**Virus-host protein-protein interactions and human disease**

Norman Goodacre1, Prajwal Devkota2, Eunhae Bae1, Stefan Wuchty2,3,4,5\*, Peter Uetz6\*

1 Division of Viral Products, Office of Vaccines Research and Review, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

2 Dept. of Computer Science, Univ. of Miami, Coral Gables, FL 33146, USA

3 Center for Computational Science, Univ. of Miami, Coral Gables, FL 33146, USA

4 Dept. of Biology, Univ. of Miami, Coral Gables, FL 33146, USA

5 Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Miami, FL 33136, USA

6 Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA

\* Corresponding authors: [wuchtys@cs.miami.edu](mailto:wuchtys@cs.miami.edu), [peter@uetz.us](mailto:peter@uetz.us)

**Keywords:** protein interaction network, phage, burden of disease, interaction databases

**Abstract**

Viruses infect their hosts by a series of protein-protein interactions (PPIs), starting with the virus binding to surface receptors, and concluding with virus assembly and egress of complete virus particles. We discuss the coverage of human viruses and how well their proteins have been studied for host-virus protein-protein interactions. Given the impossibility of reviewing the huge body of literature on virus-host interactions, we based our review primarily on data available from public databases. Notably, only 4 virus families have accumulated more than 1,000 PPIs when all their PPIs are combined, namely orthomyxoviruses (5,494), herpesviruses (5,423), papillomaviruses (3,927) and retroviruses (2,285). Thus, some viruses have been extremely well studied, with some viruses, such as HIV, having more than a hundred interactions identified for each of its proteins. We further discuss the diversity of viruses (in terms of available genome sequences) in relationship to their medical significance. In addition, we explore the reliability and physiological relevance of PPIs. While for many medically important viruses numerous genome sequences are available relatively few interactions are known (e.g. rhinoviruses). Given the broad scope of our review we discuss only a few viruses, such as Influenza, in more detail to illustrate the challenge of finding functions for numerous PPIs.

**Introduction**

Bacteria and viruses are the most important pathogens on earth. While most bacteria can be directly eliminated using antiobiotics, viruses can only be constrained in their growth, posing a challenge to the safety of the host. This is a direct consequence of the fact that viruses are composed of only nucleic acids, proteins, sometimes lipids and a few other compounds. As a consequence, the survival of viruses is almost entirely dependent on molecular protein-protein interactions (PPI) with their hosts. Furthermore, viral variation occurs rapidly, often with significant adaptation within each host. Hence, strategies for the development of safe antivirals strongly depend on precise targeting of virus-host PPIs and a deep understanding of viral biology.

In this review we related the diversity of viruses to their medical importance and then to the depths of knowledge we have about their molecular biology. As for the latter, we used our knowledge of virus-host interactions as a proxy. Given the vast body of literature about virus-host interactions, we primarily based our review on existing databases of human host-virus interactions even though they may be biased and incomplete sources of information. For instance, as medically important viruses such as HIV and Influenza have received considerable research attention we surmise that more knowledge about their interactions is available compared to other viruses. Still, some highly infectious viruses have received relatively little attention, as their spread is geographically limited and focused on a narrow range of hosts. Furthermore, their investigation may prove experimentally difficult. For example, the Zika threat is relatively recent, triggering research activity in the last few years. Although extensive sequence information from next-generation sequencing studies exists, precise knowledge of virus-host PPIs and thus potential targets for antiviral therapies may be very limited.

**Diversity and morbidity of human viruses**

Our interest in viruses is primarily driven by their impact on human health and its economic toll. Hence, we wondered how the medical and economic importance of viruses is related toresearch efforts, reflected by the number of genomes sequenced and the number of PPIs detected. Unsurprisingly, some viruses received a lot of attention. For example, both HIV and Influenza claim a large number of victims and impose a significant economic burden. In turn, viruses such as Hepatitis, MERS and SARS cost relatively few lifes, yet come with a considerably economic price tag.

Virus diversity can be measured by sequencing many virus isolates from different geographical areas. Such investigations are especially informative for RNA viruses that evolve rapidly, providing large sequence diversity. In **Table 1**, we summarized the 20 most studied virus families as a function of the number of sequenced genomes. Notably, we found more than 7,000 genomes of flaviviridae that include the Zika virus. Furthermore, we counted more than 2,000 genomes of retroviridae, including the HIV virus. Sequences of variants of flaviviridae and retroviridae appear to be highly variable, as more than 200,000 sequences in both families cannot be clustered into similar sub-groups. A similarity threshold of 98% sequence identity with the tool CD-HIT-EST was used, and cluster representatives were assumed to be the longest sequence in each 98%-similar group. Where available, complete genomes from RefSeq Viral and neighbor complete genomes were assumed to be the cluster representatives.

As another aspect of viral variability, humans are infected by a variety of different viruses. For instance, Wylie et al. found that an average of 5.5 viral genera that occurred in each of 102 healthy individuals (1). As only five body habitats were screened, including nose, skin, mouth, vagina, and stool, most people probably carry dozens of different viruses. However, only a few lead to clinical symptoms or disease. Furthermore, Poon et al. found numerous variants of Influenza A in individual human hosts. Such variants showed changing abundance over time, suggesting that the underlying viruses adapted to a changing environment. Moreover, types and populations of variants varied widely between individuals, reflecting their ability to evolve and adapt quickly to changing hosts and conditions.

