

Research article

Open Access

Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks

Marc RJ Carlson¹, Bin Zhang², Zixing Fang^{3,4}, Paul S Mischel⁵, Steve Horvath^{1,6} and Stanley F Nelson^{* 1,7}

Address: ¹Department of Human Genetics, David Geffen School of Medicine at UCLA, Gonda (Goldschmied) Neuroscience and Genetics Research Center, 695 Charles E. Young Drive South, Los Angeles, CA 90095-7088, USA, ²Rosetta Inpharmatics LLC, 401 Terry Avenue North, Seattle, WA 98109, USA, ³Cancer Prevention Institute 4100 South Kettering Blvd., Dayton, OH 45439, USA, ⁴Department of Community Health, School of Medicine, Wright State University 136 F.A. White Health Center 3640 Colonel Glenn Highway, Dayton, OH 45435, USA, ⁵Department of Pathology and Laboratory Medicine, UCLA, 10833 Le Conte Ave. Los Angeles, CA 90095, USA, ⁶Department of Biostatistics, UCLA, CHS Suite 51-236 650 Charles E. Young Dr. Los Angeles, CA 90095, USA and ⁷Department of Psychiatry, David Geffen School of Medicine, UCLA, 760 Westwood Plaza Los Angeles, CA 90095, USA

Email: Marc RJ Carlson - mrjcarlson@mednet.ucla.edu; Bin Zhang - binzhang.ucla@gmail.com; Zixing Fang - fang@CancerPreventionInstitute.org; Paul S Mischel - pmischel@mednet.ucla.edu; Steve Horvath - shorvath@mednet.ucla.edu; Stanley F Nelson* - snelson@ucla.edu

* Corresponding author

Published: 03 March 2006

Received: 04 September 2005

BMC Genomics 2006, 7:40 doi:10.1186/1471-2164-7-40

Accepted: 03 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2164/7/40>

© 2006 Carlson et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Genes and proteins are organized into functional modular networks in which the network context of a gene or protein has implications for cellular function. Highly connected hub proteins, largely responsible for maintaining network connectivity, have been found to be much more likely to be essential for yeast survival.

Results: Here we investigate the properties of weighted gene co-expression networks formed from multiple microarray datasets. The constructed networks approximate scale-free topology, but this is not universal across all datasets. We show strong positive correlations between gene connectivity within the whole network and gene essentiality as well as gene sequence conservation. We demonstrate the preservation of a modular structure of the networks formed, and demonstrate that, within some of these modules, it is possible to observe a strong correlation between connectivity and essentiality or between connectivity and conservation within the modules particularly within modules containing larger numbers of essential genes.

Conclusion: Application of these techniques can allow a finer scale prediction of relative gene importance for a particular process within a group of similarly expressed genes.

Background

Genes and their protein products carry out cellular processes in the context of functional modules and are related to each other through a complex network of interactions [1]. Understanding an individual gene or protein's network properties within such networks may prove to be as

important as understanding its function in isolation [2]. Because of this, numerous studies have focused on the large scale modeling of genomic and proteomic data. Utilizing network theory, these studies have yielded insights into biological systems. For example, both protein interaction networks and gene co-expression networks exhibit

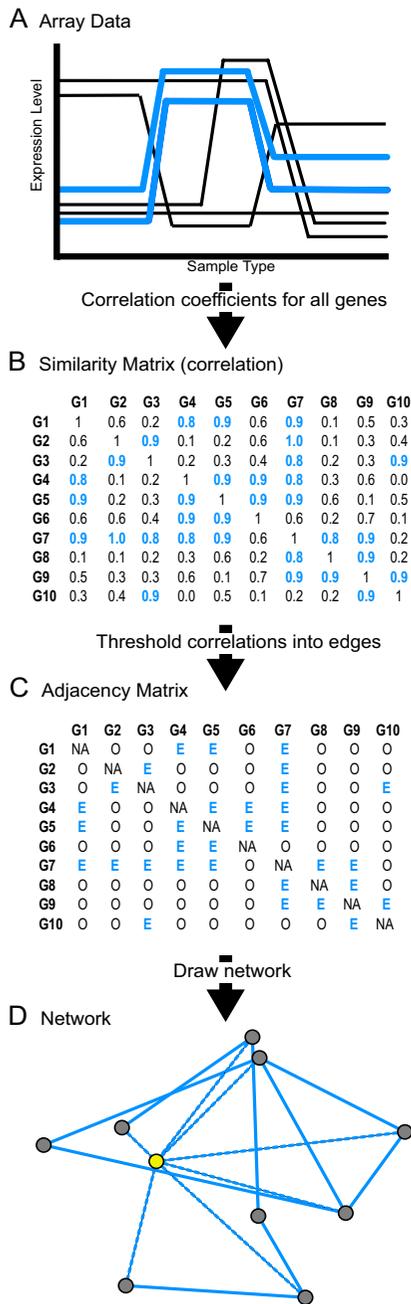


Figure 1
Generating a gene co-expression network. (A) Illustration to show how genes that are highly correlated (blue) will look across a dataset. (B-C) Flow chart for defining a gene co-expression network based on a simple Pearson correlation matrix. (D) An example visualization of a network produced by using the Fruchterman Rheingold algorithm on the data in Figure 1B and 1C. The node highlighted in yellow, is an example of the kind of highly connected node that this study shows is more likely to be an essential gene for survival.

a strong modularity reflecting functional partitioning. Both of these network types have been frequently observed as having a scale-free topology, with the existence of highly connected hub nodes [3-7].

Scale-free networks are resistant to random perturbations but sensitive to targeted removal of highly connected nodes [8]. Comprehensive efforts to determine the functional consequences of individual gene deletions in yeast provide the opportunity to study the relationship between individual gene network properties and gene deletion lethality [9]. For example, physical interaction studies in yeast have allowed comparison of connectivity to gene essentiality based on gene deletion. [10]. Typical of scale-free networks, there were few highly connected proteins within the network, and the deletion of a protein with a large number of binding partners is more likely to be lethal in yeast. Thus, the relative position of nodes within a protein interaction network is strongly affiliated with distinct biological properties of individual proteins. Similarly, analysis of unweighted gene co-expression networks have revealed a relationship between connectivity and essentiality across all genes [11].

Correlation of gene expression across a wide variety of experimental perturbations has been shown to cluster genes of similar function [12]. Since this guilt-by-association approach may lead to false positive groupings, approaches have been refined by comparing orthologs across divergent species indicating that highly conserved co-expression is a strong predictor that two genes will function in similar pathways [13-15]. This indicates that functionally related genes are under similar expression constraints. The gene co-expression networks that are based on these relationships have been shown in multiple species to be frequently scale-free and exhibit a small world architecture similar to protein interaction networks even though they are generally more strongly connected [3,14]. It is still unclear however, to what degree the network properties of individual genes within a co-expression network can predict relative gene importance for a particular process. To assess this, we have constructed three networks based on a weighted measure of connectivity [16] of correlated gene expression in yeast using three separate microarray data sets. We assessed relationships between essentiality and connectivity of each gene within the whole network. Further, we define 'modules' (groups of highly correlated genes) and determine that in some instances, the relative importance of genes within these modules can be inferred from network connectivity. We demonstrate that genes which have high connectivity (i.e. 'hub' genes) within a weighted co-expression network are significantly more likely to be essential for yeast viability. Furthermore, we demonstrate a relationship between connectivity and a measure of sequence conservation. Finally,

