5-29-2021

Hypergraph models of biological networks to identify genes critical to pathogenic viral response

Song Feng
Pacific Northwest National Laboratory
Larissa B Thackray
Washington University School of Medicine in St. Louis
Qing Tan
Washington University School of Medicine in St. Louis
Michael S Diamond
Washington University School of Medicine in St. Louis
et al

BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Hypergraph models of biological networks to identify genes critical to pathogenic viral response

Song Feng[1], Emily Heath[2], Brett Jefferson[3], Cliff Joslyn[3,4], Henry Kvinge[3], Hugh D. Mitchell[1], Brenda Praggastis[3], Amie J. Eisfeld[5], Amy C. Sims[6], Larissa B. Thackray[7], Shufang Fan[5], Kevin B. Walters[5], Peter J. Halfmann[5], Danielle Westhoff-Smith[5], Qing Tan[7], Vineet D. Menachery[8,9], Timothy P. Sheahan[8], Adam S. Cockrell[10], Jacob F. Kocher[8], Kelly G. Stratton[1], Natalie C. Heller[3], Lisa M. Bramer[1], Michael S. Diamond[7,11,12], Ralph S. Baric[8], Katrina M. Waters[1,13], Yoshihiro Kawaoka[5,14,15,16], Jason E. McDermott[1,17] and Emilie Purvine[3*]

*Correspondence: emilie.purvine@pnnl.gov
[3] Computing and Analytics Division, Pacific Northwest National Laboratory, Seattle, WA, USA
Full list of author information is available at the end of the article

Abstract

Background: Representing biological networks as graphs is a powerful approach to reveal underlying patterns, signatures, and critical components from high-throughput biomolecular data. However, graphs do not natively capture the multi-way relationships present among genes and proteins in biological systems. Hypergraphs are generalizations of graphs that naturally model multi-way relationships and have shown promise in modeling systems such as protein complexes and metabolic reactions. In this paper we seek to understand how hypergraphs can more faithfully identify, and potentially predict, important genes based on complex relationships inferred from genomic expression data sets.

Results: We compiled a novel data set of transcriptional host response to pathogenic viral infections and formulated relationships between genes as a hypergraph where hyperedges represent significantly perturbed genes, and vertices represent individual biological samples with specific experimental conditions. We find that hypergraph betweenness centrality is a superior method for identification of genes important to viral response when compared with graph centrality.

Conclusions: Our results demonstrate the utility of using hypergraphs to represent complex biological systems and highlight central important responses in common to a variety of highly pathogenic viruses.

Keywords: Systems biology, Hypergraph, Viral infection, Biological networks, SARS, MERS, Influenza, West Nile Virus, Host response, Viral pathogenesis

Background

Identifying molecular signatures critical to a biological process requires an accurate model of both the process and the biological system in which it occurs. Thus it is essential that such a model be able to represent its target with complexity commensurate with

that of the system itself, rather than presenting only a simplified view. Commonly, biological systems and processes present as complex networks of interacting entities, for example within and between genes, pathways, and complexes. Graphs are frequently used to model these interactions, but since graphs can only capture interactions between pairs of entities, they fall short in many cases and are not able to model the full complexity present in biological systems and processes.

In this paper we investigate the role that hypergraph models, as mathematical generalizations of graph models, can play in providing the necessary complexity to capture multi-way interactions in biological systems inferred from genomic expression data. We introduce a hypergraph model of this data using data thresholding, and assert that the complexity provided by our proposed hypergraphs more closely represents the systems being studied. In order to validate this assertion we introduce new average hypergraph centrality metrics and provide a comparison between the use of graph and hypergraph centrality metrics to identify genes that are critical in host responses to viral infection. Our findings show that the genes identified using our hypergraph model and centrality metrics align better with genes previously known to correlate with viral response than do genes identified using similar metrics applied to graphs or using average fold change for each gene across all experimental conditions.

Network science for high-throughput data analysis

Modern biology has been transformed by the rapid growth of technologies to measure the abundance of large numbers of biological entities over many samples simultaneously. Such high-throughput methods like transcriptomics, proteomics, metabolomics, and lipidomics allow researchers to gain unparalleled scientific insight into the mechanisms underlying biological systems. A wide range of biological questions have been addressed using such systems biology approaches including questions related to cancer, microbiomes, and infectious disease. Analysis methods for high-throughput measurements are also varied, ranging from simple statistical tests for differential abundance (between control and experimental conditions, for example), to dimensionality reduction, to machine learning, all with the aim of extracting more relevant information from the high-dimensional and often noisy measurements.

A powerful approach for modeling systems using high-throughput data is network biology. Here biological systems are modeled as graphs, with molecular entities (genes, proteins, metabolites) represented as vertices, and relationships between molecules represented as edges connecting them. Relationships between molecules are generally determined from existing knowledge of protein-protein interactions, regulatory interactions, metabolic networks, or can be inferred from high-throughput systems biology data. We and others have used networks inferred from correlation or mutual information between abundance profiles of genes and proteins to identify critical entities [1–3], integrate different data types [4–7], and represent and predict temporal dynamics in the system [8–10].

Hypergraphs for complex network models

While graph-based methods have been quite successful in the biological domain, their ability to model complex relationships amongst entities is necessarily limited. Graphs

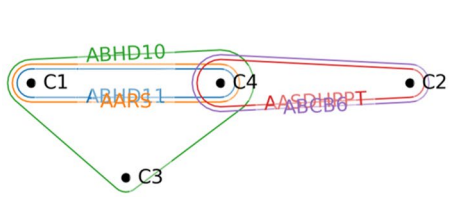
inherently model relationships (edges) between *pairs* of entities (vertices). But biological systems are replete with relationships among *many* entities, for example in protein complexes, transcription factor and microRNA regulation networks, lipid and metabolite enzyme-substrate interactions, metabolic networks, pathways, and protein function annotations. Relationships may be interactions, for example, metabolites working together in a metabolic process, or they may represent some commonality among the entities, like genes being differentially expressed in the same conditions, or regulated by the same transcription factor. In a graph model all of these multi-way relationships would be represented as groups of pairs of subunits, which would not fully capture how groups of components interact or have similar behavior.

Sometimes sets of related components are already understood, and sometimes they need to be discovered in experimental data, like high-throughput 'omics. In either event, a higher order mathematical model is needed. The mathematical object that *natively* represents multi-way interactions amongst entities is called a "hypergraph". In contrast to a graph, in a hypergraph the relationships amongst entities (still called vertices) are connected generally by "hyperedges", where each hyperedge is an arbitrary subset of vertices. Thus every graph is a hypergraph in which each hyperedge happens to have exactly two vertices. A challenge for scientists is to recognize the presence of hypergraph structure in their data, and to judge the relative value of representing them natively as hypergraphs or reducing them to graph structures.

Hypergraph models allow for higher fidelity representation of data that may contain multi-way relationships, albeit at the price of a higher complexity model. An example using a small subset of transcriptomic expression data is shown in Fig. 1. In the upper left is an expression matrix with $\log_2$-fold change values for five genes (rows) across four experimental conditions (columns). The lower left shows a hypergraph representation of the data, with each gene modeled as a hyperedge surrounding those conditions (vertices) for which the $\log_2$-fold change is greater than 2. Those cells in the expression matrix are shown in bold, distinguishing those conditions that are included in that gene's hyperedge. The upper right of Fig. 1 shows a matrix produced from one possible graph-based approach to representing these data. Here each pair of genes is related if there is at least

	C1	C2	C3	C4
AARS	**2.79**	1.65	0.88	**2.13**
AASDHPPT	0.17	**2.14**	1.76	**2.04**
ABCB6	0.68	**2.02**	0.94	**2.04**
ABHD10	**3.21**	1.88	**3.60**	**2.04**
ABHD11	**2.21**	0.78	1.68	**2.22**

	AARS	AASDHPPT	ABCB6	ABHD10	ABHD11
AARS		Y	Y	Y	Y
AASDHPPT	Y		Y	Y	Y
ABCB6	Y	Y		Y	Y
ABHD10	Y	Y	Y		Y
ABHD11	Y	Y	Y	Y	



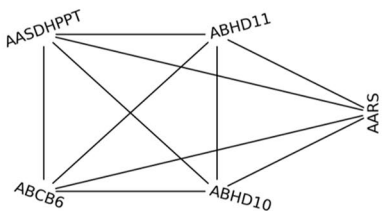


Fig. 1 Transcriptomics example comparing graphs and hypergraphs. (Upper left) $\log_2$-fold change values for 5 genes across 4 conditions. (Lower left) Visualization of a corresponding hypergraph. (Upper right) Adjacency matrix for expression data. (Lower right) Underlying graph

one condition for which both genes have $\log_2$-fold change greater than 2. This would then be interpreted as an "adjacency matrix" of a graph, which is shown in the lower right. It can be seen that this graph representation necessarily loses a great deal of information, boiling down the rich interaction structure that we know to be present to a fully connected graph on all five genes. For example, the hypergraph shows that two pairs of genes—AARS and ABHD11, AASDHPPT and ABCB6—are much more related than other pairs. This fact is not apparent in the graph model.

