frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 07 January 2014
doi: 10.3389/fgene.2013.00309

Multi-omic network signatures of disease

David L. Gibbs[1]*, Lisa Gralinski[2], Ralph S. Baric[2] and Shannon K. McWeeney[1,3]

[1] McWeeney Lab, Division of Bioinformatics and Computational Biology, Oregon Health & Science University, Portland, OR, USA
[2] Baric Lab, Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[3] McWeeney Lab, OHSU Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA

Edited by:
Xiaogang Wu, Indiana University-Purdue University Indianapolis, USA

Reviewed by:
Lifan Zeng, Indiana University, USA
Andrei Dragomir, University of Houston, USA
Anaïs Baudot, Centre National de la Recherche Scientifique, France

***Correspondence:**
David L. Gibbs, Division of Bioinformatics and Computational Biology, Oregon Health & Science University, 3181 S.W. Sam Jackson Park Rd., Portland, OR 97239-3098, USA
e-mail: gibbsd@ohsu.edu

To better understand dynamic disease processes, integrated multi-omic methods are needed, yet comparing different types of omic data remains difficult. Integrative solutions benefit experimenters by eliminating potential biases that come with single omic analysis. We have developed the methods needed to explore whether a relationship exists between co-expression network models built from transcriptomic and proteomic data types, and whether this relationship can be used to improve the disease signature discovery process. A naïve, correlation based method is utilized for comparison. Using publicly available infectious disease time series data, we analyzed the related co-expression structure of the transcriptome and proteome in response to SARS-CoV infection in mice. Transcript and peptide expression data was filtered using quality scores and subset by taking the intersection on mapped Entrez IDs. Using this data set, independent co-expression networks were built. The networks were integrated by constructing a bipartite module graph based on module member overlap, module summary correlation, and correlation to phenotypes of interest. Compared to the module level results, the naïve approach is hindered by a lack of correlation across data types, less significant enrichment results, and little functional overlap across data types. Our module graph approach avoids these problems, resulting in an integrated omic signature of disease progression, which allows prioritization across data types for down-stream experiment planning. Integrated modules exhibited related functional enrichments and could suggest novel interactions in response to infection. These disease and platform-independent methods can be used to realize the full potential of multi-omic network signatures. The data (experiment SM001) are publically available through the NIAID Systems Virology (https://www.systemsvirology.org) and PNNL (http://omics.pnl.gov) web portals. Phenotype data is found in the supplementary information. The ProCoNA package is available as part of Bioconductor 2.13.

Keywords: omics, networks, data integration, proteomics, transcriptomics, virology, biomarkers, SARS

INTRODUCTION

Statistical and computational methods are used in systems biology to infer underlying networks associated with disease (Aderem et al., 2011). Networks can be used for deriving predictive signatures of disease progression or severity, as well as helping to elucidate the underlying mechanisms (Zak and Aderem, 2009). A primary objective in systems biology is to understand the structure and connection between the diverse biological elements composing the living system, and how they dynamically change and interact in response to biologically important events, such as the host response to infection (Forst, 2006).

Single data-type signatures and biomarkers have found mixed success where many potentially useful biomarkers have not been validated (Ntzani and Ioannidis, 2003; Feng et al., 2004; Brenner and Normolle, 2007; Hughes, 2009; Bhavsar et al., 2010; Kint et al., 2010; Sturdevant et al., 2010). In virology, biomarkers could be used to predict the host response, allowing for earlier care, before the onset of extreme and damaging cytokine responses (Davey et al., 2013). The biomarker discovery process can utilize a range of different data types including genomic (DNA sequence data), transcriptomic (gene expression), proteomic (protein levels), metabolomics (metabolite levels), and prior biological knowledge such as that found in interactomics (encompassing protein-protein interactions databases).

It is thought that predictors or biomarkers utilizing multiple data types and/or exploiting the underlying network structure will prove more robust, as these more reflect the complex biology involved (Sung et al., 2012). A range of integration techniques have been suggested including machine learning methods (Lanckriet et al., 2004; Zhang et al., 2006; Daemen et al., 2008), probabilistic networks (Hartemink et al., 2002; Troyanskaya et al., 2003; Gat-Viks et al., 2006; Vaske et al., 2009), correlation networks (Adourian et al., 2008), statistical models (Nie et al., 2006; Fagan et al., 2007; Lê Cao et al., 2008; Torres-García et al., 2009), clustering techniques (Cancer Genome Atlas Network, 2012; Waters et al., 2012) and applications of spectral theory (Berger et al., 2006; Tan et al., 2009; Kim et al., 2012). To produce integrated network signatures, however, methods must be applied across extremely heterogeneous sources, which has proven difficult because of the extreme differences between data types. In particular, the integration of the transcriptome and proteome is a current challenge in omics research due to differences in dynamic range of measurements, incomplete annotation, isoform differences, and temporal effects, as several examples (Cox et al., 2005; Waters et al., 2006a,b; Cancer Genome Atlas Network, 2012).

We have developed the methods needed to explore whether a relationship exists between co-expression network models built from transcriptomic and proteomic data types, and whether this relationship can be used to improve the disease signature discovery process. This work uses publically available data from an NIAID systems biology consortium study involving infection of SARS-CoV in mice. We have developed an approach to produce integrated network signatures of disease by leveraging earlier work on co-expression transcriptome networks (Zhang and Horvath, 2005; Yip and Horvath, 2007; Mason et al., 2009; Langfelder et al., 2008, 2011; Langfelder and Horvath, 2012) and our own work in proteomic co-expression networks (Gibbs et al., 2013). The signature consisted of a bipartite module graph, connecting co-expression modules obtained from transcriptomic and proteomic data, that is constructed using significant module member overlap, correlation of eigenvector summaries, and common phenotypic associations with outcomes of interest. The functional enrichment of module sub-graphs was overlapping across data types, further offering evidence of the underlying biological network structure. This work provides a framework for multi-omic prioritization of module members for biomarker studies as well as perturbation and validation experiments (see **Figure 1**).

METHODS

EXPERIMENTAL DATA

This publically available data (https://www.systemsvirology.org, experiment SM001) was generated from 20-week-old C57/B6 mice infected with the MA15 mouse adapted SARS-CoV virus (Roberts et al., 2007). In total, transcript and proteomic expression profiles were collected for 92 mice representing four dosage levels (10^2, 10^3, 10^4, 10^5 PFU) over four time points (1, 2, 4, 7 days), including 3 mock samples per day (5 mice $*$ 4 time points $*$ 4 dose levels $+$ 3 mocks $*$ 4 time points).

The control mouse at Day 7, replicate 2, and infected mouse (PFU 10^2), Day 4, replicate 3, were removed from the study since in the transcript data, the mice clustered with the incorrect infection label.

Phenotype data quantified the pathological severity resulting from infection (see Gibbs et al., 2013 for more on the phenotype data. The phenotype data can be found in **Supplementary Table 3**). While the difference in pathology level among the viral dosages is small, there was some observed difference in the kinetics of infection. Higher doses prompted more immediate responses. The difference in pathology among the viral dosages is relatively small. In total, 15 phenotype variables were recorded including an aggregate measure called the "overall pathology score." Many of the phenotype variables are highly correlated, such as inflammation, airspace inflammation, and interstitial inflammation. Other phenotype variables include physical characteristics such as diffuse alveolar damage (DAD), debris, edema, and hyaline membranes. Day and dose are also included in the analysis.