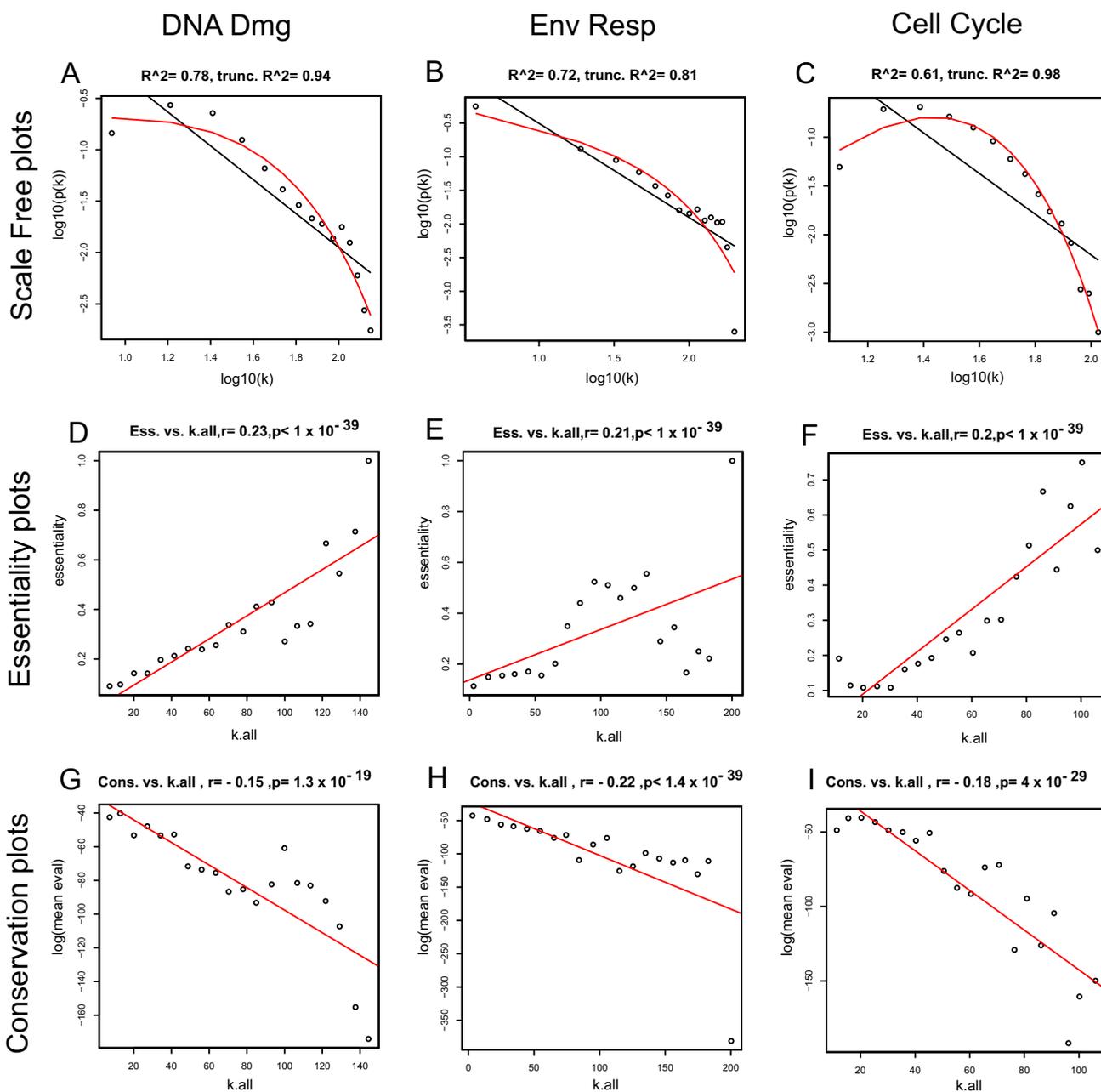


Figure 2
Global gene co-expression networks. (A-C) Log-log plots of connectivity distributions in each of the three networks drawn from DNA Damage, Environmental Response, and Cell Cycle, respectively. The linear relationship indicates a scale-free structure. (D-F) Correlation plots showing the relationship between gene group connectivity and essentiality in the same order as A-C. On the y axis of these plots is the percentage of genes determined to be essential in yeast. (G-I) Relationship between the average blastp score of a node and k for all datasets in the same order as A-C. On the y axis of these plots is the average log(e score) of genes within that bin. For plots D-I, the connectivity within each network for each gene was determined and each gene rank ordered by connectivity. 20 equal sized bins for each of the expression datasets were formed and the average connectivity of the genes in each bin plotted on the X axis

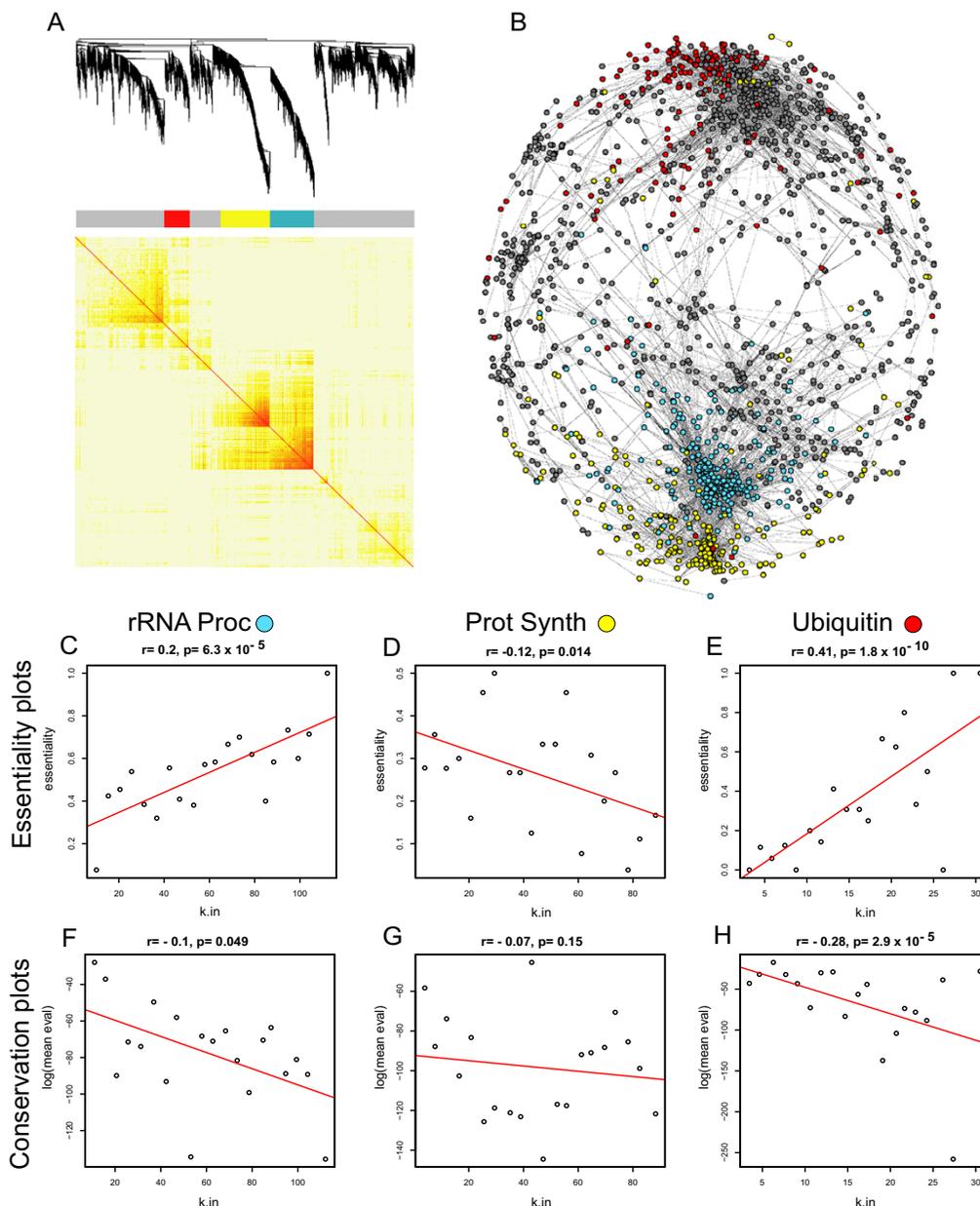


Figure 3

A co-expression network of the DNA Damage dataset. For all panels in the figure: blue color represents members of the rRNA processing module, yellow color represents members of the protein synthesis module and red color represents members of the ubiquitin pathway. (A) A hierarchical clustering of the topological overlap matrix for the DD dataset. (B) A drawn network of gene co-expression from the DD dataset. Edges were computed from the Pearson correlation coefficients. Network structure was drawn in Pajek [20]. Each gene is represented as a dot. Edges are drawn as grey lines. Colored dots belong to the module that their color indicates and grey dots indicate all other genes in the network. (C-E) Scatter plots showing the relationship between gene group connectivity and essentiality for rRNA processing, protein synthesis and ubiquitin from the DD network respectively. On the y axis of these plots is the percentage of genes determined to be essential in yeast. The number of essential genes in each module/total number of genes in each module was: $196/390 = 50.3$ (C), $122/441 = 27.7$ (D), $50/222 = 22.5$ (E). (F-H) Relationship between the average blastp score of a node and k for the same module members listed in C-E. On the y axis of these plots is the average log(e score) of genes within that bin. For plots C-H, the connectivity within each module for each gene was determined and each gene rank ordered by connectivity. 20 equal sized bins for each of the expression datasets were formed and the average connectivity of the genes in each bin plotted on the X axis.

