Published in final edited form as:

Curr Protoc Bioinformatics. 2018 March ; 61(1): 8.26.1–8.26.12. doi:10.1002/cpbi.44.

Leveraging Experimental Details for an Improved Understanding of Host-Pathogen Interactome

Mais Ammari^{1,§}, Fiona McCarthy¹, and Bindu Nanduri^{2,3}

¹School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ 85721, Phone: 520-626-2875, Fax: 520-626-5602

²Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS 39762, Phone: 662-325-5859, Fax: 662-325-1031

³College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, Phone: 662-325-5859, Fax: 662-325-1031

Abstract

An increasing proportion of curated host-pathogen interaction (HPI) information is becoming available in interaction databases. This data represents detailed, experimentally-verified, molecular interaction data which may be used to better understand infectious diseases. By their very nature, HPIs are context dependent, where the outcome of two proteins as interacting or not depends on the precise biological conditions studied and approaches used for identifying these interactions. The associated biology and the technical details of the experiments identifying interacting protein molecules are increasing being curated using defined curation standards but are overlooked in current HPI network modeling. Given the increase in data size and complexity, awareness of the process and variables included in HPI identification and curation, and their effect on data analysis and interpretation is crucial in understanding pathogenesis. We describe the use of HPI data for network modeling, aspects of curation that can help researchers to more accurately model specific infection conditions, and provide examples to illustrate these principles.

Keywords

host-pathogen interaction; interactome; network analysis; biocuration; proteomics; controlled vocabulary

1. INTRODUCTION

Despite decades of research and technological advances, infectious diseases continue to cause significant morbidity and mortality in humans and animals worldwide (Nii-Trebi, 2017). Host-microbe interactions ultimately determine the outcome of an infection, and understanding these interactions offer avenues for therapeutic intervention. Host-pathogen interactions during initiation, establishment and clearance of infection determine the ultimate outcome of an encounter between a host and a pathogenic microorganism, and are

[§]Corresponding author: mammari@email.arizona.edu.

the focus of this chapter. The complex host ecosystem also includes a multitude of interactions with the ‘microbiome’, the collective genome of the microbiota residing in different anatomical locations within the host. However, a description of host-microbiome interactions is not the focus of this chapter, for this information we point the reader to recent a review article on this topic (Thomas et al., 2017).

The ability of a pathogen to establish an infection is dependent on manipulating host cellular functions, including host defense mechanisms and replication pathways. The crosstalk between host and a pathogen during infection involves molecular interactions between a set of molecules (e.g. of proteins, nucleic acids and small molecules) which are the basis of any biological function. Molecular interactions involved in pathogenesis are characterized into intra- (host-host and pathogen-pathogen) and inter- (host-pathogen) species interactions. In the context of this chapter, we will focus and confine host-pathogen interactions (HPI) to inter-species protein-protein interactions.

1.1. Network modeling of HPIs

Since proteins function as elements within a network of their interactions, the study of their sum is greater than their individual parts, which is the key to their network analysis approaches. Protein-protein interactions are identified either by experimentation (e.g. yeast two-hybrid and pull-down assays) or *in silico* predictions; this information is then curated into databases to make it easily accessible. Interaction networks are often visualized as graphs representing relationships between nodes (proteins) and edges (connections between proteins). Subsequent analysis of these networks identifies key proteins and sub-networks which may be targeted to protect against or reduce infection. Any network analysis is fundamentally circumscribed by the starting set of interactions, and limited by sophisticated analyses that visualization tools enable. Therefore, effective network analysis requires a good understanding of how to find the best set of protein interactions and software for network analysis.

During the last decade, proteomic technologies for the identification of protein interactions lead to an exponential increase in published literature studying HPIs. The collective data from all available HPIs publications represent a more global and comprehensive model of the infectious disease under study. Despite the increase in such studies, the understanding of many infectious diseases is incomplete, even when a large amount of interactions are identified.

Data included in the network modeling vary in quality and depth of information. This is dependent upon available data in different interaction databases. Importantly, the classification of a pair of proteins as interacting or not is context dependent on experimental techniques used to identify them. The increase in curated HPI data, and the information also collected about the biological environment and technical details of the identification of this interaction, means that researchers have more opportunities to make more detailed network analyses to support their study of infectious disease. In this overview chapter, we will briefly discuss identification, annotation and analysis of HPIs. The goal of this overview is to enable researchers to do more sophisticated analyses of HPIs by better understanding the process and variables involved in a comprehensive study of host-pathogen interactomes and the

effect of data type in understanding the dynamic nature of interactomes. Using examples, we demonstrate how awareness of the increasing amount of annotated experimental detail helps answer more specific questions on the interplay between a pathogen and its host.