TAKING THE PEPTIDE-TRANSCRIPT INTERSECTION

Using the VIPER software (v3.48) (Monroe et al., 2007) peptides are matched to an Accurate Mass and Time (AMT) tag database (Zimmer et al., 2006). Details are given on the systemsvirology.org site and in Gibbs et al. (2013). Abundance measurements for 16,890 peptides mapping to 3277 proteins were recorded for 184 LC-MS runs. Taking all observed peptides, protein inference was performed using the Fido protein inference model (Serang et al., 2010; Serang and Noble, 2012). Inferred Proteins were accepted with scores above 0.95. From the total set of proteins, 691 proteins had this score or better.

Peptide data was filtered by peptide matching quality scores (given by the VIPER software) STAC ($>$0.6), UP ($>$0.5), and Peptide Prophet tag score ($>$0.9), which resulted in a matrix of 184 sample rows by 9326 peptides. Sample replicates were combined by taking the mean over replicates.

A second round of filtering was performed by assessing the quantity of missing data. Missing data are encountered when peptides are identified in a subset of samples. "Missingness filtration" involves removing any peptide with greater than X% missing data across samples. Peptides were filtered by missingness, taking peptides with not more than 20% missing data, resulting in a matrix of 90 samples by 2273 peptides. This process eliminates a large proportion of the measured peptides (86.5%), however, for statistical analysis that depends on nearly complete matrices, many peptides are unusable since they are identified in a very small number of samples.

The matched microarray data were processed using the Agilent Preprocess Bioconductor package (Lopez-romero, 2010). After considering quality measures, 31,416 probes passed probe QC flags for all replicates of at least one infected time point.

In order to focus on the dynamic relationship between the transcriptome and proteome, the two data types were subset using the intersection based upon Entrez gene IDs. Transcript probes on the gene microarray were mapped to Entrez gene IDs using annotation databases (mgug4122a.db) found in Bioconductor (2.11). Proteins in the AMT tag database were mapped to Entrez IDs and protein families using the Uniprot web service (Apweiler et al., 2004; Wu et al., 2006; Magrane and Consortium, 2011).

Networks were constructed using intersection between quality filtered peptide and transcript data. This corresponded to 90 total samples with 2205 peptides mapping to 445 Uniprot IDs (in the mass tag database) and 490 Entrez gene IDs. These gene IDs were used to select transcript probes, resulting in 814 probes that corresponded to 439 Uniprot IDs and 447 Entrez gene IDs. Between the two data sets, 437 Entrez IDs are shared. Some discrepancy was observed, since peptides are often degenerate, mapping to multiple proteins, creating a scenario in which an entry maps from a given Uniprot ID to multiple Entrez IDs. To address mapping discrepancies, after mapping to Entrez IDs, only IDs contained within the intersection were considered.

CONSTRUCTING THE INTEGRATED BIPARTITE MODULE GRAPHS

Network based integration of transcript and peptide data was performed by constructing independent co-expression networks for each data type using methods derived from WGCNA. (Langfelder and Horvath, 2008; Mason et al., 2009; Iancu et al., 2012a).

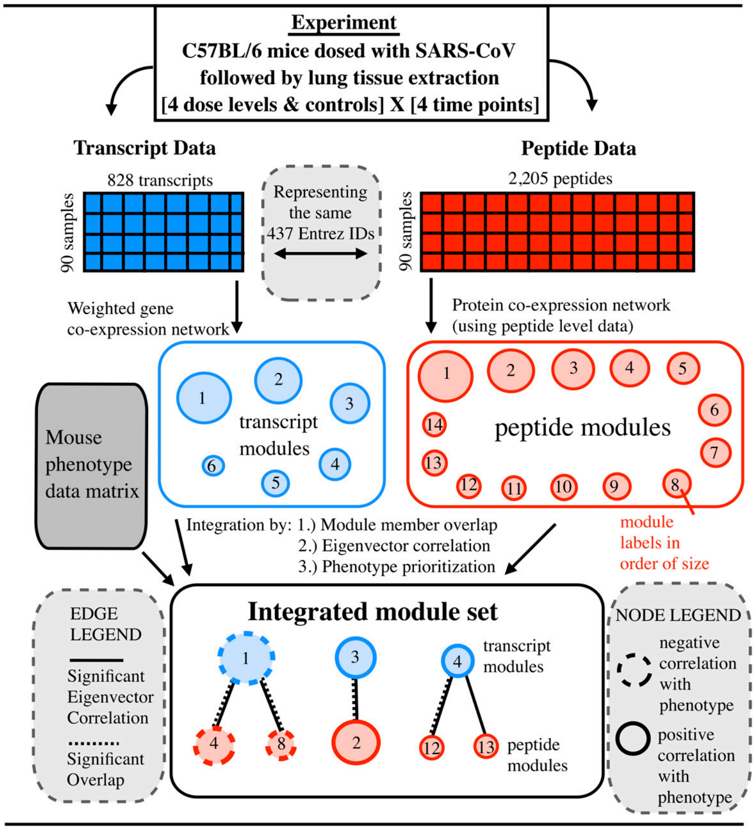

Experiment
C57BL/6 mice dosed with SARS-CoV
followed by lung tissue extraction
[4 dose levels & controls] X [4 time points]
Transcript Data
828 transcripts
90 samples
Representing the same 437 Entrez IDs
Peptide Data
2,205 peptides
90 samples
Weighted gene co-expression network
Protein co-expression network (using peptide level data)
transcript modules
peptide modules
Mouse phenotype data matrix
module labels in order of size
Integration by: 1.) Module member overlap
2.) Eigenvector correlation
3.) Phenotype prioritization
Integrated module set
transcript modules
peptide modules
EDGE LEGEND
Significant Eigenvector Correlation
Significant Overlap
NODE LEGEND
negative correlation with phenotype
positive correlation with phenotype

FIGURE 1 | Prioritization of integrated data leading to a multi-omic integrated co-expression signature for SARS-CoV infection. Given mice infected with a mouse adapted SARS-CoV virus, transcript (gene microarray) and peptide (LC-MS) data are collected from lung tissue. The data is used to construct independent co-expression networks for each data type. Using three metrics, transcript and peptide modules are compared and joined, producing a bipartite module graph. In the module graph, two kinds of edges are shown. Solid edges indicate a significant correlation between module eigenvectors. Dashed edges indicate significant overlap in terms of module membership. Node sizes correspond to the size of the module and node outlines show the direction of correlation compared to the degree of lung pathology (the overall-pathology-score).

Co-expression networks are composed of nodes connected by weighted edges. Edge weights were computed using signed, robust correlations producing a similarity matrix (Langfelder et al., 2012). The similarity matrix raised to a power, beta, produces the adjacency matrix. Beta was selected according to scale-free model criterion (R^2 describes the model fit), peptide networks had an R^2 greater than 0.8 and transcript networks greater than 0.9. The adjacency matrix is used to compute topological overlap between nodes, weighting the network edges. Topological overlap is defined as $TOM_{ij} = (l_{ij} + a_{ij})/[\min(k_i, k_j) + 1 - a_{ij}]$ where l_{ij} is the dot product on row i and column j in adjacency matrix [a] and k_i (the connectivity) is the summation of row i in adjacency matrix [a].