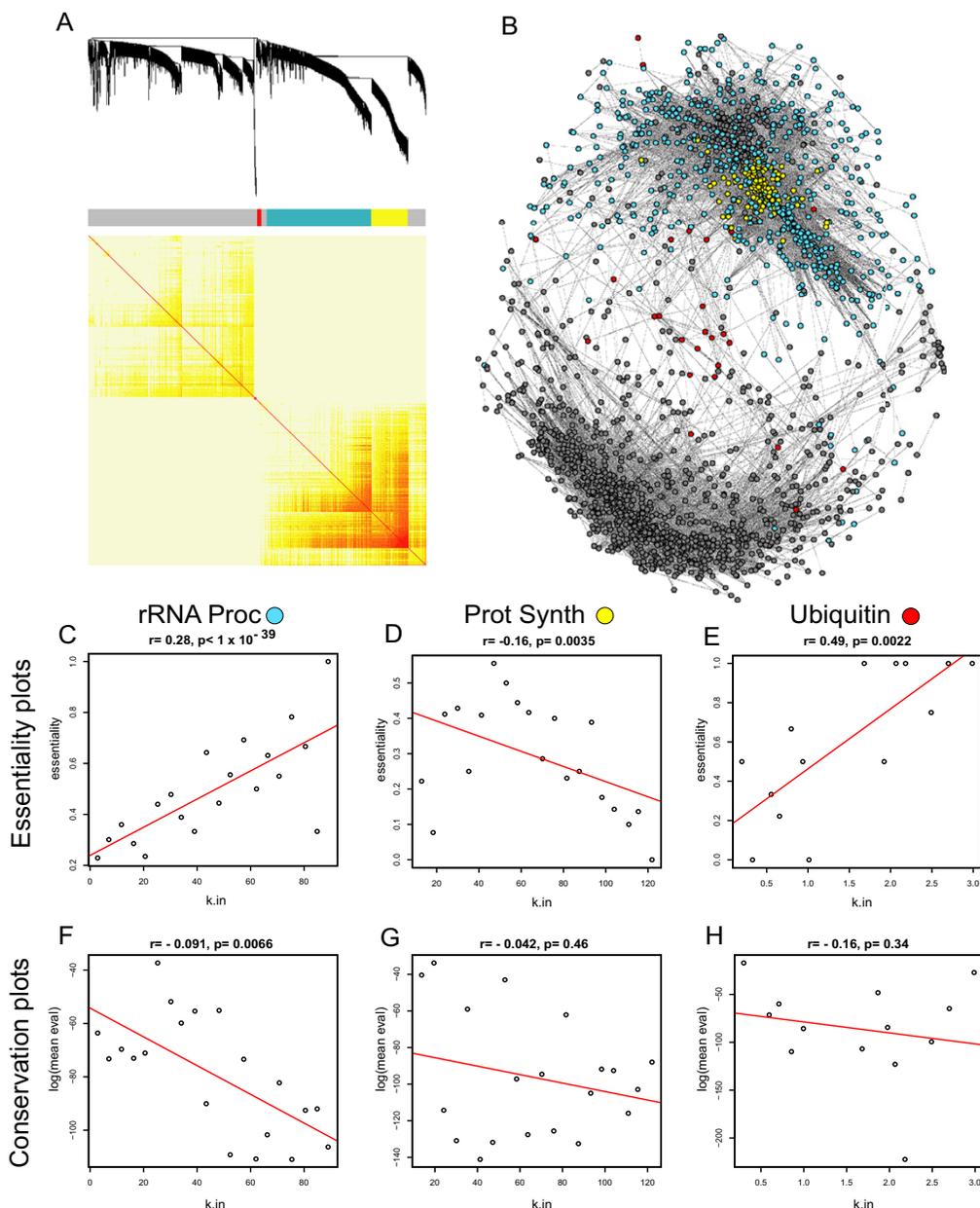


Figure 4

A co-expression network of the Environmental Response dataset. For all panels in the figure: blue color represents members of the rRNA processing module, yellow color represents members of the protein synthesis module and red color represents members of the ubiquitin pathway. (A) A hierarchical clustering of the topological overlap matrix for the ER dataset. (B) A drawn network of gene co-expression from the ER dataset. A drawn network of gene co-expression from the DD dataset. Edges were computed from the Pearson correlation coefficients. Network structure was drawn in Pajek [20]. Each gene is represented as a dot. Edges are drawn as grey lines. Colored dots belong to the module that their color indicates and grey dots indicate all other genes in the network. (C-E) Correlation plots showing the relationship between gene group connectivity and essentiality for rRNA processing, protein synthesis and ubiquitin from the ER network respectively. On the y axis of these plots is the percentage of genes determined to be essential in yeast. The number of essential genes in each module/ total number of genes in each module was: 317/929 = 34.1 (C), 98/323 = 30.3 (D), 18/37 = 48.6 (E). (F-H) Relationship between the average blastp score of a node and k for the same module members listed in C-E. On the y axis of these plots is the average log(e score) of genes within that bin. For plots C-H, the connectivity within each module for each gene was determined and each gene rank ordered by connectivity.

