Published online 25 July 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 18* ***e138***
doi: 10.1093/nar/gku678

Finding pathway-modulating genes from a novel Ontology Fingerprint-derived gene network

Tingting Qin[1], Nabil Matmati[2], Lam C. Tsoi[3], Bidyut K. Mohanty[4], Nan Gao[5], Jijun Tang[5,6], Andrew B. Lawson[7], Yusuf A. Hannun[2] and W. Jim Zheng[8,*]

[1]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA, [2]The Stony Brook University Cancer Center and the Department of Medicine, Stony Brook, NY 11794, USA, [3]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA, [4]Department of Biochemistry & Molecular Biology, Medical University of South Carolina, Charleston, SC 29425, USA, [5]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA, [6]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China, [7]Department of Public Health Science, Medical University of South Carolina, Charleston, SC 29425, USA and [8]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received July 12, 2013; Revised July 10, 2014; Accepted July 14, 2014

ABSTRACT

To enhance our knowledge regarding biological pathway regulation, we took an integrated approach, using the biomedical literature, ontologies, network analyses and experimental investigation to infer novel genes that could modulate biological pathways. We first constructed a novel gene network via a pairwise comparison of all yeast genes' Ontology Fingerprints—a set of Gene Ontology terms overrepresented in the PubMed abstracts linked to a gene along with those terms' corresponding enrichment *P*-values. The network was further refined using a Bayesian hierarchical model to identify novel genes that could potentially influence the pathway activities. We applied this method to the sphingolipid pathway in yeast and found that many top-ranked genes indeed displayed altered sphingolipid pathway functions, initially measured by their sensitivity to myriocin, an inhibitor of *de novo* sphingolipid biosynthesis. Further experiments confirmed the modulation of the sphingolipid pathway by one of these genes, *PFA4*, encoding a palmitoyl transferase. Comparative analysis showed that few of these novel genes could be discovered by other existing methods. Our novel gene network provides a unique and comprehensive resource to study pathway modulations and systems biology in general.

INTRODUCTION

Biological pathways are the *de facto* functional unit for studying biology at the systems level. Discovering new players and mechanisms of pathway modulation may not only allow us to understand the mechanisms of pathway regulation but may also provide novel targets for therapeutics. However, such a task faces many challenges: pathways are complex and consist of many components—some pathways may consist of hundreds of genes, the regulation of signaling or metabolic pathways is dynamic and involves sophisticated mechanisms, such as feed-forward and feedback loops, and genes that influence a pathway activity may not have obvious links to that pathway. These challenges make finding novel players of a pathway very difficult and are particularly acute in our efforts to understand the sphingolipid pathway. Bioactive sphingolipids play many roles in regulating biological functions, from cell-cycle progression to cancer pathogenesis. Whereas biochemical reactions by the enzymes involved in this pathway are well understood, our knowledge of its regulation is modest. Studies employing computational approaches and mathematical modeling have helped us to understand the dynamics of the pathway (1), but little is known about other genes that can influence the sphingolipid pathway activity.

Biological networks, such as protein–protein interaction (PPI) networks (2,3), genetic interaction (GI) networks (4) and co-expression networks (5,6), have been used to infer new genes that may influence pathway activities, as genes are connected in the network regardless of the pathway boundaries. The limitation of this approach is that each type of network focuses on only one specific aspect of biology, e.g. a PPI network helps us understand the disease only from the aspect of PPIs, even though many proteins that play critical

*To whom correspondence should be addressed. Tel: +1 713 500 3641; Fax: +1 713 500 3907; Email: wenjin.j.zheng@uth.tmc.edu.

roles in human disease do not directly interact with each other. To overcome this hurdle, methods have been developed to combine these networks into functional networks, such as YeastNet (7), but their performance is affected by the quality of individual data sets (e.g. PPIs have a high false-positive rate) (8).

As an alternative, recent work inferred comprehensive networks from the biomedical literature in which genes, diseases and the relationships among them were comprehensively characterized by experimental and quantitative analyses throughout the history of biomedical research. The majority of these approaches, such as the work by Jenssen *et al.* (9)), infer gene–gene relationships by identifying the co-occurrence of two genes within the biomedical literature, with extensions, such as Génie (10), in which comparative genomics approaches were employed. However, this approach is limited by the fact that 70% of PubMed abstracts only contain information about a single gene (based on a gene2pubmed file from the National Center for Biotechnology Information (NCBI). Two alternative approaches are the rule-based, or knowledge-based, approach (11) and the statistical, or machine-learning-based, approach (12). These two approaches use text-mining algorithms and natural language processing methods to extract useful information from the biomedical literature. Both methods rely on high-quality corpuses that are difficult to compile, and the error rate can be close to 35% (13). To overcome this hurdle, ontology-based methods have been developed to infer gene–gene relationships by calculating semantic similarities (14–16). However, these methods use manually curated Gene Ontology (GO) annotation of genes. Although the annotation is of high quality, it is limited and cannot take advantage of the yet-unannotated biomedical literature linked to genes. However, text-mining efforts to recognized concepts are making significant progress (17–19), but the contribution of these methods to gene annotation is still not ready for prime time compared to the manual annotation. In addition, the calculation of semantic similarity relies on the structure of GO. Whereas some aspects of GO are well developed and capture the most intricate details of biology, other aspects are still under development, leaving the biological details granular and coarse.