Groups of nodes were partitioned into subnetworks, or modules, containing (in this case) either transcripts or peptides. Modules were composed of strongly connected (high edge weights) nodes. The dynamic hybrid treecut method was used to derive subnetworks or modules (Langfelder et al., 2008), using default settings.

Modules were summarized by taking the first right-singular vector produced by singular value decomposition on expression data represented by nodes contained in the module. The module eigenvectors (MEs or module summaries) allowed us to associate modules to biological phenotypes using correlation. Modules were labeled numerically according to decreasing size, module 1 being the largest.

Although the hybrid treecut algorithm generally works well, there is no guarantee for a module's average connectivity to be greater than what is expected by random. Permutation testing was used to assess module significance by comparing the mean topological overlap of nodes within a module to the mean topological overlap of a randomly sampled set of nodes equal in size to the module being tested. This is equivalent to permuting the module labels on nodes. Ten thousand permutations were performed.

Finally, to build the bipartite module graph, the first step involved measuring the member overlap between all pairs of modules (peptide-transcript). To test the significance of overlaps, random modules were constructed, keeping the module sizes fixed (equivalent in size to our derived modules) and varying the contents. Ten thousand permutations were performed. The count of permuted overlaps larger than the observed was used as an empirical *p*-value. Significant overlaps were used as edges between modules if FDR adjusted p-values were less than 0.1. This relaxed threshold was picked to increase sensitivity.

The edges of the bipartite graph are filtered by correlation between eigenvector summaries. The connection between module eigenvectors was measured using the Pearson correlation. Using Bonferroni multiple testing correction, p-values less than 0.0006 [0.05/(14 peptide modules $*$ 6 transcript modules)] were accepted. If deemed appropriate, edges could be retained without significant overlap, as long as the annotation reflects that. The third step involved filtering edges by comparing the joined modules association to a phenotype. Edges were kept if the eigenvector-phenotype association was in the same direction and adjusted p-values were less than 0.05.

The combination of these three measures—member overlap, eigenvector correlation, and similar phenotype associations—constructed an integrated bipartite module graph that describes an integrated signature.

Functional enrichment via Gene Ontology terms was performed using the GOstats package (Ashburner et al., 2000; Falcon and Gentleman, 2007). The universe consisted of the 5521 Entrez IDs found in mass tag database, the limiting factor on peptide identifications. The conditional method was used which minimizes the correlation between GO terms. P-values were adjusted using the Benjamini and Yekutieli method (Benjamini and Yekutieli, 2001).

NAÏVE METHOD BASED ON CORRELATION FOR COMPARISON TO THE MODULE LEVEL ANALYSIS

A Pearson correlation based naïve approach was designed for comparison to the module level analysis. The approach involved computing correlations—independently for each peptide and transcript—on the overall pathology score phenotype. This produced two ranked lists for peptides and transcripts, each with both positive and negative correlations. The naïve top ranked entities were compared to the rankings within modules.

For comparison to the enrichment results, members of the ranked lists were selected from the most negative and most positive correlations separately, with size equal to the mean module sizes for the peptide and transcript networks (151 peptides and 130 transcripts). These selections were used for gene ontology enrichment using the same method as described in Constructing the Integrated Bipartite Module Graphs.

RESULTS

CHARACTERIZATION OF THE CO-EXPRESSION NETWORKS OBTAINED FROM SINGLE OMIC DATA TYPES

The transcript network consisted of 6 modules containing between 42 and 357 transcripts. The peptide network consisted of 14 modules containing between 70 and 316 peptides. The modules were labeled numerically in order of decreasing size. Each co-expression network independently showed significant Pearson correlations between the module eigenvectors and phenotype data (See **Supplementary Figures 1** and **2**). In both networks, the greatest positive correlation was found with day of infection (transcript module 3, $r = 0.84$, peptide module 10, $r = 0.70$). When inflammation related variables were considered (inflammation, airspace inflammation and interstitial septum inflammation) transcript module 3 and peptide module 2 showed the strongest correlations (inflammation, transcript module 3, $R = 0.6$, peptide module 2, $R = 0.6$). These modules also showed the strongest correlation with the overall pathology of the mice. Conversely, transcript module 1 and peptide module 4 showed the same pattern of associations with phenotype as transcript module 3 and peptide module 2, but with negative correlations. Members of these modules had abundance profiles that decreased over time. In contrast to the previous two patterns of association, transcript module 4 and peptide module 12 were more associated with the administered viral dosage instead of day of infection. The modules showed strong associations with denudation, debris, and airway pathology whereas the previous two module sets did not.

INITIAL CONSTRUCTION OF THE BIPARTITE MODULE GRAPH BY OVERLAP OF MODULE MEMBERS

Significant module member overlaps were observed between the two co-expression networks. Overlaps were represented as a count of similar Entrez gene IDs (after mapping). Using permutation testing we ascertained whether the size of the overlap was larger that what is expected by chance. Permutation test significance was defined as FDR adjusted p-values less than 0.1. By that definition, ten out of 84 possible significant module overlaps were observed, forming three distinct sub-graphs, initializing the bipartite module graph. Overlaps were quantified by taking |Intersection(A,B)|/ min[size(A), size(B)]. Transcript module 1 overlapped with peptide modules 1, 4, and 8 (overlap amounts of 0.33, 0.39, and 0.42 with FDRs 0.0, 0.08, and 0.016 respectively). Transcript module 2 overlapped with peptide modules 2 and 3 (overlap amounts of 0.38 and 0.35 with FDRs 0.016 and 0.041 respectively). Transcript module 3 overlapped with peptide modules 2, 3, 5, and 10 (overlap amounts of 0.37, 0.38, 0.31, and 0.26 with FDRs of 0.068, 0.016, 0.087, and 0.016 respectively), and lastly, transcript module 4 overlapped with peptide module 12 (overlap amount 0.11, FDR 0.080). These overlaps formed the initial edges of the bipartite module graph.

MODULE EIGENVECTOR CORRELATION CONFIRMS AND ADDS EDGES TO THE BIPARTITE GRAPH

From the 10 edges in the overlap graph, eight showed significant eigenvector correlation after Bonferroni multiple testing correction (p-values < 0.0006; See **Supplementary Figure 3**). Summarizing the results: the Pearson correlations between transcript module 1 and peptide modules 4 and 8 were 0.523 and 0.434 respectfully (p-values 1.16e-07 and 1.908e-05). The correlation between transcript module 4 and peptide modules 12 and 13 was 0.696 and 0.683 (p-values 1.159e-13 and 2.554e-14). The correlation between eigenvectors of transcript module 3 and peptide modules 2 and 10 was 0.755 and 0.801 (p-value 2.2e-16 for both). A particularly interesting case is seen with transcript module 4 and peptide module 13, where module overlap is not observed, but a strong connection between module eigenvectors is present. This encouraged us to include an additional edge between these modules. These cases are potentially very interesting connections, where connected modules are driven by unmatched hubs in the network, which might imply previously unknown interactions.

