Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

Putting genetic interactions in context through a global modular decomposition

Jeremy Bellay, Gowtham Atluri, Tina L. Sing, et al.

Genome Res. published online June 29, 2011
Access the most recent version at doi:10.1101/gr.117176.110

Supplemental Material	**http://genome.cshlp.org/content/suppl/2011/05/27/gr.117176.110.DC1.html**
P<P	Published online June 29, 2011 in advance of the print journal.
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Copyright © 2011 by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

Method

Putting genetic interactions in context through a global modular decomposition

Jeremy Bellay,[1] Gowtham Atluri,[1] Tina L. Sing,[2,3] Kiana Toufighi,[4] Michael Costanzo,[3,5,6] Philippe Souza Moraes Ribeiro,[1] Gaurav Pandey,[7] Joshua Baller,[1] Benjamin VanderSluis,[1] Magali Michaut,[3,5] Sangjo Han,[3,5] Philip Kim,[3,5] Grant W. Brown,[2,3] Brenda J. Andrews,[3,5,6] Charles Boone,[3,5,6] Vipin Kumar,[1] and Chad L. Myers[1,8]

[1]*Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota 55455, USA;* [2]*Department of Biochemistry, University of Toronto, Toronto, Ontario M5S 3E1, Canada;* [3]*The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada;* [4]*Centre for Genomic Regulation (CRG), Barcelona 08003, Spain;* [5]*Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada;* [6]*Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 3E1, Canada;* [7]*Plant and Molecular Biology Department, University of California, Berkeley, California 94720-3102, USA*

Genetic interactions provide a powerful perspective into gene function, but our knowledge of the specific mechanisms that give rise to these interactions is still relatively limited. The availability of a global genetic interaction map in *Saccharomyces cerevisiae*, covering ~30% of all possible double mutant combinations, provides an unprecedented opportunity for an unbiased assessment of the native structure within genetic interaction networks and how it relates to gene function and modular organization. Toward this end, we developed a data mining approach to exhaustively discover all block structures within this network, which allowed for its complete modular decomposition. The resulting modular structures revealed the importance of the context of individual genetic interactions in their interpretation and revealed distinct trends among genetic interaction hubs as well as insights into the evolution of duplicate genes. Block membership also revealed a surprising degree of multifunctionality across the yeast genome and enabled a novel association of *VIP1* and *IPK1* with DNA replication and repair, which is supported by experimental evidence. Our modular decomposition also provided a basis for testing the between-pathway model of negative genetic interactions and within-pathway model of positive genetic interactions. While we find that most modular structures involving negative genetic interactions fit the between-pathway model, we found that current models for positive genetic interactions fail to explain 80% of the modular structures detected. We also find differences between the modular structures of essential and nonessential genes.

[Supplemental material is available for this article.]

A genetic interaction is generally defined as multiple genetic perturbations whose combination results in a phenotype that is unexpected given the phenotypes of the individual perturbations. Genetic interactions reveal specific redundancies or dependencies within the genetic network and can provide a powerful means for functional characterization. Recently, a genome-scale study of digenic quantitative genetic interactions was completed in baker's yeast using Synthetic Genetic Array (SGA) technology (Costanzo et al. 2010). This study produced a quantitative survey of genetic interactions covering ~5.4 million double mutants, including *negative interactions*, instances where the double mutant was less fit than expected, as well as *positive interactions*, which are instances where the double mutant was healthier than expected. Because of its global coverage (~30% of all possible digenic interactions) and unbiased nature, the study revealed that genetic interactions often span seemingly disparate cellular functions and may give new insights into the global functioning of a given gene and the contexts in which the gene is utilized.

[8]Corresponding author.
E-mail cmyers@cs.umn.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.117176.110.

Despite the apparent power of genetic interactions in characterizing gene function, the principles that account for the structure of the genetic interaction (GI) network remain unclear. Like a protein-protein interaction (PPI), a genetic interaction between two genes implies a functional relationship between those genes. However, unlike the PPI network, there is no obvious functional interpretation of a single genetic interaction, either negative or positive. The existence of a genetic interaction between two genes does not imply that the two gene products interact physically, or that the genes are temporally coexpressed; it simply suggests they share some kind of functional relationship. Because the individual interactions do not have a precise functional interpretation, the larger structure and organization become essential in understanding the implications of the GI network.

The structure of the genetic interaction network and how it relates to other networks has been addressed in several previous studies (for example, St. Onge et al. 2007; Ma et al. 2008; Ulitsky et al. 2008). Generally, most previous proposals fall within the *modular hypothesis* of genetic interactions, meaning that genetic interactions can be explained through genes' membership in various types of functional "modules" in the cell, whether protein complexes, pathways, or otherwise. Two primary models have been proposed to explain the occurrence of positive and negative genetic interactions. Negative interactions are thought to arise

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

between functionally redundant pathways such that deleting any pair of genes spanning across the pathways results in a significant reduction in fitness (Kelley and Ideker 2005). In contrast, positive interactions are thought to arise *within* pathways or physical complexes due to the fact that a second deletion in an already compromised pathway or complex does not cause an additional fitness defect (Schuldiner et al. 2005; Collins et al. 2006). We will refer to these two hypothesized structures as the *negative between-pathway* model and the *positive within-pathway* model for nonessential genes, respectively (Supplemental Fig. S1). Pathways or protein complexes containing essential genes can behave differently and often exhibit negative genetic interactions within members of a pathway or protein complex, which we will refer to as the *negative within-pathway* model for essential genes (Bandyopadhyay et al. 2008; Baryshnikova et al. 2010; Costanzo et al. 2010). It should be noted that the between-pathway and within-pathway models were developed with deletion or null mutants in mind, and it is not clear how partial loss of function perturbations of essential genes should fit within these models. These models have been used successfully to map some protein complexes and pathways when genetic interactions are combined with PPI networks (Bandyopadhyay et al. 2008; Ma et al. 2008; Ulitsky et al. 2008) and to extrapolate genetic interactions from partial profiles (Qi et al. 2008).

With the publication of a substantial fraction of the yeast GI network (Costanzo et al. 2010), the various models for how genetic interactions relate to functional modules can be evaluated explicitly and globally in an unbiased fashion. These models predict specific structures that occur within the genetic interaction network. Specifically, the between- and within-pathway interactions for nonessential genes should give rise to *biclusters*—two sets of genes, either overlapping or disjoint, in which every gene from one set interacts with all genes of the other. These block structures have been observed in existing genetic interaction networks, and in fact, a simple hierarchical clustering analysis of interaction profiles will reveal a number of such structures (Supplemental Fig. S1). However, a single grouping of the genes may reveal only a small fraction of these structures, and thus, many bicluster structures will be missed. We wished to not only find the most salient block structures in the interaction network but also to take full advantage of the global nature of the current yeast network and discover *all* such block structures and therefore fully account for the role of modularity in the GI network.

The discovery of block patterns in two-dimensional data is usually referred to as *biclustering* in the biological sciences and *block modeling* within the social sciences (Doreian et al. 2005) (see, for example, Fig. 1A). A wide variety of biclustering algorithms have been proposed (particularly in the context of gene expression analysis), including the seminal method proposed by Cheng and Church, ISA (and PISA), PLAID, spectral clustering, and SAMBA (Cheng and Church 2000; Ihmels et al. 2002; Tanay et al. 2002; Kluger et al. 2003; Kloster et al. 2005; Turner et al. 2005), to name a few. These methods have proven effective and useful in the analysis of gene expression data. A recent algorithm proposed in Pu et al. (2008) was specifically designed for finding block structures in quantitative genetic interaction data. While a wide variety of approaches are represented in these algorithms, they are generally designed to find the large coherent blocks common to gene expression data and are mostly based on heuristics that begin with a randomly generated initial bicluster. Because of this, they are prone to rediscover the prominent biclusters many times, while missing smaller or more subtle patterns. We desired an approach that will exhaustively find all such biclusters to enable us to directly test the applicability of the between-pathway and within-pathway genetic interaction models. Furthermore, as has been previously observed, the methods developed in the context of gene expression data are not effective when applied to genetic interaction networks, likely due to the more discrete and fine-grained characteristics of the modular structure in these networks (Pu et al. 2008).

In the present study, we employed an approach based on an algorithm from the field of association rule mining that is guaranteed to find all biclusters of sufficient size. We applied this approach to exhaustively discover any significant block structure in the genetic interaction network enriched for either negative or positive interactions. This global decomposition of genetic interactions into discrete modular structures provided a powerful basis for assessing the connection between genetic interaction network topology and gene function. We find that genetic interaction hubs can be clearly differentiated into distinct classes based on their modular structure and that there are functional differences between modular genetic interactions and those that occur outside of modular structures. Moreover, we show that module membership provides an effective means of dissecting the functional contexts that explain a gene's genetic interactions and that the genetic interaction network contains evidence for a high degree of multifunctionality across the yeast genome. Finally, the modular structure of the GI network gives surprising evidence both for and against the traditional between-pathway and within-pathway models of genetic interactions, and we explore the structural differences between essential and nonessential genes.

Results

Data set and bicluster discovery

We used the recent genetic interaction data from Costanzo et al. (2010) in which 1712 "query" genes (including 334 conditional or hypomorphic alleles of essential genes) were screened against 3885 nonessential "array" genes, resulting in interaction scores for ~5.4 million double mutants. Interactions were filtered based on their confidence, and the resulting set of significant negative and positive interactions was used as a basis for our modular decomposition approach (see Methods).