2. TRANSLATING AN INFECTIOUS DISEASE QUESTION INTO KNOWLEDGE USING HPI DATA

HPIs often span multiple host tissues, different pathogen sub-types during infection and include diverse microenvironments. Thus, host-pathogen interactomes are dynamic in nature and vary due to differences in the type of infection and infected tissue/cell, location of interacting proteins, and the stage of infection. A comprehensive mechanistic understanding of such complex system requires modeling of host-pathogen interactome data in a systems biology framework, as described in the following four major steps [Figure 1]:

- i. Identification of HPIs using experimental approaches. Whether or not a HPI is identified depends on (1) the type of experimental assays (e.g. yeast two-hybrid) used that differ in scale and sensitivity and (2) experimental details (e.g. cell type). Such variables also have a significant impact on the outcome of an infection. Furthermore, the experimental approach used to identify HPI can impact the confidence in the identification, as some approaches are prone to more false positives.
- ii. Quality of HPI annotation accessible in different bioinformatics databases that store HPI. Biocuration of HPI from experimental data can generate detailed contextual information associated with HPI and described using structured vocabularies. However different databases may collect different contextual information as well as focusing on different host-pathogens.
- iii. Functional network analysis of available HPI data. Network analysis requires data integration and use of suitable tools to identify key components and pathways of HPI network to generate a testable biological model.
- iv. Data integration and refinement of HPI networks is complex. A model that cannot adequately describe the original infectious disease question is refined by integrating additional data generated by targeted experiments.

The first step in this process (the experimental approach used to identify HPI) determines the type of information associated with the interaction, which will impact subsequent confidence in accuracy and sensitivity. Thus, the experimental approach reflects how data is annotated, analyzed and interpreted for a given host-pathogen system. Therefore, network modeling based on HPI where the nodes and edges are information rich pertaining to detailed functions of interacting proteins enables subsequent refinements of the analysis to understand complex HPIs. The demand for such experimental data can be seen by the increase in databases (e.g. IntAct (Orchard et al., 2014) and HPIDB (Ammari, Gresham, McCarthy, & Nanduri, 2016)) that collect both experimentally verified molecular interactions and the conditions under which they occur. HPI data that can be utilized to study infections can be as simple as host protein A interacts with pathogen protein B. However, when researchers limit their infectious disease question to ‘which genes/proteins’ interact

with host/pathogen', they are not utilizing additional information associated with the HPI data for assessment of the biological context of the HPI complexity. In the following sections, we will demonstrate the importance and impact of experimental settings in each step in host-pathogen interactome modeling shown in Figure 1.

2.1. HPI identification

The challenge with identifying (and curating) HPI-related molecular interaction data is that all molecular interactions are context dependent, relying on biological (e.g. type of infection) and experimental (e.g. method or tissue type) factors. Minor changes in experimental settings can have considerable impact on the type and number of interactions that are identified, as demonstrated by the examples described in this section.

2.1.1. Impact of experimental methods on the identification of HPI—

Experimental methods commonly used for identifying protein-protein interactions are classified into four main groups; biochemical (e.g. co-immunoprecipitation), genetic (e.g. yeast two-hybrid), biophysical (e.g. surface plasmon resonance), and microscopic (e.g. confocal microscopy). For a more detailed discussion of experimental approaches used to identify interactions between pathogen and host, we direct the reader to a recent review article (Jean Beltran, Federspiel, Sheng, & Cristea, 2017).

Different experimental methods allow the detection of interactions with different sensitivities. For example, comparison of *Escherichia coli* HPIs identified using yeast two-hybrid and stable isotope labeling of amino acids (SILAC), two commonly used assays for molecular interaction identification, showed very little overlap. Using SILAC identified five host proteins which interacted with EspZ, while yeast two-hybrid identified 11 host protein interactions; only CD98 protein was common to both assays (Shames et al., 2010). In addition, each experimental method identifies false positives and false negatives. For example, host proteins identified using yeast two-hybrid assays could be between proteins that may not be present in the same cellular location *in vivo*. Moreover, pull down assays used for interaction identification could include indirect (complex) proteins that co-purify with protein of interest. However, techniques such as cross-linking strategies can enhance the chance of detecting a direct interaction *in situ*.

These examples demonstrate that the type of detection method used can affect quality, scale and confidence of the identified set of interactions. Therefore, interactions validated by more than one method are more likely to be true positive and physiologically relevant. There are multiple challenges associated with this, as the interaction space is infinite due to the variety and the complexity of pathogen and host microenvironments, coupled with inherent limitations of the assays. Nevertheless, experimental techniques are constantly changing to provide a richer, more sensitive and detailed information set that can be evaluated by researcher in considering the appropriate method to apply to the host-pathogen system under study.