**Virus-human host protein-protein interaction databases**

During the last decade numerous protein interactions between human viruses and their host cells have been mapped and more thoroughly investigated. Most of these efforts focused on relatively few viruses, such as Hepatitis C virus (2-5), Human Immunodeficiency Virus (6, 7), Influenza A virus (8), herpesviruses (9), including Epstein-Barr Virus (10, 11), but also Dengue (12) and a few others (13). As a consequence of many high-throughput screens, databases of PPIs have been filled with tens of thousands of virus-human interactions. The IMEx (International Molecular Exchange) consortium of databases has been particularly valuable, as members use a standard format (PSI-MS) for recording meta-data for PPIs. Importantly, this meta-data includes experimental evidence information. IMEx members and observers are BioGRID (23), DIP (24), IntAct (26), and MINT (27). Other generalist PPI databases include BIND (22), HPRD (25), and iRefInd (27). The Reactome database (28) also includes biochemical pathway information to facilitate understanding of PPI function. However, these generalist PPI databases are often very large and contain considerable internal redundancy as well as overlap among themselves. Therefore, web-based resources have been developed to integrate pathogen-host molecular interactions (PHI) and related data from these PPI databases (generic as well as PHI databases are shown in **Table 3**). Some PHI databases specialize on only one specific pathogen species such as HCVpro (14) and the HIV-1 Human Interaction Database at NCBI (15). A wider range of human specific viruses are covered by VirHostNet (16), VirusMentha (17), PHIDIAS (18), HPIDB (19), and PHISTO (20) which include interactions between human host and different viral and other pathogen proteins.

PHI databases of specific pathogen species may contain unique and extensive information for focused research. HCVPro (HCV interaction database) is dedicated to only HCV and catalogues protein interactions for intraviral and virus–human systems. Additionally, the database includes information on the structure and functions of HCV proteins (14). As for HIV, out of the over 17,000 HIV-1 – human PPIs reported in HIV-1db (15) as of August, 2017, fewer than 7,000 are physical interactions and the number of direct (rather than complexed) interactions is unclear. While clear evidence of direct physical interaction is provided for some PPIs (e.g. “binds”, “phosphorylates”, “cleaves”), for others the evidence is weaker. Compared to PSI-MS standards, these evidence descriptors were generally more ambiguous. Using the most restrictive set of criteria, the number of physical and direct PPIs between HIV-1 and human proteins is about 1,600. Of the additional ~ 5,400 physical interactions, the majority have no evidence for nor against a direct interaction (e.g. “interacts with”, “stabilizes”, “recruits”). While NCBI’s HIV-1DB is an invaluable resource for HIV-1 researchers, it can be difficult to “mine” physical PPIs from the entire database confidently without manual inspection or natural language processing. PHI databases for only specific pathogens were omitted from deeper analysis in this review, even though some (e.g. HCVpro) follow PSI-MS standards.

PHI databases that cover a wide range of viruses and hosts include VirHostNet (16), VirusMentha (17) and HPIDB (19). VirHostNet (Virus–Host Network) is one of the earliest pathogen-host interaction (PHI) resources that specialized in the management and analysis of integrated virus–virus, virus–human, and (human) host–host protein interaction networks coupled to their functional annotations. The recently developed tool, VirusMentha is another virus-virus and virus–host protein interaction resource and the most comprehensive viral PHI data source in terms of the total number of virus species or host organisms. In addition to viruses, HPIDB (Host–Pathogen Interaction Database) captures protein interaction data between many different pathogens and hosts. Finally, web-based PHI databases comprising all pathogen types with known interactions are PHIDIAS (20) and PHISTO (20). PHIDIAS (Pathogen–Host Interaction Data Integration and Analysis System) stores data on genome sequences, conserved domains, and gene expression data related to PHIs. In addition to data storage, PHIDIAS offers the analysis of these data. PHISTO (Pathogen-Host Interaction Search Tool) is a comprehensive PHI database including data of all pathogenic microorganisms for which experimental protein interactions with human are available. Most PHI databases, with the exception of HPIDB, have several drawbacks, however. A major drawback is the lack of standard (i.e. PSI-MS) format, making inference about the molecular details of interactions challenging. Also, the filtering criteria for collecting virus-host PPIs is often difficult or impossible to ascertain. Finally, the update cycles are irregular and often the underlying pipelines for collecting data change from one version to the next. Generally, host-virus protein interaction data in the above PHI databases are integrated mainly from other PPI databases using automatic integration tools such as PSICQUIC (21) and by manual literature curation. However, a user may desire to obtain additional interaction information and therefore may have to use PSICQUIC firsthand, bringing into question the useability of PHI databases.

PHI databases generally do not follow PSI-MS standards, with the exception of HPIDB, so the majority were excluded from deeper review. One of the aims of this review was to give estimates for the reliability of available virus-human PPIs. Therefore, experimental evidence for either direct physical or simply physical (and potentially indirect) was necessary. All relevant databases are shown in **Table 3**, with those used for more in-depth study of PPIs emphasized in **bold**. The resources used are: BioGRID, HPIDB, IntAct, and MINT.