Here, we report a novel approach to identify new players of important biological pathways based on the concept of the Ontology Fingerprint, which is the set of GO terms overrepresented in PubMed abstracts linked to a gene or disease, along with those terms' corresponding enrichment *P*-values (20). In this work, we derived a novel global yeast gene network based on Ontology Fingerprints and used the network to discover novel genes that influence sphingolipid pathway activities *in vivo*. The gene network can be applied to discover new players of biological pathways and to study other cellular functions at the systems level.

MATERIALS AND METHODS

Data sources and pre-processing

PubMed abstracts were downloaded from NCBI. GO terms and their descriptions were obtained from the GO Consortium. The links between PubMed abstracts and genes were obtained from the GO 'gene_association.goa' file (downloaded on December 20, 2009) and the NCBI 'pubmed2gene' file (December 30, 2009). Abstracts that contain GO terms and their synonyms were identified by exact string match. The mapping quality was assessed by manually evaluating a set of randomly selected abstracts that contain mapped terms. The identified PubMed-Go term pairs were then combined with those extracted from GO 'gene_association.goa' file. The abstracts containing a GO term were also labeled with its parent terms retrieved from GO hierarchy from the Web Ontology Language (OWL) file. In addition, each abstract was labeled with a mapped GO term only once regardless of how many times the term occurred. Other details can be found in the Supplementary Materials (Supplementary Table S1). Yeast genes involved in the sphingolipid metabolic pathway were manually reviewed by a group of experts in this area and deemed as known sphingolipid genes for our network analysis.

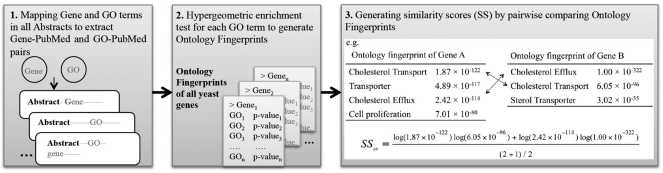

Figure 1. Using Ontology Fingerprint to quantify the biological relevance between two genes. (1) Identifying gene-PubMed and GO-PubMed relationships in all PubMed abstracts that were annotated with yeast genes; (2) the Ontology Fingerprint of each yeast gene was generated by a hypergeometric test of the enrichment of all the ontology terms in all the PubMed abstracts associated with the gene; (3) the biological relevance of Gene A and Gene B was quantified by calculating a gene–gene similarity score by comparing the two genes' Ontology Fingerprints (Equation (1)). Genome-wide pairwise comparison yielded the initial gene network.

Construct a novel yeast gene network from Ontology Fingerprints

To develop the Ontology Fingerprint derived gene network, we first employed a hypergeometric test on each pair of yeast gene and GO term to construct the Ontology Fingerprint for the genes (Figure 1, steps 1 and 2) (21). We then performed pairwise comparison of the resulting Ontology Fingerprints of genes for the whole genome to develop the network. The comparison of two genes' fingerprints generated a similarity score to indicate the extent to which the genes were biologically relevant (Figure 1, step 3). The score was calculated by Equation (1), which uses a modified version of the inner product:

$$S_{ij} = \frac{\sum_{o=1}^{O} \log(r_{io}) \log(r_{jo}) I(r_{io} < \lambda < r_{jo} < \lambda)}{\max\left\{1, \frac{1}{2}\sum_{o=1}^{O}\left[I(r_{io} < \lambda) I(r_{jo} \geq \lambda) + I(r_{io} \geq \lambda) I(r_{jo} < \lambda)\right]\right\}} \quad (1)$$

The similarity score considered the enrichment level r (i.e. *P*-value from the hypergeometric test) of ontology term o among the PubMed abstracts annotated with gene i or gene j. The denominator of the similarity score emphasized the number of overlapping ontology terms between two Ontology Fingerprints compared. Where there was a large number of non-overlapping ontology terms, the score

decreased. We also eliminated ontology terms with insignificant enrichment *P*-values that were higher than λ when calculating the similarity score of Ontology Fingerprints (see the following section for how λ was estimated). After the genome-wide pairwise comparison of Ontology Fingerprints of genes, we obtained a novel gene network in the form of a weighted, undirected graph where vertices are genes and edges are similarity scores. The network was further optimized as described in Supplementary Figure S1.

In order to generate a biologically meaningful network, we developed a Bayesian hierarchical model to estimate the similarity score threshold that indicates whether two genes are biologically relevant (Supplementary Figures S2, S3, S4 and Table S2). The threshold was utilized to separate biologically relevant gene pairs from irrelevant gene pairs, and applied to the optimized network by leaving out edges with the similarity scores lower than the threshold (Supplementary Figure S3). The properties of the resulting network were analyzed extensively (Supplemental Figures S5 and S6). We further evaluated the ability of the network to capture known biological relationships, such as PPIs or genetic interactions, and compared the performance to that of the STRING database (22) using the shortest distance (Dijkstra algorithm) (23) and shortest link measures (Supplementary Table S3).

Identify novel genes that modulate the yeast sphingolipid pathway by ranking on demand

Our sphingolipid experts handpicked 30 yeast sphingolipid genes based on current knowledge (Supplementary Table S4). We then identified 855 first neighbors of these 30 genes from the network. Out of these first neighbors, 453 genes had linked PubMed abstracts or gene descriptions that contained the 'sphingo' prefix or 'ceramide'. These genes were considered not to have novel relevance to sphingolipid pathway and were excluded from further analysis. The remaining 402 genes (candidate genes) were analyzed in the developed gene network to identify the most likely candidates to modulate the yeast sphingolipid pathways.