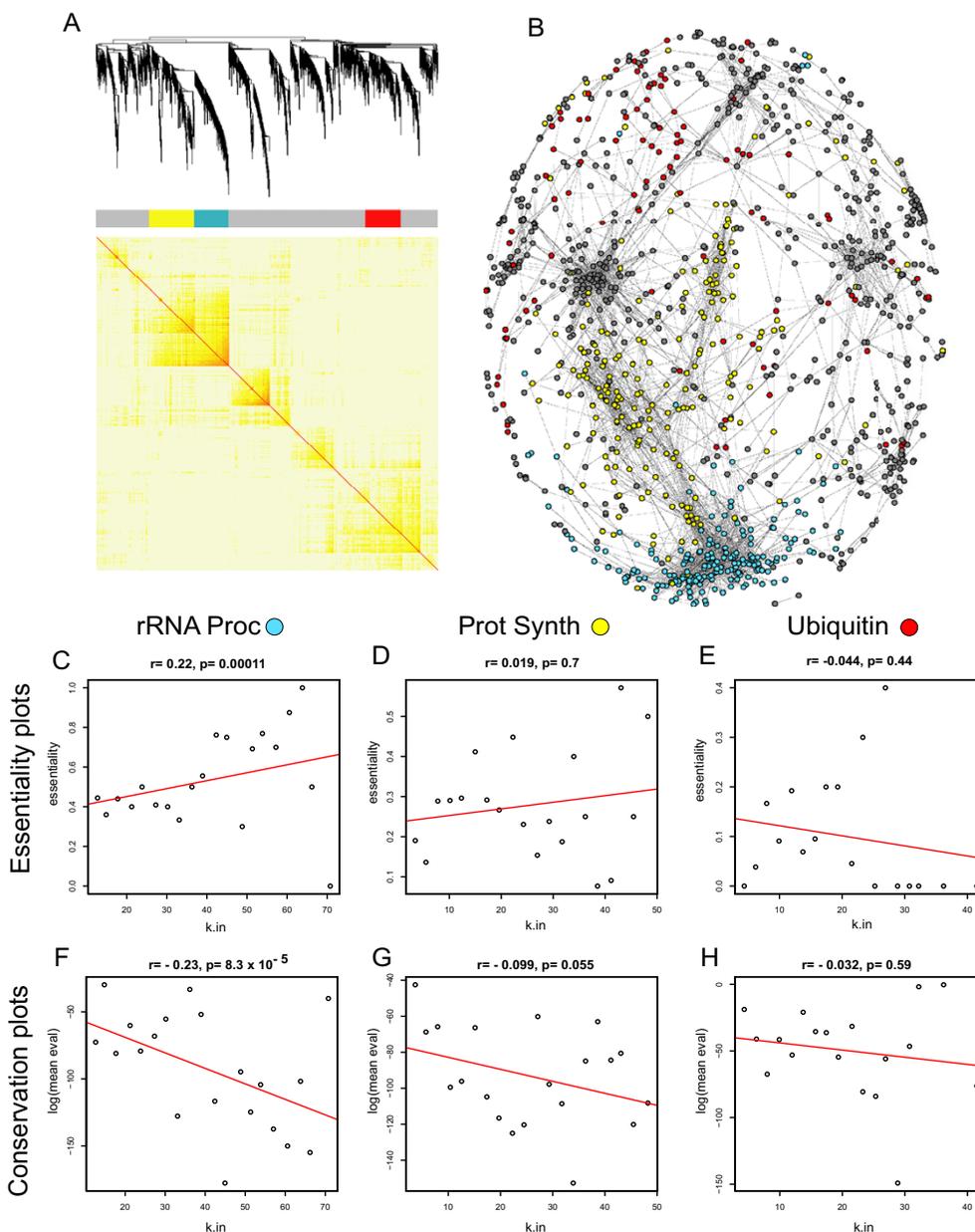


Figure 5

A co-expression network of the Cell Cycle dataset. For all panels in the figure: blue color represents members of the rRNA processing module, yellow color represents members of the protein synthesis module and red color represents members of the ubiquitin pathway. (A) A hierarchical clustering of the topological overlap matrix for the CC dataset. (B) A drawn network of gene co-expression from the CC dataset. A drawn network of gene co-expression from the DD dataset. Edges were computed from the Pearson correlation coefficients. Network structure was drawn in Pajek [20]. Each gene is represented as a dot. Edges are drawn as grey lines. Colored dots belong to the module that their color indicates and grey dots indicate all other genes in the network. (C-E) Correlation plots showing the relationship between gene group connectivity and essentiality for rRNA processing, protein synthesis and ubiquitin from the CC network respectively. On the y axis of these plots is the percentage of genes determined to be essential in yeast. The number of essential genes in each module/total number of genes in each module was: 154/300 = 51.3 (C), 105/398 = 26.4 (D), 38/312 = 12.2 (E). (F-H) Relationship between the average blastp score of a node and k for the same module members listed in C-E. On the y axis of these plots is the average log(e score) of genes within that bin. For plots C-H, the connectivity within each module for each gene was determined and each gene rank ordered by connectivity.

we show that certain critical modules are conserved from one network to another, and that in many cases it is possible to extend the relationship between connectivity and essentiality or between connectivity and conservation within given modules. Thus, analysis of gene co-expression networks provides insight into the functional importance of individual genes within modules of co-expressed genes.

Results

Network validation and properties

In order to test the relationship between gene connectivity within the co-expression network and gene essentiality, we first constructed three networks corresponding to three microarray datasets. The yeast microarray data were derived from experiments designed to analyze DNA Damage (DD) ($n = 51$), Cell Cycle (CC) ($n = 44$), and Environmental Response (ER) ($n = 151$) [17-19]. Each network was constructed as conceptually shown in Figure 1. Nodes were defined as individual genes. The co-expression network is based upon the correlation relationships between individual genes across a dataset (Figure 1A). From the microarray data, the absolute value of the Pearson correlation is the initial measure of gene co-expression similarity (Figure 1B). This co-expression similarity was then transformed into an adjacency matrix (Figure 1C) by raising the similarity to a certain power based on a scale free topology criterion as described previously [16]. The powers selected were 12 (DD), 10 (CC), and 18 (ER). These data can also be visualized as a drawn network (Figure 1D).

Each of the datasets resulted in the construction of distinct networks but with global similarities. The differences in the modularity of the three networks is demonstrated by comparing the heat maps of the topological overlap matrices (Figures 3A, 4A, and 5A). A topological overlap matrix is an indicator of the similarity of all the nodes in question and is defined in detail in the methods section of this manuscript. Approximations of all three of the networks are also drawn in Figures 3B, 4B and 5B using the Fruchterman-Reingold algorithm as implemented in Pajek [20,21] and demonstrate obvious modularity with aggregation of genes into strongly co-expressed groups.

Individual gene connections can be difficult to visualize in such complex graphs as presented. Thus, in order to probe the structure of the network, we first sought to determine characteristics of the general network topology. We considered two models. To quantify the fit to the scale free topology model or to the truncated exponential model, we used the model fitting index R-squared of the corresponding linear model involving the log transformed variables [16]. The pattern of network connectivity in these datasets approximated a scale-free topology

(Figures 2A, 2B, and 2C). These results are similar to the general trend with genetic perturbation networks and other gene co-expression networks [3,22]. However, gene co-expression networks may better fit an exponential power law [23]. For two of the three constructed yeast networks (DD and ER), the distribution of connectivity (k) of genes best fits a power law $p(k) \sim k^{-\gamma}$ and the other (CC) best fits an exponentially truncated power law $p(k) \sim k^{-\gamma} e^{-\alpha k}$ [24]. We point out that in general that regardless of the overall structure, the major trend is that the high connectivity genes are few in number, and that most genes have low connectivity (Figure 2A-2C).

Global network connectivity relationships

Network modeling in biological systems [2,8,11] stresses the relative importance of highly connected nodes. Thus, the small number of high connectivity genes within the gene co-expression networks defined here are expected to be more important. To test the relative importance of each gene across the entire network, we determined the relationship between connectivity and gene essentiality. All the genes for a network were first ranked according to their weighted connectivity and then placed into 20 equal sized bins according to their position in this ranking. The weighted connectivity of a given gene is defined as the sum of its connection strengths with all other genes in the network (see the methods section). The percentage of essential genes in each bin was then plotted against the mean connectivity in each bin (Figure 2D-2F). We also calculated the Spearman correlation coefficient based on the ranked unbinned data and found that there is a relationship between the connectivity and the proportion of essential genes in each of the three datasets ($R > 0.2$, and p value $< 10^{-39}$) for each network. While the binned plots are well suited for visualizing the results, we calculated the correlation coefficients on the raw, unbinned data to arrive at a valid p -value.

To study whether our conclusions were being unduly influenced by data outliers, we removed genes whose connectivity was more than two standard deviations away from mean connectivity but found no substantial variation in the reported relationships. Thus, there is a clear relationship between connectivity of genes, in these weighted gene co-expression networks, and gene essentiality across the whole network. These results are robust with respect to experimental conditions since the three separate datasets all have the same functional relationship between connectivity and gene essentiality. These results are consistent with a single whole network generated from gene co-expression data and using an unweighted network [11], as well as networks based on protein-protein interactions [10].

Table 1: Identifying modules that repeated from one dataset to another

Ontology term	rRNA			Prot			Ubiq		
	DNA Dmg	Env Stress	Cell Cycle	DNA Dmg	Env Stress	Cell Cycle	DNA Dmg	Env Stress	Cell Cycle
35S primary transcript processing	<i>1E-17</i>	<i>8.8E-16</i>	<i>7.5E-16</i>	0.17	0.51	0.52	0.26		0.16
processing of 20S pre-rRNA	<i>2E-32</i>	<i>5E-19</i>	<i>3.8E-32</i>	0.018	0.37		0.17		0.11
ribosomal large subunit biogenesis	<i>9.6E-13</i>	<i>0.0000071</i>	<i>3.5E-15</i>	0.38		0.38			0.62
rRNA modification	<i>2.1E-08</i>	<i>0.0043</i>	<i>0.00014</i>	0.66	0.019	0.00066			0.62
rRNA processing	<i>1.2E-25</i>	<i>1.6E-16</i>	<i>2.5E-23</i>	0.48		0.62	0.11		0.046
protein biosynthesis	0.0001	1.1E-11	0.0023	<i>5.9E-83</i>	<i>5.6E-108</i>	<i>1.4E-80</i>	0.53	0.42	0.00033
translational elongation	0.62	0.14	0.63	<i>2.1E-08</i>	<i>3E-08</i>	<i>1.7E-09</i>			0.64
translational initiation	0.19		0.025	<i>0.000017</i>	<i>1.8E-14</i>	<i>3.4E-07</i>	0.41		0.25
ubiquitin-dependent protein catabolism	0.039	0.02	0.071	0.027	0.048	0.018	<i>2.5E-25</i>	<i>7.3E-25</i>	<i>6.3E-17</i>
ATP-dependent RNA helicase activity	<i>7.2E-11</i>	<i>9.2E-08</i>	<i>4.1E-11</i>	0.39	0.62				0.62
DNA-directed RNA polymerase activity	<i>6.1E-13</i>	<i>4E-08</i>	<i>1.4E-09</i>	0.77	0.012	0.13	0.42		0.4
snoRNA binding	<i>1.3E-27</i>	<i>7.6E-17</i>	<i>2.4E-25</i>	0.042	0.51	0.76	0.42		0.16
structural constituent of ribosome	0.00012	2.2E-11	0.000078	<i>3.1E-92</i>	<i>5.3E-117</i>	<i>1.8E-95</i>	0.097	0.41	0.0021
translation initiation factor activity	0.16	0.48	0.14	<i>4.7E-08</i>	<i>5.5E-17</i>	<i>2.3E-10</i>			0.39
endopeptidase activity	0.25	0.21	0.62	0.16	0.62	0.39	<i>9.2E-21</i>	<i>2.6E-18</i>	<i>2.7E-11</i>
nucleolus	<i>1.6E-48</i>	<i>3.4E-29</i>	<i>1.3E-45</i>	0.18	0.65	0.84	0.021	0.48	0.0056
nucleus	<i>1.2E-18</i>	<i>8.3E-13</i>	<i>7.8E-10</i>	0.0053	2.1E-07	0.12	0.0048		0.033
small nucleolar ribonucleoprotein complex	<i>7.9E-20</i>	<i>2.8E-09</i>	<i>1.3E-14</i>	0.16	0.67	0.42	0.63		0.39
cytosolic large ribosomal subunit	0.03	0.0000016	0.0087	<i>2.1E-65</i>	<i>3.4E-75</i>	<i>6.1E-67</i>	0.021		0.0056
cytosolic small ribosomal subunit	0.059	0.0013	0.26	<i>1.5E-41</i>	<i>1.2E-47</i>	<i>4.6E-42</i>	0.073		0.03
ribosome		0.39	0.3	<i>2.7E-09</i>	<i>6.2E-14</i>	<i>3.7E-09</i>			0.62

