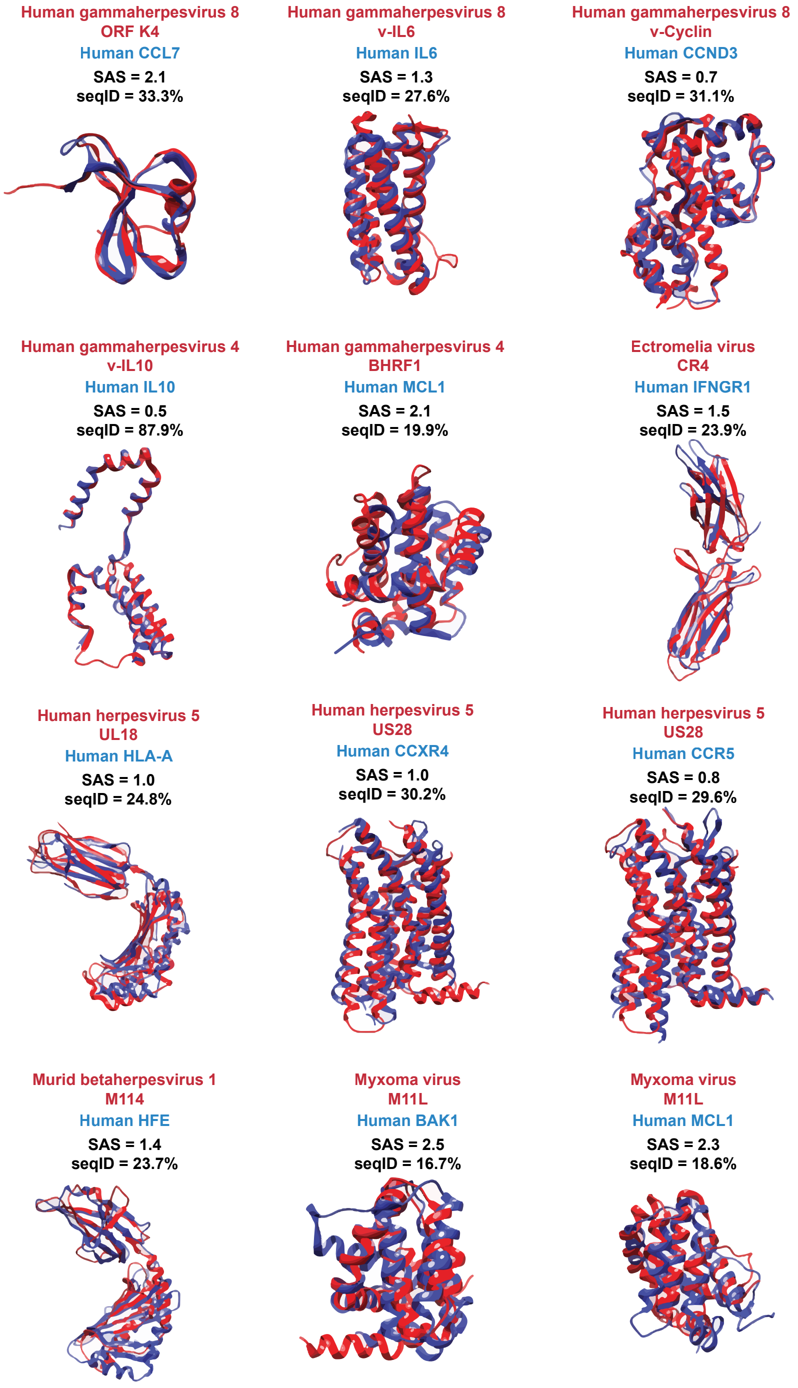


Figure S2I Virome-wide scan rediscovers known viral-host structural mimics. Shown are examples of known viral-host structural mimics (Fernandez et al., 2000, Franzosa and Xia, 2011, Elde and Malik, 2009, Li et al., 1997, Kvansakul et al., 2007, Burg et al., 2015) recapitulated with structural neighbor search. Viral and host proteins are colored in red and blue respectively. SAS is shown in black.