2.1.2. Impact of experimental conditions on the identification of HPI—When identifying HPI using experimental methods, it is important to consider the details regarding the infection that can impact protein interactions. These include, but are not limited to, cell

type, pathogen type, host/pathogen protein selected for identifying HPI, protein sequence mutation and protein tags. Tags are peptide sequences genetically attached onto a recombinant protein for various purposes including purifying a protein from their biological source. For example, *E. coli* effector protein EspZ protein immunoprecipitates with host protein TIM17b only when the C-terminal of EspZ has an affinity tag, and not when the tag is in the N-terminal (Shames, Croxen, Deng, & Finlay, 2011). Likewise, a tag expressed as either an N- or C-terminal fusion of Ebola Virus VP24 protein identified 48 and 51 interacting human proteins, respectively, of which 40 proteins that are common to both N and C terminal fusions (Garcia-Dorival et al., 2014). Adding to this complexity, pathogen genotypes and isolates can contribute to different pathology during infection; and interaction differences between different isolates explain differences in infectivity and pathogenesis. For example, a study using mass spectrometry and human embryonic kidney 293 cell type showed two different sets of interactions between Hepatitis C virus (HCV) genotype 1a isolate H and genotype 1b isolate Con1. Additionally, there are common and specific patterns of invasion and proliferation of a pathogen within different hosts and cell types. Thus, using a relevant model system that accounts for specificity of pathogen infection can be critical when identifying HPI. In addition, experimental design using two different cell types can identify a different set of interactions. For example, a binding study of HCV 3' X RNA element showed differences among interactions with proteins from human Huh7.5, hepatocytic clone B, and 293t cell types analyzed by mass spectrometry (Park et al., 2013). Multiple *in vitro* and *in vivo* experimental systems are now available to generate HPI models.

It is also worth noting that using experimental method and conditions can be used identify certain aspect of the pathogen biology. For example, for a given pathogen type, different pathogen proteins are expected to have different number of interactions. A global study to identify HPI for 67 Kaposi's sarcoma-associated herpesvirus (KSHV) proteins using mass spectrometry under otherwise identical conditions, identified 564 KSHV-human interactions with different numbers of interactions (ranging from 1 to 71) for each of the viral proteins (Davis et al., 2015). Furthermore, a mutation in the binding region of the interacting pathogen protein with the host can negatively affect the HPI identified and the role in pathogen life cycle. For example, a mutant HCV core protein expressing a defective JAK-binding motif was found to be impaired in its ability to produce infectious viruses with no effect on the viral genome replication. This suggests the essential requirement of HCV core-JAK protein interaction for efficient production of infectious viruses and the potential of using core-JAK interaction in anti-HCV therapy (Lee, 2013). Several infection models using mutant pathogens that can impact infection severity and outcome are usually required to validate and produce a predictive model of HPIs.

In summary, the experimental method and conditions can impact the identification of HPI. Therefore, there is a need for the documentation and use of such data in host-pathogen interactome studies.

2.2. HPI annotation: biocuration and databases

There are many different databases that provide molecular interaction data but only a few contain HPIs, and as such it may be difficult to find all the interactions for the pathogen of interest from a single database source. Moreover, HPIs available in these databases could include manually curated (from experimental data reported in literature) and computationally predicted HPI, and researchers should be aware of the distinction between these two curation approaches to help them determine which HPI to include in their analysis.

The increasing number of available publications about HPIs provides a complex search problem for the researcher. In addition, such publications contain a diverse set of experimental design and techniques which require understanding of the type of HPI data they generate. Thus, experimental data is not readily accessible for HPI network analysis until it is extracted and entered into databases that enable the systematic capture of a molecular interaction data compliant with community standards for data sharing. The consistent, structured format of HPI annotations can easily enable users to identify, filter and import the data into software for network analysis.

The increase in databases collecting molecular interactions (including HPI data) led to the development of the Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC), an interface for integrating and identifying sets of molecular interactions (Aranda et al., 2011). Each of the 36 databases (as of July 2017) that are available in PSICQUIC includes interaction data that can be categorized into predicted, manually curated, specialized (i.e. interactions for a specific system of interest), archival (i.e. integrated data with no focus of any specific organism) generic (i.e. contains interactions without restriction to species) or a combination of more than one type. Searching PSICQUIC using a single query (such as the pathogen taxon identifier) will identify all available interactions, including pathogen-pathogen and host-pathogen, from all participating databases, allowing researchers to rapidly access data from multiple sources. In addition, the PSICQUIC query can identify interactions supported by multiple experiments and publications. However, the depth and quality of the annotations varies widely between independent databases, which can cause several problems including downloading and combining data from multiple resources that differ in the use of identifiers, description of method used, and quality assessment. Therefore, many of the databases included in PSICQUIC use a standardized data format required to provide a more comprehensive, high confident, accessible dataset.