Our understanding of human-virus PPIs is highly biased towards a few well-studied viruses. For instance, only 5 viruses have more than 1,000 physical interactions listed in , namely influenza A with 3,746, Epstein-Barr Virus (EBV) with 3,163, HIV-1 with 2,540, herpesvirus 8 with 1,643, and hepatitis C with 1,082. The human papilloma viruses totaled 4,645 interactions across 29 different species. Given the abundance of human host-viral interactions we collected protein interactions for a variety of viral families using the databases in **Table 3 (bold)**. Evidence for direct interactions of these top viruses varied, but was notably high for herpesvirus 8 (1,623 direct /1,643 physical) and HIV-1 (2,365 direct /2,540 physical) (**Table 4**). Numbers that follow are for physical interactions, total. Notably, we found 5,957 interactions with human host proteins involving proteins of orthomyxoviridae. Such a set of interactions is mostly dominated by interactions that occurred between proteins of the human host and the influenza A virus. Similarly, we found 4,645 interactions of papillomaviridae, where most interactions were experimentally determined using the Herpes simplex virus. While our set provided 3,748 interactions of orthomyxoviridae, with virtually all interactions involving the influenza A virus,, 2,998 retroviral interactions were mostly provided by the HIV-1 virus. Pooling interactions of other virus families with ≥ 100 PPIs, we obtained 3,578interactions. In addition, we obtained 462 interactions for virus families with fewer than 100 PPIs (not shown). In **Fig. 1**, we summarize the sets of human proteins that were targeted by different virus families and their substantial overlap. While these numbers are roughly similar, they represent vastly different genome sizes and virus diversity within families. For instance, HIV encodes only about 10 proteins with more than 100 interactions per protein. EBV, by comparison, encodes about 85 proteins leading to “only” 20 interactions per protein on average. Before we can even begin to interpret these interactions, we need to ask if it is biologically meaningful or even possible if a protein has >100 interactions (see below). It is also important to note that the extent of known human-virus PPIs often differs from the extent of known viral sequences. This is particularly true for recent viral outbreaks, such as Zika, and is largely due to the availability of next-gen sequencing.

**How reliable are published virus-host interactions?**

One means by which we have attempted to answer the question of reliability of publicly-available virus-host PPIs is through filtering for specific evidence of direct physical interaction, rather than just physical interaction. Around 50% of all virus-human PPIs may be indirect (**Table 3**, bold). Although for certain source databases the proportion of direct interactions was less than half (IntAct, MINT), for other the proportion was significantly higher, at around 70% (BioGRID, DIP). However, the reliability of the experimental techniques for detecting direct interactions also needs to be considered.

Only a few studies exist that systematically validated human-virus interactions for their biochemical or even physiological validity.

Among the first attempts to validate human virus-host interactions was our study of KSHV-human interactions (29). We predicted homologous interactions from experimental data *in S. cerevisiae, Caenorhabditis elegans*, and *Drosophila melanogaster* whenever at least one of them had interacting orthologs. Although this is a somewhat far-fetched approach, we predicted 20 interactions between 8 KSHV and 20 human proteins. Nineteen of these 20 virus-host interactions were tested by CoIP and an unexpectedly large percentage (13 out of 19, or 68%) were confirmed.

Zhang et al. found 109 interactions among 33 Vaccinia and ~160 human proteins of which 27 were tested by GST pull-downs (30). 17 of these were confirmed, which translates to a 63% validation rate. While these numbers appear to be rather high, only certain subsets were selected and thus do not necessarily represent an unbiased validation rate of a complete Y2H data set.

Khadka et al. (12) screened all 10 proteins of Dengue virus against a human liver Y2H library and found 139 interactions involving 8 Dengue and 105 human proteins. 33 interactions were detected by two or more Dengue protein fragments. 16 out of 23 tested interactions were also confirmed by an independent split-luciferase assay. 23 human genes were tested by siRNA assays for their effect on viral replication and 12 of them were found to be required for Dengue replication. Interestingly, over 40% of the human proteins reported in this study to interact with DENV proteins have been implicated in the life cycles of at least one other virus, with the greatest overlap with proteins linked to HCV infection.

For comparison, 95 human proteins were identified in yeast two-hybrid screens with INFV and all were tested for their ef- fects on INFV replication in similar siRNA experiments (Shapira et al. 2009). Of these, three were required for INFV replication and eight exerted a negative effect, for a total of 11 (12%) proteins that affected INFV replication (Shapira et al. 2009). Similarly, in large-scale siRNA screens for host factors affecting viral replication, the average hit rate was 1% (Brass et al. 2008, 2009).

While it is possible that many interactions found in high-throughput screens are physiological, this is unlikely. Unfortunately, it is difficult to prove that an interaction has no physiological role, as for viruses even small differences may result in significant survival rates over many generations. In addition, given the high mutation rate of many viruses, small differences may expand to larger ones once more mutations accumulate and add further (small) advantages.

**How many interactions does a virus require?**

The sheer number of host-virus interactions that have been found for many viruses may suggest that we have identified most if not all interactions. Yet, we know from other studies that interaction screens likely contain a large number of false positives (see above). Therefore, we wonder how many interactions does a virus realistically need or use to infect an organism. How do we identify the physiological interactions among those that have been found overall? Unfortunately, such estimates, that especially compare multiple viruses, currently do not exist, assuming that viruses of similar proteome size may use similar numbers of PPIs. While hard to answer in human viruses, bacteriophages may serve as a model: 50 years of research have identified about 30 host-virus interactions between *E. coli* and phage lambda, which encode ~4,000 and 73 proteins, respectively (31). Assuming that most host-virus interactions in this system have been identified, a large-scale analysis of *E. coli*-lambda interactions revealed 62 interactions in a high-confidence set (among a raw data set of 631 PPIs total) (32). However, of the 62 high-confidence PPIs only two were previously known to be physiological, while the role of the other 60 remains unknown. We surmise that lambda is unusual as compared to other phage in that many lambda proteins are processed during maturation and thus interactions are more difficult to detect. Protein processing seems to be less common among other phages, such as T7, whose 55 proteins are known to be involved in only 15 interactions with its host (31). Both T7 and lambda have about 30-40 interactions between virus proteins, which are easier to detect and possibly more abundant than in human viruses, given the more elaborate virus structure in these tailed phage when compared to the often simple-structured human viruses. Although the role of bacteriophages lambda and T7 as meaningful references for human viruses is unclear the latter includes many cases with substantial proteolytic processing, such as Hepatitis C Virus, in which the polyprotein prevents many interactions that are found when the proteolytic fragments are used for interaction screens (REF: Flajolet).