Although graphs and graph theory dominate network science applications and methods [11], hypergraphs are well-known objects in mathematics and computer science. They have a history of use in a range of applications [12–14], and are seeing increasingly wide adoption [4, 15–18]. In the biological literature we have seen hypergraphs used to model gene and protein interaction networks, pathways, and metabolic networks as derived from a variety of data types. In many of these cases the authors derive hypergraphs from an underlying graph, rather than directly from data. For example, Chitra built a hypergraph model based on an existing graph model of gene interaction networks [19]. They adapt the PageRank algorithm to hypergraphs in order to study disease-gene prioritization, and find that for monogenic diseases hypergraph PageRank noticeably outperforms graph PageRank. Tran studies protein function prediction building a graph from a similarity matrix derived from gene expression data [20] and then applying soft clustering to this graph to produce a hypergraph. Function prediction using this hypergraph is then shown to be superior to predictions based on graphs. Protein-protein interaction networks are studied by Klamt *et al.* using graph algorithms to find sets of independent elements or tightly connected elements [13]. In those three papers the authors infer a hypergraph from a graph structure rather than directly from data.

Ramadan *et al.* use hypergraphs to model the yeast proteome, where proteins are vertices and complexes are hyperedges [21], and apply an algorithm that finds tightly connected vertices to identify the core proteome. Finally, Zhou and Nakhleh study the claim that metabolic networks are hierarchical and small-world [22]. While this claim comes from a graph model of the networks, Zhou and Nakleh instead model the metabolic networks of *E. coli* as a hypergraph and show that the claimed hierarchy and scaling properties are not supported. This result in particular conveys a critical message: when biological interactions are simplified into pairwise relationships and modeled using a graph, they can exhibit very different structure than when their true complexity is modeled using a hypergraph. Because of this structural variance, conclusions drawn based on the graph could provide misleading results. Although the data we consider are different, our method is similar to these last two papers in that we build hypergraphs directly from biological data rather than inferring a hypergraph from a standard graph model of the data. We have not observed researchers building hypergraphs directly from 'omics data, as we will in this paper.

Modeling host response in viral pathogenesis

Viral infection causes a response in the host cells in which the expression of a variety of cell systems are up- or down-regulated. The pathogenesis of the infection is reflected in the signature of host responses elicited by each virus. Host response to viral infection has been extensively studied for decades, yet the root mechanisms of why some infections

are severe and some are not remain poorly understood. However, high-throughput molecular approaches offer a way to discover novel host response genes, proteins, and pathways that contribute to the systems-level development of pathogenesis. A major advantage of such a systems biology approach to pathobiology is the ability to identify novel, key elements of a biological process, such as which regulators are involved in critical processes. High-throughput profiling methods (e.g. transcriptomics) provide powerful tools for examining how entire systems respond to different perturbations such as acute disease. Network reconstruction provides the opportunity to utilize all available data and is a critically important tool for representing complex sets of interactions [23].

In this paper we develop and explore a new *hypergraph* model (see "Hypergraph representations and centrality metrics" section) of host response using transcriptomics data from viral infection by five highly pathogenic viruses in a number of biological systems (see "Data acquisition and processing" section). We found that gene rankings computed using an average hypergraph centrality were highly enriched for known immune and infection-related genes. While rankings derived from graphs constructed using other traditional computational biology techniques applied to the same infection data also resulted in rankings enriched for critical genes, we demonstrate that our hypergraph-based metrics yield superior enrichment results. These results highlight the usefulness of our hypergraph model for exploring mechanisms of virus pathobiology.

Results

By analyzing the curated omnibus transcriptomic data set described in "Data acquisition and processing" section from cells infected with five different viruses and their mutants using both graph and hypergraph approaches, we illustrate the advantages of applying our hypergraph approach to uncover the underlying molecular signatures and mechanisms common across host response to viral infection broadly.

Hypergraph and graph structure

We create hypergraphs from transcriptomics $\log_2$-fold change data calculated from gene expression levels of infected experiments relative to time-matched uninfected mock experiments

Formally defined in "Hypergraph representations and centrality metrics" section, in our hypergraphs hyperedges represent *genes* and vertices represent *conditions*. The vertex representing condition X is contained in the hyperedge representing gene G if gene G is significantly perturbed, either up- or down-regulated, for condition X. The entire hypergraph represents $n = 179$ experimental conditions (vertices) and $m = 7{,}782$ genes (edges). A small subset of highly connected hyperedges (that is, genes with a large core of common conditions), is shown in Fig. 2.

Distributions of fundamental hypergraph statistics can illuminate some of the complex interaction structure present in the data. Figure 3a shows that the distribution of the sizes of the hyperedges (that is, the number of conditions a gene is significantly perturbed in) is roughly power-law, sometimes referred to as "heavy tailed". This means that there are many genes (1,247 of them) significantly perturbed in only one condition and relatively few genes significantly perturbed in many conditions, with a maximum number of conditions for a single gene on the order of 100. The six largest hyperedges, with

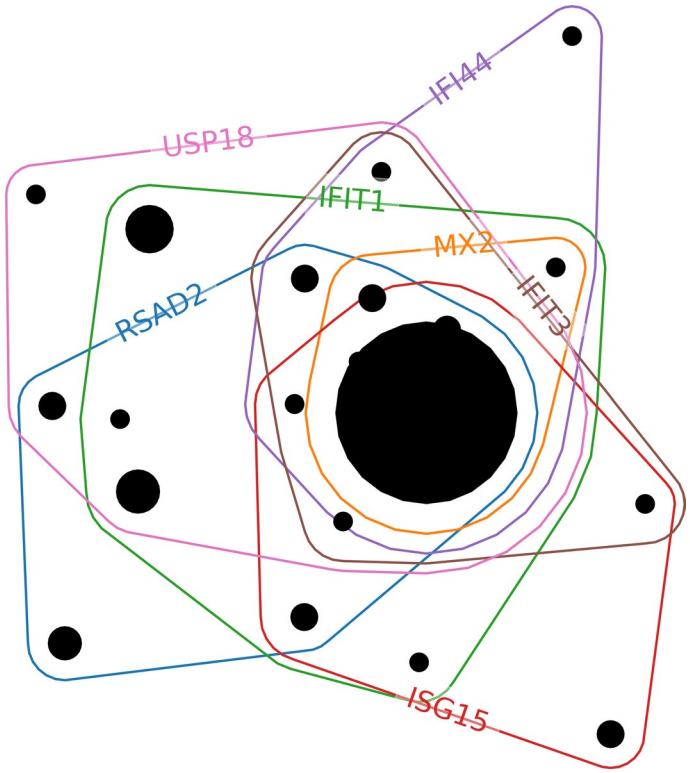


Fig. 2 Small connected subset of the condition/gene hypergraph. Hyperedges are genes, and black circles indicate groups of vertices (conditions), with larger circles indicating larger groups

sizes greater than 100 in increasing order, correspond to the genes ISG15, IL6, ATF3, RSAD2, USP18, and IFIT1. All of these genes are part of the interferon response, a critical pathway in response to viral infections [24, 25].