The Protein Standard Initiative-Molecular Interaction (PSI-MI) Consortium provides molecular interaction data using standardized files, including a detailed XML file and a simplified tab-delimited MITAB file (Kerrien et al., 2007). A common controlled vocabulary was developed to describe biocurated data such as interaction type, experimental techniques, and molecular features. The consistent way of capturing the interaction data enables searching, filtering, combining, downloading, and uploading data for visualization and analysis. For example, users can use the interaction type (indicates proximity of the interacting proteins) to assess the evidence for interactions, i.e. direct (indicating that the two proteins are in close contact) versus association (indicating the use of experimental methods that identify a loss complex). The International Molecular Exchange (IMEx) Consortium curates data using these controlled vocabulary rules and co-ordinates curation efforts to

ensure that each publication is curated once by participating databases (Orchard et al., 2012). The systematic annotation of molecular interaction data to IMEx standards can be done using the European Bioinformatics Institute (EBI) IntAct interface (Orchard et al., 2014). IMEx databases are part of PSICQUIC and include 11 databases, of which IntAct, The Host-Pathogen Interaction DataBase (HPIDB) (Ammari et al., 2016), MINT (Orchard et al., 2014) and UniProtKB (The UniProt, 2017) are examples of IMEx databases with HPIs. For more information regarding standard data formats and databases for molecular interactions see review (Orchard et al., 2014).

The IMEx database, HPIDB (<http://www.agbase.msstate.edu/hpi/main.html>), provides manually annotated HPI data focusing on agricultural pathogens and related species. To provide a comprehensive HPI set, HPIDB integrates HPI data from other PSICQUIC resources and also provides predicted interactions based upon sequence similarity to experimentally annotated datasets. All HPIDB data can be downloaded in a PSI-MITAB 2.5 format directly and detailed manual annotations can be accessed in multiple formats from PSICQUIC.

Detailed annotation of molecular interactions includes information regarding interacting proteins (e.g. accession, species), the method used to identify the interaction (e.g. assay type, interaction type), cellular context (e.g. the protein's biological role (enzyme or enzyme target) and experimental role (e.g. bait or prey)), Gene Ontology (GO) (Ashburner et al., 2000), protein modifications (e.g. tags, interacting regions and mutations) and expression systems (e.g. tissue/cell type). Availability of this detailed information allows researchers to:

- i. Interpret the HPI networks to derive practical information from these analyses. For example, curation can provide interacting domains that mediate the interaction, and molecular alterations (such as mutations) that influence the interaction. This information allows the identification of regions to target for interaction disruption.
- ii. Evaluate the reliability of an interaction based on the confidence values. Confidence for a HPI can be based on curation information about the interaction type (i.e. physical, association, direct, co-localization etc.), type of experimental assay, and whether the interaction is validated, and all these factors can be used to assess the quality and to filter the reported HPI. Interactions detected in different studies and/or using different methods increase the confidence. Scoring based on this information is available for all databases in the IMEx consortium, derived using an implementation of MIscore (Villaveces et al., 2015) which can be used for filtering or ranking of data.
- iii. Data merging and comparison by the use of consistent information.

2.3. Network analysis and interpretation of HPIs

HPIs can trigger a number of cellular processes, including host defenses against a pathogen and pathogen evasion of host defense. The balance between these two processes ultimately determines the infection phenotype. Thus, considering both the pathogen characteristics and the understanding of functional consequence of interactions within the host and pathogen is

essential for studying infections. Functional network analysis to identify key host proteins targeted by pathogens and also functional processes controlled by host proteins interacting with pathogens is the most common method used to gain systems-level understanding of HPI dynamics. Pathogen proteins bind specific host targets, especially highly connected proteins (referred to as “hubs”) involved in specific pathways in the host network. The importance of each of the pathogen protein during infection is directly related to their number of interactions with the host. Pathogen and host proteins with high connectivity during infection could be important therapeutic targets, and may differ among different pathogen types/isolates and host species. In contrast to intra-species interactions, HPis require analysis of host changes in perspective to the pathogen infection type. HPI network analysis allows an improved understanding of how each of the pathogen proteins contributes not only to the pathogen survival in the host during infection but also host cellular processes involved in pathogenesis.

A variety of tools have been developed to help researchers analyze and visualize their data within the context of biological networks. Tools can vary with respect to the size of networks supported, ability to compare different data sets, functional enrichment background and type, and the species they support. Researchers typically require a combination of several tools to identify proteins of interest, construct a network, and visualize/analysis network for a comprehensive view of data. To date, Cytoscape (and its Apps) is a commonly used, powerful toolkit that provides a wide range of functions for network visualization and analysis (Shannon et al., 2003; Su, Morris, Demchak, & Bader, 2014). Cytoscape allows customization and interactive exploration of network entities and their associated data, which includes detailed annotations. Cytoscape Apps allow analysis of network topology and modules, as well as doing GO and pathway functional enrichment analysis. Researchers can query and import molecular interactions from PSICQUIC into Cytoscape using the “import network from public databases” option. This approach allows researchers to rapidly identify an integrated set of HPI interactions and do sophisticated network analyses.