*For each module in each dataset analyzed, an EASE analysis [25] was performed by using Fisher's exact test and the resulting p-values were placed into the table based on the prevalence of the GO (Gene Ontology) term being searched for. The most significant enrichments for a given term are indicated in italics

Yeast co-expression networks fit an evolutionarily plausible model in which network connection distribution is a natural consequence of gene duplication, mutation and deletions [13]. The emergence of approximate power-law distribution (scale free topology) is intimately linked to the growth of the network in which new nodes are preferentially attached to highly connected already established nodes [2,8]. A direct consequence of this network growth model is that high connectivity 'hub' genes are more likely to be highly conserved across different species. Stated another way, within a gene co-expression network the oldest nodes would be more likely than the newer nodes to be in hub positions. With this in mind, we examined the relationship between connectivity and the relative evolutionary sequence conservation of a node (as evidenced by the average best blastp hit log(e score) against *N. crassa*, *C. elegans*, *D. melanogaster*, *H. sapiens*, and *M. musculus*). High connectivity genes are plotted in Figure 2G–2I relative to the average of the blastp hit in four other eukaryotic genomes. As before, we calculated the Spearman correlation for the ranked, but still unbinned, data

and found that there is a relatively weak but highly significant correlation indicating that high connectivity genes are more evolutionarily conserved ($R = -0.15$, p value = 1.3×10^{-19} , for the DD network; $R = -0.22$, p value < 1.4×10^{-39} for ER network; and $R = -0.18$, p value = 4.0×10^{-29} for the CC network). Our results were robust with respect to the removal of outlier genes (i.e. genes whose connectivity was more than two standard deviations away from mean connectivity). This relationship is similar to that reported for gene perturbation networks [22] and in evolutionarily conserved gene co-expression relationships [14].

Identification of modules

In order to identify modules, we performed hierarchical clustering on the topological overlap matrix derived from these networks as described in the methods section. In doing so, we identified modules that involve conserved processes for rRNA processing, protein synthesis, and ubiquitination (Table 1) using an EASE analysis [25]. One prediction about network constructions is that certain crit-

ically important modules ought to be repeatedly detected in independent datasets. Within the network modules (groups of high connectivity genes within the network) produced by the different datasets for gene ontology enrichment, 3 distinct modules that were common to the 3 different data sets were detected (Table 1 and Figure 6). Comparing the modules that had an enrichment in rRNA processing, protein synthesis and the ubiquitin pathway terms across the DD, CC and ER datasets resulted in a highly significant overlap of 215 genes (hypergeometric distribution $p = 3.1e-173$). Similarly, the protein synthesis module also had significantly more overlap than expected by chance in which 172 genes were present in all three modules ($p = 5.6e-99$). The ubiquitin module had the least overlap of common genes (12 genes, $p = 3.4e-07$) probably because of its smaller size in general.

These three processes are well known to be fundamental processes and have been described before in great detail, and so we decided to investigate the nature of these overlapping regions further. Gene members were selected that were shared within similar functional modules from the three datasets and searched for in the KEGG [26] database. In the ubiquitin pathway, 9 of the overlapping 12 genes were found in KEGG, and all of these were found to be in the same proteasome structure. Since the overlap for this module was unusually small due in part to the very small size of the ER dataset's ubiquitin module, we also looked at the overlap just of the CC and DD ubiquitin modules. Of the 57 genes present in both the CC and DD modules, only 14 were present in the KEGG database. Of those 14 members, all of them were in the same proteasome struc-

ture. For the Protein Biosynthesis module, 135 of 172 overlapping members were represented in the KEGG database, and of these, 72 out of a possible 86 were components of the KEGG annotated yeast ribosome. Notably, if the 172 overlapping members were first rank ordered by their average network connectivity, and the bottom quartile removed, then 100% of the remaining members are part of this same yeast ribosome structure, which shows that in this overlap set the high connectivity genes were in fact the members of the yeast ribosome. Finally, the overlap of genes detected in the rRNA modules overlap identified purine and pyrimidine synthesis and RNA polymerase 3 pathways. These pathways dominate this list with 30 out of the 49 identified matches to the KEGG database. Unlike the previous two results however, multiple pathways are found that represent this module. The three network modules discussed all contain large protein complexes. This is in agreement with the recent discovery that there is often a correlation between genes that are expressed together and proteins that bind to each other [27]. However, there is no reason to believe that modules in co-expression networks will always correspond to protein complexes since our module detection method should find any set of genes with high pair-wise correlations. Any experimental perturbation that gives rise to a highly correlated gene set is expected to result in a corresponding network module.

Intramodular network connectivity relationships

Having identified 3 functionally similar modules in the three different datasets, we assessed whether these same modules contain the same information as the entire net-

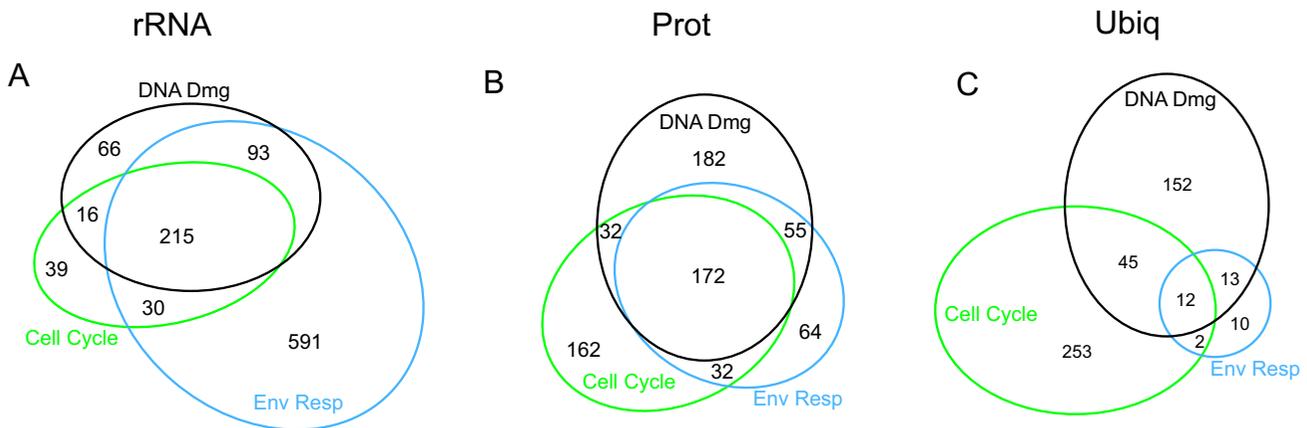


Figure 6
Module members retrieved from very different datasets are conserved. For all panels in the figure: black color represents modules retrieved from the DD dataset, blue color represents modules retrieved from the ER dataset, and green color represents modules retrieved from the CC dataset. In each panel A-C a different module is examined for overlap of its members in a Venn diagram. In each diagram the relative overlap of the various modules has been diagrammed as different amounts of area with the count represented for each zone by a number that quantifies the precise number of matches.

work. Is the connectivity of a gene within the identified modules correlated with the likelihood that a gene was essential by deletion analysis and also whether it was conserved (Figures 3, 4, 5)? To test connectivity correlations within these modules, we define the intramodular connectivity (k_{in}) as the sum of all connectivity measures only with genes within the same module. Similar to that performed for the total gene connectivity within the whole network, binned values were plotted to observe the possible relationships (Figures 3C–H, 4C–H, and 5C–H). To evaluate significance, Spearman correlation coefficients for the ranked but unbinned data were calculated for relationship of k_{in} to essentiality and sequence conservation. Interestingly, there was a general trend to similar correlations as that seen within the whole network. The correlation with sequence conservation was generally weak: The ubiquitin module formed from the DD dataset and the rRNA processing modules in the ER and CC were the most correlated with k_{in} . There was a correlation of k_{in} and essentiality in all three datasets within the rRNA Processing module (Figures 3C, 4C and 5C), and for two of the datasets (DD and ER but not CC) within the ubiquitin module, and a trend towards a weak anticorrelation within the protein synthesis module (Figures 3D, 4D, and 5D). Thus, some of the relationships between connectivity and essentiality or sequence conservation and not detected when the analysis is restricted within the modules.