**What are the protein targets of human viruses?**

The investigation of different host-virus protein interaction interfaces reveal the functional classes of host proteins that different viruses commonly target through their host-virus protein interactions. In particular, viruses prefer human host proteins that were involved in cell cycle regulation, signaling, nuclear transport and cell trafficking as well as transcription and translation functions (2, 10, 29, 33, 34). However, as well, reminding us that each viral family uses a different strategy to invade a human host cell (35, 36).

The abundance of such virus-host interaction data prompted topological analysis of networks thus obtained. Navratil et al. described a human infectome network (HIN) that linked 416 viral proteins to 1,148 human proteins through 2,099 manually curated virus-host PPIs (37). In fact, 32% of these cellular proteins are targeted by more than one virus protein. A similar fraction, 28% of these cellular targets interact with proteins from more than one virus.

Clearly, virus proteins attack a relatively small number of human proteins that are relevant for their replication. More specifically, these human targets appear to be highly connected: the mean degree of these targets was 38 vs. 10 in non-targeted proteins (37). Even among highly connected proteins (k>5) in the human interactome, the degree of virus targets was twice as large as those of non-targeted proteins. Note that only 50% of all human proteins were known to interact with other proteins in the 2011 human proteome, and 50% of these interacting proteins were interacting with only one other protein. In independent analyses, it was found that viral proteins preferably target human host proteins that are involved in a large number of interactions (10, 33, 34, 38-44). As the number of interactions is a local measure of centrality, other more global measures of centrality were considered as well. In particular, betweenness centrality measures how many connections go through a particular protein in a network when proteins were connected by their shortest path. Indeed, various viruses target human host proteins with high betweenness centrality. As a corollary, such central proteins also have significantly shorter paths to other proteins (10, 33, 34, 38-44), as well as participate in a higher number of pathways and protein complexes (45). Recently, the focus of modern network research has shifted to the determination of nodes that allow the control of a network (46, 47). Explain what “control” means. Notably, such controlling genes were enriched with essential genes and disease genes, and they appeared in regulatory interactions (48, 49). Furthermore, they also played a role as targeted and required genes of viral infections (50, 51). As a consequence, such centrality measures and the determination of control nodes allow the computational prediction of potential viral targets based on topological measures (41, 42, 52-54).

**How are protein-protein interactions related for viruses with multiple hosts?**

Most viruses are capable of infecting multiple hosts, at least closely related host species. However, some viruses infect quite distantly related species, such as human and bird influenza. One might expect that such “generalist” viruses enjoy an evolutionary advantage due to their potential to access more hosts. However, in most cases adaptive mutations for one host bring about antagonistic pleiotropy, or a decrease in fitness for replicating within another host, with a possible exception when a host’s population tends to fluctuate widely (Elena et al., 2009). Some multi-host viruses may require multiple hosts as part of their life cycle (notably arborviruses such as dengue) while others require only host (while maintaining their promiscuity). For obligate multi-host viruses, a phenotypic change specific to each host is often developed, which increases the chances of transmission. The hosts themselves are often very different, in terms of pH, temperature, and cell type. Thus, it is likely that PPIs from obligate multi-host viruses vary more than do PPIs from optional multi-host viruses.

Arboviruses (arthropod-borne viruses) are among the most burdensome multi-host viruses for humans. A good example is dengue. The DenvInt database (Dey and Mukhopadhyay, 2017) catalogues PPIs between the 10 dengue proteins, of which 7 are non-structural, and both human and mosquito proteins. The dengue-human network consists of 535 interactions between 10 dengue and 335 human proteins, while the dengue-mosquito network consists of 249 interactions between 10 dengue and 140 mosquito proteins. The dengue-human and dengue-mosquito networks are similar at the higher degrees nodes, with NS5, E and NS3 being the most highly-connected dengue proteins. However, the mosquito network is both smaller and more highly-clustered, with the C protein sharing a very similar set of partners with E. No meta-analysis has yet been performed with DenvInt to compared the functions of the human and mosquito targets. However, Mairang et al. (2013) suggested that NS3 and NS5 may be involved in down-regulation of innate host defenses via Toll-like receptor and unfolded protein pathways in both human and mosquito. The importance of host immune suppression may be greater for obligate multi-host viruses, because it is important for the viral titer to become sufficiently high that transmission (often via blood) can be successful (Ebel, 2017).

Even with extensive recombination, as among the “chromosomes” of influenza viruses, many viruses are known to jump from animals to humans (REF: Mandl, Vijaykrishna). This often leads to unusually severe outbreaks [REF]. The host environments don’t vary as widely for viruses like influenza A as for arboviruses, therefore the PPIs are likely to be more similar [REF]. Crossover mutations occur predominantly in hemagglutinin and neuraminidase, which allow for entry and exit, respectively.

**The virus interactome-diseasome connection**

It has been long known that some viruses are involved in diseases not typically associated with infection. For instance, up to 20% of cancers may be caused by viruses such as papilloma or certain herpesviruses (55). Navratil et al used a list of virus targets and compared it to a list of 1,729 human genetic disease-related proteins (from OMIM) and found that 13% of human virus targets are also associated with at least one human disease (37). That is, a human protein interacting with a virus protein is twice as likely to be involved in a disease than a non-target. Most of the diseases found in this study were related to cancer or neurodegenerative diseases. Surprisingly, type 1 diabetes was also associated with virus infection, as were autoimmune diseases in general. The latter may not be surprising, given that many virus infections elicit a strong immune reaction.

An independent study came to a similar conclusion: Gulbahce et al. analyzed the connection between Epstein-Barr-Virus (EBV), human papilloma virus (HPV) and disease (56). However, these authors not just used PPIs but also metabolic networks and regulatory interactions (**Fig. 4**). Using U.S. Medicare patient medical history data derived from 13 million patients, Gulbahce et al. found that many diseases are often associated with viral infection, including EBV or HPV. For instance, HPV patients have 15.7 and 2.7 times increased chance of developing retina and bladder cancer but also a higher risk of Fanconi anemia (56). It is clear that PPIs are a key to understanding the connections between virus infection and non-infectious disease but the mechanisms remain often elusive as long as we have not validated the myriad of PPIs involved.