2.3.1. Using curated details for detailed analysis—Selecting the most comprehensive set of HPis and doing network analysis based on this set (as described above) is a common approach that can provide the researcher with a useful overview. For host-pathogen pairs where detailed, curated data is available the application of this additional information to the analysis can provide more detailed and specific information about the interplay between a pathogen and its host. We demonstrate this using HCV curated HPI analysis as an example.

Hepatitis C virus (HCV) is an RNA virus that causes cirrhosis and hepatocellular carcinoma in humans. The rapidly growing number of experimental and annotated HCV-host protein interactions enabled multiple network-based analyses of viral infection (de Chassey et al., 2008; Han, Niu, Wang, & Li, 2016; Poortahmasebi et al., 2016). Despite these efforts HCV continues to be a major public health challenge. Recently, the HCV-host protein interactions were annotated with detailed experimental data using controlled vocabularies. Therefore, we chose HCV interactions as an example to demonstrate the impact of experimental details on data use, comparison and filtering. We will first summarize existing HCV-host protein

interactions and the associated detailed annotations, and then discuss how detailed network analysis reveals aspects of HCV infection.

2.3.1.1 Overview of annotated HCV-host interactome data: A comprehensive set of HCV-host protein interaction data was obtained by searching PSICQUIC using HCV taxon identifier '11103'. Figure 2 details the steps to download data from PSICQUIC. Our search identified 5,629 interactions (as of July 2017), including HCV-HCV, HCV-host, and HCV-small molecule interactions. Of these interactions, 2,037 are HCV-host protein interactions, with 60% deemed non-redundant based on interacting proteins. HCV-host protein interactions are manually curated from 230 publications and include experimental details regarding the interactions from multiple databases such as HPIDB, IntAct, MINT and UniProtKB. Of the 2,037 interactions, 84% are identified by one experimental method and 90% are assigned with one interaction type. Two hybrid and co-immunoprecipitation are the most common methods used for identifying HCV-host protein interactions [Figure 3A], where association is the most common interaction type [Figure 3B]. The number and type of methods and thus the interaction types identified reflect the MIscore for the interaction; where the higher the MIscore number, the higher the confidence in the interaction [Figure 3C]. This distribution of the data can change as more biological testing and subsequent curation is completed, and could vary for different host-pathogen systems. Any of the curated details about the interactions such as the detection method, interaction type, and confidence score can be used by researchers to filter interactions for network analyses, depending on their analysis goal for different host-pathogen pairs.

Since pathogens, including HCV, have fewer genes compared to their hosts, infection involves interactions between a few pathogen gene products and the more complex pool of host proteins. One network analysis strategy is to identify pathogen proteins with high connectivity with host proteins to be examined for their therapeutic potential. Genome-wide studies of HCV gene products indicated that viral protein NS3 had one of the highest host connectivity, while NS4B protein had the lowest (de Chasse et al., 2008; Germain et al., 2014). Therefore, we investigated the distribution of known interactions for each HCV gene product. Our collective data of annotated interactions (from multiple assays and settings) shows that HCV interactions involve 1,004 host proteins and 13 HCV gene products from different HCV genotypes/isolates. Figure 4 shows the number of interactions, publication and interacting host for each of HCV gene product. We note that the number of curated viral protein interactions is dependent upon not only the number of interactions the protein is capable of, but also the number of published reports (available for curation), HCV genotype/isolate under study, and the type of method used.

2.3.1.2. Interactome analysis of HCV infection: The *in vivo* target cell of HCV is the primary human hepatocyte. However, *in vitro* systems use a range of systems, so detailed curation includes information on the type of host cells/organisms/systems used to identify interactions. Example of cells used in HCV studies are human hepatoma cell lines huh7, and its derivative huh7.5. Huh7.5 cells supports high levels of sub genomic HCV replication due to a mutation (Thr-55-Ile) in Retinoic acid-inducible gene I (RIG-I) protein which impairs interferon signaling (Liu & Gale, 2010). Thus the molecular basis for HCV persistence

requires control of host RIG-1 signaling by virus. HCV-host protein interactions identified to date consist of 194 host proteins identified from huh7 cells and 133 proteins identified from huh7.5 cells, with 24 proteins identified from both cell lines. Our analysis of these host proteins indicates that there are differences in the associated pathways/GO terms for these host proteins from two different cell lines. Huh7 proteins interacting with HCV are involved in cancer, apoptosis, immune defense response, and cell cycle functions, while huh7.5 proteins are enriched for protein folding, localization, and transport. While it is not surprising that different cell lines have different molecular interactions with HCV proteins during infection, the difference in fundamental biological processes upon HCV infection can be expected to affect HCV studies in these lines, and researchers should carefully consider this aspect of HCV biology. Although, HCV infection of human hepatocyte is an area of intense research focus, how these interaction differences correspond to *in vivo* infection has not yet been determined.