Given a weighted gene network, we ranked genes based on their similarity to a set of seed genes from the same functional class, i.e. ranking on demand (24). In this process, the 402 candidate genes identified above were ranked in decreasing order by their overall connectivity to the 30 sphingolipid genes as measured by a Total Score (*TotalS*), i.e. the sum of gene–gene similarity scores between a candidate sphingolipid gene and all 30 known sphingolipid genes (Equation (2)):

$$TotalS_i = \sum_{j=1}^{30} SS_{ij} \quad (2)$$

where i is the ith candidate gene, j is the jth sphingolipid genes and SS_{ij} is the similarity score between the ith candidate gene and the jth sphingolipid gene. $TotalS_i$ is the Total Score of candidate gene i. This Total Score captures both the candidate gene's connectivity to the known sphingolipid genes and the similarity of Ontology Fingerprints between the candidate gene and the sphingolipid genes. We hypothesized that the higher the gene's ranking, the more likely the gene's involvement in the sphingolipid pathway. This hypothesis was tested by several measures. We performed the leave-one-out cross-validation to evaluate whether the network can identify genes belonging to the known pathways (Supplementary Figure S7). We also compared the ability of the network to prioritize candidate genes to other similar programs (Supplementary Figure S8). Furthermore, we experimentally tested by examining whether the proportion of genes that influenced the sphingolipid pathway out of the top-ranking genes was greater than that out of the bottom-ranking genes, as described in the following section.

Experimental validation of novel yeast sphingolipid genes

Strains and growth conditions. All strains used in this study are listed in Supplementary Table S5. Cells were grown on standard Difco YPD solid or liquid medium. Myriocin was purchased from Sigma (Sigma-Aldrich, St. Louis, MO, USA. Catalog #M1177). We used two concentrations of myriocin (750 and 1000 ng/ml) in liquid and solid medium. Phytosphingosine and dihydrosphingosine were purchased from Avanti Polar Lipids (catalog #860499P and #860498P, respectively) and used in liquid medium at 0.3 μM each. Myriocin, phytosphingosine and dihydrosphingosine were dissolved in methanol.

Testing the sensitivity or resistance to the de novo sphingolipid synthesis inhibitor myriocin. We selected top 30 candidate genes as the test gene set (Supplementary Table S6). As a control, we also selected 15 random samples from the 25% lowest-ranked candidate genes and 15 random samples among background genes (genes with no connection to any of the 30 sphingolipid genes in the network) (Supplementary Table S7). Deletion strains of these 60 genes were obtained from a BY4741 gene-deletion library (Invitrogen) (Supplementary Table S5). Followed by spot test (25), myriocin was added to medium at 0, 750 or 1000 ng/ml. As controls, the mutant *LCB2* (*Lcb2-DAmP*, purchased from Thermo Scientific, Open Biosystems Cat# YSC5094–99851666) and wild type (WT) BY4741 were used.

Rescue of pfa4Δ strain by phytosphingosine and dihydrosphingosine in the medium. The growth of WT and *pfa4*Δ strains in YPD liquid medium was measured at 30°C at 18, 24 and 48 h under six different treatment conditions: methanol (solution control), myriocin (750 ng/ml), phytosphingosine (0.3 μM), dihydrosphingosine (0.3 μM), myriocin (750 ng/ml) + phytosphingosine (0.3 μM) and myriocin (750 ng/ml) + dihydrosphingosine (0.3 μM).

Method comparison

We evaluated four other comparable networks and methods for their ability to predict the experimentally validated novel genes that influenced the sphingolipid pathway activity *in vivo*. PPI networks (26) and GI networks (27) are the two most commonly used biological networks and have been used to prioritize genes for diseases and pathways. Another yeast network that can serve this purpose is the yeast probabilistic functional gene network developed by Lee *et al.* (7,28). In addition, genes can be prioritized by

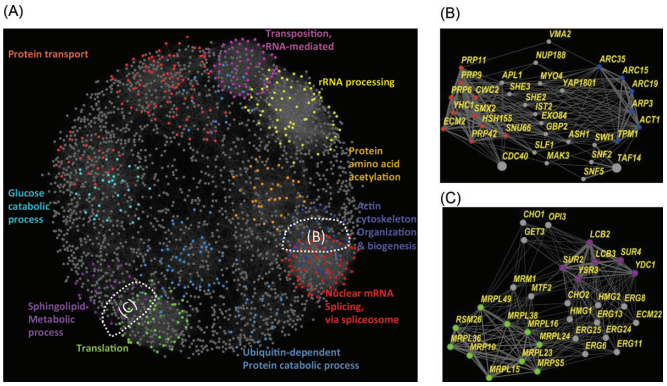

Figure 2. Network structure and its functional relevance in the novel yeast gene network. (A) The largest component of the yeast gene network with edges weighted by top 5th percentile of similarity scores. Ten clusters identified by the MCODE algorithm were highlighted by distinct colors and annotated by the most enriched GO terms, indicating the correlation between subnetwork structure and biological functions. (B) The partly zoomed-in interactions between clusters enriched for 'nuclear mRNA splicing, via spliceosome' (in red) and 'actin cytoskeleton organization and biogenesis' (in blue) among the yeast global gene network; *CDC40* and *TAF14* were annotated by large nodes. (C) The partly zoomed-in interactions between clusters enriched for 'sphingolipid metabolic processes' (in purple) and 'translation' (in green) among the yeast global gene network.