Characterizing modules that contain essential hub genes

The relationship between intramodular connectivity and essentiality is not necessarily universal for any module in any dataset. Whether or not a module contains meaningful information for predicting essentiality depends on the expression data (biological perturbations) that were analyzed. One can view the correlation of k_{in} and essentiality as a test of the modules importance. The key issue is that in order for a module to demonstrate this relationship, there must be a sufficient number of essential genes within the module. As some modules have substantially fewer essential genes than others, there is little power to detect any relationship with k_{in} within those modules. We calculated the Spearman correlation between the proportion of essential genes within a module and the correlation of k_{in} with essentiality and show that there is a strong positive correlation ($r = 0.48$). Since in actual applications the goal will be to find unknown 'essential' genes, it would be preferable to have another metric of 'essential' gene enrichment. In this regard, it is interesting to note that of the three modules studied in the 3 datasets (9 total), there were 4 that had a correlation of k_{in} with sequence conservation ($p < 0.05$). Of these 4 modules, all 4 were shown to have a strong relationship of k_{in} with essentiality in the same modules. Of the 5 without a correlation ($p > 0.05$), only one showed a correlation

between k_{in} and essentiality (DD, ubiquitin module). Thus, the detection of a relationship between intramodular connectivity and sequence conservation may predict those modules with a relationship between k_{in} and essentiality.

Discussion

Our results extend the inferences made from protein and metabolic networks to the more prolific datasets available from gene expression data. These results are made possible due to broad surveys of expression and the comprehensive characterization of the yeast genome with deletion mutants. In contrast to gene expression data, which are based on the assessment of cells in many different states of perturbation, protein interaction measurements are determined in only a limited set of conditions which by nature only capture a subset of all possible protein-protein interactions. Others [22,28,29] have constructed gene perturbation networks by looking at the expression changes when individual genes are eliminated or in relation to transcription factor expression. Microarray based surveys of gene expression allow detailed surveys of gene expression across a wide variety of experimental perturbations. From these data a strong correlation of gene expression for two genes implies that both genes act within a common functional group and are under similar transcriptional control. Since these networks are not based on direct measurements of protein binding/interaction, the presence of two genes that are highly correlated has a different interpretation than that derived from a protein interaction network. In a co-expression network, two closely tied nodes would be highly correlated but not necessarily direct binding partners. In spite of this key difference, numerous other studies have indicated that genes that are correlated with each other across a modest range of experiments tend to function in similar ways and that this functional categorization is strengthened when gene co-expression exists across large evolutionary distances [12,14].

Here, we use a more general weighted gene co-expression network construction method for network construction that can be applied generically to microarray data using simple correlational analyses of individual genes [16]. Advantages of weighted networks over unweighted networks based on dichotomizing the correlation matrix include a) that the continuous nature of the gene co-expression information is preserved and b) that the results of weighted network analyses are highly robust with respect to the choice of the parameter β , [16]. We show that there is a strong correlation with the weighted gene co-expression connectivity measure and gene essentiality and sequence conservation. Our analysis demonstrates that individual conserved modules are identifiable within a network. Moreover, the presence of these modules,

which were detected in three independent networks, illustrates that these networks have a structure to them which matches known cellular biology. We have also shown that when sufficient information is available, these correlation modules can provide further information about which of their members are more likely to be important for a given process. The ability to assess whether or not a module is complete enough to be informative using a test like either of the ones employed here will be important for validating the merit of less understood modules in the future.

Our study extends the strong links between the network properties of a gene within co-expression networks and the global importance of each gene to the yeast cell, and may have implications for targeted drug therapy in more complex biological systems like cancer. For instance, one can infer from an analysis of the single gene deletion data that single targets for inhibition can be partially predicted from expression data. However, in biological systems an optimal effect is frequently achieved by inhibiting multiple specific targets. Can we extrapolate network positions and relationships to infer effective combinations of targeted therapeutics? Yeast is an excellent model for establishing the rules for applying network theories to practical predictions. For instance, will double deletions of high connectivity genes within the network which reside in the same module have substantially greater negative consequences to the yeast than double deletions of low connectivity genes? Alternatively, will genes which are connecting distinct modules, and serving as key connections between integrators prove to be better disruption targets? The modeling of such proposed interactions in yeast is attractive because of the relative tractability of yeast studies and the availability of large scale datasets. Technologies like synthetic genetic arrays which automate double deletion screens or chemical genomic profiling provide an expanding capability to survey the effects of multiple target inhibition or deletion [30-32]. However, even with these powerful methods, only a small fraction of the millions of possible combinations will be assayed. For instance in a substantial recent survey of double deletions only 0.02% of all possible double deletions were tested for synthetic lethality [33]. Thus, network based modeling may be a useful adjunct for the design of double knockout experiments or combination inhibition experiments to permit a more efficient search for synthetic lethals. An interplay between predictive models and targeted large scale inhibition studies may be quite useful for surveying pathway analysis in yeast. In the long term, such studies may also be critical in the development of therapies for diseases like cancer.

Conclusion

In conclusion, our study demonstrates that weighted gene co-expression networks can be used to predict both the

likelihood of essentiality and the likelihood of conservation for their gene members within the entire network. Furthermore, we show that these networks can be subdivided into identifiable modules, which themselves can often contain enough individual information to indicate which genes are more likely to be crucial for a particular process. We have indicated a means for testing whether this is true or not. Thus, we have illustrated how this process can be applied to common microarray data for the purpose of predicting both functional assignment as well as the ranking of gene significance in a particular process with which it is found to be associated. We believe that these kinds of predictions will be helpful for identifying gene targets for therapeutics and for directing multi-drug therapies.

Methods

Data sources

Yeast microarray data were obtained from the *Saccharomyces* Genome Database in ratio format [34]. In each case, microarray based gene expression analysis was previously reported to have been performed for the analysis of DNA Damage [18], Cell Cycle [17], and Environmental Response [19]. Not all of the data obtained from each datasets listed was used in these experiments. Because missing data is especially confounding for our analysis, arrays where greater than 5% of the gene expression information was missing were removed from our analysis. The exact number of arrays chosen to represent each dataset was: DNA Damage (n = 51), Cell Cycle (n = 44), and Environmental Response (n = 151). Essential genes were determined from the *Saccharomyces* Genome Deletion Project [35]. Genomic data for conducting the blastp searches were obtained from the following locations: *Neurospora* genome data (version 7) was obtained from the Broad Institute [36,37]. *Drosophila* genome data was obtained from the Berkeley *Drosophila* genome project [38], *C. elegans* genome data was obtained from the wormbase database [39], while both the human and mouse genomes were obtained from the UCSC genome browser [40]. The blastp algorithm was run locally.

Weighted Network construction

In gene co-expression networks, each gene corresponds to a node. The neighbors of a node i are the nodes that are connected to the node i . Two genes are connected by an edge with a weight indicating the connection strength. A gene co-expression network can be represented by an adjacency matrix $A = [a_{ij}]$, where a_{ij} is the weight of a connection between two nodes i and j . For all of the networks considered in this paper, the connectivity equals the sum of connection weights being considered.