In order to study the role of modularity within the GI network, we wished to discover *all* significant biclusters, each of which should indicate interactions spanning within or across two functional modules. To this end, we developed an approach utilizing the Apriori algorithm (Agrawal et al. 1993) from the field of association rule mining to exhaustively discover all biclusters. We then used a nonparametric statistical assessment (described in the following paragraph) to filter out biclusters that could be explained by the degree distribution alone. This approach, which we call XMOD (eXhaustive MOdular Discovery), is guaranteed to discover all complete bipartite graphs of sufficiently large size and significance as designated by the user.

We applied XMOD to negative and positive interactions separately. To account for bipartite graph structures potentially arising by chance, we randomized the genetic interaction network by switching edge targets, thus randomizing the network structure while preserving the degree distribution (see Methods; Supplemental Fig. S2). Even with randomized edges, the existence of a few highly connected genes (i.e., network hubs) was sufficient to produce numerous biclusters by chance. To filter out biclusters that

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

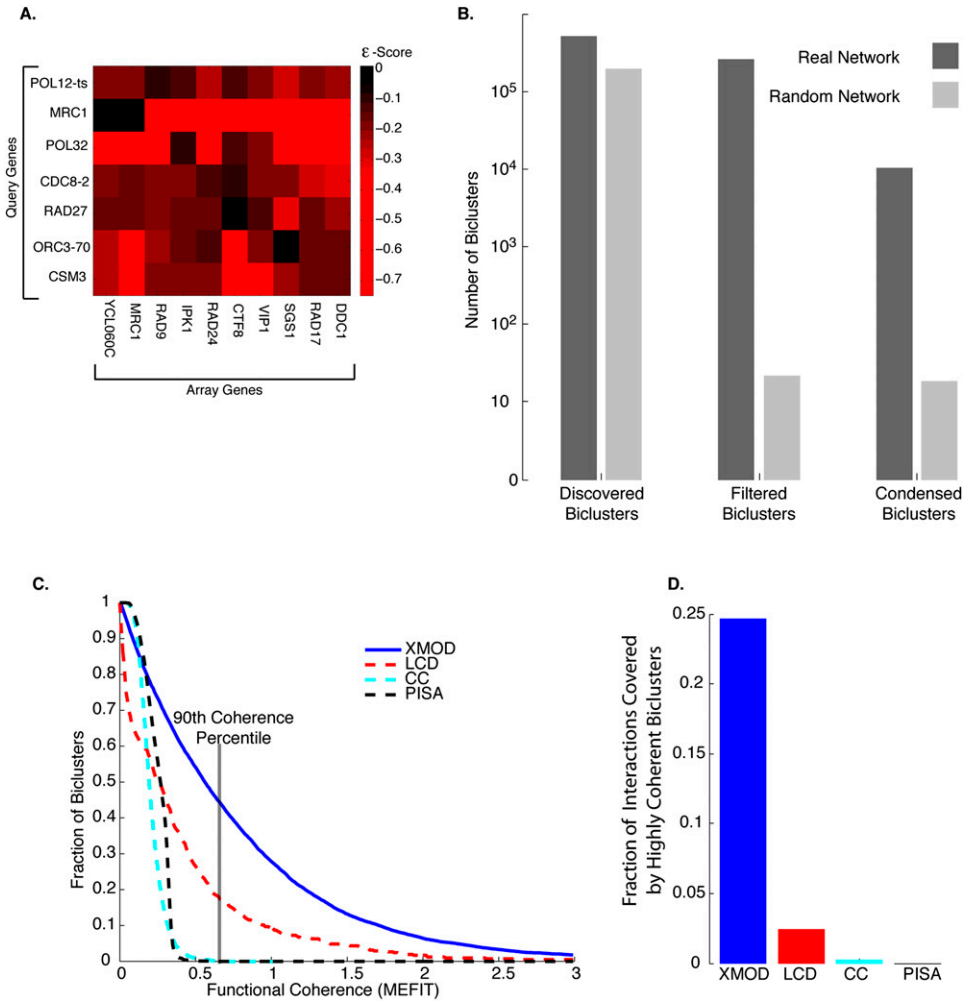

Figure 1. (*A*) A bicluster from the SGA genetic interaction network—an enrichment for (negative) genetic interactions between a set of query genes and a set of array genes. This bicluster was found by combining XMOD biclusters using MCL (see Methods). (*B*) A comparison of biclusters found on the real SGA network and the average number of biclusters found on randomized networks. While a significant number of biclusters are found on the random networks (Discovered Biclusters), statistical filtering leaves, on average, 20 random biclusters versus ~200,000 real biclusters (Filtered Biclusters). After controlling for redundancy, there were ~10,000 real biclusters (Condensed Biclusters) but again only ~20 random biclusters on average. (*C*) A plot of the proportion of biclusters that meet or exceed a given functional relatedness score. XMOD condensed blocks is solid blue, LCD (Pu et al. 2008) is dashed red, CC (Cheng and Church 2000) is dashed cyan, and PISA (Ihmels et al. 2002; Kloster et al. 2005) is dashed black. We also compare the methods after removing overlapping clusters from all methods in Supplemental Figure S7. (*D*) The percentage of interactions covered by biclusters whose functional relatedness score was in the top 90th percentile of the MEFIT network [the line is marked in (*C*)].

are likely to arise simply due to the degree distribution (and thus not biologically meaningful), we calculated a score for each bicluster—the product of the probabilities of each edge occurring independently conditioned on the degrees of the two interacting genes (see Methods). We assumed that the interactions in biologically relevant biclusters should not be independent of each other and therefore have a lower score than biclusters that randomly occur due to the degree distribution. Using the scores of blocks generated from 10 random networks as a null distribution, we selected score cutoffs that resulted in an estimated false discovery rate of ~0.1% for negative interaction biclusters and 1% for positive interaction biclusters (Fig. 1B; Supplemental Figs. S3, S4; see Methods for details). This resulted in a final set of 256,502 negative biclusters and 2194 positive biclusters. After removing biclusters with >40% overlap (see Methods for details), we are left with 10,459 negative biclusters and 615 positive distinct biclusters (as compared to 20 negative, on average, and 6 positive for the random networks). In the following, we will refer to this set of biclusters as "condensed", while referring to the full set as "filtered". Both sets of biclusters are available for browsing at the following website: http://csbio.cs.umn.edu/XMOD.

Over half of genes screened in Costanzo et al. (2010) appear in at least one bicluster (2571 out of 4458); 2462 are contained in negative biclusters, and 1139 are contained in positive biclusters. The mode of the size distribution of the negative biclusters is 6 query genes by 5 array genes, while the mode of the positive biclusters is 4 query genes by 3 array genes. The distribution of sizes of the biclusters can be seen in Supplemental Figure S5.

Due to its basis in an association rule mining framework, the XMOD algorithm is guaranteed to find all biclusters of sufficient size and significance, which is not typical of other biclustering algorithms. Indeed, in comparison to other biclustering methods, XMOD outperforms existing approaches both in terms of the functional coherence of identified biclusters and overall network

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

coverage of the biclusters. We evaluated the functional coherence of the modules represented on the array and query sides of each bicluster generated by XMOD by using a global functional network inferred from a large number of gene expression data sets (Huttenhower et al. 2006) (see Methods). Even when controlling for redundancy, biclusters found by XMOD exhibit more functional coherence than other biclustering methods (Fig. 1C), a result which held when we controlled for redundancy among the other methods as well (Supplemental Fig. S6). These results were similar for another measure of functional coherence, Gene Ontology (GO) semantic similarity (Supplemental Fig. S7). Additionally, functionally coherent biclusters found by XMOD cover an order of magnitude more interactions than other previous methods. For example, 25% of interactions (either negative or positive) are contained in at least one XMOD bicluster with a coherence score in the top 90th percentile of the score when scored in the MEFIT network, while <3% are contained in similarly coherent biclusters produced by the next closest method (Fig. 1D).

We explored how missing values affect the coverage of genetic interactions by biclusters by introducing random missing values into the SGA network (Supplemental Fig. S7). By randomly changing 10% of measured SGA scores to missing values, XMOD covered 70% of the interactions that were covered on the full, unperturbed network. Thus, the interactions covered by XMOD biclusters represent a lower bound of the interactions contained in modular structures.

A modular decomposition of the genetic interaction network

The exhaustive nature of the network decomposition provided by XMOD allows for an investigation into which interactions exist as parts of larger modular structures and which exist as individual interactions. We first asked what portion of the genetic interactions observed in the global yeast network are part of a larger modular structure and therefore "covered" by a bicluster in the filtered set and which interactions remain uncovered or "isolated" (Fig. 2A). Out of 85,714 negative interactions at the chosen cutoff (see Methods), 49,983 (58%) were contained in a bicluster structure, suggesting that a sizable fraction of negative genetic interactions should be interpreted in the context of a larger modular structure. A smaller proportion of the positive interactions were contained in biclusters (6802 out of 35,858, or 19%).