On the other hand the vertex degree distribution is the number of edges a particular vertex is contained in (that is, the number of genes significantly perturbed for each condition (vertex)). This is shown as a histogram in Fig. 3b, and it has a very different shape than the edge size distribution. There are relatively few conditions with small numbers (less than 200) or large numbers (more than 500) of significantly perturbed genes. The most common number of significantly perturbed genes for a condition is between 400 and 500. This peak is likely an artifact of how we choose when a vertex is contained within an edge. The degree of a condition vertex is the number of genes that are significantly perturbed for that condition. By our procedure these are genes with z-score higher than 2 and p-value less than 0.05. If we only used the z-score threshold, and our fold change data are normally distributed for each condition, then we would expect that 5% of the genes would have z-score greater than 2. There are 9,760 genes in our data and 5% of that would be 488 genes, which is roughly where the peak is. The skewness and additional modes of the distribution of degrees are likely due to the addition of the p-value condition.

Finally, Fig. 3c shows another power-law distribution, this time of the size of pairwise edge intersections, or in other words, the number of conditions that pairs of genes are both significantly perturbed within. We see that there are many pairs of genes that have

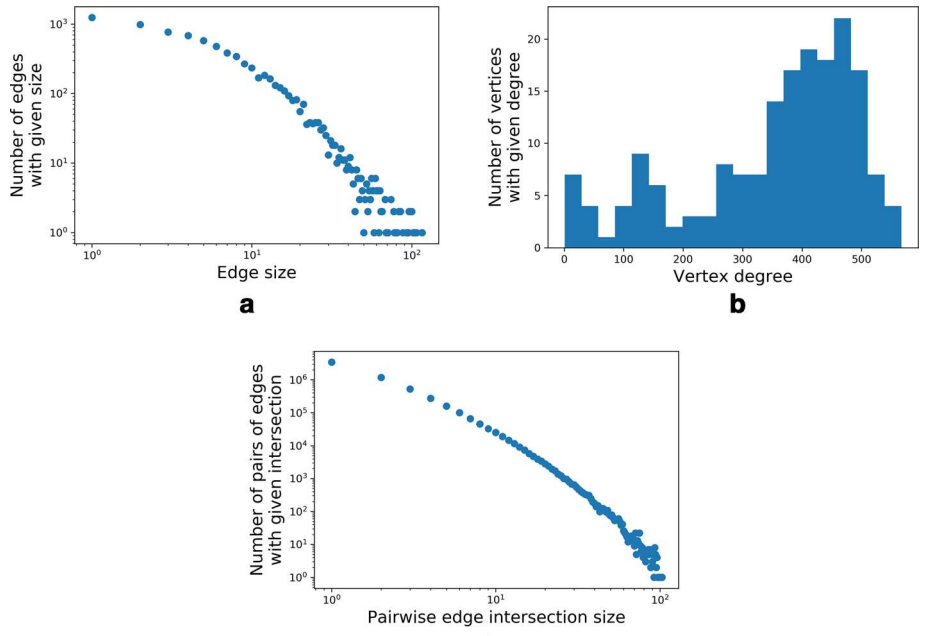


Fig. 3 Distributions of simple hypergraph statistics of our hypergraph. **a** Edge size distribution. Each edge represents a gene, the size of the edge is the number of conditions in which the gene is significantly perturbed. **b** Vertex degree histogram. Each vertex represents a condition, the degree of a vertex is the number of genes significantly perturbed in that condition. **c** Pairwise edge intersection size distribution. Each edge represents a gene. The intersection between two edges indicates the set of conditions that both genes are significantly perturbed in

few conditions in common and only a few pairs of genes that have many conditions in common, again with a maximum on the order of 100. The pair of genes with largest intersection is not surprisingly the two largest edges, IFIT1 and USP18, with 103 conditions in common. Interestingly, IFIT1 and USP18 are both well-established interferon response genes, with IFIT1 strongly promoting interferon activity, and USP18 serving to dampen the response [26].

In order to compare our hypergraph approach to more common graph approaches we employed the CLR graph methodology that we have used previously to enrich for important genes in a network [1, 2, 27]. The CLR algorithm was run on the matrix of transcriptomics $\log_2$-fold change values using parameters spline $= 3$ and bins $= 10$ (as used in the original CLR manuscript [28]) and the resulting matrix was filtered for all mutual information values ≥ 2. With this approach, any two genes with mutual information above the threshold have similar expression profiles and form a graph edge.

In comparing the CLR graph to our hypergraph we find distinct differences, indicating that the structures are capturing different relationships about sometimes different sets of genes. First, the set of genes present in the hypergraph is a subset of the genes present in the CLR graph, meaning that any gene that is shown to be significantly perturbed in at least one condition (i.e., present in the hypergraph) has mutual information ≥ 2 with at least one other gene (i.e., present in the CLR graph), but not necessarily *vice versa*. Comparing the edges present in the CLR graph to the hyperedge intersections present

in the hypergraph, of the nearly 3.7 million edges in the CLR graph, 2.4 million of them are between genes that are in the hypergraph. Roughly 1.1 million of these are present as hyperedge intersections in the hypergraph while 1.3 million do not have a corresponding hyperedge intersection. The remaining 1.3 million edges have one or both endpoints in the set of genes that are not present in the hypergraph. Moreover, there are a total of nearly 6 million nonempty pairwise hyperedge intersections (gene interactions) in the hypergraph, indicating that the hypergraph is expressing additional structure and relationships among genes that the CLR graph does not capture. Finally, for each gene hyperedge we compute the number of other gene hyperedges that it intersects. These values are only loosely correlated with the CLR graph vertex degrees (number of other genes with mutual information ≥ 2), with Pearson correlation 0.25. This indicates that not only are there additional connections in the hypergraph, as observed above, there is not a linear relationship between the number of CLR graph connections for a gene and the hypergraph connections for a gene. In other words, one cannot infer the hypergraph from the CLR graph, as it represents fundamentally different higher order relationships specifically over the set of genes with significant perturbation.

Gene importance rankings

Previous studies using graph approaches with similar viral data have demonstrated that network measures like betweenness centrality could be used to identify critical genes [1]. In the present work we hypothesize that extensions of these common graph metrics to a hypergraph, as defined in "Hypergraph representations and centrality metrics" section, can be leveraged to improve upon this prior work. In particular, we hypothesize that, as has been shown in the graph setting, high s-centrality is correlated with the gene being more important in host response to pathogenic viruses.

We calculate average s-betweenness centrality, $\overline{BC}_s(g)$, and average harmonic s-closeness centrality, $\overline{HCC}_s(g)$, for all genes (hyperedges) in the hypergraph for $1 \leq s \leq 50$. These average centralities change as s increases only for gene hyperedges that overlap in s vertices with at least one other gene hyperedge. In our hypergraph, for $s = 50$ less than 1% of hyperedges remain connected to at least one other hyperedge with that overlap size, resulting in very little change in the centrality values. Both of these average s-centrality computations provide a numerical value for each gene that can be used to rank the genes from most important (high centrality) to least (low centrality).

To serve as a simpler, but still hypergraph-based, comparison we created another ranked list using hyperedge size for each gene, i.e., the number of conditions the gene was significantly perturbed in. Larger edge sizes indicate more conditions in which the gene was significantly perturbed and therefore the gene is potentially more important to host response.

To compare our hypergraph centrality ranking approach to the CLR graph approach we use the NetworkX graph analytics Python package [29] to calculate vertex degree, betweenness centrality, and harmonic closeness centrality of the CLR association graphs.

To provide a simple baseline for comparison a final ranked list was computed directly from the $\log_2$-fold change table without using any graph structure. For each gene we computed its average absolute value of $\log_2$-fold change and ranked the genes from

highest to lowest average. Higher values mean the gene is more likely to be highly perturbed from the mock-infected samples in many conditions.

In the Additional file 1 we provide all gene rankings for average s-betweenness centrality and average harmonic s-closeness centrality, hyperedge size, CLR graph betweenness, CLR graph closeness, CLR graph degree, and average fold change.