Our analysis follows a general framework for construction of gene co-expression networks [16]. To transform the co-

expression measure (here the absolute value of the Pearson correlation matrix) into measures of pair wise connection strengths, one can make use of an adjacency function. The choice of the adjacency function determines whether the resulting network will be weighted (soft thresholding) or unweighted (hard thresholding). A widely used adjacency function is the signum function which implements 'hard' thresholding involving the threshold parameter tau. Specifically, $a_{ij} = \text{Ind}(|\text{cor}(x_i, x_j)| > \tau)$ where the indicator function takes on the value 1 if the condition is satisfied and 0 otherwise. Hard thresholding using the signum function leads to intuitive network concepts (e.g., the node connectivity equals the number of direct neighbors), but it may lead to a loss of information. For instance, if the threshold has been set to 0.8, there will be no connection between two genes if their correlation equals 0.79. To avoid the disadvantages of hard thresholding, Zhang and Horvath [16] proposed a 'soft' power adjacency function: $a_{ij} = |\text{cor}(x_i, x_j)|^\beta$ with the single parameter beta. The power adjacency function preserves the continuous nature of the gene co-expression information and, leads to results that are highly robust with respect to the choice of beta.

Zhang and Horvath [16] also proposed a criterion for choosing the power beta of the adjacency function which is based on the fact that despite significant variation in their individual constituents and pathways, metabolic networks have been found to display approximate scale free topology. The linear regression model fitting index R^2 can be used to quantify how well a network satisfies a scale-free topology. While the scale free model has only 1 parameter (γ), the truncated exponential model allows for 2 parameters γ and α . Empirically, we find that the two parameters of the truncated exponential model provide too much flexibility in curve fitting so that the R-squared values are always high [16]. Thus, we focus on the scale free topology model fitting index in this application and emphasize the major trend that the high connectivity genes are few in number, and that most genes have low connectivity (Figure 2A-2C). There is a natural trade-off between maximizing scale-free topology model fit (R^2) and maintaining a high mean number of connections: parameter values that lead to an R^2 value close to 1 may lead to networks with very few connections. These considerations motivated the scale-free topology criterion. To choose the parameters of an adjacency function: Only those parameter values that lead to a network satisfying scale-free topology at least approximately were considered (e.g. signed $R^2 > 0.80$). When considering the signum and power adjacency functions, we find the relationship between R^2 and the adjacency function parameter (tau or beta) is characterized by a saturation curve. In our applications, we use the first parameter value where saturation is reached as long as it is above 0.8. An $R^2 > .80$ results in networks that may only approximate the scale

free property as is highlighted by the Cell Cycle network in this manuscript.

For each dataset, we first selected the top 4000 most varying transcripts based on the standard deviation. The network adjacency between 2 genes i and j was defined as a power of the correlation coefficient between the corresponding gene expression profiles x_i and x_j , i.e. $a_{ij} = |\text{cor}(x_i, x_j)|^\beta$ where the power beta is chosen using the scale free topology criterion. From this function, the 3000 most-connected genes were selected for further detailed network analysis. Using the 4000 most varying transcripts was not very restrictive since the original data are derived from yeast arrays and therefore only contained only about 6000 genes in total. Many of these genes needed to be removed due to high rates (>40%) of missing data for a gene across a data set. By definition, module genes are highly connected with the genes of their module (i.e. module genes tend have relatively high connectivity). Thus, for the purpose of module detection, restricting the analysis to the most connected genes does not lead to major information loss for the key points of this presentation. However, there may be applications where genes with relatively low connectivity are biologically interesting so that gene filtering based on connectivity would lead to information loss. The adjacency matrix was also visualized as a graph and the number of topological properties of this graph examined. Important network derived concepts include: hub (a node with many strong connections to other nodes), topological overlap (the interconnectivity between two nodes), modules (tightly connected subsets of the network), connectivity (the sum of the adjacencies, i.e. $k_i = \sum_{u \neq i} a_{iu}$) and intramodular connectivity (the sum of the adjacencies within a module). One potential drawback of soft thresholding is that the network cannot be visualized directly because the number of relationships to display becomes intractable. A soft adjacency matrix specifies pair wise connection strengths and all nodes are connected. To visualize the network with Pajek, one needs to threshold the connection strengths. The resulting dichotomized adjacency matrix can be considered as an unweighted network approximation of the weighted network. Therefore, to make a visual representation, only the strongest correlations (0.9 or greater) were drawn in these renderings. Subsequent to this, networks were then visualized with Pajek [20,41].

Definition of Topological Overlap and Modularity

Topological overlap of two nodes (genes) reflects their relative interconnectivity. For a network represented by an adjacency matrix $A = [a_{ij}]$, $a_{ij} \in [0,1]$, a well-known formula for defining topological overlap was given by [4], and it can be extended to weighted networks[16]:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (1)$$

where, $l_{ij} = \sum_{u \neq i, j} a_{iu} a_{uj}$ denotes the number of nodes to

which both i and j are connected, and k is the number of connections of a node, with $k_i = \sum_{u \neq i} a_{iu}$ and $k_j = \sum_{u \neq j} a_{ju}$.

Since $a_{ij} \in [0,1]$, we find that

$$l_{ij} = \sum_{u \neq i, j} a_{iu} a_{uj} \leq \min\left(\sum_{u \neq i, j} a_{iu}, \sum_{u \neq i, j} a_{uj}\right) = \min\left(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{ju}\right) - a_{ij}$$

. Thus, $l_{ij} + a_{ij} \leq \min\left(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{ju}\right)$. Since $1 - a_{ij} \geq 0$, we

find that ω_{ij} is a number between 0 and 1. There are 2 reasons for adding $1 - a_{ij}$ to the denominator in the topological overlap matrix: 1) in this form, the denominator can never be 0 and 2) for an *unweighted* network, one can show that $\omega_{ij} = 1$ only if the node with fewer links satisfies two conditions: (a) all of its neighbors are also neighbors of the other node, i.e. it is connected to all of the neighbors of the other node and (b) it is linked to the other node.

In contrast, $\omega_{ij} = 0$ if i and j are unlinked and the two nodes don't have common neighbors. Further, the topological overlap matrix is symmetric, i.e. $\omega_{ij} = \omega_{ji}$, and its diagonal elements are set to 0 (i.e. $\omega_{ii} = 0$). The rationale for considering this similarity measure is that nodes that are part of highly integrated modules are expected to have high topological overlap with their neighbors [4].

It is important to note that the definition of topological overlap matrix (equation 1) was taken from the methods supplement of reference [4] since there is a typo in the primary text. Different definitions of the topological overlap matrix are possible, but we have found empirically that many of these possible generalizations lead to very similar modules [42].

Modules are groups of nodes that have high topological overlap. Module identification is based on the topological overlap matrix $\Omega = [\omega_{ij}]$ defined as (1). To use it in hierarchical clustering, it is turned into a dissimilarity measure

by using the standard approach of subtracting it from one (i.e. the topological overlap based dissimilarity measure is defined by $d_{ij}^{\omega} = 1 - \omega_{ij}$).

In networks involving few nodes, modules can easily be identified by inspecting the network, but for large networks involving hundred of nodes, it is useful to generate a topological overlap matrix (TOM) plot, which we review in the following. A TOM plot provides a "reduced" view of the network that allows one to visualize and identify network modules. The TOM plot is a color-coded depiction of the values of $[d_{ij}^{\omega}]$, described above.

Based on the TOM based dissimilarity matrix we can use hierarchical clustering to discriminate one module from another. We used a dynamic cut-tree algorithm for automatically and precisely identifying modules in hierarchical clustering dendrogram [43]. The algorithm makes use of the internal structure in a dendrogram and clips branches of the dendrogram. The algorithm is based on an adaptive process of cluster decomposition and combination and the process is iterated until the number of clusters becomes stable. No claim is made that our module construction method is optimal and future research should compare different method for cluster identification. Each module represents a group of genes with similar expression profiles across the samples and the expression profile pattern is distinct from those of the other modules. In each case, the modules were then assessed for enrichment in ontology terms based on Fisher's exact test as described in [25]. These modules and their relative connectivities were then separated out for further analysis.

Based on the network modules, we can define the intramodular connectivity (k_{in}) for each node i (i.e., the number of neighbors in the same module as node i). The connectivity of node i across the whole network is denoted as k_{all} .