Génie, a literature-based gene prioritization approach (10). We used the 30 known sphingolipid genes as input to evaluate whether these networks or methods can identify the genes discovered by Ontology Fingerprint approach as significantly related to the sphingolipid pathway. For the PPI, GI and yeast functional networks, we used the same algorithm to detect relevant genes as we did for Ontology Fingerprint derived networks. For Génie, we used their optimized algorithms to identify genes relevant to the known sphingolipid genes. Detailed methods can be found in Supplementary Figures S9 and S10.

RESULTS

Developing a novel global gene network in yeast

By performing pairwise comparisons of the Ontology Fingerprints of 5446 yeast genes, we identified 7 586 754 gene–gene connections weighted by the resulting similarity scores. These genes and their connections constituted a novel global gene network for yeast (Figure 2).

Unlike networks that depend on a single biological feature, such as PPI, our network summarized all the biological aspects captured in GO and the literature. As a result, our network is dense and captures a wide range of biological relationships. The extent of such relationships is indicated by the weighted edges (similarity scores), which allowed us to trim the dense network by omitting insignificant connections. These insignificant connections were evaluated by a biologically meaningful threshold of the similarity score estimated by a Bayesian hierarchical model (Supplementary Figure S2 and Table S2). Any connection between two genes with a score above this threshold was comparable to two genes belonging to the same known biological pathway with a connection in the pathway. The resulting network consisted of 5445 genes and 528 581 edges, and was capable of capturing known biological relations such as the protein–protein and genetic interactions as compared to widely used STRING database (Supplementary Table S3).

Evaluating the network properties

The Ontology Fingerprint-derived gene network exhibited the robustness properties of typical biological networks: the distribution of the node degree followed the power law, indicating a scale-free network; the clustering coefficient plot also demonstrated the modularity of the network, where the average clustering coefficient was negatively correlated with the node degree (Supplementary Figure S5).

We further explored the network modularity and found that the structure of the network had strong connections with biological functions. We selected the largest network component, which contained 2511 genes and 24 156 edges weighted by the top 5th percentile of similarity scores, and identified 10 major functional clusters that were annotated by distinct colors (Figure 2A). We found that clusters close to each other in the network were functionally related. For example, the cluster enriched for 'nuclear mRNA splicing, via spliceosome' and the cluster enriched for 'actin cytoskeleton organization and biogenesis' were close to each other in the network (Figure 2A). This finding is supported by the work of Dahan *et al.*, which showed that *CDC40* controls cell-cycle progression through the splicing of the *TAF14* gene. *TAF14* mutants exhibit defects in actin cytoskeletal organization and have several morphological aberrations, which correlate with cytoskeletal imperfection (29). Some of the connections between these two clusters within the global gene network are illustrated in detail in Figure 2B. These connections were mediated through a group of genes (in gray), including *CDC40* and *TAF14* (shown as large nodes).

As another example, the cluster enriched for 'sphingolipid metabolic processes' was close to the cluster enriched for 'translation function' in the network (Figure 2A), and some of their connections within the global gene network are shown in Figure 2C. The functional relevance of these two clusters is supported by the fact that sphingolipid bases are required for translation initialization during heat stress in yeast (30) and by the role of sphingolipids in regulating the formation of P Bodies (31). Our analyses indicate that the Ontology Fingerprint-derived novel gene network can be used not only for inferring functional relevance between genes but also for examining interconnections among different modules—a useful approach to investigating cellular functions at a systems level.

Identify pathway-modulating genes from the Ontology Fingerprint-derived gene network

The structure–function relationship of our network allows us to use the subnetwork structure to discover novel pathway-modulating genes. We tested this idea by performing leave-one-out cross-validation using 107 KEGG pathways as the standard. The comparison shows that our method can successfully assign genes to their corresponding pathways with an average AUC of 0.92 (Supplementary Figure S7). We also compared our method with the

Table 1. Experimentally validated sphingolipid pathway-modulating genes identified by different methods. Our Ontology Fingerprint-derived gene network uniquely identified 14 novel myriocin-sensitive or resistant genes, of which only a small portion could be discovered by other methods. See Supplementary Figures S9 and S10 for details

	Methods				
	Ontology Fingerprint-derived Network	Genie	YeastNet	PPI	GI
Number of novel sphingolipid pathway-modulating genes	14	2	1	1	3
Percentage of novel sphingolipid pathway-modulating genes (%)	100	14.29	7.14	7.14	21.43

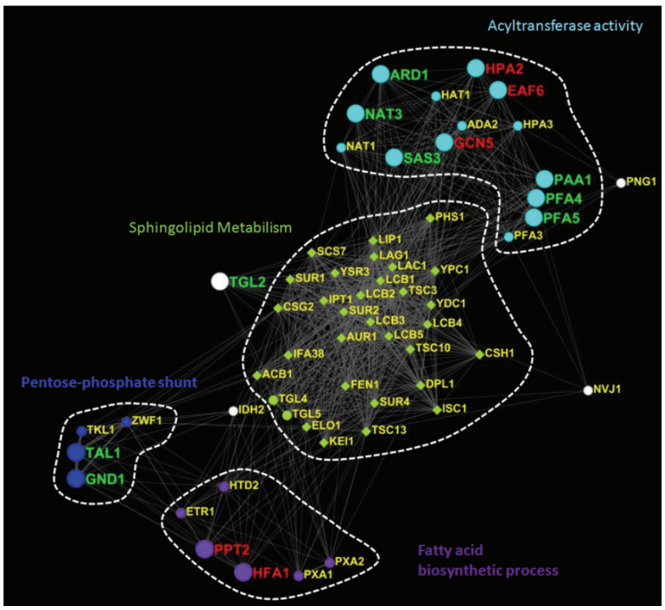

Figure 3. Analysis of candidate genes influencing the yeast sphingolipid pathway. In the subnetwork of 30 known yeast sphingolipid genes (diamonds) and the top 30 candidate genes (circles), four functional clusters circled by dashed lines were identified and labeled with the most enriched GO terms. The larger nodes denote genes that were identified to be sensitive (in green) or resistant to myriocin (in red). The nodes outside of the dashed lines were the genes that fail to form clusters by MCODE.