**Virus interactions with the host metabolome**

There is increasing evidence that viruses not just highjack the host replication machinery, but also the host metabolic machinery. For instance, Adenovirus 5 proteins E4ORF1 and E4ORF6 co-immunoprecipitate with MYC in the nucleus, probably by directly interacting with the cancer protein MYC. While MYC has diverse effects on numerous target genes that it regulates, E4ORF1 induces MYC to activate a subset of glycolytic targets (viruses with a deletion of the E4 protein are defective for inducing glycolysis). Thai et al. (57) conclusively demonstrated that adenovirus induced glycolysis generates metabolites for increased nucleotide biosynthesis in infected cells.

It is not surprising that viruses manipulate host metabolism to generate more nucleotides and other compounds that they need for replication (58). However, in most cases it remains unclear whether interactions of virus proteins with host enzymes directly or indirectly reprogram metabolism.

**Conclusions and outlook**

Protein-protein interactions are at the core of any virus infection, hence a detailed understanding of such interactions is critical for understanding viral diseases but also critical for the development of new drugs. For a small group of viruses we have so much interaction data that it has become nearly impossible to evaluate their physiological significance any time soon, given that most databases do not document detailed follow-up studies. By contrast, for many viruses of lesser medical importance only few interactions are known and those are not sufficient to understand the infection mechanism.

Viruses evolve much quicker than their hosts, especially in RNA viruses, not the least because they can produce numerous (variant) virus particles by the time the host can mount an immune response. Thus viruses can also adapt their host-virus interactions faster than a host population can react by mutating its target proteins, although an individual immune system is usually able to fight off an infection (59).

With a rapidly growing number of human viruses, especially resulting from microbiome studies, we will identify many more viruses in humans. In the course of these studies, we will also find many more commensal viruses which do interact with their human host but may actually be beneficial, and potentially even help us to fight other pathogens and parasites.

**Conflicts of interest**

None.

**References**

1. Wylie KM*, et al.* (2014) Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol* 12:71.

2. de Chassey B*, et al.* (2008) Hepatitis C virus infection protein network. *Mol Syst Biol* 4:230.

3. Tripathi LP*, et al.* (2010) Network based analysis of hepatitis C virus core and NS4B protein interactions. *Mol Biosyst* 6(12):2539-2553.

4. Dolan PT*, et al.* (2013) Identification and comparative analysis of hepatitis C virus-host cell protein interactions. *Mol Biosyst* 9(12):3199-3209.

5. Ngo HT, Pham LV, Kim JW, Lim YS, & Hwang SB (2013) Modulation of mitogen-activated protein kinase-activated protein kinase 3 by hepatitis C virus core protein. *J Virol* 87(10):5718-5731.

6. Gautier VW*, et al.* (2009) In vitro nuclear interactome of the HIV-1 Tat protein. *Retrovirology* 6:47.

7. Jager S*, et al.* (2011) Global landscape of HIV-human protein complexes. *Nature* 481(7381):365-370.

8. Shapira SD*, et al.* (2009) A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 139(7):1255-1267.

9. Fossum E*, et al.* (2009) Evolutionarily conserved herpesviral protein interaction networks. *PLoS Pathog* 5(9):e1000570.

10. Calderwood MA*, et al.* (2007) Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci U S A* 104(18):7606-7611.

11. Forsman A, Ruetschi U, Ekholm J, & Rymo L (2008) Identification of intracellular proteins associated with the EBV-encoded nuclear antigen 5 using an efficient TAP procedure and FT-ICR mass spectrometry. *J Proteome Res* 7(6):2309-2319.

12. Khadka S*, et al.* (2011) A physical interaction network of dengue virus and human proteins. *Mol Cell Proteomics* 10(12):M111 012187.

13. Pichlmair A*, et al.* (2012) Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* 487(7408):486-490.

14. Kwofie SK, Schaefer U, Sundararajan VS, Bajic VB, & Christoffels A (2011) HCVpro: hepatitis C virus protein interaction database. *Infect Genet Evol* 11(8):1971-1977.

15. Ako-Adjei D*, et al.* (2015) HIV-1, human interaction database: current status and new features. *Nucleic Acids Res* 43(Database issue):D566-570.

16. Guirimand T, Delmotte S, & Navratil V (2015) VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 43(Database issue):D583-587.

17. Calderone A, Licata L, & Cesareni G (2015) VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res* 43(Database issue):D588-592.

18. Xiang Z, Tian Y, & He Y (2007) PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol* 8(7):R150.

19. Kumar R & Nanduri B (2010) HPIDB--a unified resource for host-pathogen interactions. *BMC Bioinformatics* 11 Suppl 6:S16.

20. Durmus Tekir S*, et al.* (2013) PHISTO: pathogen-host interaction search tool. *Bioinformatics* 29(10):1357-1358.

21. Aranda B*, et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* 8(7):528-529.

22. Alfarano C*, et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 33(Database issue):D418-424.

23. Chatr-Aryamontri A*, et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 41(Database issue):D816-823.

24. Salwinski L*, et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32(Database issue):D449-451.

25. Keshava Prasad TS*, et al.* (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res* 37(Database issue):D767-772.

26. Orchard S*, et al.* (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42(Database issue):D358-363.

27. Licata L*, et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40(Database issue):D857-861.

28. Croft D*, et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res* 42(Database issue):D472-477.

29. Uetz P*, et al.* (2006) Herpesviral protein networks and their interaction with the human proteome. *Science* 311(5758):239-242.