We first hypothesized that genetic interactions that are isolated would be more likely to be experimental false positives. Indeed, we found some evidence for this; the false discovery rate (FDR) for interactions covered by biclusters is significantly lower than that on isolated interactions. Using small-scale validation experiments for individual interactions that were performed in Tong et al. (2004), we estimated a false discovery rate of 17.6% on covered interactions, which was less than one-third the FDR on the isolated interactions (54%). Taking these FDRs into account, we estimate the coverage of true negative interactions by biclusters to be closer to ~75%. Only 19% of positive interactions are contained in biclusters, indicating perhaps that the positive interactions contain more false positives, but we also note the possibility that positive interactions occur more frequently among modules and complexes too small to be detected by our method (the minimum bicluster size we could reliably recover was 3×3).

Despite the apparent increase in the false positive rate among isolated interactions, we found evidence suggesting that many of the isolated negative interactions are indeed real but simply functionally distinct from modular interactions. For example, we found a striking enrichment for duplicate gene pairs among the isolated negative interactions; of 106 negatively interacting duplicates, 80 appear in the isolated set, which is far above the expected rate ($P < 2 \times 10^{-15}$, binomial test). The tendency of duplicate interactions to appear as isolated interactions is consistent with the expectation that they are cases of individual redundancy, not redundancy between larger gene modules [see, for example, Musso et al. (2008); VanderSluis et al. (2010)].

We were intrigued that a significant fraction of the duplicate pairs exhibiting negative genetic interactions with each other appeared as parts of larger modular structures (biclusters). Strikingly, the duplicates with interactions in these modular structures were significantly more divergent in terms of sequence identity ($P < 0.038$, Wilcoxon rank-sum test; Fig. 2B). These two distinct classes of negative interactions among duplicate pairs suggest that redundancy among duplicate pairs may arise from at least two different functional scenarios. In the first case, duplicates with a high degree of functional similarity specifically compensate for the loss of one another, while in the second case they appear to have diverged into entirely different functional modules. In the latter scenario, the duplicate gene pair is not unique in their negative genetic interaction; all gene pairs spanning the two modules exhibit interactions. It is interesting to note that there are also striking functional differences in the duplicates in each of these different classes. Duplicates that compensate each other on an individual basis are enriched for ribosomal proteins ($P < 0.04$, hypergeometric CDF), while the duplicates with evidence for modular divergence are enriched for protein transport and localization ($P < 10^{-3}$, hypergeometric CDF). In addition, duplicates with isolated interactions are enriched for whole genome duplicates ($P < 4 \times 10^{-4}$, hypergeometric CDF). The relative enrichment for isolated pairs among duplicates persists after removing the "ribosome" GO term ($P < 10^{-6}$, binomial test). Without ribosomal genes, the median sequence identity of covered pairs is lower than isolated pairs (47.5 versus 51), but this difference is no longer significant ($P > 0.3$, Wilcoxon rank-sum test).

Using modularity to dissect the topological properties of the genetic interaction network

Genetic interaction hubs have been shown to be important for cellular processes from a variety of perspectives (Costanzo et al. 2010). For example, based on the recent global map of genetic interactions, Costanzo et al. report that genetic interaction hubs are highly enriched for pleiotropic and multifunctional genes. We speculated that, based on our decomposition of genetic interactions into isolated or modular structures, we might find different classes of genetic interaction hubs. By simply finding the proportion of genetic interactions covered by biclusters for a given gene, we derived a measure of "profile structure," i.e., the fraction of a given gene's interaction profile that was contained within biclusters of the filtered set.

On the set of all screened genes, profile structure is generally well correlated with genetic interaction degree ($\rho = 0.76$; $P < 10^{-10}$). However, profile structure demonstrates surprisingly different properties than degree, particularly when we restrict the analysis to genes in the top 10th percentile of genetic interactions (genetic interaction "hubs"). For example, hubs exhibit a wide variety of different levels of profile structure (Fig. 3A). Profile structure shows significant correlation with multifunctionality (i.e., the number of distinct GO annotations) ($\rho = 0.27$; $P < 10^{-7}$; Fig. 3B), while there is no significant correlation between multifunctionality and degree

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

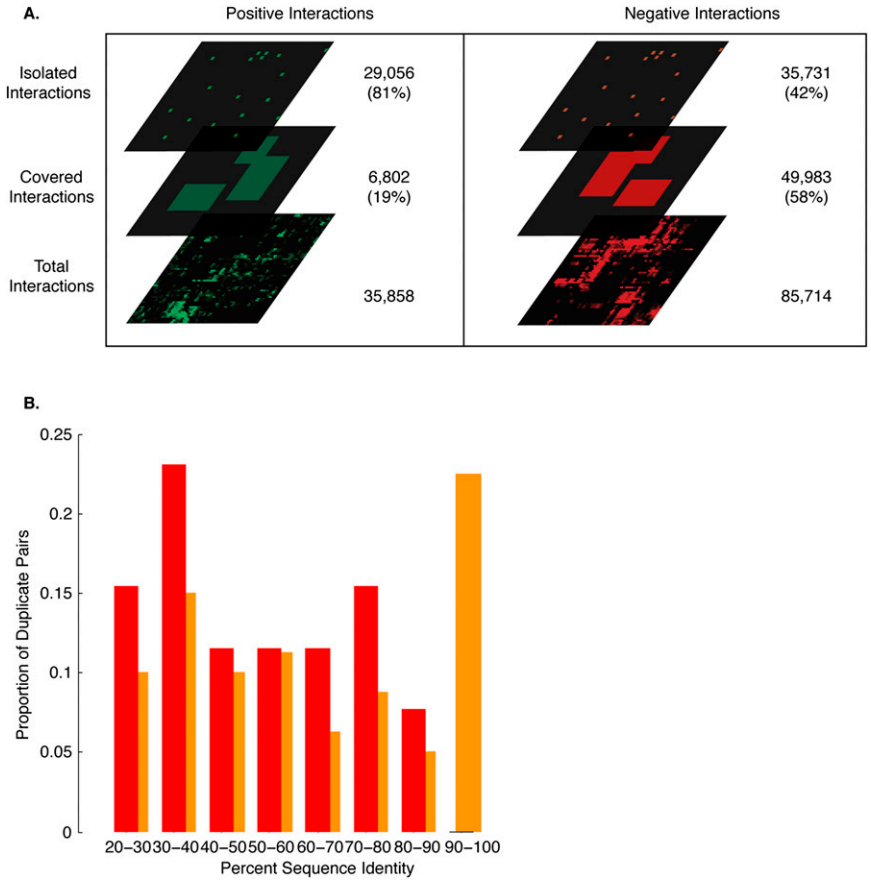

Figure 2. (*A*) The coverage of genetic interactions by XMOD biclusters. The interactions are divided into "covered" interactions that occur within a bicluster and "isolated" interactions that occur outside of a bicluster. Fifty-eight percent of negative interactions are covered compared to only 19% of positive interactions. (*B*) A plot of the proportion of negatively interacting duplicate pairs [either "isolated" (orange) or "covered" (red)] against the amino acid sequence identity of the duplicate pair.

($P > 0.13$) on the top 10th percentile hubs. Profile structure is also a better predictor of the number of chemical genetic interactions (Hillenmeyer et al. 2008) ($\rho = 0.37$; $P < 3 \times 10^{-8}$) than is degree ($\rho = 0.12$; $P > 0.08$) on interaction hubs. Finally, profile structure is also well correlated with flexible protein disorder (Bellay et al. 2011), which has been associated with transient physical interactions and signaling.

One might infer from these observations that profile structure is simply a truer measure of gene "importance" on hubs than is interaction degree, but we find that, while degree has a strong correlation with fitness defect ($\rho = 0.38$; $P < 10^{-12}$; Fig. 3C), profile structure is actually negatively correlated with fitness defect on the hubs ($\rho = -0.23$; $P < 10^{-5}$). In addition, degree is well correlated with phenotypic capacitance [a measure of pleiotropy described by Levy and Siegal (2008)] ($\rho = 0.23$; $P < 10^{-5}$), while profile structure has no significant correlation ($P > 0.6$). Finally, negative degree is correlated with positive genetic interaction degree ($\rho = 0.35$; $P < 10^{-11}$), while profile structure shows no significant correlation ($P > 0.5$).

We hypothesize that these seemingly paradoxical results indicate that genetic interaction hubs (and possibly genetic interactions in general) are the result of two distinct types of phenomena—those that arise due to specific genetic buffering between functional modules, and those that are a result of more general instability. The hubs whose interaction profiles are composed of many different modular interactions truly reflect functional versatility and, furthermore, that the specific functions of these *structured hubs* are directly buffered by other modules. On the other hand, a predominance of unstructured genetic interactions is not indicative of direct genetic buffering, and these *unstructured hubs* may indicate more indirect aggravation phenotypes. This difference is reflected in the fitness defects caused by a single hub deletion in each class; the structured hubs apparently are healthier single mutants, reflecting the fact that they are buffered despite their widespread functionality, while the unstructured hubs tend to exhibit fitness defects even when independently deleted.