other state of the art approaches in gene prioritization, and found that our method performed better (Figure S8). We used this approach to identify a subnetwork containing the 30 known yeast sphingolipid genes (Supplementary Table S4) and the top 30 non-sphingolipid genes (candidate genes) based on their total connectivity to known sphingolipid genes in the network and the fact that these genes had no known published connection to the pathway (Supplementary Table S6). The resulting subnetwork (Figure 3) was further investigated by clustering analysis. We evaluated different network clustering algorithms in identifying biologically meaningful clusters and found that the MCODE algorithm performed well (Supplementary Figure S6). We then used the MCODE algorithm to identify four major functional modules represented by their most enriched GO terms: 'sphingolipid metabolism', 'acyltransferase activity', 'fatty acid biosynthetic process', and 'pentose-phosphate shunt' (Figure 3, Supplementary Figure S6).

As shown in Figure 3, the 30 known sphingolipid genes in a diamond shape were clustered together in a single module enriched with 'sphingolipid metabolism'. More than half of the candidate genes (16 out of 30) were clustered in the module enriched with 'acyltransferase activity', implying that this function might be highly relevant to the yeast sphingolipid pathway. The other two clusters, with five and six candidate genes, respectively, also provided insight into the cellular functions relevant to the yeast sphingolipid pathway. These functional possibilities were revealed from our network analysis despite the lack of publications linking these genes to the sphingolipid pathway—a finding that represents a significant advantage of this approach over methods based on the co-occurrence of genes in the literature.

To understand how the identified genes could modulate sphingolipid pathway, we analyzed the ontology terms contributed to the sphingolipid pathway–candidate gene connections. Because our method built upon the literature and ontology, we were able to trace back to see what ontology terms contributes to the connections between the known sphingolipid genes and the identified candidate genes. Supplementary Figure S11 shows two heat maps that illustrate GO contribution to the association between 30 known sphingolipid genes and top 30 candidate genes. The Ontology Fingerprint of PFA4 was generated from 22 publications linked to the gene and consists of 42 GO terms, compared to only 4 PubMed abstracts and 10 GO terms from the GO database. Among all the GO terms in the fingerprint, 'protein palmitoylation' contributes the most to the connection between PFA4 and sphingolipid genes. While 'protein palmitoylation' is a function of PFA4, many sphingolipid genes also connected to this function in several different ways. For example, Palmitoyl-CoA is the starting material of the *de novo* sphingolipid biosynthesis, and some of the enzymes involved in sphingolipid pathway may also be palmitoylated. While this piece of information may provide some hint of how PFA4 might influence the activity of sphingolipid pathway, there is no explicit connection between PFA4 and sphingolipid genes in any publication, demonstrating the unique ability of the Ontology Fingerprint method in inferring implicit relationships.

Experimental assessment of sphingolipid pathway modulation by predicted novel genes

The novel connections we identified were not found in published results, indicating their potential novel relevance to the sphingolipid pathway. Therefore, we evaluated whether

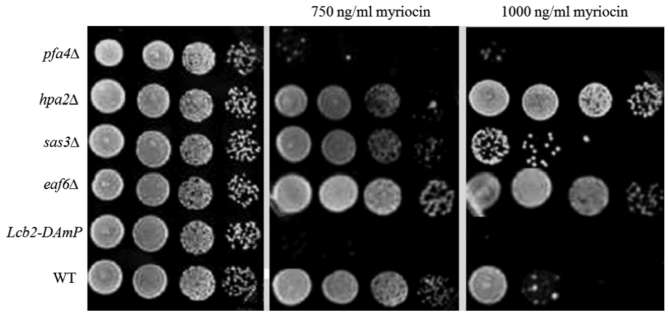

Figure 4. Sphingolipid pathway activity altered by novel genes predicted from the Ontology Fingerprint-derived gene network. Spot test of *pfa4*Δ, *hpa2*Δ, *sas3*Δ, *eaf6*Δ, *Lcb2-DAmP* (positive control that was sensitive to myriocin) and WT strains with myriocin treatment at 0, 750 or 1000 ng/ml. Compared with WT, *pfa4*Δ and *sas3*Δ showed elevated sensitivity to myriocin, whereas *hpa2*Δ and *eaf6*Δ were resistant, all indicating altered sphingolipid pathway activity *in vivo*.

these genes influenced the sphingolipid pathway *in vivo* using myriocin. Myriocin specifically inhibits serine palmitoyltransferase, which catalyzes the first step in *de novo* sphingosine biosynthesis. Whereas normal cells can survive lower concentrations of myriocin, changes in myriocin sensitivity indicate alterations in sphingolipid pathway activity (32).

We investigated the sensitivity of the deletion strains of the 29 non-essential candidate genes (partly shown in Figure 4, with the 30th gene being essential) and found that 14 of them exhibited altered myriocin sensitivity (Figure 3, Supplementary Tables S8 and S9). In contrast, myriocin sensitivity changed in only 4 out of 29 deletion strains of those genes with the least connectivity to known sphingolipid genes (control strains, Supplementary Table S8). Because the sphingolipid pathway is complex and not all pathway changes can be captured by altered myriocin sensitivity, we cannot draw any conclusion about those genes whose deletion strains had sensitivities similar to the WT. Nevertheless, our network analysis effectively identified many genes that influenced the sphingolipid pathway, as measured by myriocin sensitivity (one-tailed Fisher's exact test comparing the proportion of altered sensitivity in the most versus the least connected genes resulted in a P-value = 0.005, indicating a significant difference at the level of 0.01).