In addition to variations in the cell lines used for HCV studies, different HCV isolates/genotypes are also used for HPI identification. HCV variants span six major genotypes (Nakano, Lau, Lau, Sugiyama, & Mizokami, 2012) that display phylogenetic heterogeneity, differences in infectivity and interferon sensitivity. Since biological variations in disease expression, persistence, and response to treatment is influenced by genotype variability of the infecting pathogen, differences in interactions with host proteins are also expected. Analysis of curated HCV interactions indicates that the most common isolates used in HCV-host protein interactions studies are H77 (genotype 1a), JFH-1 (genotype 1b) and Con1 (genotype 2a). Compared to HCV genotype 1a and 2a, genotype 1b infection is known to be associated with more severe complications. Currently, curated data for JFH-1 consists of 426 HPI while 527 HPI are associated with Con1. While there are a total of 729 host interacting proteins from these two genotypes, only 52 host proteins are experimentally confirmed to interact with both genotypes. Functional enrichment of host proteins interacting with either genotype identifies few similarities, and these are likely to be processes fundamental to HCV infection. Genotype 2a isolate JFH-1 shows extensive targeting of host proteins involved in cholesterol and lipid metabolism. Interaction with lipoproteins was previously shown to facilitate HCV replication and assembly and to correlate with the infectivity of genotype 2a but not 1a. However, genotype 1b interactions focus on host proteins involved in NF- κ B signaling, and apoptosis processes. This demonstrates that different interactions with the host may be driven by the pathogen genotype and should be considered when performing any data analysis.

Curated experimental detail includes linking GO annotations to interaction partners, and this information can be used to inform network analyses. For example, a 2003 study demonstrated that HCV NS5B interacts with endogenous nucleolin, that results in relocation of nucleolin to colocalize with NS5B in the perinuclear region (Hirano et al., 2003). Therefore, the detailed annotation will have NS5B annotated to perinuclear (GO:0097038), nucleolin to nucleolus (GO:0005730), and the NS5B-nucleolin interaction is associated with GO:0044220 'host cell perinuclear region of cytoplasm' term [Table 1]. Another annotation from the same publication shows that when the membrane-anchoring domain of NS5B is deleted, the NS5B-nucleolin interaction occurs in the nucleus (GO:0005634 'nucleus'). This finding strongly suggests that two regions in NS5B are required for interaction and

localization of host nucleolin. In addition to localization, GO terms associated with HCV-host protein interactions include: GO:0042326 ‘negative regulation of phosphorylation’, GO:0019068 ‘virion assembly’ and GO:0019034 ‘viral replication complex’. This functional information specific to interactions context that enable researchers develop more detailed, testable hypotheses about molecular mechanisms of pathogenesis.

Higher connectivity and multi-functionality of HCV proteins implies that the virus should encode proteins with multiple small binding regions to be able to bind multiple cellular proteins. Detailed annotations include information about these binding regions when this information is available. Using the feature identifiers supplied by the UniProt database, allows the IMEx curator to accurately capture the form of the protein used in the assay. This of particular importance in the annotation of the interactions of viral proteins; where one gene in the viral genome may encode multiple proteins. Around 30% of the HCV HPIs have binding and/or mutation region information for HCV viral products. Similarly to the Table 1 example, binding/mutation regions can be accessed in the HPI downloaded file. For example, annotation would show that the amino acid region 1–76 of the HCV core protein from genotype 2a bind three different host proteins; ddx3x, tripb and no38. In another example, HCV NS5A interaction with host protein, plin3, is through amino acid 1–14, where a single mutation in amino acid 9 disrupts the interaction with plin3 and significantly decreases HCV RNA replication. Biologists studying HPIs can access this additional binding information by checking the features information associated with each interaction; this type of detailed information provides researchers with specific information for targeted binding and/or mutation studies to block viral replication.