The distinction between structured and unstructured hubs is reminiscent of date and party hubs defined on protein interaction networks and coexpression data [debated in a number of papers, including Han et al. (2004) and Batada et al. (2006)]. In short, party hubs are hypothesized to be proteins that are simultaneously expressed with their interaction partners, while date hubs are differentially expressed from their interaction partners and therefore thought to act in multiple cellular locations and times. In this spirit, we used a weighted functional network based on coexpression across a large collection of expression studies (Huttenhower et al. 2006) to develop an *expression incoherence score* (EIS) as a measure of how well a gene's functionally related neighbors are themselves coexpressed (see Methods; a higher score indicates greater regulatory independence). Profile structure is correlated with EIS on hubs ($\rho = 0.18$; $P < 1 \times 10^{-3}$; Fig. 3B), and this trend persists after controlling for degree ($\rho = 0.17$; $P < 2 \times 10^{-3}$). On the other hand, controlling for profile structure destroys the correlation between degree and EIS ($P > 0.1$). This indicates that structured hubs are coexpressed with genes in a wide variety of functions, while the coexpression partners of unstructured hubs tend to be coexpressed more often as single functional units. Along these lines, we found that profile structure among query genes could be used to differentiate between singlish and multi-interface protein interactions hubs (Kim et al. 2006) ($P < 0.01$, KS-test), while there was no significant difference in interaction degree ($P > 0.05$, KS-test). Singlish hubs have few interfaces and therefore are presumed to bind their partners in a temporally disparate fashion, while multi-interface hubs can bind their partners simultaneously, again suggesting that structured hubs tend to participate in many diverse functions more frequently than unstructured hubs.

Extracting functional context from bicluster membership

Beyond identifying distinct classes of genetic interaction hubs, our modular decomposition provides a direct summary of functional contexts to which a particular gene contributes. Thus, our exhaustive network decomposition provides an effective basis for exploring

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

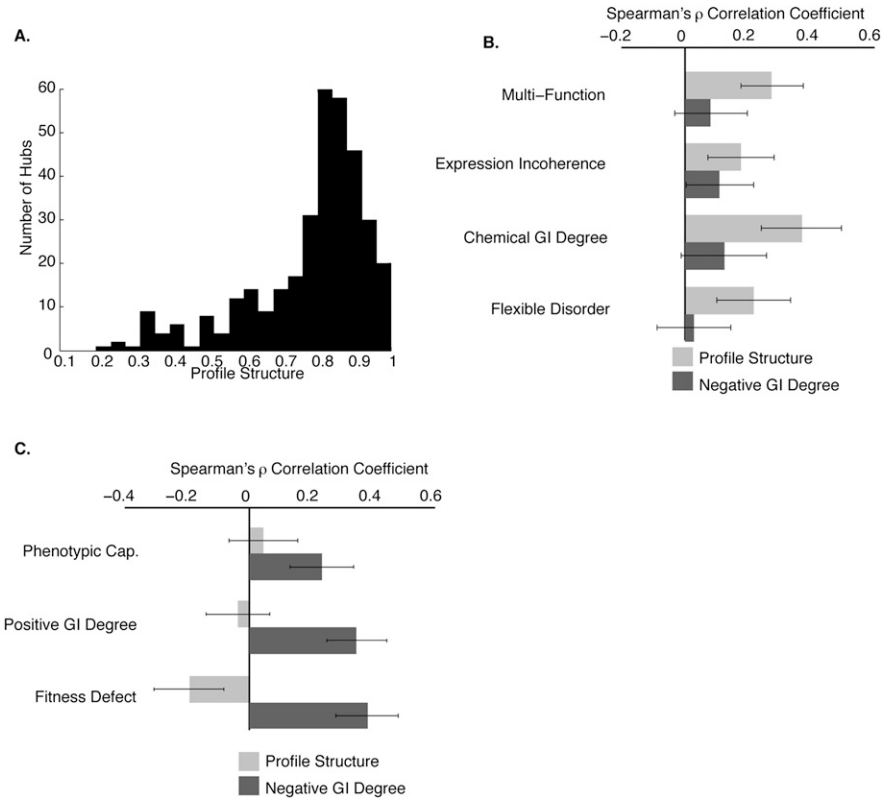

Figure 3. (*A*) A histogram of profile structure among hubs. (*B*) The Spearman correlation coefficients of profile structure and negative GI degree against gene properties that are better correlated with profile structure than degree. Error bars represent 95% confidence intervals and were derived through bootstrapping. (*C*) The Spearman correlation coefficients of profile structure and negative GI degree against gene properties better correlated with negative GI degree than profile structure. Error bars represent 95% confidence intervals and were derived through bootstrapping.

module memberships of each gene, as well as assessing the reuse of genes across many different modules. Given the fact that most of our biclusters represent interactions between functionally coherent groups (see Fig. 1C, for example), we assigned a putative functional association to each bicluster by finding the significant Gene Ontology enrichments (biological process) among the set of genes appearing with the candidate gene on the query side of the bicluster (see Methods; Fig. 4A). Applying this procedure to all genes and totaling the unique enrichments among the set of condensed biclusters for each gene reveals a surprisingly high degree of associated functions or contexts across the genome (Fig. 4B). Interestingly, 74 genes had associations with more than 20 distinct processes, reflecting widespread reuse of genes in a large number of different functional contexts. Some of the genes with the highest number of associations had annotations that reflected their involvement in disparate processes, but there are many examples where the genetic interaction data suggest prevalent multifunctionality that is not reflected in the current GO annotations. Many of them are, however, supported by independent functional genomic data, suggesting the genetic interaction biclusters are indeed capturing real functional modules (Supplemental Fig. S9). It is important to note that our results do not necessarily suggest a large number of different biochemical or molecular functions for each protein, which are ultimately limited by the protein's structure. Instead, these associations revealed by each bicluster may reflect different contexts in which the same molecular function is reused.

One particularly interesting example is the gene *VIP1*, which encodes one of two yeast inositol pyrophosphate synthases required for synthesis of hexakisphosphate (IP_6) and heptakisphosphate (IP_7) (Fig. 4C; Mulugu et al. 2007). *VIP1* synthesizes a form of IP_7 that is distinct from the isomers produced by Kcs1, the other yeast inositol pyrophoshate kinase (Mulugu et al. 2007). While *KCS1* has been implicated in many different cellular processes (Bennett et al. 2006), the role for *VIP1* is less well defined. Despite the pleiotropic nature of these important signaling molecules, Vip1 has thus far only been annotated to the "inositol heptakiphosphate kinase activity" GO term. Insights into the physiological role of Vip1 were based on studies of Asp1, its functional ortholog in the fission yeast, *Schizosaccharomyces pombe* (Mulugu et al. 2007). These studies suggested that, similar to Asp1, the inositol pyrophosphate kinase activity of Vip1 was important for maintaining cellular integrity, morphology, and interactions with the ARP complex. Consistent with the multifunctional nature of inositol pyrophosphate synthases, we identified *VIP1* associated with several otherwise nonoverlapping biclusters, enriched for a total of 13 distinct functional annotations (Fig. 4C).

One of the enrichments obtained from the biclusters suggested a previously unappreciated role for *VIP1* in DNA replication and repair (Fig. 4C). Using biclusters combined with Markov Clustering (MCL) (see Methods; Fig. 1A), we found this association also included another inositol phosphate signaling enzyme, Ipk1. We explored a role for Vip1 and Ipk1 in DNA synthesis and repair by examining *VIP1* and *IPK1* sensitivity to the DNA alkylating agent, methyl methanesulfonate (MMS). Strains lacking either *VIP1* or *IPK1* exhibited reduced growth in the presence of MMS (Fig. 4D). Furthermore, flow cytometric analysis revealed that the slow growth phenotype was due to impaired progression through S phase in the presence of MMS, as opposed to general MMS toxicity, as *vip1*Δ and *ipk1*Δ cells progress more slowly through S phase in the presence of MMS, taking longer to reach a 2C DNA content than do wild-type cells (Fig. 4E). Together, these data suggest a role for *VIP1* and *IPK1* in DNA replication and repair and confirm functional predictions stemming from our clustering analysis. Importantly, analysis of genetic interactions using the method described in our study support a pleiotropic role for *VIP1* in polarity and DNA replication.

"Between-within" distinctions in the GI network—the surprising abundance of positive bicliques and negative clique-like structures

The between-pathway model of negative interactions implies that negative genetic interactions should occur in "bicliques" (biclusters in which the genes on the query and array sides do not overlap), while under the within-pathway model, positive interactions should occur in "cliques" (biclusters in which the query and array