ADDITIONAL EDGE CONFIRMATION WITH SHARED PHENOTYPE ASSOCIATIONS

The bipartite module graph can be further modified by comparing independent module associations to a phenotype of interest. Strong correlation between eigenvectors typically brings a shared correlation to sample phenotypes due to the similar vector structures. However, correlation is not transitive, which explains why this should be accounted for, because if two module eigenvectors correlate, it does not mean the two modules both correlate to a given phenotype.

Considering only the overall pathology score phenotype, transcript module 3 and peptide module 2 remained connected. Transcript module 4 and peptide modules 12 and 13 also demonstrated strong shared phenotype associations. Transcript module 1 and peptide modules 4 and 8 shared negative correlations with the overall pathology. These associations reinforced the bipartite graph structure. Edges that did not reflect this similarity in phenotype associations were removed.

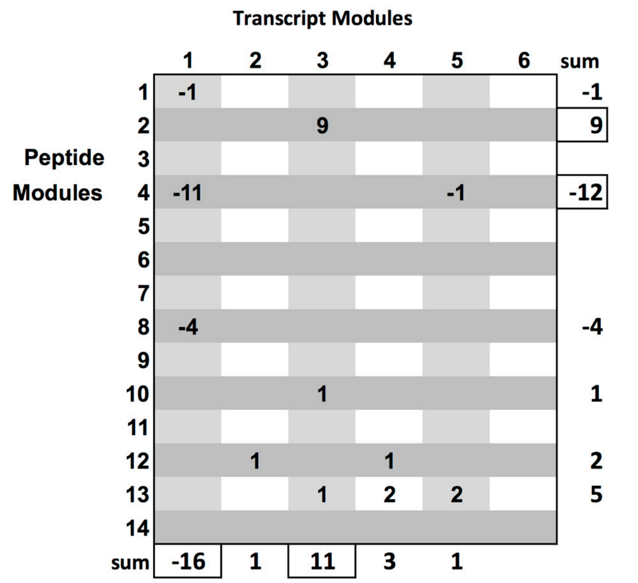
The rich set of phenotypes was used to prioritize the bipartite sub-graphs. We briefly describe an algorithm to prioritize module sub-graphs, given a set of phenotypes: initialize an n by m matrix where n is the number of transcript modules and m is the number of peptide modules. For each phenotype, the maximum and minimum (i.e., negative correlation) correlating modules from each data type are found. For each pair of modules in the matrix, and for each phenotype, if a pair of modules is maximum, +1 is added to the matrix element corresponding to this pair, and if the module pair has the minimum correlation, a −1 is added to the matrix position corresponding to the pair (see **Figure 2** and Supplementary_Network_Results). In some cases, a module could arrive at a final summed score of 0 by alternatively winning both negative and positive correlations. Therefore, it might be necessary to keep scores strictly positive. This would return the maximum score by magnitude, without regard to direction, as an alternative prioritization procedure. After the prioritization routine, the set of biologically relevant multi-omic modules is ranked, providing a clear path toward targeted, downstream, analysis.

MODULE SUB-GRAPHS SHOW TEMPORAL TRENDS

The peptides and transcript expression profiles, within a module sub-graph, showed two types of temporal patterns. The expression response either varied with time or with viral dosage. The patterns are made clear after collapsing the eigenvector summaries by day (see **Figure 3**).

The first row of **Figure 3** shows the abundance peaks of transcript module 3 and peptide module 2. Baseline abundance is observed on day 1 followed by abundance increases over the course of infection.

The second type of trend is exemplified by transcript module 1 with peptide modules 4 and 8, as well as transcript module 4 with peptide modules 12 and 13. This trend showed a response pattern associated with viral dosage rather than time. This is clearly seen in the high dose column. In summary, we have found two different patterns of module response, one by time (increasing over time) and the other by dosage.



		Transcript Modules						
		1	2	3	4	5	6	sum
Peptide Modules	1	-1						-1
	2			9				9
	3							
	4	-11				-1		-12
	5							
	6							
	7							
	8	-4						-4
	9							
	10			1				1
	11							
	12		1		1			2
	13			1	2	2		5
	14							
	sum	-16	1	11	3	1		

FIGURE 2 | Module graph prioritization by examination of the relationship of module-pairs to all phenotypic variables. Clear patterns show transcript module 1 and peptide module 4 with the bulk of maximum negative correlations and transcript module 3 and peptide module 2 with the bulk of maximum positive correlations with the phenotype. This matrix clearly provides ranking on sub-graphs.

SUB-GRAPHS SHARE FUNCTIONAL ANNOTATIONS

After sub-graph prioritization, we were interested in what biological entities were most central (i.e., correlated with the module eigenvector) in the network modules. By filtering module members by centrality, we have a metric by which to rank and prioritize the module members. Taking the ten most central module elements and using Uniprot web services, we examined the associated protein families. For transcript module 1 and peptide module 4, shared families included the Caveolin family, GST superfamily, Mu family, Cu-Zn superoxide dismutase family and transcript module 4 contained members of the aldehyde dehydrogenase family. These protein families are associated with metal binding proteins, interactions with DNA and engaging in changes to acetylation patterns. Putative caveolin-binding sites have been observed in SARS-CoV encoded proteins, and aldehyde dehydrogenases have been found to have a role in infection (Cai et al., 2003; Cinatl et al., 2004).

In transcript module 3, central members included proteins from the CAP family, Histone H1/H5 family, Histone H2A family, and the intermediate filament family, while for peptide module 2 protein families included the GTP-binding elongation factor family, the EF-Tu/EF-1A subfamily, the heat shock protein 90 family, the Histone H1/H5 family, and the DEAD box helicase family (eIF4A subfamily). Both histones and elongation factors have previously been associated with SARS-CoV infection (Reghunathan et al., 2005).

Transcript module 4 and peptide modules 12 and 13 were enzyme-driven and were associated with Serpin family members, which are protease inhibitors. This is potentially important since