Blastp

The translated sequences of the entire yeast genome were each blasted against the entire translated sequences of the Human, Mouse, Fly, *Neurospora* and *C. elegans* genomes. In each case, the best hit was determined using blastp, and the $\log(e \text{ score})$ scores were averaged across the 5 different comparisons in a manner similar to what was published by [22]. These average blastp scores were then averaged for each point in (Figure 2H–2K, 3F–3H, 4F–4H, 5F–5H). This method is not necessarily an optimal measure of gene conservation, and it is used here simply as an approximate measure of gene conservation, that is inde-

Table 2: Summary of tests for network predictive power

Module name	DNA Dmg		Env Stress		Cell Cycle	
	Ess Test	Cons Test	Ess Test	Cons Test	Ess Test	Cons Test
Entire Network	0.23 1.0E-39	-0.15 1.3E-19	0.21 1.0E-39	-0.22 1.4E-39	0.2 1.0E-39	-0.18 4.0E-29
rRNA proc	0.2 6.3E-5	-0.1 0.049	0.28 1.0E-39	-0.091 0.0066	0.22 0.00011	-0.23 8.3E-5
Prot Synth	-0.12 0.014	-0.07 0.15	-0.16 0.0035	-0.042 0.46	0.019 0.7	-0.099 0.055
Ubiq	0.41 1.8E-10	-0.28 2.9E-5	0.49 0.0022	-0.16 0.34	-0.044 0.44	-0.032 0.59

*Spearman correlation and associated p-values for each network and module tested

pendent of annotation. Future research should study the advantages and disadvantages of using different measures for gene conservation in network analysis.

Analysis of module overlap

We used a hyper-geometric p-value to evaluate the significance of the overlap of the module genes across the 3 data sets. Suppose N_1 , N_2 , N_3 out of N genes are in rRNA processing module of the DNA Damage data, the Environmental Response data, and the Cell Cycle data, respectively. We then denote by 'overlap' the number of genes in the rRNA module that overlap in the 3 data sets.

To arrive at an upper limit of the p-value, we first computed pair wise p-values as follows. Denote as p_{12} the p-value that by randomly selecting N_2 genes from a pool of N genes with N_1 module 1 genes, \geq overlap selected genes are in the rRNA-processing module in both data sets. This p-value can easily be computed using the hypergeometric distribution (dhyper function in R). N_1 and N_2 were then swapped to obtain a second p-value, which we denoted as p_{21} . The maximum of those two values can be used to assess the significance of overlap between data sets 1 and 2. Similarly, we computed a p-value of overlap between data sets 1 and 3, and data sets 2 and 3. The minimum of the 3 resulting p-values provided an upper limit for assessing the significance of overlap between all 3 data sets, i.e. the final p-value was computed as $p = \min(\max(p_{12}, p_{21}), \max(p_{13}, p_{31}), \max(p_{23}, p_{32}))$

Data and software

Prepared Data and R code are available online [44].

Abbreviations

CC, cell cycle; DD, DNA damage; ER, environmental response

Authors' contributions

MC and SFN conceived of and developed the project, and wrote the final manuscript. MC analyzed the data, participated in the construction of all the software tools used in this study and drafted the manuscript and figures. SH, BZ authored the majority of the R code needed for develop-

ing these weighted networks, developing network measures, and providing crucial statistical guidance for this project. ZF helped with the statistical analysis. PM and SFN both contributed key ideas and suggestions as the project progressed. All authors participated in the development of the ideas presented in this manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors extend their thanks to the SGD for providing data for this analysis. Also, special mention should be made of Allen Day, Ally Perlina and Andy Yip for their input during the analysis of this project. The work was supported by a grant from the NCI (U01 CA881273) as well as the Henry Singleton Brain Tumor Program, the Elliott Family Foundation in memory of Neal Elliott, the Ziering Family Foundation in memory of Sigi Ziering, Accelerate Brain Cancer Cure, and Art of the Brain. MRJC was supported by the UCLA Tumor Cell Biology Training Program, funded by NIH/ National Cancer Institute 5T32CA09056. This work was also supported by the UCLA DNA Microarray Facility and the NINDS/NIMH Microarray Consortium.

References

- Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**:C47-52.
- Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**:101-113.
- van Noort V, Snel B, Huynen MA: **The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model.** *EMBO Rep* 2004, **5**:280-284.
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-654.
- Bergmann S, Ihmels J, Barkai N: **Similarities and differences in genome-wide expression data of six organisms.** *PLoS Biol* 2004, **2**:E9.
- Carter SL, Brechbuhler CM, Griffin M, Bond AT: **Gene co-expression network topology provides a framework for molecular characterization of cellular state.** *Bioinformatics* 2004, **20**:2242-2250.
- Albert R, Jeong H, Barabasi AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406**:378-382.
- Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachet S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S,

- Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelm J, Winzler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M: **Functional profiling of the *Saccharomyces cerevisiae* genome.** *Nature* 2002, **418**:387-391.
10. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41-42.
11. Provero P: **Gene networks from DNA microarray data: centrality and lethality.** *arXiv:cond-mat/0207345* 2002, **2**.
12. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863-14868.
13. van Noort V, Snel B, Huynen MA: **Predicting gene function by conserved co-expression.** *Trends Genet* 2003, **19**:238-242.
14. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**:249-255.
15. Quackenbush J: **Genomics. Microarrays--guilt by association.** *Science* 2003, **302**:240-241.
16. Zhang B, Horvath S: **A General Framework for Weighted Gene Co-Expression Network Analysis.** *Statistical Applications in Genetics and Molecular Biology* 2005, **4**.
17. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273-3297.
18. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.** *Mol Biol Cell* 2001, **12**:2987-3003.
19. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11**:4241-4257.
20. Batagelj VMA: **Pajek - Program for Large Data Analysis.** *Connections* 1998, **21**:47-57.
21. Fruchterman TMJ, Reingold EM: **Graph Drawing by Force-Directed Placement.** *Software - Practice and Experience* 1991, **21**:1129-1164.
22. Featherstone DE, Brodie K: **Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network.** *Bioessays* 2002, **24**:267-274.
23. Pereira-Leal JB, Audit B, Peregrin-Alvarez JM, Ouzounis CA: **An exponential core in the heart of the yeast protein interaction network.** *Mol Biol Evol* 2005, **22**:421-425.
24. Csanyi G, Szendroi B: **Fractal-small-world dichotomy in real-world networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2004, **70**:16122.
25. Hosack DA, Dennis GJ, Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4**:R70.
26. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.
27. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.** *Nat Genet* 2001, **29**:482-486.
28. Rung J, Schlitt T, Brazma A, Freivalds K, Vilo J: **Building and analysing genome-wide gene disruption networks.** *Bioinformatics* 2002, **18 Suppl 2**:S202-10.
29. Wagner A: **Estimating coarse gene network structure from large-scale gene perturbation data.** *Genome Res* 2002, **12**:309-315.
30. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CWV, Bussey H, Andrews B, Tyers M, Boone C: **Systematic genetic analysis with ordered arrays of yeast deletion mutants.** *Science* 2001, **294**:2364-2368.
31. Haggarty SJ, Clemons PA, Schreiber SL: **Chemical genomic profiling of biological networks using graph theory and combinations of small molecule perturbations.** *J Am Chem Soc* 2003, **125**:10543-10545.
32. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW: **Chemogenomic profiling: identifying the functional interactions of small molecules in yeast.** *Proc Natl Acad Sci U S A* 2004, **101**:793-798.
33. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C: **Global mapping of the yeast genetic interaction network.** *Science* 2004, **303**:808-813.
34. **Saccharomyces Genome Database.** [<http://db.yeastgenome.org/cgi-bin/SGD/expression/expressionConnection.pl>].
35. **Saccharomyces Genome Deletion Project.** [http://www-sequence.stanford.edu/group/yeast_deletion_project/data_sets.html].
36. **Version 7 Neurospora genome data from the Broad Institute.** [http://www.broad.mit.edu/annotation/fungi/neurospora_crassa_7/index.html].
37. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B: **The genome sequence of the filamentous fungus *Neurospora crassa*.** *Nature* 2003, **422**:859-868.
38. **Drosophila genome project.** [<http://www.fruitfly.org/sequence/download.html>].
39. **Wormbase database.** [<ftp://ftp.wormbase.org/pub/wormbase/>].
40. **UCSC genome browser.** [<http://genome.ucsc.edu/downloads.html>].
41. **Pajek Algorithm.** [<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>].
42. **Generalized Topological Overlap Matrix and its Applications.** [<http://www.genetics.ucla.edu/labs/horvath/GTOM/>].
43. **Dynamic Tree Cut Algorithm.** [<http://www.genetics.ucla.edu/labs/horvath/binzhang/DynamicTreeCut/>].
44. **Prepared Data and R Code.** [<http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/MarcCarlson/>].

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