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

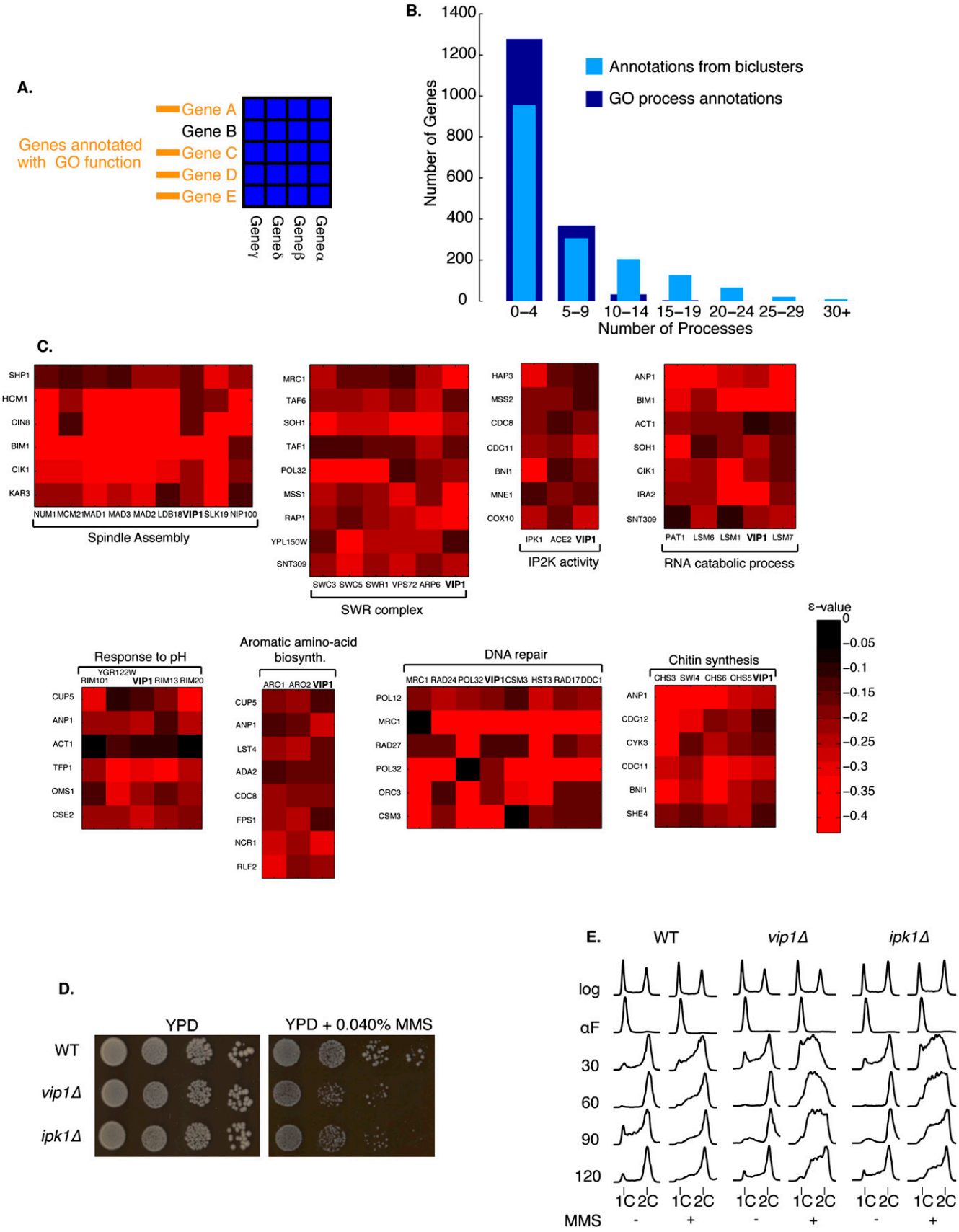

Figure 4. (Legend on next page)

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

genes have significant overlap). Our method for exhaustive discovery of bicluster structures allowed for an unbiased and explicit evaluation of the current models for explaining modularity in genetic interaction networks. To evaluate these models, we divided the condensed biclusters into "bicliques" (biclusters with no overlap between query and array genes) and "quasi-cliques" (biclusters with significant overlap between query and array genes hereafter referred to as q-cliques), based on the overlap in gene sets (see Methods for details). The between-pathway model of negative interactions predicts a large number of bicliques among structures containing negative interactions, while the within-pathway model for positive interactions predicts that positive interactions should largely occur within q-cliques.

Surprisingly, we found that the within-pathway model for positive interactions explained only a small fraction of the modular structure we observed. Specifically, of the 572 positive biclusters that were classed as either q-clique or biclique, only 18% of them (104) were q-cliques as predicted by the within-pathway model, while the remaining 82% were bicliques (Fig. 5A), suggesting positive interactions were frequently spanning across distinct functional modules. There were proportionally fewer instances of q-clique structures among the negative interaction biclusters; of the 8532 total negative biclusters, 9% (762) represented q-cliques, while the remaining 91% were biclique structures. Thus, while the canonical within-pathway model explains a larger portion of positive biclusters than negative biclusters, most positive and negative biclusters appear to fit the between-pathway model.

Given the surprising prevalence of positive bicliques, which indicate positive interactions spanning across functional modules, we searched for further explanation for these modular structures by considering other measures of functional relatedness within the gene sets that define the array side of the biclusters. Positive q-cliques were highly enriched for both protein interactions and coexpression in comparison to positive bicliques (Fig. 5B), leading us to believe that most positive q-cliques were, in fact, examples of interactions occurring within pathways or protein complexes. In contrast, positive bicliques exhibited a much lower degree of coexpression and protein interactions on average, which is consistent with the hypothesis that they are likely not connecting closely related genes in the same protein complex or specific pathway. Interestingly, several experimentally confirmed examples of genetic suppression spanning across protein complexes were recently detailed in Baryshnikova et al. (2010), suggesting these structures are, in fact, more common than previously appreciated.

We next turned our attention to the distribution of negative biclusters identified by our approach. In contrast to the canonical understanding of positive genetic interactions, the current between-pathway model for negative interactions does appear to explain a large percentage (91%) of the identified modular structures. However, the interactions composing the negative q-cliques, which comprise a surprising 9% of the total negative biclusters (762 clusters in total), appear to be highly distinct from those supporting the between-pathway model. Previous studies, including our own (Bandyopadhyay et al. 2008; Baryshnikova et al. 2010), have noted negative interactions within protein complexes containing essential genes. Indeed, there is an enrichment of protein interactions involving essential genes among negatively interacting q-cliques, but they only explain a small minority of the q-cliques that are actually observed. For example, only 7% of genes appearing on both sides of negative q-cliques also appear together in an essential protein complex [as defined in Baryshnikova et al. (2010)]. Surprisingly, negative q-cliques are substantially depleted for protein-protein interactions when compared to negative bicliques (Fig. 5B). This is exactly the opposite of the trend observed in the positive biclusters and suggests that many negative q-cliques are neither explained by protein complexes nor by the between-pathway model. Despite their lack of enrichment for protein-protein interactions, genes within negative q-clique structures did show a tendency toward coexpression when compared with bicliques, based on an integration of expression data sets across a variety of environmental conditions (Fig. 5B; Huttenhower et al. 2006). Thus, while the negative q-cliques do not appear to represent single protein complexes or pathways, they appear to represent genes that are functionally related and expressed in a coordinated fashion. Among genes that appear on both sides of a negative bicluster, we found a striking enrichment for processes supporting regulation of the cell, including cell cycle ($P < 10^{-10}$, hypergeometric CDF), chromosome segregation ($P < 10^{-5}$, hypergeometric CDF), and several DNA replication and repair checkpoints (see Supplemental Table S1). A similar set of annotations was observed in Costanzo et al. (2010) among genes with significantly more negative interactions than positive, and indeed, we found that genes that appear on both sides of negative biclusters also have a heightened ratio of negative to positive interactions compared to other genes ($P < 10^{-6}$, Wilcoxon rank-sum test). We hypothesize that these negative interaction q-cliques represent a distinct cellular phenomenon from the usual between-pathway compensation associated with negative genetic interactions.

The set of genetic interactions produced in Costanzo et al. (2010) include 334 essential genes as queries in the form of 214 temperature sensitive (TS) alleles and 120 DAmP alleles. The role played by essential genes in the genetic interaction network has been speculated upon, but the expected structure of the genetic interactions of "partial knockouts" like TS and DAmP alleles is unclear. Both sets of essentials were only used as queries and therefore cannot drive the formation of q-cliques in our study. They appear in many biclusters, and 165 negative and 42 positive condensed biclusters have only essential genes as queries. To ascertain the nature of between- and within-pathway interactions for essential genes, we considered the ratio of the bicluster memberships $\frac{\#\text{q-cliques}}{\#\text{q-cliques}+\#\text{bicliques}}$ for every query gene (Fig. 5C). Deletion mutants appear more often in positive q-cliques (27.2%) than in negative q-cliques (6%), while the reverse was true for TS alleles, which appeared more often in negative q-cliques (11%)

Figure 4. (*A*) Functional enrichment for each gene is derived from significant functional enrichments of other genes that appear in the same biclusters (see Methods for details). (*B*) A histogram of the number of processes associated with each gene as determined by the distinct functional enrichments of the biclusters of which that gene is a member (light blue) and the number of GO annotations currently given to that gene (dark blue). (*C*) Some example biclusters that contain the gene *VIP1*. *VIP1* only has one GO annotation, but appears in a variety of highly enriched functional contexts. (*D*) *vip1*Δ and *ipk1*Δ exhibit reduced viability in the presence of MMS. Serial 10-fold dilutions of *vip1*Δ and *ipk1*Δ were spotted onto YPD or YPD + 0.040% MMS and incubated at 30°C for 3 d. (*E*) *vip1*Δ and *ipk1*Δ exhibit DNA replication defects in the presence of MMS. Cells were arrested in G_1 and released synchronously into the cell cycle in either the presence or absence of 0.035% MMS. Histograms represent the cell cycle distribution after release from G_1 arrest ± MMS for the specified times. Positions of cells with 1C and 2C DNA contents are indicated.