3. SUMMARY

These examples demonstrate that HPI datasets are complex, incomplete and heavily influenced by experimental conditions. Curation of interactions to capture experimental details and biological context of the interaction enables comprehensive insights into pathogenesis at a molecular level. Assisting researchers to understand how best to navigate and apply this information will ensure that future approaches for controlling pathogens are more effective. Due to the level of complexity and multidimensionality of HPIs, integration of the more detailed data into network analysis will require resources that provide a collection of interactive methods that support exploration of data sets by adjusting parameters to see how they affect the information being presented; such resources are starting to emerge. In addition to analysis, the availability of comprehensive curated HPIs can be key for supporting design of new experiments, identifying therapeutic targets, and improving interaction prediction methods. Based on this, it is fundamental to increase annotated interactions and the availability of contextual information that supports any given interaction in publications for a broad range of pathogen-host systems.

Acknowledgments

The authors thank the IntAct team for their support with IMEx annotations. This work was supported by the US Department of Agriculture National Institute of Food and Agriculture [Competitive Grant no. 2015-67015-23271] and the National Institutes of Health [COBRE P20GM103646-05S1].

References

- Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host-pathogen interactions. Database (Oxford). 2016; 2016doi: 10.1093/database/baw103
- Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, ... Hermjakob H. PSICQUIC and PSIScore: accessing and scoring molecular interactions. Nat Methods. 2011; 8(7):528–529. DOI: 10.1038/nmeth.1637 [PubMed: 21716279]
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, ... Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25–29. DOI: 10.1038/75556 [PubMed: 10802651]
- Davis ZH, Verschuere E, Jang GM, Kleffman K, Johnson JR, Park J, ... Glaunsinger BA. Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for late genes. Mol Cell. 2015; 57(2):349–360. DOI: 10.1016/j.molcel.2014.11.026 [PubMed: 25544563]
- de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaue S, ... Lotteau V. Hepatitis C virus infection protein network. Mol Syst Biol. 2008; 4:230.doi: 10.1038/msb.2008.66 [PubMed: 18985028]
- Garcia-Dorival I, Wu W, Dowall S, Armstrong S, Touzelet O, Wastling J, ... Hiscox JA. Elucidation of the Ebola virus VP24 cellular interactome and disruption of virus biology through targeted inhibition of host-cell protein function. J Proteome Res. 2014; 13(11):5120–5135. DOI: 10.1021/pr500556d [PubMed: 25158218]
- Germain MA, Chatel-Chaix L, Gagne B, Bonneil E, Thibault P, Pradezynski F, ... Lamarre D. Elucidating novel hepatitis C virus-host interactions using combined mass spectrometry and functional genomics approaches. Mol Cell Proteomics. 2014; 13(1):184–203. DOI: 10.1074/mcp.M113.030155 [PubMed: 24169621]
- Han Y, Niu J, Wang D, Li Y. Hepatitis C Virus Protein Interaction Network Analysis Based on Hepatocellular Carcinoma. PLoS One. 2016; 11(4):e0153882.doi: 10.1371/journal.pone.0153882 [PubMed: 27115606]
- Hirano M, Kaneko S, Yamashita T, Luo H, Qin W, Shirota Y, ... Murakami S. Direct interaction between nucleolin and hepatitis C virus NS5B. J Biol Chem. 2003; 278(7):5109–5115. DOI: 10.1074/jbc.M207629200 [PubMed: 12427757]
- Jean Beltran PM, Federspiel JD, Sheng X, Cristea IM. Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases. Mol Syst Biol. 2017; 13(3): 922.doi: 10.15252/msb.20167062 [PubMed: 28348067]
- Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, ... Hermjakob H. Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. 2007; 5:44.doi: 10.1186/1741-7007-5-44 [PubMed: 17925023]
- Lee C. Interaction of hepatitis C virus core protein with janus kinase is required for efficient production of infectious viruses. Biomol Ther (Seoul). 2013; 21(2):97–106. DOI: 10.4062/biomolther.2013.007 [PubMed: 24009866]
- Liu HM, Gale M. Hepatitis C Virus Evasion from RIG-I-Dependent Hepatic Innate Immunity. Gastroenterol Res Pract. 2010; 2010:548390.doi: 10.1155/2010/548390 [PubMed: 21274284]
- Nakano T, Lau GM, Lau GM, Sugiyama M, Mizokami M. An updated analysis of hepatitis C virus genotypes and subtypes based on the complete coding region. Liver Int. 2012; 32(2):339–345. DOI: 10.1111/j.1478-3231.2011.02684.x [PubMed: 22142261]
- Nii-Trebi NI. Emerging and Neglected Infectious Diseases: Insights, Advances, and Challenges. Biomed Res Int. 2017; 2017:5245021.doi: 10.1155/2017/5245021 [PubMed: 28286767]
- Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, ... Hermjakob H. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014; 42(Database issue):D358–363. DOI: 10.1093/nar/gkt1115 [PubMed: 24234451]
- Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, ... Hermjakob H. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods. 2012; 9(4): 345–350. DOI: 10.1038/nmeth.1931 [PubMed: 22453911]