Comparison of rankings

To ascertain whether our hypergraph rankings are more highly enriched for genes known to be important in host response to viral infection, we gathered three distinct sets of genes: 1) all genes associated with the Gene Ontology (GO) term "immune response" (GO:0006955), downloaded from amigo.geneontology.org, referred to as 'IR' hereafter, 2) interferon-stimulated genes gathered from interferome.org (http://www.interferome.org/interferome/search/searchGene.jspx), referred to as 'ISG', and 3) a set of human proteins known to be targets of pathogens acquired from Dyer, et al. [30], referred to as 'PT'. Although this is a limited set in terms of number of targets, it represents a set collected from a wide number of pathogens, both viral and bacterial, and is a conservative set for assessing the performance of our method and making comparisons between different approaches and parameters. In Table 1 we show the size of each gene set (along the diagonal) and the sizes of each pairwise intersection of gene sets (off the diagonal). Since our data encompasses a wide variety of virus types and infection systems, general immune-related sets were deemed suitable for our purpose.

In order to measure the performance of our rankings, we applied gene set enrichment analysis (GSEA) [31] to each of our gene rankings (average hypergraph s-centralities, hyperedge size, CLR centralities, CLR vertex degree, and mean fold-change) using the three immune-related sets as target gene sets. The GSEA score of a ranked list, computed for a specific gene set, quantifies how concentrated the gene set is at the extremal values of the list. A high GSEA score means the gene set is concentrated at the top of the list while a low (highly negative) score indicates that the gene set is concentrated towards the bottom of the list. A score closer to zero means that the gene set is more uniformly distributed throughout the ranked list. The significance, or p-value, of an observed enrichment score, ES, is assessed by comparing it with a set of ES_0 randomized scores.

Figure 4 shows GSEA scores for all rankings and for all three target gene sets. We note the following conclusions:

Table 1 (Diagonal, bold) Size of each gene set; (off-diagonal, non-bold) size of the pairwise intersections of the gene sets

	IR	ISG	PT
IR	**1,202**	250	297
ISG	250	**1,071**	152
PT	297	152	**906**

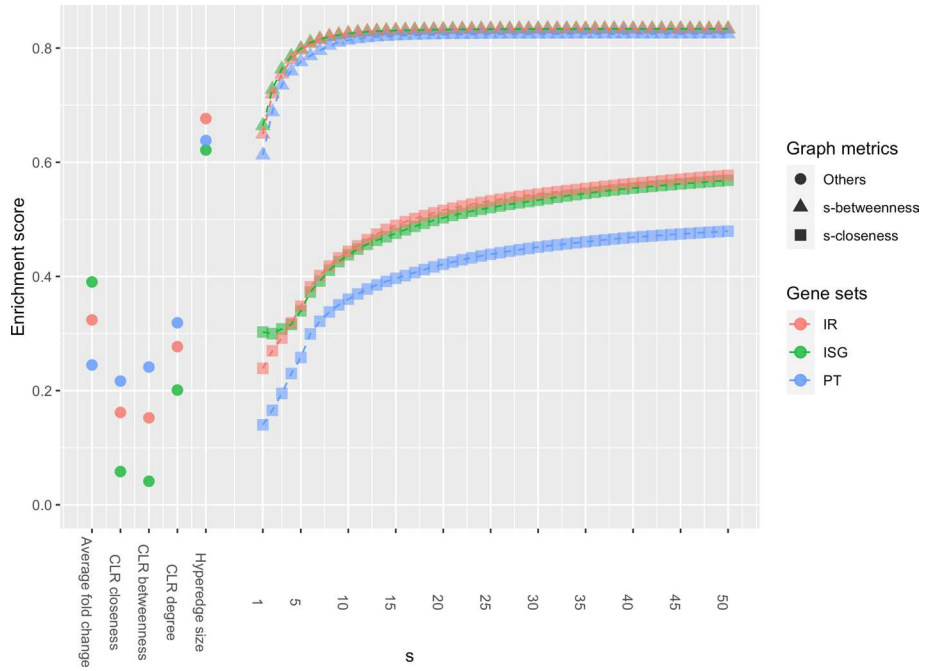


Fig. 4 Enrichment scores of gene sets using average *s*-betweenness and *s*-harmonic closeness metrics. All results are significant with $p < 0.05$

- Both average *s*-centrality metrics for most *s* values, as well as hyperedge size, showed much higher enrichment than lists derived from CLR graphs and average fold change.
- But average *s*-betweenness enrichment was universally higher than average harmonic *s*-closeness enrichment, suggesting that these two measurements are capturing fundamentally different behavior within hypergraphs, and that average *s*-betweenness appears to be more effective at capturing genes that are important in host responses to viral infection.
- Both centrality enrichment results (betweeness and closeness) improve significantly when larger *s* values are taken into account, indicating that when higher order interactions are considered, they become more powerful in identifying important genes. Although the maximum intersection between two hyperedges is 103, Fig. 4 indicates that by $s = 20$ our best performing measure, average *s*-betweenness centrality, plateaus with only minimal increase in enrichment score as *s* increases further.

As noted in "Hypergraph representations and centrality metrics" section we also constructed hypergraphs using *z*-score thresholds of 3, 4, and 5 and computed GSEA scores for their average *s*-betweeness and average harmonic *s*-closeness rankings. Versions of Fig. 4 for $z = 3$, 4, and 5 are included in the Additional file (see Additional files 2–4: Fig. S1–Fig. S3). The choice of *z*-score values did not change the conclusion that GSEA scores increase with *s* values, or that rankings derived from the hypergraph have higher GSEA scores than those derived from the CLR graph. However, with smaller hypergraphs (from large *z* thresholds), the *p*-values of GSEA increase

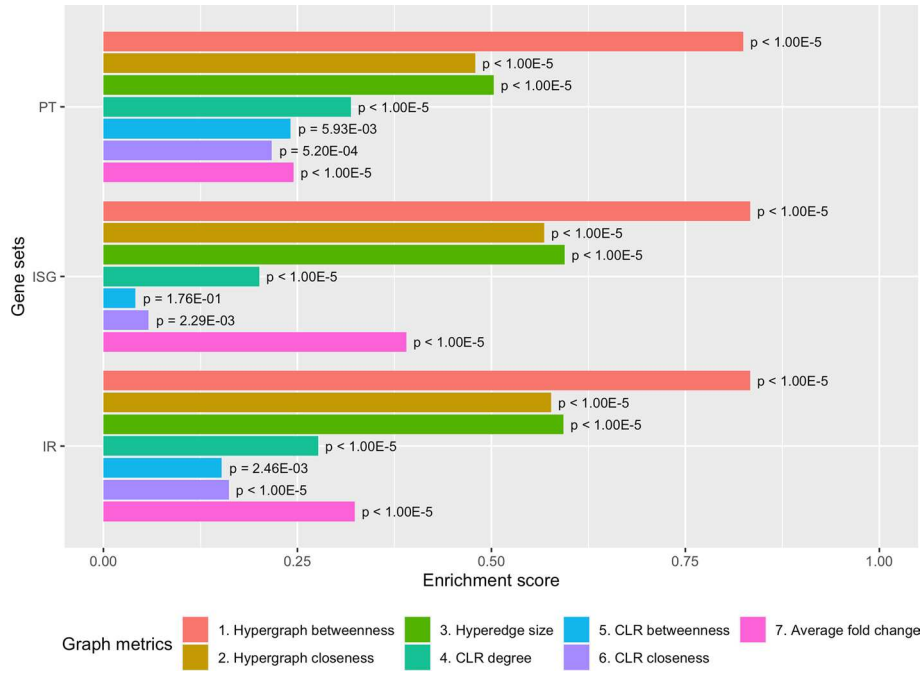


Fig. 5 Comparison between different hypergraph and graph metrics

(see Additional files 5–8 Fig. S4–Fig. S7). Therefore, $z = 2$ is an adequate z-score threshold to balance high GSEA score with low p-value of the score, under a permutation test.

A summary visualization of our results is shown in Fig. 5 taking the rankings for both average s-betweenness and harmonic s-closeness for the highest level of $s = 50$ as the representative hypergraph centrality rankings. We compare those with the five other rankings and again see that average s-betweenness centrality outperforms all other measures. While average harmonic s-closeness centrality outperforms all graph measures it is outperformed by the simple hyperedge size ranking. The p-values, nearly all significantly less than 0.05, are shown in the same plot at the end of the bars. These results demonstrate that average s-betweenness, but not necessarily average harmonic s-closeness, considers the complexity of the hypergraph and provides superior performance over graph metrics with regards to identifying biologically important genes. This aligns with prior work in which betweenness centrality computed for vertices in a graph identifies important genes in a network [1, 2, 27, 32].