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

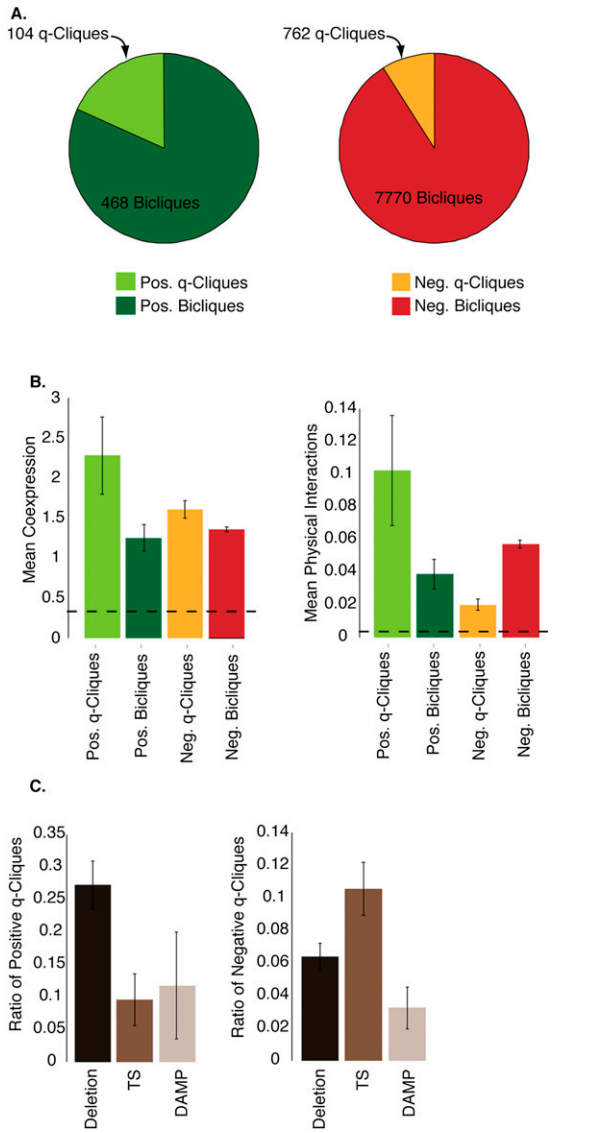

Figure 5. (*A*) The proportion of q-cliques and bicliques for positive and negative interactions. Nine percent of negative biclusters are q-cliques, while 18% of positive biclusters are q-cliques. (*B*) The *left* bar plot shows the mean of the mean coexpression of array genes within positive q-cliques, positive bicliques, negative q-cliques, and negative bicliques. The dashed line is the mean MEFIT score between all gene pairs. The *right* plot shows the mean of the mean number of physical interactions among the array genes of the positive q-cliques, positive bicliques, negative q-cliques, and negative bicliques. The dashed line is the background rate of PPIs between gene pairs. (*C*) The ratio of q-cliques to q-cliques + bicliques for different types of query alleles. Deletion mutants appear in a high proportion of positive q-cliques, while TS mutants appear in more negative q-cliques. All error bars are bootstrapped 95% confidence intervals.

than in positive q-cliques (10%). DAmP alleles seem to appear rarely in either positive or negative q-cliques (12% and 3%, respectively), probably because they have fewer interactions than TS alleles.

To understand the reason for this dichotomy between deletion mutants and TS alleles, it is important to note that the interactions we are observing occur between TS alleles and deletions of nonessential genes, never between TS alleles themselves, since these have not yet been screened on a large scale (Costanzo et al. 2010). Furthermore, the q-cliques that contain TS alleles are classified as q-cliques only because there are deletions that overlap on the query and array side (since the TS alleles can only be present on one side). In fact, when we consider biclusters whose support depends only on deletion mutants (size at least 3×3, not including TS or DAmP alleles), we find that there are proportionally as many negative bicliques (13%) (Supplemental Fig. S10) and similar enrichments of coexpression and protein-protein interactions (Supplemental Figs. S11–S12) as observed on the complete data set. This suggests that the difference we observed is more likely a result of the types of processes that contain TS alleles, not a consequence of the fact that TS alleles represent partial loss of function mutations rather than deletions. In this light, our results suggest that processes that contain essential genes (and thus TS alleles) tend to exhibit the negative q-clique structures more frequently than the positive q-clique structures. This is consistent with earlier observations that reported that protein complexes containing essential genes often exhibit negative interactions between the nonessential components as opposed to positive genetic interactions (Bandyopadhyay et al. 2008; Baryshnikova et al. 2010).

The difference in q-clique frequency between deletion and TS alleles also provides insight about the nature of positive q-cliques. Specifically, these are most often nonessential cellular components that depend on the presence of all genes involved and whose loss of function incurs a slight fitness deficit. Thus, there are typically no essential genes associated with these modules (and thus, no TS alleles), which may explain the reduction in positive q-cliques for TS alleles. Instead, the positive interactions for TS alleles more often connect across functionally distinct pathways. Interestingly, Baryshnikova et al. (2010) actually confirmed several cases of genetic suppression (extreme positive interactions) involving TS alleles that spanned across protein complexes, which we expect is typical of TS alleles' positive interactions, based on our analysis.

Interactive online software for browsing biclusters

Both the full and condensed set of biclusters are available at: http://csbio.cs.umn.edu/XMOD/.

Discussion

Despite the long history of combining genetic perturbations to gain insight into gene function, the interpretation of individual genetic interactions remains challenging. This is, to a large extent, a consequence of their usefulness; they capture broad, often nonlocal functional relationships, and this can make them hard to interpret. The approach we have taken here is to use the structural context of the network to better understand the broader functional setting in which a particular interaction occurs. We can ask if a given interaction occurs between two modules, within a module, or outside of any modular structure. Once the interaction is placed in a functional context, we can employ other known systems characteristics (protein-protein interactions, expression studies, sequence data) to further explain the interaction.

This structural context approach to genetic interaction interpretation relies critically on an exhaustive cataloging of all modular structures of the network. Here, we have employed techniques from association rule mining that are guaranteed to find all biclusters in the genetic interaction network above a certain size or quantitative coherence. This class of techniques is relatively rare in application to biological data sets to date, and there are a number of tradeoffs in using association rule mining. First, the data is dis-

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

cretized, therefore neglecting the resolution provided by SGA. Second, the Apriori algorithm only discovers complete bipartite graphs, so missing values can break large structures into numerous smaller structures. This results in wide spread redundancy in the discovered patterns, with the same pattern being discovered with slight variations numerous times. Because Apriori patterns are often small and redundant, post-processing is often required if one hopes to discover entire complexes or pathways (we provide such an analysis below using MCL). However, we show that despite the large numbers, biclusters produced with this approach are more functionally coherent than those of other biclustering algorithms and provide vastly higher coverage than previous methods. Moreover, the biclusters retain these properties after simple post-processing (in our case, the elimination of sufficiently overlapping biclusters). A slightly more sophisticated method of combining the biclusters would be valuable for module and pathway discovery but is beyond the scope of the present study. For example, using MCL to cluster the biclusters based on overlap (see Methods; Fig. 1A), we found that *VIP1* and *IPK1* both interact coherently with DNA repair pathways as was confirmed in Figure 4, D and E. Further investigation of strategies for effectively summarizing the large collection of biclusters would be a fruitful direction.

Through the use of our exhaustive method, we are able to determine what modular structures each interaction belongs to, and if the interaction belongs to any modular structure at all. By distinguishing between isolated and modular interactions, we found that duplicates that have high sequence and functional similarity are almost always without larger modular structure, indicating that they buffer each other at an individual gene level. The interactions of more diverged duplicate pairs (in terms of amino acid sequence) reside in larger modular structures, indicating that, while they may still compensate for each other, they can only do so at a modular level. This has clear implications for models of duplicate evolution and suggests that these two classes of negative genetic interactions between duplicate pairs should be treated differently.

The biclusters allowed us to explain and explore the elusive topic of multifunctionality and pleiotropy on a global scale. Through bicluster membership, we are able to differentiate two trends that lead to genetic interaction hubs, which we refer to as "structured" and "unstructured" hubs, that imply that genetic interaction hubs acquire their interactions for at least two distinct reasons—through involvement and buffering of multiple processes and contexts, or through causing a general instability in the cell. By counting functional enrichments among the blocks in which each gene appears, we observed a striking prevalence of multifunctionality across the genome. This has two important implications. First, it provides a global, unbiased assessment of gene reuse across modules, a fundamental systems-level property that is likely to have implications in organisms beyond yeast. Second, the actual biclusters associated with each pleiotropic gene actually point to the specific functional contexts in which it participates. Consequently, we explored *VIP1*, an inositol pyrophosphate kinase, which was known to be widely pleiotropic but whose specific role in these functions was poorly characterized. Based on a specific clustering with genes involved in DNA replication and repair, we were able to experimentally confirm one of these roles by demonstrating MMS sensitivity and abnormal cell cycle progression. The remaining functional roles for *VIP1* suggested by the many modules will require further experimental validation, and these and many other modules corresponding to other genes are publicly available for browsing at http://csbio.cs.umn.edu/XMOD/xmod.html.