30. Zhang L*, et al.* (2009) Analysis of vaccinia virus-host protein-protein interactions: validations of yeast two-hybrid screenings. *J Proteome Res* 8(9):4311-4318.

31. Hauser R*, et al.* (2012) Bacteriophage protein-protein interactions. *Adv Virus Res* 83:219-298.

32. Blasche S, Wuchty S, Rajagopala SV, & Uetz P (2013) The protein interaction network of bacteriophage lambda with its host, Escherichia coli. *J Virol* 87(23):12745-12755.

33. Dyer MD, Murali TM, & Sobral BW (2008) The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog* 4(2):e32.

34. Wuchty S, Siwo G, & Ferdig MT (2010) Viral organization of human proteins. *PLoS One* 5(8):e11796.

35. Brito AF & Pinney JW (2017) Protein-Protein Interactions in Virus-Host Systems. *Front Microbiol* 8:1557.

36. Dix A, Vlaic S, Guthke R, & Linde J (2016) Use of systems biology to decipher host-pathogen interaction networks and predict biomarkers. *Clin Microbiol Infect* 22(7):600-606.

37. Navratil V, de Chassey B, Combe CR, & Lotteau V (2011) When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Syst Biol* 5:13.

38. Durmus Tekir SD & Uelgen K (2013) Systems biology of pathogen-host interaction: Networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era. *Biotechn. J.* 8(1):85-96.

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

40. Korth MJ, Tchitchek N, Benecke AG, & Katze MG (2013) Systems approaches to influenza-virus host interactions and the pathogenesis of highly virulent and pandemic viruses. *Semin Immunol* 25(3):228-239.

41. Nourani E, Khunjush F, & Durmus S (2015) Computational approaches for prediction of pathogen-host protein-protein interactions. *Front Microbiol* 6:94.

42. Dyer MD, Murali TM, & Sobral BW (2011) Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect Genet Evol* 11(5):917-923.

43. Durmus Tekir S, Cakir T, & Ulgen KO (2012) Infection Strategies of Bacterial and Viral Pathogens through Pathogen-Human Protein-Protein Interactions. *Front Microbiol* 3:46.

44. Halehalli RR & Nagarajaram HA (2015) Molecular principles of human virus protein-protein interactions. *Bioinformatics* 31(7):1025-1033.

45. Mariano R, Khuri S, Uetz P, & Wuchty S (2016) Local Action with Global Impact: Highly Similar Infection Patterns of Human Viruses and Bacteriophages. *mSystems* 1(2).

46. Ishitsuka M, Akutsu T, & Nacher JC (2016) Critical controllability in proteome-wide protein interaction network integrating transcriptome. *Scientific reports* 6:23541.

47. Nacher JC & Akutsu T (2014) Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. *J. Compl. Networks* 2(4):394-412.

48. Wuchty S (2014) Controllability in protein interaction networks. *Proc Natl Acad Sci U S A* 111(19):7156-7160.

49. Khuri S & Wuchty S (2015) Essentiality and centrality in protein interaction networks revisited. *BMC bioinformatics* 16:109.

50. Wuchty S, Boltz T, & Kucuk-McGinty H (2017) Links between critical proteins drive the controllability of protein interaction networks. *Proteomics*.

51. Vinayagam A*, et al.* (2016) Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci U S A* 113(18):4976-4981.

52. Mariano R & Wuchty S (2017) Structure-based prediction of host-pathogen protein interactions. *Curr Opin Struct Biol* 44:119-124.

53. Murali TM, Dyer MD, Badger D, Tyler BM, & Katze MG (2011) Network-based prediction and analysis of HIV dependency factors. *PLoS Comput Biol* 7(9):e1002164.

54. Tastan O, Qi Y, Carbonell JG, & Klein-Seetharaman J (2009) Prediction of interactions between HIV-1 and human proteins by information integration. *Pac Symp Biocomput*:516-527.

55. Morales-Sanchez A & Fuentes-Panana EM (2014) Human viruses and cancer. *Viruses* 6(10):4047-4079.

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

57. Thai M*, et al.* (2014) Adenovirus E4ORF1-induced MYC activation promotes host cell anabolic glucose metabolism and virus replication. *Cell Metab* 19(4):694-701.

58. Miyake-Stoner SJ & O'Shea CC (2014) Metabolism goes viral. *Cell Metab* 19(4):549-550.

59. Christiaansen A, Varga SM, & Spencer JV (2015) Viral manipulation of the host immune response. *Curr Opin Immunol* 36:54-60.

60. Boppana SB & Fowler KB (2007) Persistence in the population: epidemiology and transmisson. *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, eds Arvin A, Campadelli-Fiume G, Mocarski E, Moore PS, Roizman B, Whitley R, & Yamanishi KCambridge).

61. Burrel S*, et al.* (2017) Ancient Recombination Events between Human Herpes Simplex Viruses. *Mol Biol Evol* 34(7):1713-1721.

62. Johnston C, Gottlieb SL, & Wald A (2016) Status of vaccine research and development of vaccines for herpes simplex virus. *Vaccine* 34(26):2948-2952.

63. Looker KJ*, et al.* (2015) Global and Regional Estimates of Prevalent and Incident Herpes Simplex Virus Type 1 Infections in 2012. *PLoS One* 10(10):e0140765.

64. Szucs TD, Berger K, Fisman DN, & Harbarth S (2001) The estimated economic burden of genital herpes in the United States. An analysis using two costing approaches. *BMC Infect Dis* 1:5.

65. Menzies NA*, et al.* (2011) The cost of providing comprehensive HIV treatment in PEPFAR-supported programs. *AIDS* 25(14):1753-1760.