In order to compute the average s-centrality measures we first separately computed s-centralities $BC_s(e)$ and $HCC_s(e)$ for each gene edge s. We also computed enrichment scores for these ranked lists. However, we do not include these in our overall comparison because while some s-values produced rankings that were more enriched than others we did not see any trend in performance that would lead us to choose an optimal single s value (e.g., enrichment scores are not unimodal as a function of s). This led us to consider the average s-centralities, in order to take advantage of all intersection levels simultaneously. In future work it may be worth reexamining single s values, or removing

single s values from the average, to understand the relative importance of each s to the averaging.

Discussion

We draw attention to two primary observations of interest in our results. First is the observation that s-betweenness centrality consistently outperforms s-closeness centrality. At first glance this seems surprising since betweenness and closeness calculated on the CLR graph have comparable performance. While the explanation for this result is the subject of ongoing investigation, we observe that these two types of centrality are measuring significantly different properties. For both graphs and hypergraphs high harmonic closeness centrality indicates that on average a gene is close to many other genes, while high betweenness centrality means that a gene is on many short paths between other genes. Sometimes these two notions coincide, as seems to be the case in the CLR graph, but there are cases in which they do not. For example, a gene that may be more on the periphery, i.e., not on many (short) paths, could still be very close to a central core. Being on few short paths this gene would have very low betweenness. However, since it is close to a central core it could have high closeness score.

This seems to be the case in the hypergraphs we are studying. Since there are many conditions that have a lot of significantly perturbed genes (see Fig. 3b) there is likely a large central core in the hypergraph that increases the closeness scores for peripheral genes, and perhaps all genes. Indeed we have observed that for small values of s the s-closeness values do not correlate with edge size, however for large s values the s-closeness scores do tend to correlate with edge size. This likely means that for low s values the closeness is somehow washed out by this central core and any variability we see is not significant. In contrast, for s-betweenness we see a correlation between edge size and betweenness at all s values. However, even though s-betweenness is correlated to edge size for all s values its enrichment score is still much larger than that for edge size and so seems to be capturing something more significant about hypergraph structure.

This difference between closeness and betweenness may also be related to the nature of the large gene expression data set used in our study. Since both mouse and human-based gene expression data were included in the hypergraph some genes may serve as bridges between different regions of the hypergraph (e.g. predominantly human regions vs predominantly mouse regions). Genes that are truly important in host response to viral infection would be important across species and more effectively brought to light by the betweenness measure that tends to highlight elements occupying bridge-like positions in the hypergraph. Thus, betweenness centrality may be most useful for identifying critical elements when heterogeneous data sets are analyzed.

Our second observation is that our average s-betweenness centrality significantly outperforms established graph centrality techniques. This is entirely in keeping with our expectation, as the purpose of hypergraphs is to capture the complex, multi-way interactions present in a system that are beyond the ability of graphs to model. Thus where betweenness centrality has been used in prior studies to identify important biological features the application of hypergraph s-betweenness may promote discovery of additional features of interest. While the finding that hypergraph betweenness represents a new tool for identifying critical hypergraph elements is an exciting contribution of this

study, it also presents an additional immediate benefit: genes highly ranked by hypergraph betweenness that do *not* appear in any of our target gene sets represent potentially novel discoveries of genes central to viral infection. One good example of this is the ZZZ3 gene, which appears in position 4 out of 7,782 in the average hypergraph betweenness ranking, but does not appear in any of the IR, ISG or PT gene sets. ZZZ3 is part of the histone reader ATAC complex, which scans the state of histone modification and contributes to gene activation/repression mechanisms [33]. No known connection between virus infection and ZZZ3 exists, but it may serve a critical role in regulating gene expression in response to general infection.

Similarly, EPHX1, GDF15, and DUSP1 were not included in the three gene sets and ranked 29, 30 and 33, respectively. These genes are identified as an epoxide detoxification component, a stress responsive cytokine and a stress-responsive phosphatase, respectively. These roles may be related to virus-induced stress in host cells, but the specific mechanisms involved are yet to be elucidated. More exploration of these and other highly ranked genes is the subject of future work for us.

Conclusion

The work we present in this paper is similar to much of the work surveyed in our literature review in that we show the value of hypergraphs over traditional graph analysis of biological data. However, our work differs from these prior studies in a number of ways. First, our hypergraphs are built natively from transcriptomics data rather than based on existing graph models of systems. Although still capturing some multi-way complexities, hypergraphs inferred from graphs may include some induced interactions not actually present in the system that is being modeled. Creating hypergraphs natively from the data avoids this imputation. Other papers we surveyed do create hypergraphs natively from other types of data, but rather than applying centrality measures instead study more structural features like highly connected vertices.

Previous work [1, 2, 27, 32] had demonstrated that graph metrics can be used to identify important genes in association graphs, and so we set out to determine if hypergraphs provided an improvement over graphs. To assess performance of (hyper)graphs derived from our large viral infection gene expression data set, we identified three gene sets related to virus/pathogen infection and performed an enrichment analysis of our ranked lists compared to these gene sets. While the sets were partially overlapping they represented relatively distinct aspects of viral infection in general. Our results show that average s-betweenness, but not necessarily average harmonic s-closeness, was a useful metric that is able to identify key genes in a comprehensive gene expression data set. While average harmonic s-closeness does outperform both CLR centrality measures, CLR degree, and average fold change, it does not exceed the performance of a simple ranking according to hyperedge size, which does not require the full hypergraph structure to calculate. On the other hand, ranking based on average s-betweenness outperformed all other metrics.

The hypergraphs we created used samples from a wide range of viruses, strains, cell types, and time since infection. In future work we plan to apply this measure to compare critical genes in viral response across differing sample features. For example, we will split our hypergraph based on pathogenicity (high vs. low), cell type or host, and time since

infection (early vs. late). Comparing the critical genes across these different hypergraphs may allow us to discover previously unknown indicators of viral infection for early detection or severity determination. Other future work we plan to pursue includes considering other hypergraph constructions, other data types, and hypergraph algorithms to identify highly connected vertices. We plan to combine transcriptomics with proteomics and other 'omics measurements to understand whether hybrid hypergraphs yield better results or if the inclusion of more data washes out the complexities.

Methods

Data acquisition and processing

Microarray datasets collected from 2014 to 2017 and available from the Gene Expression Omnibus (GEO) were gathered and compiled as described below. GEO accession IDs: GSE80059, GSE86533, GSE69027, GSE76600, GSE80697, GSE69945, GSE68945, GSE72008, GSE65575, GSE79458, GSE86528, GSE100496, GSE81909, GSE86530, GSE100504, GSE106523, GSE86529, GSE100509, GSE108594, GSE77193, GSE77160, GSE78888, GSE33267, GSE37827, GSE48142, GSE33266, GSE49262. While details of experimental systems and conditions can be gathered from individual accessions from GEO, we list the infection conditions here:

Ebola Virus	(Wild type and two mutants) in human hepatocyte cells.
Influenza Virus	H7N9 (Wild type and two mutants) in human lung epithelial cells, H1N1 (wild type) in human lung epithelial cells, H1N1 (wild type) in mouse lung, H5N1 (wild type and one mutant) in mouse lung, H7N9 (wild type and two mutants) in mouse lung.
MERS-coronavirus	(Wild type and four mutants) in human lung epithelial cells, (wild type only) in *ex vivo* human epithelial cells, in *ex vivo* human lung fibroblasts, and *ex vivo* human lung microvascular endothelial cells.
SARS-coronavirus	(Wild type and four mutants) in human lung epithelial cells and mouse lung.
West Nile Virus	(Wild type and one mutant) in mouse cerebral cortex, mouse cerebellum and mouse lymph node.

Raw microarray data was processed for background correction, quantile normalization and summarization using the limma package for R (available on Bioconductor) to derive a single normalized intensity value per probe. We utilized a conservative approach and decided to only remove samples (arrays) that showed obvious evidence for failed hybridization or damaged arrays. These types of occurrences are easily detected by inspecting PCA plots and expression heatmaps. Specifically, we looked at PCA plots and expression heatmaps of the 500 most variable genes according to the coefficient of variation. This manual approach resulted in preserving as much data as possible.