- Park IW, Ndjomou J, Wen Y, Liu Z, Ridgway ND, Kao CC, He JJ. Inhibition of HCV replication by oxysterol-binding protein-related protein 4 (ORP4) through interaction with HCV NS5B and alteration of lipid droplet formation. *PLoS One*. 2013; 8(9):e75648.doi: 10.1371/journal.pone.0075648 [PubMed: 24069433]
- Poortahmasebi V, Poorebrahim M, Najafi S, Jazayeri SM, Alavian SM, Arab SS, ... Amiri M. How Hepatitis C Virus Leads to Hepatocellular Carcinoma: A Network-Based Study. *Hepat Mon*. 2016; 16(2):e36005.doi: 10.5812/hepatmon.36005 [PubMed: 27148389]
- Shames SR, Croxen MA, Deng W, Finlay BB. The type III system-secreted effector EspZ localizes to host mitochondria and interacts with the translocase of inner mitochondrial membrane 17b. *Infect Immun*. 2011; 79(12):4784–4790. DOI: 10.1128/IAI.05761-11 [PubMed: 21947777]
- Shames SR, Deng W, Guttman JA, de Hoog CL, Li Y, Hardwidge PR, ... Finlay BB. The pathogenic *E. coli* type III effector EspZ interacts with host CD98 and facilitates host cell prosurvival signalling. *Cell Microbiol*. 2010; 12(9):1322–1339. DOI: 10.1111/j.1462-5822.2010.01470.x [PubMed: 20374249]
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, ... Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003; 13(11):2498–2504. DOI: 10.1101/gr.1239303 [PubMed: 14597658]
- Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics*. 2014; 47:8–24. 1311–24. DOI: 10.1002/0471250953.bi0813s47 [PubMed: 25199793]
- The UniProt C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017; 45(D1):D158–D169. DOI: 10.1093/nar/gkw1099 [PubMed: 27899622]
- Thomas S, Izard J, Walsh E, Batich K, Chongsathidkiet P, Clarke G, ... Prendergast GC. The Host Microbiome Regulates and Maintains Human Health: A Primer and Perspective for Non-Microbiologists. *Cancer Res*. 2017; 77(8):1783–1812. DOI: 10.1158/0008-5472.CAN-16-2929 [PubMed: 28292977]
- Villaveces JM, Jimenez RC, Porras P, Del-Toro N, Duesbury M, Dumousseau M, ... Hermjakob H. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database (Oxford)*, 2015. 2015; doi: 10.1093/database/bau131

SIGNIFICANCE STATEMENT

Host-pathogen interactions (HPIs) are a specialized type of molecular interactions that occur between a host and a bacterium, virus, or other microorganism that can cause disease. Since those interactions define pathogenesis, our ability to effectively model HPIs improves our ability to identify approaches for mitigating the effects of disease. Molecular techniques development have enabled the elucidation of individual HPIs and increased the amount of HPI information in the scientific literature. Subsequent development of standardized files for interchanging not only the curated interaction details but relevant biological and technical information that impact these interactions, have enabled the accumulations of accessible, high quality HPI data. We discuss aspects and impact of HPIs identification and contextual information on network analysis. The ability to expertly access and model HPIs will facilitate infectious disease studies.

Figure 1.
Steps and factors affecting the outcome of a host-pathogen interactome modeling study.

Figure 2.

Detailed steps for downloading molecular interactions from PSICQUIC. More information on search tips and interaction clustering are provided on PSICQUIC 'help' page.

Figure 3.

Overview of HCV-host protein interaction data. HCV-host interactions can be sub-classified by (A) detection methods, (B) interaction types and (C) MI score (ranging from 0–1, with 1 being the highest confidence score). Data collected as of July 2017.

Figure 4. Distribution of HCV-host interactome data by gene products. For each HCV gene product we show the number of publications, curated interactions and number of identified interacting host proteins. Data collected as of July 2017.

Table 1

A simplified part of the curated detail data for NS5B-nucleolin interactions.

Alias(es) interactor A	Alias(es) interactor B	Interaction detection method(s)	Publication Identifier(s)	Interaction type(s)	Confidence value(s)	Interaction Xref(s)	Feature(s) interactor B
nucl_human	P27958-pro_0000037577 (NS5B)	Confocal microscopy	12427757	Colocalization	intact-miscore:0.46	go:"GO:0044220"(host cell perinuclear region of cytoplasm)	green fluorescent protein tag:?-?
nucl_human	p27958-pro_0000037577 (NS5B)	Confocal microscopy	12427757	Colocalization	intact-miscore:0.46	go:"GO:0005730"(nucleous)	green fluorescent protein tag:?-? necessary binding region:208-214 necessary binding region:500-506 sufficient binding region:1-570

The “?-?” indicates that the region of the protein sequence is not identified.