66. Johnson NP & Mueller J (2002) Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bull Hist Med* 76(1):105-115.

67. Merson MH, O'Malley J, Serwadda D, & Apisuk C (2008) The history and challenge of HIV prevention. *Lancet* 372(9637):475-488.

68. Molinari NA*, et al.* (2007) The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* 25(27):5086-5096.

69. Patel MK*, et al.* (2016) Progress Toward Regional Measles Elimination - Worldwide, 2000-2015. *MMWR Morb Mortal Wkly Rep* 65(44):1228-1233.

70. Wong JB, McQuillan GM, McHutchison JG, & Poynard T (2000) Estimating future hepatitis C morbidity, mortality, and costs in the United States. *Am J Public Health* 90(10):1562-1569.

71. Wang H & Collaborators GBoD (2016) Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 388(10053):1459-1544.

72. Maynard JE (1990) Hepatitis B: global importance and need for control. *Vaccine* 8 Suppl:S18-20; discussion S21-13.

73. Ott JJ, Stevens GA, Groeger J, & Wiersma ST (2012) Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity. *Vaccine* 30(12):2212-2219.

74. Keshavarz K*, et al.* (2015) Economic burden of hepatitis B virus-related diseases: evidence from iran. *Hepat Mon* 15(4):e25854.

75. Fischer M (2016) Zika virus epidemiology update.

76. Tang XC*, et al.* (2014) Identification of human neutralizing antibodies against MERS-CoV and their role in virus adaptive evolution. *Proc Natl Acad Sci U S A* 111(19):E2018-2026.

77. Oberholtzer K (2004) *Learning from SARS: Preparing for the Next Disease Outbreak--Workshop Summary.* (National Academies Press).

78. Lee J-W & McKibbin WJ (2004) Estimating the global economic costs of SARS. *Learning from SARS: preparing for the next disease outbreak: workshop summary.,* Institute of Medicine (US) Forum on Microbial Threats., (National Academies Press, WAshinton, DC).

79. Simasek M & Blandino DA (2007) Treatment of the common cold. *Am Fam Physician* 75(4):515-520.

80. Bartsch SM, Lopman BA, Ozawa S, Hall AJ, & Lee BY (2016) Global Economic Burden of Norovirus Gastroenteritis. *PLoS One* 11(4):e0151219.

81. Robilotti E, Deresinski S, & Pinsky BA (2015) Norovirus. *Clin Microbiol Rev* 28(1):134-164.

**Tables**

**Table 1**. **The 20 best-studied viruses (by number of genomes sequenced)**. Sequence numbers as of July, 2016. Clustered sequenced were clustered at ≥98% sequence identity). U/C = un-/ clustered. Genome data from Genbank.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Baltimore Class | Family name | Sequences (unclust.) | Disease (examples) | U/C | Total complete genomes | Complete genomes (clustered ) |
| III (dsRNA) | reoviridae | 65870 | Rare diarrhea | 5.50 | 31945 | 5803 |
| IV (+ssRNA) | flaviviridae | 225112 | Zika | 3.88 | 7837 | 2019 |
| VII (dsRNA-RT) | hepadnaviridae\* | 78558 | hepatitis | 3.72 | 7248 | 1946 |
| II (ssDNA) | geminiviridae | 13158 | --- | 2.77 | 6421 | 2316 |
| IV (+ssRNA) | picornaviridae | 85636 | Cold etc | 2.30 | 3447 | 1500 |
| VI (ssRNA-RT) | retroviridae | 716088 | AIDS etc | 1.37 | 2890 | 2103 |
| animi (ssDNA) | circoviridae | 7838 | --- | 4.99 | 2706 | 542 |
| V (-ssRNA) | phenuiviridae | 4139 | Rift Valley fever | 4.37 | 1678 | 384 |
| IV (+ssRNA) | coronaviridae | 19164 | SARS | 4.84 | 1549 | 320 |
| IV (+ssRNA) | potyviridae | 16115 |  | 1.82 | 1536 | 843 |
| I (dsDNA) | papillomaviridae | 17847 | Warts, cancer | 3.80 | 1364 | 359 |
| I (dsDNA) | polyomaviridae | 8604 | Rare cancers | 7.79 | 1277 | 164 |
| V (-ssRNA) | filoviridae | 2165 | Ebola | 34.03 | 1259 | 37 |
| IV (+ssRNA) | togaviridae | 8924 | rubella | 9.04 | 1239 | 137 |
| V (-ssRNA) | pneumoviridae | 22578 | Cold-like | 20.18 | 1231 | 61 |
| II (ssDNA) | nanoviridae | 3110 | --- | 4.20 | 1183 | 282 |
| IV (+ssRNA) | caliciviridae | 32405 | gastroenteritis | 3.67 | 1072 | 292 |
| V (-ssRNA) | paramyxoviridae | 29726 | measles | 3.08 | 1008 | 327 |
| IV (+ssRNA) | bromoviridae | 4677 | (plants) | 1.99 | 764 | 384 |
| V (-ssRNA) | arenaviridae | 2639 | e.g. Lassa fever | 1.62 | 758 | 469 |

Statistics of viruses known to infect humans, genomes sequenced, genetic diversity?

\* Hepadnaviruses have an RNA intermediate and thus are not strict DNA viruses.

**Table 2**. **Human disease burden by viruses.** Infections include infected number of people while morbidity and mortality include those that get sick or die, respectively. Cost is the economic damage of these viral diseases from hospitalization or lost work time. Unless otherwise indicated, figures are yearly.