Recently, genetic interactions have been interpreted in terms of the structure of the gene ontology (Michaut et al. 2011), and it was found that most monochromatic genetic interaction patterns within GO terms can be explained by protein complexes. Our biclusters provide another perspective for addressing this question as they also represent monochromatic groups of genes, only defined using a data-driven, rather than a GO annotation-driven, approach. We repeated an analysis similar to that described in Michaut et al. (2011) by counting the number of our biclusters (either bicliques or quasi-cliques) that were explained by protein complexes. We found that only 2344 negative condensed blocks (out of a total of 10,549) have a majority of array genes annotated to a single protein complex, and 8933 of the condensed blocks would have been found, not counting genes annotated to the same complex multiple times, indicating that only a minority of our structures can be explained by protein complexes. This represents an interesting contrast to the earlier findings, suggesting that there is evidence for monochromatic structure outside of protein complexes, but that GO annotations are simply too coarse to capture these specific modular structures. Indeed, we found that the genetic interaction network appears to contain a great deal of structure not yet annotated in GO, as discussed above (Fig. 4B). Future large-scale experimental validation of specific modules appearing in biclusters would be of significant interest to improve the resolution of function annotations.

The actual modular structure of the genetic interaction network is quite different from what was previously anticipated based on the canonical within-pathway model of positive genetic interactions. In fact, most positive interactions do not exist within single modules or protein complexes but instead appear to span across modules. Furthermore, there are a surprising number of within-module negative interactions (q-cliques) that do not appear to be associated with protein complexes, a phenomenon which has not been previously described. These show clear enrichment for chromosome segregation and cell cycle processes, which suggests that this form of genetic redundancy may play a unique role in the context of these functions. For example, the strong enrichment for processes such as chromosome segregation makes it tempting to speculate that these interactions may arise due to general sensitivity to perturbation in fragile systems that depend intricately on, for example, the balancing of tension forces. Such processes might produce a general functional neighborhood of sensitivity to double perturbations, as opposed to the typically more direct between-module characteristic observed for other negative genetic interactions. Goldstein previously speculated that redundancy of this nature may be under evolutionary selection due to the demand for high-fidelity in processes like chromosome segregation (Goldstein 1993). A single deletion reduces the fidelity but leaves cells able to divide, while two deletions reduce fidelity below a threshold necessary for cell viability. Further investigation of this apparently distinct form of negative genetic interaction is worthwhile.

Finally, we find that deletion alleles and TS alleles exhibit strikingly different modular memberships, with deletion alleles appearing in more within-pathway positive structures and TS alleles appearing in more within-pathway negative structures. Although beyond the scope of the current study, it is of interest to analyze the modular structures associated with interactions between combinations of TS alleles. Such networks do not yet exist on a large scale, but based on the results presented here, we anticipate interactions between combinations of partial loss-of-function alleles will exhibit distinct structural characteristics.

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

Methods

Selection of confident positive and negative interactions

We use the genetic interaction data set available in Costanzo et al. (2010). All interaction (ε) scores between −0.1 and 0.1 were set to zero. In addition, scores with *P*-value > 0.05 were also set to zero. This left 85,714 negative interactions and 35,858 positive interactions. The criterion for positive and negative interactions is slightly stricter than the recommended ε > |0.08| for purposes of algorithmic efficiency.

Association rule mining

We used the Apriori algorithm as described in Agrawal et al. (1993), for which an implementation is available at http://www.borgelt.net/apriori.html. We ran Apriori on the binary set of positive interactions and on the set of negative interactions described above. In the negative interactions, computational concerns forced us to restrict our search to biclusters that contained at least six query genes and three array genes, or three query genes and seven array genes. On the positive interactions, the matrix was sparse enough to allow us to find biclusters that had at least three query and three array genes.

Randomizing the genetic interaction network

We randomized the genetic interaction network while preserving the number of interactions of each gene but randomizing the edge targets. Thus, for each type of interaction (positive and negative), we randomly selected two edges and switched the array target genes (Supplemental Fig. S2). Genes were not allowed to have an edge with themselves. The randomization was run until each edge had been randomly reassigned. For each set of biclusters generated from real data, we generated 10 sets of biclusters from randomized networks.

Filtering random biclusters

To differentiate between biclusters that represented biological phenomena from those that simply resulted from the degree distribution, we compared biclusters found on the real network to those found on the randomized network using the following score for each bicluster:

$$s = \prod_{i,j} \left(\frac{\text{degree}(G_i)}{\#\text{Query Genes}}\right)\left(\frac{\text{degree}(G_j)}{\#\text{Array Genes}}\right),$$

where G_i (or G_j) is a gene on the array (query) side of a bicluster, and the index i (j) ranges over the array (query) genes of the bicluster, and #Query Genes (#Array Genes) are the total number of query (array) genes screened (i.e., possible interaction partners). The score s reflects the probability of a bicluster occurring if each edge occurrence only depends on the degrees of the two interacting genes independent of the other edges in the bicluster. Presumably, in biclusters occurring due to biological phenomena, the edges should not be independent, and therefore biological biclusters should have a lower s score than random blocks. For each bicluster consisting of q query genes and a array genes, we compared its score with the scores of biclusters of the same size generated from random networks. We took the s scores on the random biclusters to be the null distribution and took biclusters found on real data that achieved a 0.0001 empirical *P*-value, i.e., biclusters whose s score was in the bottom 0.01th percentile of the s scores of the random blocks of the same size. We found that 50% of the real negative biclusters and 6% of real positive biclusters have scores below the 0.01 percentile of biclusters of the same size from the random networks. This resulted in 256,502 negative biclusters and 2194 positive biclusters. In comparison, there was an average of 123,000 negative biclusters found on the random networks, but, on average, only 22 of these were below the 0.0001 empirical *P*-value.

Removing overlap from biclusters

For certain evaluations, we considered a set of biclusters with little or no overlap. To accomplish this, we first arranged the biclusters in descending order by area. Then, beginning with the first bicluster A, we remove all biclusters whose area overlap with A was greater than 0.4, where overlap between biclusters A and B was calculated using the following formula:

$$\text{Overlap} = \frac{|\text{Row}(A) \cap \text{Row}(B)| \times |\text{Col}(A) \cap \text{Col}(B)|}{\min(\text{Area}(A), \text{Area}(B))}.$$

We proceeded to the next largest remaining bicluster and removed all smaller biclusters that have >40% overlap with that bicluster and continued in this fashion until all biclusters were considered.

Evaluation of functional coherence

To evaluate functional coherence, we used the MEFIT functional network (Huttenhower et al. 2006) as a measure of functional similarity. The MEFIT network is based only on coexpression data and includes no genetic interaction data sets. For each bicluster (B) with m row genes (R) and n column genes (C), the functional coherence for the bicluster was defined as the following:

$$\text{Functional Coherence (MEFIT)} = \min\left(\frac{\sum_{i\neq j}\text{MEFIT}(\text{R}_i,\text{R}_j)}{m}, \frac{\sum_{i\neq j}\text{MEFIT}(\text{C}_i,\text{C}_j)}{n}\right).$$

In addition to MEFIT, we also calculated a similarity score based on GO coannotation using the semantic similarity (SEMSIM) described in Lin (1998). This functional coherence score was defined as the following:

$$\text{Functional Coherence (SEMSIM)} = \min\left(\frac{\sum_{i\neq j}\text{SEMSIM}(\text{R}_i,\text{R}_j)}{m}, \frac{\sum_{i\neq j}\text{SEMSIM}(\text{C}_i,\text{C}_j)}{n}\right).$$

Other biclustering methods

Cheng and Church

We used the method as described in Cheng and Church (2000) under a variety of parameter values and possible cutoffs in the data. We display the best result, which was with parameter settings of δ = 0.05 and σ = 1.01.

PISA

We used the implementation available in Kloster et al. (2005). Again, we tried a number of parameters and settled on $t_g = 4$ and 100 runs.

LCD

We used the method available in Pu et al. (2008), with $r = 0.001$ and 100 runs. Again, we considered several r values, and this provided the best performance.

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

Statistical references

Correlation

All correlation coefficients used in this study are Spearman's ρ rank correlation coefficient [see, for example, Hollander and Wolfe (1999)].

KS–test

The two sample Kolmogorov-Smirnov test [see, for example, Hollander and Wolfe (1999)].

Wilcoxon rank–sum test

The Wilcoxon rank-sum test as described in Hollander and Wolfe (1999).

GO enrichment

All GO term enrichment was performed using a hypergeometic CDF using Bonferroni correction for multiple hypothesis testing.

Other datasets utilized in this analysis

Duplicate genes

We used the set of duplicate genes provided in VanderSluis et al. (2010). Sequence identity we determined with BLAST (Altschul et al. 1990).

Protein–protein interactions

We combined the high-throughput data sets available in the AP/MS studies of Gavin et al. (2006) and Krogan et al. (2006) and the Y2H study of Yu et al. (2008). The AP/MS data set was taken from BioGRID, and the Y2H interactions were taken from the reported Y2H-Union.

Single mutant fitness

We used the fitness scores available in Costanzo et al. (2010).

Multifunctionality

We summed the number of GO annotations from the non-redundant GO term set provided in Myers et al. (2006).

Flexible disorder

We used the percent flexible disorder as defined in Bellay et al. (2011).