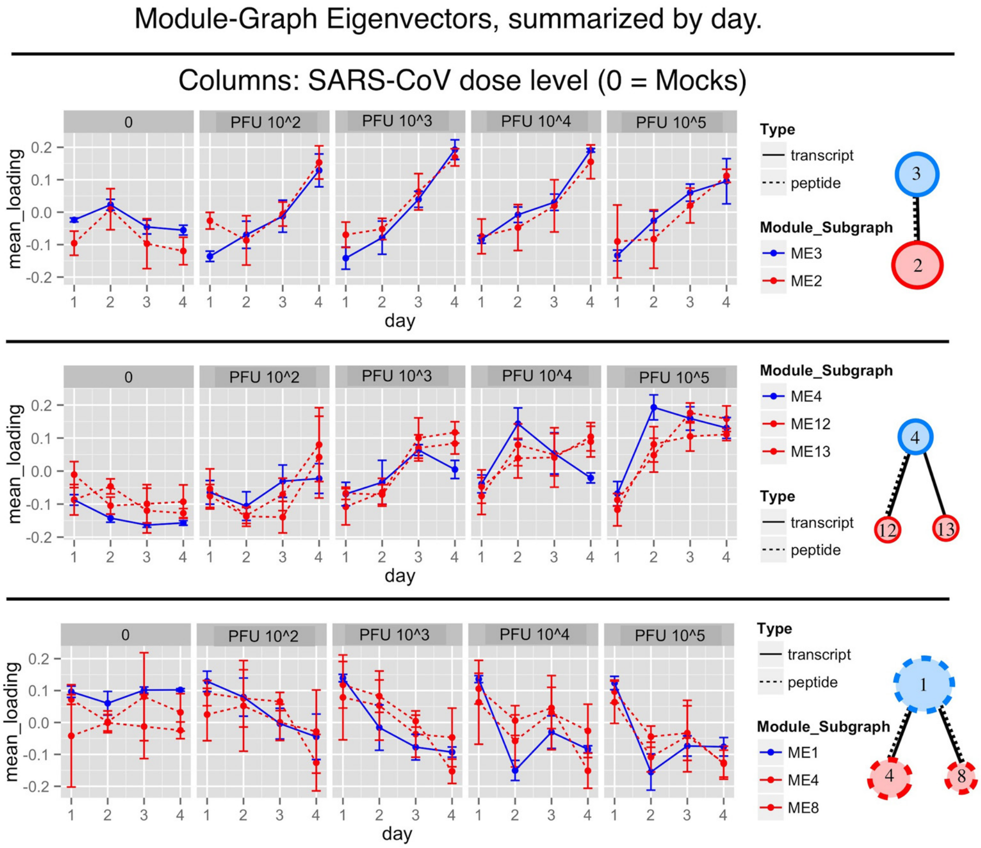


FIGURE 3 | Overlapping eigenvector summaries show similar response patterns observed across transcript and peptide modules. The blue lines show collapsed transcript module eigenvectors plotted over days post infection. Red dotted lines show the collapsed peptide module eigenvectors. The top row shows the module sub-network of transcript module 3 and peptide module 2. The middle row shows transcript module 4 and peptide modules 12 and 13. The bottom row shows transcript module 1 with peptide modules 4 and 8. There is clear evidence of a shared response between transcript and peptide modules, demonstrating a multi-omic signature. See the supplementary tables for lists of the drivers and enriched functional categories.

it was recently reported that—along with Serpin1—the coagulation and urokinase pathways are activated during infection. (Gralinski et al., 2013). This is closely related to lung pathology involving disseminated small vessel thromboses in the lungs (Ng et al., 2004a,b; de Lang et al., 2007). This module sub-graph could be used for further examining the systems level connection between omics and pathology.

Examination of significant GO terms shows largely similar trends compared to protein family annotations (all adjusted *p*-values are Bonferroni adjusted *p*-values). For transcript module 1 and peptide modules 4 and 8, the most significant overlapping GO terms include processes involving actin filament processes ("actin filament-based movement," transcript module 1, adj. *p*-value 5.51e-05, "actin cytoskeleton organization," peptides modules 4 and 8, adj. *p*-values 5.84e-02 and 4.11e-03 respectively), component assembly ("protein complex assembly," transcript module 1, adj. *p*-value 1.09e-02, "cellular component assembly," peptides module 4, adj. *p*-values 3.93e-07). It has been observed that SARS-CoV infection induces structural changes involving actin reorganization (Ng et al., 2004a,b; Surjit et al., 2004).

Transcript module 3 and peptide module 2 have a number of overlapping enriched GO terms including "nucleosome assembly" (transcript module 3, adj. *p*-value 9.66e-05, peptides module 2, adj. *p*-value 2.37e-05), and "protein-DNA complex subunit organization" (transcript module 3, adj. *p*-value 2.63e-03, peptide module 2, adj *p*-value 2.58e-04). Also, shared terms include "cellular component assembly" (transcript module 3, adj. *p*-value 5.48e-05, peptides module 2, adj. *p*-value 5.00e-05), and "cellular macromolecular complex assembly" (transcript module 3, adj. *p*-value 5.80e-05, peptides module 2, adj. *p*-value 4.71e-09). These functional associations again point to structural changes (Reghunathan et al., 2005).

In transcript module 4, peptide module 12 and peptide module 13, enriched GO terms associated with the regulation of processes. In particular, the "negative regulation of endopeptidase" (transcript module 1, adj. *p*-value 5.59-02, peptides modules 12 and 13, adj. *p*-values 8.36e-15 and 4.76e-06 respectively). Although, the transcript module is not highly significant when considering Bonferroni adjusted *p*-values, the GO term overlap between modules is strong here. Clearly endopeptidases, hydrolases, and cytokines have important roles in SARS-CoV infection (Loureiro and Ploegh, 2006). Cystatins, one of the represented protein families, has been proposed as a potential therapeutic target (Leung-Toung et al., 2006).

Overall, each set of integrated modules was overlapping in its functional annotation. Similarity in annotation between connected modules of different data types adds further evidence of true biological connection.

INTEGRATION OF MODULE SUB-GRAPHS LEADS TO RICHER RESULTS

A motivating use case for this approach was to develop a framework for integrated, network-based prioritization of targets for perturbation and validation. We compared the module level results to those attained by use of a naïve correlation-based method.

The naïve results took the form of two ranked lists for each data type. In the lists, there were both negative and positive correlations to the overall pathology phenotype. After mapping peptides and transcripts to Entrez IDs, correlation to the overall pathology phenotype, and the correlation between data types was compared (See **Supplementary Figure 4**). While some peptide-transcript pairs showed both strong correlation to the phenotype and strong correlation across data types, 33.1% of peptide-transcript pairs were essentially uncorrelated (across data type correlation, $-0.1 < r < 0.1$), and 4.6% of peptide-transcript pairs were anti-correlated ($r < -0.3$).

The naïve top ranked peptides and transcripts were not necessarily the most central within a given module. In transcript module 1, the most central (by correlation to the module eigenvector) was ranked 4th in the list of positive correlations. The next two top ranked module transcripts were not in the naïve top 10. On the peptide side, in module 4, the top ranked peptide was ranked 10th in the naïve list. This is due to the fact that the module construction is independent of any phenotype measurement. Module structure is a result of entities sharing a pattern of expression, rather than sharing a correlation with some external measurement.

The naïve results showed a reduction in significance compared to the module level tests. Four sets were taken from the ranked lists: 151 positively correlated peptides (PCP), 151 negatively correlated peptides (NCP), 130 positively correlated transcripts (PCT), and 130 negatively correlated transcripts (NCT). The results are listed in **Supplementary Table 1**. The PCP set of 151 peptides resulted in four significant GO terms, with Bonferroni adjusted *p*-values ranging from 2.67e-03 to 2.55e-02. The NCP set showed no significant GO term enrichment after Bonferroni adjustment. On the other hand, the PCT set of 130 transcripts showed 16 significant GO terms, with Bonferroni adjusted *p*-values in the range of 9.14e-11 to 2.24e-02. The NCT set showed three significant GO terms, Bonferroni *p*-value range of 2.61e-04 to 1.29e-02.

Two of the four significant GO terms found in the PCP set were also found in the PCT set ("protein polymerization" and "cellular macromolecular complex assembly"). However, one GO term, "cellular response to interleukin-4" (adj. *p*-value 2.67e-03), was not found in the enrichment results for both the naïve transcript sets and the peptide modules.

From the PCT set, eight of ten of the most significant GO terms were also found in the bipartite module graph results. The two that were not included were "regulation of actin filament length" (adj. *p*-value 1.07e-04) and "regulation of actin filament polymerization" (adj. *p*-value 3.96e-04). Enrichment in the NCT set showed one term that was also found in the bipartite module graph, and two terms, including "xenobiotic metabolic process" (adj. *p*-value 7.77e-03) and "response to xenobiotic stimulus" (adj. *p*-value 1.29e-02), which were not.