Need to add dengue virus, hRSV or rabies virus

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Virus (class)** | **Infections** | **Morbidity** | **Mortality** | **Cost** | **Refs** |
| HSV-1/2 | 3.7 B / ~700M | 3M/yr (US) | low | $540M (US) | (60-64) |
| HIV-1/2 | 36M (world) | 2.1M/yr ww | 25M total1 | $13.7B (US) | (65) |
| Influenza | >30M (US)3 | 100-600K (US)4 | 50M 19182 | $10-90B | (66-68) |
| Measles | >20 M (ww) | 250k ww | 140-500k5 | $3-7B (US) | (69) |
| Hepatitis C | 60-120M (ww) | 4M | 500k ww | $10B 6 | (70, 71) |
| Hepatitis B | 248 M ww/yr, ~2.5 B ww total | 350M ww total | 600k ww | $1B (US) | (72-74) |
| Zika | 740k S Amer | >2,6k 7 | low | $18B ww | (75) |
| MERS-CoV | 2067 ww total | 1179 ww total | 720 ww total | $15-20B | (76) |
| SARS-CoV |  | 8098 ww total | 774 ww total | $40B ww | (77, 78) |
| Common cold (rhinovirus) | 1B Cold/year (US) | 10-40% of common colds | low | $20B (US) | (79) |
| Norovirus (gastroenteritis) | 19-21M (US); 685M (ww) | 699M ww | 570-800 (US); 200K children ww  219K ww | $4.2B (indirect); $60.3B total ww | (80, 81) |

1 globally, since 1981. 2 Spanish flu of 1918. 3 30 million outpatient visits. 4 100-600 thousand hospitalizations. 5 The death rate is decreasing, from 535,000 deaths in 2000 to 139,300 deaths in 2010. 6 $10·7 billion in direct medical expenditures in the USA for HCV-related disease from 2010 to 2019. 7 cases of microcephaly. K,M,B = thousand, million, billion, WW = worldwide, SA = South America.

**Table 3: Overview of host-pathogen and other protein-protein interaction databases that provide human-virus protein interactions. PHI-PPIs were drawn from databases shown in bold.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **database** | **database type** | **pathogens** | **webpage** | **Physical PPIs\* (March, 2018)** | **Direct PPIs\*\* (March, 2018)** | **Ref.** |
| HCVPro | PHI | only HCV | http://www.cbrc.kaust.edu.sa/hcvpro/ | 618 | 565 | (14) |
| HIV-1 @NCBI | PHI | only HIV | https://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/ | 6,824 | 1,594 | (15) |
| PHIDIAS | PHI |  | http://www.phidias.us | \*\*\* | \*\*\*\* | (18) |
| PHISTO | PHI |  | http://www.phisto.org | - | - | (20) |
| **HPIDB** | **PHI** |  | **http://www.agbase.msstate.edu/hpi/main.html** | **19,681** | **10,628** | (19) |
| VirHostNet | PHI |  | http://virhostnet.prabi.fr | 28,814 | \*\* | (16) |
| **VirusMentha** | **PHI** |  | **http://virusmentha.uniroma2.it** | **10,692** | **5,863** | (17) |
| DenHunt | PHI |  | http://proline.biochem.iisc.ernet.in/DenHunt/) | 1,064 | 682 |  |
| DenvInt | PHI |  | https://denvint.000webhostapp.com | 784 | 784 |  |
| **BioGRID** | **PPI** |  | **https://thebiogrid.org/** | **2,427** | **1,936** |  |
| **DIP** | **PPI** |  | **http://dip.mbi.ucla.edu/dip/** | **519** | **430** |  |
| **IntAct** | **PPI** |  | **https://www.ebi.ac.uk/intact/** | **14,282** | **5,934** |  |
| **MINT** | **PPI** |  | **https://mint.bio.uniroma2.it/** | **6,400** | **2,530** |  |

\* “Physical PPIs” refers to PPIs for which there is experimental evidence of a physical interaction, but absence of evidence for a direct interaction using PSI-MS controlled vocabularies

\*\* “Direct PPIs” refers to PPIs for which there is experimental evidence of a physical, direct interaction using PSI-MS controlled vocabularies

\*\*\* PPI information and evidence requires manual extraction or text-mining



**\***\*\*\* evidence of direct interactions available through PSICQUIC

**Table 4: Number of host-virus protein-protein interactions of major human virus families. Interaction numbers are pooled from BioGRID, DIP, HPIDB, IntAct, MINT, and VirusMentha.**

|  |  |  |
| --- | --- | --- |
| **viral family** | **# virus-human PPIs (physical / direct)** | **representative virus-human PPIs (physical / direct)** |
| herpesviridae | 5957/3570 | herpesvirus 4 / epstein-barr (3,163/1,049); herpesvirus 8 (1,643/1,623)e |
| papillomaviridae | 4645 | papillomavirus types 1a,3,5,6,6b,8,9,11,16,18,32,33,39 (4,275/2,649) |
| orthomyxoviridae | 3748/953 | influenza A (3,746/952) |
| retroviridae | 2998 | hiv-1 (2,540/2,365); primate t-lymphotropic virus 1 (254/240) |
| flaviviridae | 1475 | hepatitis C (1,082/802); dengue (535/535) |
| paramyxoviridae | 665 | measles (481/445); nipah henipavirus (133/2) |
| adenoviridae | 451 | adenovirus types 2,5,12 (378/211) |
| pneumoviridae | 270 | respiratory synctial virus a2 (262/258) |
| poxviridae | 247 | vaccinia virus (190/47); variola virus (18/18) |
| filoviridae | 177 | ebola virus (154/11); marburg virus (23/0) |
| polyomaviridae | 165 | macaca mulatta polyomavirus 1 (79/65); jc polyomavirus (41/1); human polyomavirus 1 (39/0) |
| hepadnaviridae | 128 | hepatitis b (127/111) |



****

**Figure 1: Venn diagram of sets of interactions between proteins of different virus families and the human host.**