Expression incoherence score

This measure is one minus the clustering coefficient calculated on the MEFIT combined network (Huttenhower et al. 2006), where edges are gene pairs with an interaction score greater than 2 (~95th percentile). Define $E(N_i,N_j) = 1$ if there is an edge between N_i and N_j, and = 0, otherwise. Thus, the expression incoherence for a gene G with n neighbors $\{N_i\}$ is:

$$1 - \frac{\sum_{1 \le i < j \le n} E(N_i, N_j)}{\frac{n(n-1)}{2}}.$$

Singlish and multi–interface hubs

Singlish- and multi-interface hubs are an updated version of the set described in Kim et al. (2006) with the Structural Interaction Network (SIN) recently updated with *i*Pfam corresponding to Pfam release 21.0 (http://ipfam.sanger.ac.uk/), 2295 yeast pdb files (http://www.pdb.org), and 82,650 physical interactions in BioGRID 2.06 (http://www.thebiogrid.org).

Associating GO terms to genes using biclusters

We used the non-redundant set of GO terms found in Myers et al. (2006) to prevent multiple annotations of what is essentially the same process to a gene. For each bicluster, the array genes were tested for GO term enrichment with a *P*-value cutoff of 0.01 using Bonferroni correction for genes that appear in multiple biclusters.

Drug sensitivity assay

Strains were grown in 5 mL of YPD overnight at 30°C, serially diluted 10-fold, and spotted onto YPD plates either lacking or containing 0.040% MMS (Sigma-Aldrich). Plates were incubated at 30°C for 3 d.

Cell synchronization

Strains were grown in 35 mL of YPD at 30°C to an OD_{600} of 0.3–0.4. Alpha factor was added to 2.5 μM. After 2.5 h of incubation, cells were harvested, washed once with YPD, and released from G1 arrest by resuspension in YPD or YPD containing 0.035% MMS. Cultures were sampled at the indicated times after release for flow cytometry.

Flow cytometry

Cells were harvested, fixed in 70% ethanol, washed with water, and incubated in 0.5 mL of 0.2 mg/mL RNaseA in 50 mM Tris (pH 8.0) for 2 h at 37°C. Samples were then resuspended in 0.5 mL of 50 mM Tris (pH7.5) containing 2 mg/mL proteinase K, incubated at 50°C for 40 min, and resuspended in 0.5 mL FACS buffer (200 mM Tris (pH 7.5), 200 mM NaCl, and 78 mM $MgCl_2$.). A 0.1 mL sample of cells was then added to 0.5 mL of 50 mM Tris (pH 7.5) containing 2× Sytox Green (Molecular Probes). The samples were sonicated briefly and analyzed using a Becton Dickinson FACSCalibur.

Distinguishing quasi-cliques and bicliques

We defined q-cliques as biclusters in which there was at least a 20% overlap in the query and array genes that were screened on both the array and query side in Costanzo et al. (2010). Bicliques are biclusters with no overlap of array and query genes.

Combining biclusters with the Markov Clustering algorithm

Post-processing the biclusters with the Markov Clustering (MCL) algorithm is an effective means of summarizing the biclusters to allow for the creation for larger biclusters that often demonstrate structures that cannot be found in the purely positive or negative biclusters. We used the MCL software implementation from Van Dongen (2008). First, we constructed a matrix of the fraction of area overlaps of all biclusters found through the RCB method using the following formula for biclusters A and B:

$$\text{Overlap} = \frac{|\text{Row}(A) \cap \text{Row}(B)| \times |\text{Col}(A) \cap \text{Col}(B)|}{\min(\text{Area}(A), \text{Area}(B))}.$$

We then clustered the matrix with MCL using a cutoff of 0.66, with suggested parameters and a granularity of 7. This produces a set of clusters of biclusters, which is used to combine the biclusters. Finally, the biclusters are filtered for >40% redundancy, where they are first sorted by size and removed starting with the smallest biclusters.

Downloaded from genome.cshlp.org on June 29, 2011 - Published by Cold Spring Harbor Laboratory Press

Acknowledgments

J.B., B.V., and C.L.M. are partially supported by a seed grant from the University of Minnesota Interdisciplinary Informatics program, a seed grant from the Minnesota Partnership for Biotechnology and Medical Genomics program, a grant from the National Institutes of Health (1R01HG005084-01A1), and the National Science Foundation (DBI 0953881). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. J.B., M.C., C.L.M., B.J.A., and C.B. are also supported by a grant from the National Institutes of Health (1R01HG005853-01). G.W.B. was supported by the Canadian Institutes of Health Research grant MOP-84292, and T.L.S. received support from a Natural Sciences and Engineering Council of Canada Post-Graduate Scholarship.

References

Agrawal R, Imielinski T, Swami A. 1993. Mining association rules between sets of items in large databases. *Proc ACM SIGMOD Int Conf Manag Data* **22:** 207–216.

Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. 2008. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* **4:** e1000065. doi: 10.1371/journal.pcbi.1000065.

Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, Toufighi K, Youn J, Ou J, San Luis B, Bandyopadhyay S, et al. 2010. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat Methods* **7:** 1017–1024.

Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz B, Hurst LD, Tyers M. 2006. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biol* **4:** e317. doi: 10.1371/journal.pbio.0040317.

Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews B, Boone C, Bader G, Myers C, Kim P. 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* **12:** R14. doi: 10.1186/gb-2011-12-2-r14.

Bennett M, Onnebo S, Azevedo C, Saiardi A. 2006. Inositol pyrophosphates: Metabolism and signaling. *Cell Mol Life Sci* **63:** 552–564.

Cheng Y, Church GM. 2000. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* **8:** 93–103.

Collins SR, Schuldiner M, Krogan NJ, Weissman JS. 2006. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* **7:** R63. doi: 10.1186/gb-2006-7-7-r63.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. *Science* **327:** 425–431.

Doreian P, Batagelj V, Ferligoj A. 2005. *Generalized blockmodeling*. Cambridge University Press, Cambridge, UK.

Gavin A, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636.

Goldstein LSB. 1993. Functional redundancy in mitotic force generation. *J Cell Biol* **120:** 1–3.

Han JJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, et al. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430:** 88–93.

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St. Onge RP, Tyers M, Koller D, et al. 2008. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science* **320:** 362–365.

Hollander M, Wolfe D. 1999. *Nonparametric statistical methods*, 2nd ed. Wiley-Interscience, Hoboken, NJ.

Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. 2006. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22:** 2890–2897.

Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. 2002. Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31:** 370–377.

Kelley R, Ideker T. 2005. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23:** 561–566.

Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314:** 1938–1941.

Kloster M, Tang C, Wingreen N. 2005. Finding regulatory modules through large-scale gene-expression data analysis. *Bioinformatics* **21:** 1172–1179.

Kluger Y, Basri R, Chang JT, Gerstein M. 2003. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Res* **13:** 703–716.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440:** 637–643.

Levy SF, Siegal ML. 2008. Network hubs buffer environmental variation in *Saccharomyces cerevisiae*. *PLoS Biol* **6:** e264. doi: 10.1371/journal.pbio.0060264.

Lin D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pp. 296–304. Morgan Kaufmann Publishers, Burlington, MA.

Ma X, Tarone AM, Li W. 2008. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS ONE* **3:** e1922. doi: 10.1371/journal.pone.0001922.

Michaut M, Baryshnikova A, Costanzo M, Myers CL, Andrews BJ, Boone C, Bader GD. 2011. Protein complexes are central in the yeast genetic landscape. *PLoS Comput Biol* **7:** e1001092. doi: 10.1371/journal.pcbi.1001092.

Mulugu S, Bai W, Fridy PC, Bastidas RJ, Otto JC, Dollins DE, Haystead TA, Ribeiro AA, York JD. 2007. A conserved family of enzymes that phosphorylate inositol hexakisphosphate. *Science* **316:** 106–109.

Musso G, Costanzo M, Huangfu M, Smith AM, Paw J, San Luis B, Boone C, Giaever G, Nislow C, Emili A, et al. 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res* **18:** 1092–1099.

Myers C, Barrett D, Hibbs M, Huttenhower C, Troyanskaya O. 2006. Finding function: Evaluation methods for functional genomic data. *BMC Genomics* **7:** 187. doi: 10.1186/1471-2164-7-187.

Pu S, Ronen K, Vlasblom J, Greenblatt J, Wodak SJ. 2008. Local coherence in genetic interaction patterns reveals prevalent functional versatility. *Bioinformatics* **24:** 2376–2383.

Qi Y, Suhail Y, Lin Y, Boeke JD, Bader JS. 2008. Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res* **18:** 1991–2004.

Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, et al. 2005. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123:** 507–519.

St. Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G. 2007. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39:** 199–206.

Tanay A, Sharan R, Shamir R. 2002. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18:** S136–S144.

Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303:** 808–813.

Turner HL, Bailey TC, Krzanowski WJ, Hemingway CA. 2005. Biclustering models for structured microarray data. *IEEE/ACM Trans Comput Biol Bioinformatics* **2:** 316–329.

Ulitsky I, Shlomi T, Kupiec M, Shamir R. 2008. From E-MAPs to module maps: Dissecting quantitative genetic interactions using physical interactions. *Mol Syst Biol* **4:** 209. doi: 10.1038/msb.2008.42.

Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. *SIAM J Matrix Anal Appl* **30:** 121–141.

VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol* **6:** 429. doi: 10.1038/msb.2010.82.

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, et al. 2008. High-quality binary protein interaction map of the yeast interactome network. *Science* **322:** 104–110.

Received October 28, 2010; accepted in revised form May 24, 2011.