We focused our further experimental validation on *PFA4*—the deletion strain that showed the most sensitivity to myriocin. This elevated sensitivity indicates that deleting the *PFA4* gene may aggravate the deficiency of biosynthesis of yeast sphingolipids. To test this hypothesis, we examined the ability of two key sphingolipids, phytosphingosine and dihydrosphingosine, to rescue the growth defect of the *PFA4* deletion strain (*pfa4*Δ) in response to myriocin. The growth of WT strains was similar in the presence of vehicle (methanol), myriocin, exogenous sphingolipids or any combination thereof (Figure 5A–C). In contrast, the growth of the *pfa4*Δ strain was dramatically repressed by myriocin (Figure 5A). Importantly, this repression in *pfa4*Δ was almost completely recovered by adding either of the two sphingolipids (Figure 5B and C). These results strongly suggest that the elevated myriocin sensitivity of the *pfa4*Δ strain is

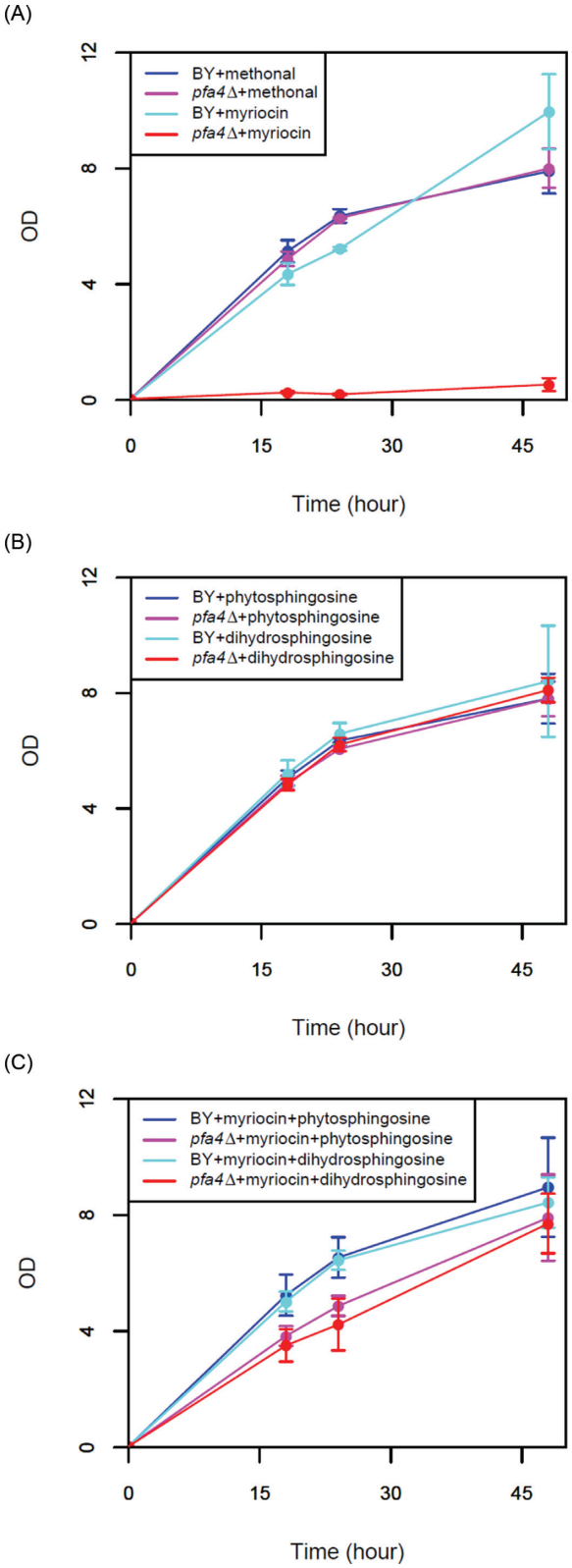

Figure 5. Sphingolipid rescue of *pfa4*Δ on myriocin media indicates an altered sphingolipid pathway *in vivo* in *pfa4*Δ cells. (A) Growth curves of WT (BY) and *pfa4*Δ show that *pfa4*Δ could not grow in myriocin, as measured by the optical density (OD) of the cell culture. (B) Phytosphingosine or dihydrosphingosine did not affect the growth of WT (BY) or *pfa4*Δ under normal growth conditions. (C) Both phytosphingosine and dihydrosphingosine rescued *pfa4*Δ from elevated myriocin sensitivity.

a consequence of an altered sphingolipid pathway. Overall, these experiments indicate that our gene network can be used to identify novel genes influencing a particular pathway.

Compare Ontology Fingerprint to other methods in identifying pathway-modulating genes

We used four other comparable networks and methods but were only able to identify few of the 14 novel genes discovered by Ontology Fingerprint-derived gene network (Table 1, Supplementary Figures S9 and S10). We believe the main reason for this performance difference comes from the comprehensiveness of the Ontology Fingerprints, which summarize different aspects of biology through ontology and literature. Our results indicate that Ontology Fingerprint-derived gene network is unique and can identify many novel pathway-modulating genes that other existing methods cannot.