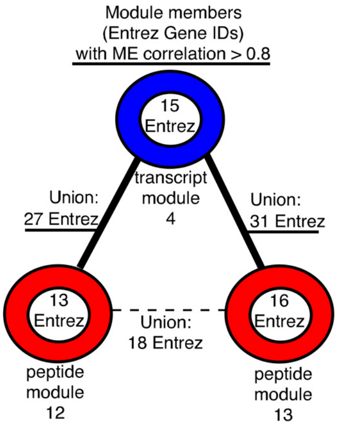
The module level organization provided more significance in enrichment tests compared to the naïve results. In peptide modules, the most significant results were found in peptide module 7 (not part of the module-graph) for GO terms "cellular component biogenesis" (adj. *p*-value 1.94e-17), "cellular macromolecular complex assembly" (adj. *p*-value 2.47e-17) and "nucleosome assembly" (adj. *p*-value 8.29e-12) where 15 of 36 Entrez IDs were present. In the NCP results, where the expression profile is negatively correlated with pathology, no GO terms were significantly enriched, whereas in the module-graph, there were sixteen significantly enriched GO terms.

On the module level, highly ranked members (by module centrality), showed little overlap across data types in terms of shared Entrez IDs, even in the presence of significant module membership overlap (**Supplementary Tables 1** and **2**). Regardless, we found that significant GO term enrichment was observed to be overlapping across the data types. By utilizing the module sub-graph, the most central members were functionally compared, producing potentially novel connections to investigate. An example was found in the module sub-graph including transcript module 4 and peptide modules 12 and 13 (See **Figure 4**). The most central members ($r > 0.8$) enriched the same GO terms, and in taking the union across modules, each data type brings unique, and functionally related, members to the analysis.

DISCUSSION

In this work, a new strategy for data integration has been developed that leverages existing network inference methodology for transcriptomics and more recent extensions in proteomics. The method was applied to infection time course data to examine both the individual networks and the bipartite module graph with regard to phenotypic correlation and functional enrichment, as well as to provide integrated prioritization across data types.

Information needed to construct the bipartite module graph included membership overlaps, summary eigenvector correlations and common correlations to phenotype. The strong correlations between module eigenvectors highlights the biological organization observed across data types. Given that a large portion of peptides and transcripts, mapping to the same Entrez ID, are essentially uncorrelated, it is striking to observe the modularity in our inferred networks, and the strong connection between a subset of the modules. Effectively, this can provide the foundation of a true multi-omic signature of SARS-CoV viral infection that may have relation to other viral respiratory infections as well. Additionally, this provides a perspective on modularity in the



GOBPID	GO Term	Transcript Module 4	Peptide Module 12	Peptide Module 13	Union All
GO:0010951	Neg. Reg. of Endopeptidase	3	9	7	10
GO:0051336	Regulation of hydrolase	5	10	8	14
GO:0006500 /	Proteolysis	5	.	.	.
GO:0030162	Regulation of Proteolysis	.	8	6	8

FIGURE 4 | An increase in the number of entities mapped to GO terms using the union of module members within a module sub-graph. Enrichment was performed using only the module members with correlations to the module eigenvector greater than 0.8, these GO terms were all in the top 5 most significant after GO term enrichment within each module, and, after taking the union, they were still in the top 5 most significant. Adjusted *p*-values for all modules can be found in **Supplementary Table 1**.

proteome and its relationship to the module structure of the transcriptome.

Initially modeling each data type individually offers a high degree of flexibility in the analysis. In co-expression network construction, one has choices about correlation type or metric for association between nodes in the network, how or whether to scale the associations, and how to cluster for module discovery. Additionally, there are questions of normalization and missing data that can affect on the downstream network structure. These parameters can be separately tuned in order to produce optimal independent networks on data types, which can be used to produce integrated module graphs. Generating well-formed, independent networks should improve the odds of successful integration since they each more accurately reflect the underlying biology. Without treating different data sources independently, it is possible that the inherent noise found in biological data would obscure patterns linking data types.

In the correlation of expression profiles across data types, there was a large degree of uncorrelated peptides and transcripts, which has been previously observed (Ghazalpour et al., 2011). Additionally, there are peptides and transcripts, which are both correlated to phenotype, but are anti-correlated across data types. The apparent disconnect between data types makes interpretation difficult. The naïve ranked list is expected to be less ordered compared to within-module rankings. This is seen in the fact that for NCP, the naïve method returned no significant enrichment results, while the module level results did show significance. Compared to the naïve ranking approach, the module level analysis avoided the problem of uncorrelated peptides and transcripts, by connecting modules using a set of metrics, rather than simple Entrez ID mapping. Additionally, the module organization returned considerably more significant enrichment results, and also showed more functional overlap across data types.

One of the most difficult aspects of data integration can be the annotation of highly heterogeneous data sets, connecting the transcript and peptide to their correct source gene for instance. For transcripts, this annotation is more straightforward since microarray probes have been designed specifically to avoid degeneracy among genes and have relatively good documentation. On the other hand, given a peptide, it can be quite difficult to determine what gene it ultimately resulted from. Our knowledge of the proteome is still rapidly expanding, directly affecting our peptide-transcript integration solution. As proteomics technologies improve, however, the intersection between the proteome and transcriptome will continue to grow, improving our integrated models, and our understanding of the cell.

In this work, we have provided a strategy for integrated analysis in order to shed light on complex biology. With these methods, it is possible to learn novel and biologically relevant information about the relationship between the host and pathogen, but more generally between the transcriptome and the proteome. This work should prove to be platform independent, allowing the use of RNAseq (see Iancu et al., 2012b) or other forms of proteomic data. This can then be used to inform systems-level prioritization for the subsequent perturbation and validation experiments, allowing the full realization of systems based approaches.

AUTHOR CONTRIBUTIONS

David L. Gibbs designed the methods, analyzed the results, developed the code, and wrote the manuscript. Lisa Gralinski and Ralph S. Baric are responsible for all aspects of generating the SARS data. Shannon K. McWeeney designed methods, contributed to the analysis and edited the manuscript.

FUNDING

This work was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services [5U54AI081680, U19AI100625] and National Library of Medicine [3T15LM7088-18S1].

ACKNOWLEDGMENTS

The authors would like to thank Arie Baratt and the Pacific Northwest National Lab assistance in peptide identification. The Baric Lab was invaluable in providing discussions regarding the phenotypic data they generated, along with helpful discussion regarding the integration strategy with respect to prioritization. David Maier provided extremely useful insight with regard to the data integration strategies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2013.00309/abstract

Supplementary Table 1 | Shows the top ten enriched GO terms for each module sub-graph. The tables correspond to the module sub-graphs in **Figures 1**, **3**, and **4**. Also contains the GO enrichment results for the Naïve comparison.

Supplementary Table 2 | Shows the top ten entities by correlation with the module summaries (ME, module eigenvector). The second column shows the correlation with the Overall-total-score phenotype describing the level of lung pathology. The tables correspond to the module sub-graphs in **Figures 1**, **3**, and **4**.