The data in this form, as it is available from the above GEO repositories, was used for further compendium construction. Differential expression analysis with linear models in limma [34] was used to identify fold changes and p-values for significance of changes, as well as adjusted p-values to account for multiple testing. Since datasets were collected using multiple microarray platforms, merging gene identifiers using probe IDs was not possible, so genes were matched at the gene symbol level instead. Individual genes are

often represented by multiple probes, so only the most significantly changed probe among those representing a single gene was retained in each dataset. This provided a way to match data rows across experiments, and resulted in a compendium matrix of genes common to all datasets with 9,760 rows, with each row representing a single gene.

Hypergraph representations and centrality metrics

Formally, a **hypergraph** is a structure $\langle V, E\rangle$, with $V = \{v_j\}_{j=1}^{n}$ a set of vertices, and $E = \{e_i\}_{i=1}^{m}$ a family of hyperedges with each $e_i \subseteq V$. Hyperedges can come in different sizes, $|e_i|$, possibly ranging from the singleton $\{v\} \subseteq V$ (distinct from the element $v \in V$) to the entire vertex set V. A hyperedge $e = \{v_1, v_2\}$ where $|e| = 2$ is the same as a graph edge and so it follows that all graphs are hypergraphs, specifically identified as being "2-uniform". Where clear from context we may use the terms edge and hyperedge interchangeably.

We construct a hypergraph from transcriptomics data using a threshold approach, much like the example in Fig. 1. Again, vertices v_j will represent individual biological or experimental "conditions" (e.g., mouse lung cells treated with a strain of Influenza virus and sampled at 8 h) and hyperedges e_i represent genes. Thus for us, a hyperedge e_i is a gene i that includes a collection of conditions j as its vertices v_j. For each condition, we transform the $\log_2$-fold change values (relative to uninfected mock) for all of the genes into absolute value z-scores. Then, the vertex representing condition X is contained in the hyperedge representing gene G if the absolute value z-score for G in X is greater than or equal to 2 and the adjusted p-value for that $\log_2$-fold change measurement is less than 0.05. Since transcriptomics $\log_2$-fold change values tend to be normally distributed for each condition across all genes a z-score transformation is a reasonable way to get all conditions onto the same scale before applying a threshold. The specific thresholds on z-score and p-value were chosen as commonly used in the field, and in exploring other z-score thresholds we have verified limited sensitivity to them. We note that using a higher z-score threshold results in smaller hypergraph, generated from the genes that change more dramatically.

In this way, hyperedges correspond to genes, and indicate the groups of conditions in which that gene is both highly perturbed (either up or down) from the mock infected control condition, and for which that perturbation is statistically significant. We say that the gene is "significantly perturbed" in the condition. Unlike hypergraph models of pathways or metabolic reactions, hypergraphs constructed from high-throughput data do not necessarily represent actual biological interactions but rather capture relationships based on similar behavior among entities.

It is important to point out that this method to construct a hypergraph using thresholds on absolute value of z-score and p-value is a specific case of a flexible framework we propose for how hyperedges can be formed from 'omics data. Applying other thresholds will result in different hypergraph models of the same data, to potentially answer different questions. For example, in order to understand the relationship and behavioral similarity among up-regulated genes one might consider a gene hyperedge to contain those conditions for which the gene has high raw (as opposed to absolute value) z-score or $\log_2$-fold change, as in the Fig. 1 example. One could also form edges from conditions for which a gene has a highly negative z-score or fold

change, to explore the structure of down-regulated genes. We chose a threshold on the absolute value of z-score in this paper as an attempt to understand genes which are perturbed at all in response to viral infection.

We recognize that this formulation is fundamentally different from a typical graph approach to systems biology data. One such example of a graph approach is context likelihood of relatedness (CLR) in which genes that show similar expression patterns across *all* conditions, as measured by mutual information, are linked together [28]. Our approach to constructing hypergraphs from the data can be seen as having greater sensitivity and flexibility since it allows similarity between genes to be assessed across any number of conditions (as quantified by the size of the overlap of their hyperedges) rather than requiring assessment across all conditions, as in the mutual information calculation used to define CLR edges. In "Gene importance rankings" section we provide a comparison between our hypergraph and the CLR graph formed from the data.

Another difference between our hypergraph formulation and typical graph approaches is that in graph approaches vertices represent genes and edges indicate some relationship between genes such as interaction or expression correlation. Our motivation for swapping the roles of vertices and edges is for the sake of clarity in our description of hypergraph centrality measures below. Moreover, as a technical matter, each hypergraph H determines another one, called its "dual" H^*, formed exactly by swapping the roles of vertices and edges [35]. Therefore, the dual to our hypergraph formulation has the more traditional form with genes as vertices, but the description of hypergraph centrality in this setting would be less intuitive.

As in graphs, the way in which hyperedges connect vertices in complex patterns is central to the study of hypergraphs. While many hypergraph topological measures are available, either as generalizations of graph measures to account for multi-way interactions or as native hypergraph-only measures, our focus in this paper is applying generalizations of graph centrality measures to hypergraphs built from transcriptomics data to identify important genes. In order to define these hypergraph centrality measures we must first introduce the notions of a hypergraph walk and distance [36]. Given two hyperedges $e,f \in E$, an s **-walk** between e and f is a sequence of hyperedges $e_0, e_1, \ldots, e_k$ such that $e_0 = e$, $e_k = f$, and $s \leq |e_i \cap e_{i+1}|$ for all $0 \leq i \leq k-1$. An s-walk with $k+1$ edges has **length** k. In other words, an s-walk is any sequence of edges, not necessarily of minimal length, such that pairwise intersections between neighboring edges have size at least s. Note that a graph walk is a 1-walk. We note that one could define a hypergraph s-walk to be between vertices rather than hyperedges, as is typically done in a graph. But as above, for the sake of clarity in defining centrality measures we use this edge-based definition.

Continuing to follow Aksoy *et al.* [36], for a fixed $s > 0$, we define the s **-distance** $d_s(e,f)$ between two edges $e,f \in E$ as the shortest length of the possibly many s-walks between them. If there is no s-walk between two edges then the s-distance is infinite. Aksoy *et al.* also define a number of network science methods generalized from graphs to hypergraphs, including vertex degree, diameter, and clustering coefficients. This work will use their generalization of betweenness centrality and harmonic closeness centrality to hypergraphs using the stratification parameter s.

- The s **-betweenness centrality** of an edge e is

$$BC_s(e) := \sum_{f \neq e \neq g \in E} \frac{\sigma_{fg}^s(e)}{\sigma_{fg}^s}$$

where σ_{fg}^s is the total number of shortest s-walks from edge f to edge g and $\sigma_{fg}^s(e)$ is the number of those shortest s-walks that contain edge e.

- The **harmonic** s **-closeness centrality** of an edge e is the reciprocal of the harmonic mean of all distances from e:

$$HCC_s(e) := \frac{1}{|E_s| - 1} \sum_{\substack{f \in E_s \\ f \neq e}} \frac{1}{d_s(e,f)}$$

where $E_s = \{e \in E : |e| \geq s\}$. We may refer to this as s-closeness in this paper, although elsewhere in the literature this term refers to a slightly different concept where the harmonic mean is replaced with the arithmetic mean.

Intuitively, harmonic s-closeness centrality captures the extent to which a given hyperedge is close in s-distance to other hyperedges. In order to have high harmonic s-closeness a hyperedge must have small s-distance to all (or most) other hyperedges. s-Betweenness, on the other hand, identifies bottlenecks in a hypergraph. A hyperedge with high s-betweenness has many shortest s-walks pass through it. In comparison, the original formulation of betweenness and harmonic closeness centrality in the setting of graphs has the s-distance and number of s-paths replaced simply by graph distance and shortest path.

In order to take into account multiple s values simultaneously in our analysis we developed the *average s -betweenness centrality* and *average harmonic s -closeness centrality*, averaging the centrality values defined in [36] across a range of s values. Our average s-centralities are defined as

$$\overline{BC}_s(e) = \frac{1}{s} \sum_{i=1}^{s} BC_i(e), \qquad \overline{HCC}_s(e) = \frac{1}{s} \sum_{i=1}^{s} HCC_i(e).$$

Computing average s-centralities for each hyperedge provides a ranked list of hyperedges from most central (high value) to least central. All hypergraph construction, metric calculations, and visualizations were performed using the Python hypergraph library HyperNetX (https://github.com/pnnl/HyperNetX).

Some conventional approaches to infer graph structures from high-throughput data use correlated gene expression patterns to build connections. In this context, a gene with high degree (i.e. a hub) has similar expression behavior to many other genes, implicating it as a potential master regulator of gene expression. A gene with high betweenness (i.e. a bottleneck) on the other hand, bridges two regions of the graph indicating that it spans two different behavioral profiles. Genes in this position are potentially involved in causing a transition from one response pattern to another. Thus hubs and bottlenecks may represent master gene expression regulators of two different varieties. Previous work by our group and others has shown that graph

vertices in hub and bottleneck positions are significantly enriched for genes critical to the process under study [2, 27, 37, 38]. Given these prior results and biological relevance of centrality in the setting of graphs, we hypothesized that hypergraph average s-betweenness (and potentially average s-harmonic closeness) will have similar biological relevance. We are the first group to apply *hypergraph* centralities to genomic data, after we first introduced them in [36].

Abbreviations

BC: Betweenness centrality; CLR: Context likelihood of relatedness; ES: Enrichment score; GEO: Gene Expression Omnibus; GO: Gene Ontology; GSEA: Gene set enrichment analysis; HCC: Harmonic closeness centrality; IR: Immune Response gene set; ISG: Interferon-stimulated gene set; PCA: Principal component analysis; PT: Pathogen target gene set.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04197-2.

Additional file 1. This Microsoft Excel file contains the ranked lists of genes that were compared in Figure 4. Each ranked list is in its own sheet in the Excel file: hypergraph betweenness (average s-betweenness centrality for $1 \leq s \leq 50$ is in one sheet), hypergraph closeness (average harmonic s-closeness centrality for $1 \leq s \leq 50$ is in another sheet), hyperedge size, CLR degree, CLR betweenness, CLR closeness, and average fold change. The ranked lists are of different length because some genes were not included in the hypergraph or graph based on the (hyper) graph construction criteria.

Additional file 2. These figure is analogous to Figure 4 for additional z-score threshold z.

Additional file 3. These figure is analogous to Figure 4 for additional z-score threshold z.

Additional file 4. This figure is analogous to Figure 4 for additional z-score threshold z.

Additional file 5. This figure shows the p-values for the GSEA enrichment scores (shown in Figure 4) for z-score theshold z.

Additional file 6. This figure shows the p-values for the GSEA enrichment scores (shown in Figure S1) for z-score theshold z.

Additional file 7. This figure shows the p-values for the GSEA enrichment scores (shown in Figure S2) for z-score theshold z.

Additional file 8. This figure shows the p-values for the GSEA enrichment scores (shown in Figure S3) for z-score theshold z.

Acknowledgements

The authors thank Dr. Tony Chiang and the anonymous referees for their helpful comments to improve writing and discussions of biological relevance.

Authors' contributions

SF1 performed the enrichment analysis, provided biological subject matter expertise, and co-wrote the manuscript. BP and EH performed the centrality calculations. BJ, CJ, and HK developed hypergraph methodology and co-wrote the manuscript. JM, and HM provided biological subject matter expertise and co-wrote the manuscript. AE, AS, LT, SF2, KW1, PH, DW, QT, VM, TS, AC, JK, KS, NH, LB contributed to generation, curation, and analysis of data. MD, RB, KW2, YK supervised generation and analysis of data. EP developed hypergraph methodology, supervised the hypergraph analysis team, and wrote the manuscript. All authors have read and approved the final manuscript.

Funding

Work in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (Grant No. 73639), a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. The experimental research and data generation were supported by the National Institute of Allergy and Infectious Diseases under Grant Number U19AI106772, including an administrative supplement (Ebola) and pilot award (MERS), and Contract Number HHSN272200800060C.

Availability of data and materials

The datasets generated during and/or analysed during the current study are available in the Gene Expression Omnibus (GEO) repository at the accession IDs listed in the paper, https://www.ncbi.nlm.nih.gov/geo/. The computed hypergraph and graph rankings are available as supplementary material. See ***Additional Files*** below.

Declarations

Ethics approval and consent to participate

Human lungs were obtained under protocol 03-1396, which was approved by the University of North Carolina at Chapel Hill Biomedical Institutional Review Board, and donors gave informed consent. 'Omics data collection studies for animal samples performed at UNC Chapel Hill were performed in animal biosafety level 3 facilities and were conducted under protocols approved by the Institutional Animal Care and Use Committee at UNC Chapel Hill (IACUC protocol #16-251) according to guidelines set by the Association for the Assessment and Accreditation of Laboratory Animal Care and the U.S. Department of Agriculture. All animal experiments and procedures performed at UW-Madison were approved by the UW-Madison School of Veterinary Medicine Animal Care and Use Committee under relevant institutional and American Veterinary Association guidelines. West Nile virus work in mice was carried out in strict accordance with the recommendations in the ***Guide for the Care and Use of Laboratory Animals*** of the National Institutes of Health. The protocols were approved by the Institutional Animal Care and Use Committee at the Washington University School of Medicine (Assurance number A3381-01).

Consent for publication

Not applicable

Competing interests

RB has ongoing unrelated collaborations and/or sponsored research agreements with Moderna, VaxArt, Eli Lily, Pfizer, Takeda and Ridgeback Biosciences. MD is a consultant for Inbios, Vir Biotechnology, and Fortressa Biotech and on the Scientific Advisory Boards of Moderna and Immunome. The Diamond laboratory has received unrelated funding support in sponsored research agreements from Moderna, Vir Biotechnology, Kaleido, and Emergent BioSolutions.

Author details

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA. [2]Department of Mathematics, University of Illinois, Urbana-Champaign, IL, USA. [3]Computing and Analytics Division, Pacific Northwest National Laboratory, Seattle, WA, USA. [4]Systems Science Program, Portland State University, Portland, OR, USA. [5]Department of Pathobiological Sciences, School of Veterinary Medicine, Influenza Research Institute, University of Wisconsin-Madison, 575 Science Drive, 53711 Madison, WI, USA. [6]Signature Science and Technology Division, Pacific Northwest National Laboratory, Richland, WA, USA. [7]Department of Medicine, Washington University School of Medicine, 63110 Saint Louis, MO, USA. [8]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [9]Department of Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX, USA. [10]KNOWBIO LLC., Durham, NC 27703, USA. [11]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. [12]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA. [13]Department of Comparative Medicine, University of Washington, Seattle, WA, USA. [14]Division of Virology, Department of Microbiology and Immunology, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan. [15]ERATO Infection-Induced Host Responses Project, Saitama 332-0012, Japan. [16]Department of Special Pathogens, International Research Center for Infectious Diseases, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan. [17]Department of Molecular Microbiology and Immunology, Oregon Health and Science University, Portland, OR, USA.