DISCUSSION

We developed a novel gene network from the Ontology Fingerprint to infer new players in the yeast sphingolipid pathway. Compared to other experiment-based approaches to construct biological networks, such as PPI (2,33,34), GI (4), and co-expression (5,35–38), our approach is not limited to a single aspect of biological function, such as PPI, and is less labor-intensive. Compared to similar literature-based methods (39,40) and GATACA (https://gataca.cchmc.org/gataca/. Last accessed 3/26/2014), our method takes advantage of high-quality biomedical GO data and, most importantly, infers gene–gene relationships that do not exist in the literature (not co-occurrence dependent). Compared to other ontology-based methods (41–43), our approach does not rely on manually curated ontology annotation of genes. Although it is of high quality, the annotation is still in progress and has not comprehensively covered a majority of biological knowledge. In contrast, the Ontology Fingerprint-based approach uses all the ontology terms to survey the entire biomedical literature linked to all genes in the genome to develop an objective yet comprehensive ontology profile. After careful calibration with the existing knowledge, the gene network derived from the Ontology Fingerprint is comprehensive, biologically meaningful and capable of directing us to new pathway players that are not explicitly described in the existing biomedical literature, as demonstrated by our results and comparison to other methods through the course of the method development (Supplementary Table S10).

Our method built upon PubMed abstracts linked to genes. One challenge is that as more high-throughput methods are used, a single publication might link to many genes and create non-specific connections between genes and GO terms. We avoided this pitfall by limiting our inclusion of PubMed abstracts to those that linked to no more than 100 genes.

Another challenge we encountered is the mapping of GO terms. As GO is developed to capture precise meaning of gene functions, many labels of the GO terms are long and detailed, and are not used typically in literature. Therefore, many GO terms may not be able to find exact match in literature, and many researchers are working on the named entity recognition performance against GO (17,18,44). On the other hand, the amount of GO terms we can recover from the literature is sufficient for us to build comprehensive fingerprints that capture enough meaningful features. This is reflected in the ability of Ontology Fingerprints to identify genes for biological pathways. However, future work in providing better mapping could improve the performance of the Ontology Fingerprints.

As our network is derived from literature, to what extend a gene is studied and published in literature could influence the gene–gene relationships in our network. Two approaches could help to alleviate this issue: first, we could impose a minimum requirement for the number of papers published for a gene. If the gene is rarely studied, we will not include it in the network due to the lack of sufficient abstracts to generate high-quality Ontology Fingerprint; second, the enrichment analysis evaluates the overrepresentation of the GO terms in the paper linked to a gene. Even if a gene has few publications, the overrepresentation assessment can still identify these terms from these abstracts.

Even though our Ontology Fingerprint-derived gene network performed well in identifying novel genes modulating sphingolipid pathway, the method could be further improved in several ways. For example, while current Ontology Fingerprint uses only GO dated back to 2009, including up-to-date GO and other bio-ontology could significantly improve the current network as these ontologies could capture additional biological information. Another improvement could stem from the development of novel network analysis methods, such as community detection algorithms in social media research. These algorithms could potentially improve the discovery of novel genes for biological pathways. Finally, many relationships are used to develop Ontology Fingerprints and there could exist false-positive relationships, e.g. gene to PubMed abstracts. Removing these false positives, such as applying Name Entity Recognition methods to identify genes in the PubMed abstracts, could help to improve the performance of Ontology Fingerprints.

With quantified relevance among genes, our network adds a unique dimension to biological networks. While this work demonstrated that the network could be used to infer implicit relationships among genes and novel pathway players, it can also be applied for many other novel analyses. These potential applications can expand our existing knowledge of biological pathways to identify novel drug targets and biomarkers for prognostic and diagnostic purposes.

AVAILABILITY

The Ontology Fingerprint of yeast genes and the derived gene network can be accessed at www.ontologyfingerprint.org.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr. Michael Boehnke for helpful discussions at the early stage of the project, and the reviewers for their extensive suggestions for improving the manuscript.

FUNDING

PhRMA Foundation [Research Starter Grant to W.J.Z.]; National Institutes of Health (NIH) [1 R56 LM010680, R01GM063265–09S1, P20 RR017677–10, 5P20RR017696–05 to W.J.Z.; R01GM063265 to Y.A.H.]. NIH/National Library of Medicine (NLM) [5-T15-LM007438–02 to L.C.T.]. NIH/National Institute of General Medical Sciences (NIGMS) [T32GM074934 07 to T.Q.]. Funding for open access charge: 1 R56 LM010680 NIH/NLM/NIGMS.

Conflict of interest statement. None declared.