Supplementary Table 3 | The mouse phenotype for experiment SM001, for both mocks and infected mice. Pathological variables are given along with viral dosages (Dose), day of tissue sampling post infection (Day), Mouse ID (RepName).

Supplementary Figure 1 | Pearson correlations between transcript module eigenvector summaries and mouse phenotypes. In rows, each module is labeled as ME_followed by the module ID number.

Supplementary Figure 2 | Pearson correlations between peptide module eigenvector summaries and mouse phenotypes. In rows, each module is labeled as ME_ followed by the module ID number.

Supplementary Figure 3 | Pearson correlations between peptide module eigenvector summaries and transcript module summaries. Each module is labeled as ME_ followed by the module ID number.

Supplementary Figure 4 | Points are composed of peptide-transcript pairs, where each entity maps to the same Entrez ID. The x and y axis are correlations with the overall pathology phenotype, while the size and hue indicates correlation across the data types. Blue indicates a positive correlation while red shows a negative correlation.

Transcript_Module_Contents. A large table with columns: Probe ID (the transcript probe), Module (the assigned module), MM1 to MM6 (correlations with module eigenvectors), p.MM1 to p.MM6 (*p*-values for each correlation), and then correlations with phenotype variables, and *p*-values for phenotype correlations.

Peptide_Module_Contents. A large table with columns: Peptide ID, Module (the assigned module), MM1 to MM14 (correlations with module eigenvectors), p.MM1 to p.MM14 (*p*-values for each correlation), and then correlations with phenotype variables, and *p*-values for phenotype correlations.

Supplementary_Network_Results. A brief account of peptide and transcript networks in terms of module sizes and module-significance permutation tests. Also shown are module overlap sizes with permutation test results, and the pairwise module prioritization results for each phenotype.

REFERENCES

Aderem, A., Adkinsb, J. N., Ansongb, C., Galaganc, J., Kaisera, S., Korthf, M. J., et al. (2011). A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *MBio* 2, e00325–10. doi: 10.1128/mBio.00325-10

Adourian, A., Jennings, E., Balasubramanian, R., Hines, W. M., Damian, D., Plasterer, T. N., et al. (2008). Correlation network analysis for data integration and biomarker selection. *Mol. Biosyst.* 4, 249. doi: 10.1039/b708489g

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32, 115D–119D. doi: 10.1093/nar/gkh131

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Gene Ontol. Consortium. Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1214/aos/1013699998

Berger, J. A., Hautaniemi, S., Mitra, S. K., and Astola, J. (2006). Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 3, 2–16. doi: 10.1109/TCBB.2006.10

Bhavsar, A. P., Auweter, S. D., and Finlay, B. B. (2010). Proteomics as a probe of microbial pathogenesis and its molecular boundaries. *Future Microbiol.* 5, 253–265. doi: 10.2217/fmb.09.114

Brenner, D. E., and Normolle, D. P. (2007). Biomarkers for cancer risk, early detection, and prognosis: the validation conundrum. *Cancer Epidemiol. Biomarkers Prev.* 16, 1918–1920. doi: 10.1158/1055-9965.EPI-07-2619

Cai, Q. C., Jiang, Q. W., Zhao, G. M., Guo, Q., Cao, G. W., and Chen, T. (2003). Putative caveolin-binding sites in SARS-CoV proteins. *Acta Pharmacol. Sin.* 24, 1051–1059.

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Cinatl, J. Jr., Hoever, G., Morgenstern, B., Preiser, W., Vogel, J. U., Hofmann, W. K., et al. (2004). Infection of cultured intestinal epithelial cells with severe acute respiratory syndrome coronavirus. *Cell. Mol. Life Sci.* 61, 2100–2112. doi: 10.1007/s00018-004-4222-9

Cox, B., Kislinger, T., and Emili, A. (2005). Integrating gene and protein expression data: pattern analysis and profile mining. *Methods* 35, 303–314. doi: 10.1016/j.ymeth.2004.08.021

Daemen, A., Gevaert, O., De Bie, T., Debucquoy, A., Machiels, J. P., De Moor, B., et al. (2008). Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *Pac. Symp. Biocomput.* 166–177.

Davey, R. T., Lynfield, R., Dwyer, D. E., Losso, M. H., Cozzi-Lepri, A., Wentworth, D., et al. (2013). The association between serum biomarkers and disease outcome in influenza A(H1N1)pdm09 virus infection: results of two international observational cohort studies. *PLoS ONE* 8:e57121. doi: 10.1371/journal.pone.0057121

de Lang, A., Baas, T., Teal, T., Leijten, L. M., Rain, B., Osterhaus, A. D., et al. (2007). Functional genomics highlights differential induction of antiviral pathways in the lungs of SARS-CoV–infected macaques. *PLoS Pathog.* 3:e112. doi: 10.1371/journal.ppat.0030112

Fagan, A., Culhane, A. C., and Higgins, D. G. (2007). A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics* 7, 2162–2171. doi: 10.1002/pmic.200600898

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* 23, 257–258. doi: 10.1093/bioinformatics/btl567

Feng, Z., Prentice, R., and Srivastava, S. (2004). Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. *Pharmacogenomics* 5, 709–719. doi: 10.1517/14622416.5.6.709

Forst, C. V. (2006). Host-pathogen systems biology. *Drug Discov. Today* 11, 220–227. doi: 10.1016/S1359-6446(05)03735-9

Gat-Viks, I., Tanay, A., Raijman, D., and Shamir, R. (2006). A probabilistic methodology for integrating knowledge and experiments on biological networks. *J. Comput. Biol.* 13, 165–181. doi: 10.1089/cmb.2006.13.165

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., et al. (2011). Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* 7:e1001393. doi: 10.1371/journal.pgen.1001393

Gibbs, D. L., Baratt, A., Baric, R. S., Kawaoka, Y., Smith, R. D., Orwoll, E. S., et al. (2013). Protein co-expression network analysis (ProCoNA). *J. Clin. Bioinforma* 3, 11. doi: 10.1186/2043-9113-3-11

Gralinski, L. E., Bankhead, A. 3rd., Jeng, S., Menachery, V. D., Proll, S., Belisle, S. E., et al. (2013). Mechanisms of severe acute respiratory syndrome coronavirus induced acute lung injury. *MBio* 4, e00271–13. doi: 10.1128/mBio.00271-13

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* 437–449.

Hughes, V. (2009). Markers of dispute. *Nat. Med.* 15, 1339–1343. doi: 10.1038/nm1209-1339

Iancu, O. D., Darakjian, P., Malmanger, B., Walter, N. A., McWeeney, S., and Hitzemann, R. (2012a). Gene networks and haloperidol-induced catalepsy. *Genes Brain Behav.* 11, 29–37. doi: 10.1111/j.1601-183X.2011.00736.x

Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012b). Utilizing RNA-Seq data for de novo co-expression network inference. *Bioinformatics* 28, 1592–1597. doi: 10.1093/bioinformatics/bts245

Kim, D., Shin, H., Song, Y. S., and Kim, J. H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Inform.* 45, 1191–1198. doi: 10.1016/j.jbi.2012.07.008

Kint, G., Fierro, C., Marchal, K., Vanderleyden, J., and De Keersmaecker, S. C. (2010). Integration of "omics" data: does it lead to new insights into host-microbe interactions? *Future Microbiol.* 5, 313–328. doi: 10.2217/fmb.10.1

Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput.* 300–311.