Received: 1 October 2020 Accepted: 13 May 2021
Published online: 29 May 2021

References

1. McDermott JE, Mitchell HD, Gralinski LE, Eisfeld AJ, Josset L, Bankhead A, Neumann G, Tilton SC, Schäfer A, Li C, et al. The effect of inhibition of PP1 and TNF*α* signaling on pathogenesis of SARS coronavirus. BMC Syst Biol. 2016;10(1):93. https://doi.org/10.1186/s12918-016-0336-6.
2. Mitchell HD, Eisfeld AJ, Stratton KG, Heller NC, Bramer LM, Wen J, McDermott JE, Gralinski LE, Sims AC, Le MQ, Baric RS, Kawaoka Y, Waters KM. The role of EGFR in influenza pathogenicity: multiple network-based approaches to identify a key regulator of non-lethal infections. Front Cell Dev Biol. 2019;7:200. https://doi.org/10.3389/fcell.2019.00200.
3. Tran VD, Sperduti A, Backofen R, Costa F. Heterogeneous networks integration for disease gene prioritization with node kernels. Bioinformatics. 2020;36(9):2649–56. https://doi.org/10.1093/bioinformatics/btaa008.
4. Adourian A, Jennings E, Balasubramanian R, Hines WM, Damian D, Plasterer TN, Clish CB, Stroobant P, McBurney R, Verheij ER, Bobeldijk I, van der Greef J, Lindberg J, Kenne K, Andersson U, Hellmold H, Nilsson K, Salter H, Schuppe-Koistinen I. Correlation network analysis for data integration and biomarker selection. Mol Biosyst. 2008;4(3):249–59. https://doi.org/10.1039/B708489G.
5. Diamond DL, Syder AJ, Jacobs JM, Sorensen CM, Walters KA, Proll SC, McDermott JE, Gritsenko MA, Zhang Q, Zhao R, Metz TO, Camp DG, Waters KM, Smith RD, Rice CM, Katze MG. Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics. PLoS Pathog. 2010;6(1):56. https://doi.org/10.1371/journal.ppat.1000719.
6. Maier TV, Lucio M, Lee LH, VerBerkmoes NC, Brislawn CJ, Bernhardt J, Lamendella R, McDermott JE, Bergeron N, Heinzmann SS, Morton JT, Gonzalez A, Ackermann G, Knight R, Riedel K, Krauss RM, Schmitt-Kopplin P, Jansson JK. Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome. MBio. 2017. https://doi.org/10.1128/mBio.01343-17.

7. McClure RS, Wendler JP, Adkins JN, Swanstrom J, Baric R, Kaiser BLD, Oxford KL, Waters KM, McDermott JE. Unified feature association networks through integration of transcriptomic and proteomic data. PLoS Comput Biol. 2019. https://doi.org/10.1371/journal.pcbi.1007241.
8. Lempp M, Farke N, Kuntz M, Freibert SA, Lill R, Link H. Systematic identification of metabolites controlling gene expression in *E. coli*. Nat Commun. 2019;10(1):4463. https://doi.org/10.1038/s41467-019-12474-1.
9. McDermott JE, Jarman K, Taylor R, Lancaster M, Shankaran H, Vartanian KB, Stevens SL, Stenzel-Poore MP, Sanfilippo A. Modeling dynamic regulatory processes in stroke. PLoS Comput Biol. 2012. https://doi.org/10.1371/journal.pcbi.1002722.
10. McDermott JE, Oehmen CS, McCue LA, Hill E, Choi DM, Stockel J, Liberton M, Pakrasi HB, Sherman LA. A model of cyclic transcriptomic behavior in the cyanobacterium Cyanothece sp. ATCC 51142. Mol Biosyst. 2011;7(8):2407–18. https://doi.org/10.1039/c1mb05006k.
11. Barabási A-L. Network science. UK: Cambridge University Press; 2016.
12. Iacopini I, Petri G, Barrat A, Latora V. Simplicial models of social contagion. Nat Commun. 2019;10:2485. https://doi.org/10.1038/s41467-019-10431-6.
13. Klamt S, Haus U-U, Theis F. Hypergraphs and cellular networks. PLoS Comput Biol. 2009;5(5):56. https://doi.org/10.1371/journal.pcbi.1000385.
14. Patania A, Petri G, Vaccarino F. The shape of collaborations. EPJ Data Sci. 2017;6(1):18. https://doi.org/10.1140/epjds/s13688-017-0114-8.
15. Javidian MA, Wang Z, Lu L, Valtorta M. On a hypergraph probabilistic graphical model. Ann Math Artif Intell. 2020. https://doi.org/10.1007/s10472-020-09701-7.
16. Joslyn CA, Aksoy S, Callahan TJ, Hunter L, Jefferson B, Praggastis B, Purvine EA, Tripodi IJ. Hypernetwork science: from multidimensional networks to computational topology. In: International conference on complex systems (ICCS 2020). 2020. https://arxiv.org/abs/2003.11782.
17. Leal W, Restrepo G. Formal structure of periodic system of elements. Proc R Soc A. 2019. https://doi.org/10.1098/rspa.2018.0581.
18. Minas M. Hypergraphs as a uniform diagram representation model. In: Proceedings of the 6th international workshop on theory and applications of graph transformations. Berlin: Springer; 1998. p. 281–95. https://doi.org/10.1007/978-3-540-46464-8_20.
19. Chitra U. Random walks on hypergraphs with applications to disease-gene prioritization. PhD thesis, Brown University. 2017.
20. Tran L. Hypergraph and protein function prediction with gene expression data. 2012. arXiv preprint arXiv:1212.0388.
21. Ramadan E, Tarafdar A, Pothen A. A hypergraph model for the yeast protein complex network. In: 18th International parallel and distributed processing symposium. 2004. p. 189. https://doi.org/10.1109/IPDPS.2004.1303205.
22. Zhou W, Nakhleh L. Properties of metabolic graphs: biological organization or representation artifacts? BMC Bioinf. 2011. https://doi.org/10.1186/1471-2105-12-132.
23. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol. 2010. https://doi.org/10.1038/nrmicro2419.
24. McDermott JE, Vartanian KB, Mitchell H, Stevens SL, Sanfilippo A, Stenzel-Poore MP. Identification and validation of ifit1 as an important innate immune bottleneck. PLoS ONE. 2012;7(6):36465.
25. Menachery VD, Eisfeld AJ, Schafer A, Josset L, Sims AC, Proll S, Fan S, Li C, Neumann G, Tilton SC, Chang J, Gralinski LE, Long C, Green R, Williams CM, Weiss J, Matzke MM, Webb-Robertson BJ, Schepmoes AA, Shukla AK, Metz TO, Smith RD, Waters KM, Katze MG, Kawaoka Y, Baric RS. Pathogenic influenza viruses and coronaviruses utilize similar and contrasting approaches to control interferon-stimulated gene responses. MBio. 2014;5(3):01174–14. https://doi.org/10.1128/mBio.01174-14.
26. Basters A, Knobeloch K-P, Fritz G. Usp18-a multifunctional component in the interferon response. Biosci Rep. 2018;38(6):56.
27. Mitchell HD, Eisfeld AJ, Sims AC, McDermott JE, Matzke MM, Webb-Robertson B-JM, Tilton SC, Tchitchek N, Josset L, Li C, et al. A network integration approach to predict conserved regulators related to pathogenicity of influenza and sars-cov respiratory viruses. PLoS ONE. 2013. https://doi.org/10.1371/journal.pone.0069374.
28. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):56. https://doi.org/10.1371/journal.pbio.0050008.
29. Hagberg A, Swart P, Chult DS. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos. 2008.
30. Dyer MD, Murali TM, Sobral BW. The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog. 2008;4(2):56. https://doi.org/10.1371/journal.ppat.0040032.
31. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
32. Kim EY, Ashlock D, Yoon SH. Identification of critical connectors in the directed reaction-centric graphs of microbial metabolic networks. BMC Bioinf. 2019;20(1):328. https://doi.org/10.1186/s12859-019-2897-z.
33. Mi W, Zhang Y, Lyu J, Wang X, Tong Q, Peng D, Xue Y, Tencer AH, Wen H, Li W, et al. The ZZ-type zinc finger of ZZZ3 modulates the ATAC complex-mediated histone acetylation and gene activation. Nat Commun. 2018. https://doi.org/10.1038/s41467-018-06247-5.

34. Li C, Bankhead A, Eisfeld AJ, Hatta Y, Jeng S, Chang JH, Aicher LD, Proll S, Ellis AL, Law GL, et al. Host regulatory network response to infection with highly pathogenic h5n1 avian influenza virus. J Virol. 2011;85(21):10955–67.
35. Berge C. Hypergraphs: combinatorics of finite sets, vol. 45. Elsevier; 1984.
36. Aksoy SG, Joslyn C, Marrero CO, Praggastis B, Purvine E. Hypernetwork science via high-order hypergraph walks. EPJ Data Sci. 2020;9(1):16. https://doi.org/10.1140/epjds/s13688-020-00231-0.
37. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3(4):59.
38. McDermott JE, Taylor RC, Yoon H, Heffron F. Bottlenecks and hubs in inferred networks are important for virulence in salmonella typhimurium. J Comput Biol. 2009;16(2):169–80.

Publisher's Note