REFERENCES

1. Alvarez-Vasquez,F., Sims,K.J., Cowart,L.A., Okamoto,Y., Voit,E.O. and Hannun,Y.A. (2005) Simulation and validation of modelled sphingolipid metabolism in Saccharomyces cerevisiae. *Nature*, **433**, 425–430.
2. Drewes,G. and Bouwmeester,T. (2003) Global approaches to protein-protein interactions. *Curr. Opin. Cell. Biol.*, **15**, 199–205.
3. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
4. Tong,A.H., Lesage,G., Bader,G.D., Ding,H., Xu,H., Xin,X., Young,J., Berriz,G.F., Brost,R.L., Chang,M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
5. de la Fuente,A., Brazhnik,P. and Mendes,P. (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.*, **18**, 395–398.
6. de Hoon,M.J., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2003) Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations. *Pac. Symp. Biocomput.*, **2003**, 17–28.
7. McGary,K.L., Lee,I. and Marcotte,E.M. (2007) Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome Biol.*, **8**, R258.
8. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.
9. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.
10. Fontaine,J.F., Priller,F., Barbosa-Silva,A. and Andrade-Navarro,M.A. (2011) Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.*, **39**, W455–W461.
11. Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1999**, 60–67.
12. Ray,S. and Craven,M. (2005) Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformat.*, **6** , (Suppl 1), S18.
13. Rodriguez-Esteban,R., Iossifov,I. and Rzhetsky,A. (2006) Imitating manual curation of text-mined facts in biomedicine. *PLoS Comput. Biol.*, **2**, e118.
14. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
15. Sheehan,B., Quigley,A., Gaudin,B. and Dobson,S. (2008) A relation based measure of semantic similarity for Gene Ontology annotations. *BMC Bioinformat.*, **9**, 468.
16. Pesquita,C., Faria,D., Bastos,H., Ferreira,A.E., Falcao,A.O. and Couto,F.M. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformat.*, **9**, (Suppl 5), S4.
17. Funk,C., Baumgartner,W. Jr, Garcia,B., Roeder,C., Bada,M., Cohen,K.B., Hunter,L.E. and Verspoor,K. (2014) Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformat.*, **15**, 59.
18. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
19. Bada,M., Eckert,M., Evans,D., Garcia,K., Shipley,K., Sitnikov,D., Baumgartner,W.A. Jr, Cohen,K.B., Verspoor,K., Blake,J.A. *et al.* (2012) Concept annotation in the CRAFT corpus. *BMC Bioinformat.*, **13**, 161.
20. Tsoi,L.C., Boehnke,M., Klein,R. and Zheng,W.J. (2009) *International Conference on Biomedical Ontology*, University at Buffalo, NY.
21. Tsoi,L.C., Boehnke,M., Klein,R.L. and Zheng,W.J. (2009) Evaluation of genome-wide association study results through development of ontology fingerprints. *Bioinformatics*, **25**, 1314–1320.
22. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
23. Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.
24. Lippert,C., Ghahramani,Z. and Borgwardt,K.M. (2010) Gene function prediction from synthetic lethality networks via ranking on demand. *Bioinformatics*, **26**, 912–918.
25. Matmati,N., Kitagaki,H., Montefusco,D., Mohanty,B.K. and Hannun,Y.A. (2009) Hydroxyurea sensitivity reveals a role for ISC1 in the regulation of G2/M. *J. Biol. Chem.*, **284**, 8241–8246.
26. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
27. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
28. Lee,I., Li,Z. and Marcotte,E.M. (2007) An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS One*, **2**, e988.
29. Dahan,O. and Kupiec,M. (2004) The Saccharomyces cerevisiae gene CDC40/PRP17 controls cell cycle progression through splicing of the ANC1 gene. *Nucleic Acids Res.*, **32**, 2529–2540.
30. Meier,K.D., Deloche,O., Kajiwara,K., Funato,K. and Riezman,H. (2006) Sphingoid base is required for translation initiation during heat stress in Saccharomyces cerevisiae. *Mol. Biol. Cell*, **17**, 1164–1175.
31. Cowart,L.A., Gandy,J.L., Tholanikunnel,B. and Hannun,Y.A. (2010) Sphingolipids mediate formation of mRNA processing bodies during the heat-stress response of Saccharomyces cerevisiae. *Biochem. J.*, **431**, 31–38.
32. Daquinag,A., Fadri,M., Jung,S.Y., Qin,J. and Kunz,J. (2007) The yeast PH domain proteins Slm1 and Slm2 are targets of sphingolipid signaling during the response to heat stress. *Mol. Cell. Biol.*, **27**, 633–650.
33. Russell,R.B., Alber,F., Aloy,P., Davis,F.P., Korkin,D., Pichaud,M., Topf,M. and Sali,A. (2004) A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.*, **14**, 313–324.
34. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
35. Zak,D.E., Pearson,R.K., Vadigepalli,R., Gonye,G.E., Schwaber,J.S. and Doyle,F.J. 3rd. (2003) Continuous-time identification of gene expression models. *Omics*, **7**, 373–386.
36. Dasika,M.S., Gupta,A. and Maranas,C.D. (2004) A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. *Pac. Symp. Biocomput.*, **2004**, 474–485.
37. Troyanskaya,O.G. (2005) Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform.*, **6**, 34–43.
38. Li,S., Wu,L. and Zhang,Z. (2006) Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics*, **22**, 2143–2150.

39. Iossifov,I., Rodriguez-Esteban,R., Mayzus,I., Millen,K.J. and Rzhetsky,A. (2009) Looking at cerebellar malformations through text-mined interactomes of mice and humans. *PLoS Comput. Biol.*, **5**, e1000559.
40. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
41. Chen,J.L., Liu,Y., Sam,L.T., Li,J. and Lussier,Y.A. (2007) Evaluation of high-throughput functional categorization of human disease genes. *BMC Bioinformat.*, **8**, (Suppl 3), S7.
42. Du,P., Feng,G., Flatow,J., Song,J., Holko,M., Kibbe,W.A. and Lin,S.M. (2009) From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, **25**, i63–i68.
43. Kang,B.Y., Ko,S. and Kim,D.W. (2010) SICAGO: semi-supervised cluster analysis using semantic distance between gene pairs in Gene Ontology. *Bioinformatics*, **26**, 1384–1385.
44. Mao,Y., Van Auken,K., Li,D., Arighi,C.N. and Lu,Z. (2014), *Overview of the Gene Ontology Task at BioCreative IV*, Database, 2014, Accepted.