Langfelder, P., Castellani, L. W., Zhou, Z., Paul, E., Davis, R., Schadt, E. E., et al. (2012). A systems genetic analysis of high density lipoprotein metabolism and network preservation across mouse models. *Biochim. Biophys. Acta* 1821, 435–447. doi: 10.1016/j.bbalip.2011.07.014

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Langfelder, P., and Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* 46, 1–17.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Comput. Biol.* 7:e1001057. doi: 10.1371/journal.pcbi.1001057

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Lê Cao, K. A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* 7. doi: 10.2202/1544-6115.1390

Leung-Toung, R., Zhao, Y., Li, W., Tam, T. F., Karimian, K., and Spino, M. (2006). Thiol proteases: inhibitors and potential therapeutic targets. *Curr. Med. Chem.* 13, 547–581. doi: 10.2174/092986706776055733

Lopez-romero, P. (2010). Agi4x44Preprocess. Available online at: http://www.bioconductor.org/packages/2.12/bioc/html/Agi4x44PreProcess.html

Loureiro, J., and Ploegh, H. L. (2006). Antigen presentation and the ubiquitin−proteasome system in host–pathogen interactions. *Adv. Immunol.* 92, 225–305. doi: 10.1016/S0065-2776(06)92006-9

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011, 1–13. doi: 10.1093/database/bar009

Mason, M. J., Fan, G., Plath, K., Zhou, Q., and Horvath, S. (2009). Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics* 10:327. doi: 10.1186/1471-2164-10-327

Monroe, M. E., Tolić, N., Jaitly, N., Shaw, J. L., Adkins, J. N., and Smith, R. D. (2007). VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* 23, 2021–2023. doi: 10.1093/bioinformatics/btm281

Ng, L. F., Hibberd, M. L., Ooi, E. E., Tang, K. F., Neo, S. Y., Tan, J., et al. (2004a). A human *in vitro* model system for investigating genome-wide host responses to SARS coronavirus infection. *BMC Infect. Dis.* 4:34. doi:10.1186/1471-2334-4-34

Ng, M. L., Lee, J. W., Leong, M. L., Ling, A. E., Tan, H. C., and Ooi, E. E. (2004b). Topographic changes in SARS coronavirus-infected cells at late stages of infection. *Emerg. Infect. Dis*. 10, 1907–1914. doi: 10.3201/eid1011.040195

Nie, L., Wu, G., and Zhang, W. (2006). Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: a multiple regression to identify sources of variations. *Biochem. Biophys. Res. Commun*. 339, 603–610. doi: 10.1016/j.bbrc.2005.11.055

Ntzani, E. E., and Ioannidis, J. P. A. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 362, 1439–1444. doi: 10.1016/S0140-6736(03)14686-7

Reghunathan, R., Jayapal, M., Hsu, L. Y., Chng, H. H., Tai, D., Leung, B. P., et al. (2005). Expression profile of immune response genes in patients with Severe Acute Respiratory Syndrome. *BMC Immunol.* 6:2. doi: 10.1186/1471-2172-6-2

Roberts, A., Deming, D., Paddock, C. D., Cheng, A., Yount, B., Vogel, L., et al. (2007). A mouse-adapted SARS-coronavirus causes disease and mortality in BALB/c mice. *PLoS Pathog.* 3:35. doi:10.1371/journal.ppat.0030005

Serang, O., MacCoss, M. J., and Noble, W. S. (2010). Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* 9, 5346–5357. doi: 10.1021/pr100594k

Serang, O., and Noble, W.S. (2012). Faster mass spectrometry-based protein inference: junction trees are more efficient than sampling and marginalization by enumeration. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 809–817. doi: 10.1109/TCBB.2012.26

Sturdevant, D. E., Virtaneva, K., Martens, C., Bozinov, D., Ogundare, O., Castro, N., et al. (2010). Host–microbe interaction systems biology: lifecycle transcriptomics and comparative genomics. *Future Microbiol.* 5, 205–219. doi: 10.2217/fmb.09.125

Sung, J., Wang, Y., Chandrasekaran, S., Witten, D. M., and Price, N. D. (2012). Molecular signatures from omics data: from chaos to consensus. *Biotechnol. J.* 7, 946–957. doi: 10.1002/biot.201100305

Surjit, M., Liu, B., Jameel, S., Chow, V. T., and Lal, S. K. (2004). The SARS coronavirus nucleocapsid protein induces actin reorganization and apoptosis in COS-1 cells in the absence of growth factors. *Biochem. J.* 383, 13–18. doi: 10.1042/BJ20040984

Tan, C. S., Salim, A., Ploner, A., Lehtiö, J., Chia, K. S., and Pawitan, Y. (2009). Correlating gene and protein expression data using Correlated Factor Analysis. *BMC Bioinformatics* 10:272. doi: 10.1186/1471-2105-10-272

Torres-García, W., Zhang, W., Runger, G. C., Johnson, R. H., and Meldrum, D. R. (2009). Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics* 25, 1905–1914. doi: 10.1093/bioinformatics/btp325

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. U.S.A.* 100, 8348–8353. doi: 10.1073/pnas.0832373100

Vaske, C. J., House, C., Luu, T., Frank, B., Yeang, C. H., Lee, N. H., et al. (2009). A Factor graph nested effects model to identify networks from genetic perturbations. *PLoS Comput. Biol.* 5:e1000274. doi: 10.1371/journal.pcbi.1000274

Waters, K. M., Liu, T., Quesenberry, R. D., Willse, A. R., Bandyopadhyay, S., Kathmann, L. E., et al. (2012). Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data. *PLoS ONE* 7:e34515. doi: 10.1371/journal.pone.0034515

Waters, K. M., Pounds, J. G., and Thrall, B. D. (2006a). Data merging for integrated microarray and proteomic analysis. *Brief. Funct. Genomic. Proteomic.* 5, 261–272. doi: 10.1093/bfgp/ell019

Waters, K. M., Singhal, M., Webb-Robertson, B. J., Stephan, E. S., and Gephart, J. M. (2006b). Breaking the high-throughput bottleneck: new tools help biologists integrate complex datasets. *Sci. Comput.* 23, 22–26.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., et al. (2006). The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, 34, D187–D191. doi: 10.1093/nar/gkj161

Yip, A. M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 8:22. doi:10.1186/1471-2105-8-22

Zak, D. E., and Aderem, A. (2009). Systems biology of innate immunity. *Immunol. Rev.* 227, 264–282. doi: 10.1111/j.1600-065X.2008.00721.x

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4. doi: 10.2202/1544-6115.1128

Zhang, X., Lu, X., Shi, Q., Xu, X. Q., Leung, H. C., Harris, L. N., et al. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 7:197. doi:10.1186/1471-2105-7-197

Zimmer, J. S., Monroe, M. E., Qian, W. J., and Smith, R. D. (2006). Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.* 25, 450–448. doi: 10.1002/mas.20071

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 September 2013; paper pending published: 21 October 2013; accepted: 19 December 2013; published online: 07 January 2014.

Citation: Gibbs DL, Gralinski L, Baric RS and McWeeney SK (2014) Multi-omic network signatures of disease. Front. Genet. ***4****:309. doi: 10.3389/fgene.2013.00309*

This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics.