BIOINFORMATICS

Vol. 24 ISMB 2008, pages i330–i338
doi:10.1093/bioinformatics/btn160

Assessing the functional structure of genomic data

C. Huttenhower[1,2] and O.G. Troyanskaya[1,2,*]

[1]Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540 and [2]Lewis Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544, USA

ABSTRACT

Motivation: The availability of genome-scale data has enabled an abundance of novel analysis techniques for investigating a variety of systems-level biological relationships. As thousands of such datasets become available, they provide an opportunity to study high-level associations between cellular pathways and processes. This also allows the exploration of shared functional enrichments between diverse biological datasets, and it serves to direct experimenters to areas of low data coverage or with high probability of new discoveries.

Results: We analyze the functional structure of *Saccharomyces cerevisiae* datasets from over 950 publications in the context of over 140 biological processes. This includes a coverage analysis of biological processes given current high-throughput data, a data-driven map of associations between processes, and a measure of similar functional activity between genome-scale datasets. This uncovers subtle gene expression similarities in three otherwise disparate microarray datasets due to a shared strain background. We also provide several means of predicting areas of yeast biology likely to benefit from additional high-throughput experimental screens.

Availability: Predictions are provided in supplementary tables; software and additional data are available from the authors by request.

Contact: ogt@princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The technological developments of the past several decades have driven a continuing expansion of our understanding of molecular biology and a similar expansion in the analysis techniques applied to this data. In particular, genome-scale assays for coexpression (Eisen *et al*., 1998; Spellman *et al*., 1998), genetic interactions (Giaever *et al*., 2002; Tong *et al*., 2004), physical interactions (Gavin *et al*., 2002; Ho *et al*., 2002), protein localization (Huh *et al*., 2003) and regulatory networks (Harbison *et al*., 2004; Zhu and Zhang, 1999) have all opened up new opportunities for computational data mining that have been richly explored. Data such as these have been used in a variety of machine learning and other computational contexts (Franke *et al*., 2006; Jansen *et al*., 2003; Karaoz *et al*., 2004; Lee *et al*., 2004; Troyanskaya *et al*., 2003).

As the amount of available genome-scale data has continued to increase, it has become possible to ask higher level questions about the systems-level functional associations between entire pathways and processes. These associations represent the complex interplay between linked biological processes: DNA replication and mitosis are distinct cellular processes, for example, but they are functionally associated in their biological goals (cell division), regulation and genetic participants. Understanding this network of associations between processes is a critical link between functional relationships at the single-gene level and phenotypes at the organismal level.

By deriving an understanding of large-scale functional structure based directly on genome-scale datasets, we also gain an understanding of the data itself. An examination of the pathways and processes perturbed by whole-genome experiments allows those experimental results to be described in terms of their functional activity. For example, microarrays performed under conditions of heat shock and oxidative stress might both show functional activity related to an environmental stress response; this similarity of functional activity reveals biological commonalities between otherwise disparate experiments. By combining these two lines of inquiry—functional associations between processes and functional similarities between datasets—we gain insight into unexpected relationships in existing data, and we can direct experimenters to biological areas that are currently unexplored. All of these analyses deal with the high-level functional structure of genome-scale data and biological processes, which allows us to answer increasingly complex questions using the ongoing flood of high-throughput data.

We present such an analysis of functional associations among 141 biological processes and over 180 datasets (spanning >950 publications, >2300 microarray conditions, and several thousand interaction, localization and sequence-based data) in *Saccharomyces cerevisiae*, where a functional association entails co-operation, coregulation or other interaction between pathways and processes to perform a cellular task. These associations are derived by examining functional relationships between many individual genes, which are in turn predicted in a process-specific, probabilistic manner from heterogeneous data integration. This provides a global view of the functional structure of biological processes in yeast, including the degree of data-driven associations between processes, the experimental cohesiveness of gene behavior within each process, and the coverage of individual biological processes by currently available data. Likewise, we obtain measures of functional activity within each dataset—that is, which biological processes are covered by a dataset, independently of experimental platform. This high-level functional analysis technique is not specific to yeast and is extensible to any organism with a sufficiently large body of experimental data.

This analysis of functional structure produces a number of findings useful for guiding future experimental efforts and further computational studies. Specifically, we provide maps of data-driven associations between biological processes and of similar functional activities among datasets. By examining associations between processes, we observe several biological processes that could

[*]To whom correspondence should be addressed.

benefit from additional high-throughput data coverage, including *ion homeostasis and transport* and *mitochondrion organization*. We also highlight biological processes likely to be performed by currently uncharacterized genes (e.g. *autophagy*). Similar functional activities among datasets demonstrate commonalities in several large microarray studies and consistency between protein localization, synthetic lethality and protein–protein interaction screens. These similarities also expose specific biological relationships, such as a subtle effect due to strain background we discovered in three otherwise diverse microarray datasets. All of these relationships are fundamentally driven by similarities in gene and protein response across hundreds of datasets, and this high-level analysis of such large-scale functional structure is valuable for guiding future experimentation and in understanding systems-level associations among biological processes.

2 METHODS

In summary, we analyzed the large-scale structure of functional relationship networks predicted based on Bayesian integration of genomic data. Functional associations between biological processes from the Gene Ontology (GO; Ashburner *et al*., 2000) were derived by further integration and analysis of these networks in a context-sensitive manner. Functional activity information for each dataset was calculated during the integration process, and this was used to further characterize functional similarities between datasets. The resulting process/process, process/dataset and dataset/dataset association networks were mined for subgraphs and interactions of high weight. All network visualization was performed using Graphviz from AT&T (Gansner and North, 2000).

2.1 Data collection and gold standard generation

2.1.1 Data collection The data employed in this study is a union of that from Hibbs *et al*. (2007) and Myers and Troyanskaya (2007). Non-expression data includes pairwise physical and genetic interaction data from a variety of databases (Alfarano *et al*., 2005; Stark *et al*., 2006), protein localization (Huh *et al*., 2003), and sequence and TFBS similarities (Harbison *et al*., 2004; SGD, 2006). Pairwise interaction data were represented as binary presence/absence values; where applicable, interaction profile similarities were calculated between genes from binary data using an inner product. For details, see Myers and Troyanskaya (2007).

Expression data was collected from ∼80 publications comprising ∼120 datasets and ∼2300 conditions as described in Hibbs *et al*. (2007) and initially processed as described in Huttenhower *et al*. (2006). Datasets containing fewer than four experiments were initially merged, creating a merged microarray set that was subsequently processed identically to the remainder of the datasets. Each of these was converted from expression values to gene pair similarity scores using Pearson correlation normalized using Fisher's z-transform (David, 1949) and subsequently z-scored:

$$Fisher(g_i,g_j)=\frac{1}{2}\log\left(\frac{1+\rho(g_i,g_j)}{1-\rho(g_i,g_j)}\right) \tag{1}$$

$$z(g_i,g_j)=\frac{Fisher(g_i,g_j)-\mu_f}{\sigma_f} \tag{2}$$

That is, the Fisher's transformed score between any two genes g_i and g_j is a transformation of their Pearson correlation ρ, and the final similarity between two genes $z(g_i, g_j)$ is the pair's Fisher score minus the mean Fisher score μ_f divided by the Fisher score SD σ_f (both over all gene pairs).

After z-scoring, each expression dataset was quantized using the binnings $(-\infty,-1.5)$, $[-1.5, -0.5)$, $[-0.5, 0.5)$, $[0.5, 1.5)$, $[1.5, 2.5)$, $[2.5, 3.5)$, $[3.5, \infty)$; these represent steps of 1 SD in z-score space. Mutual information was calculated between the resulting sets of discrete values, and any pairs of datasets sharing $>15\%$ of the possible information were merged by averaging z-scores. PISA (Kloster *et al*., 2005) modules (a biclustering algorithm) were also calculated for the expression data collection and transformed into pairwise scores for our analysis by counting the number of times each pair of genes coclustered after 500 iterations. These biclusters offered an orthogonal analysis of the microarray data capable of providing different information than the normalize correlation scores.

2.1.2 Gold standard generation To perform supervised learning, we generate a gold standard of known functionally related and unrelated gene pairs. Biological processes of interest were selected from the GO (Ashburner *et al*., 2000) using a method based on Myers *et al*. (2006). The standard developed in Myers *et al*. (2006) is specific to *S.cerevisiae*; using a similar voting method and polling six biologists, a set of 433 GO terms were selected for this study to be experimentally informative independent of organism. Of these 141 have at least 10 gene annotations in *S.cerevisiae*, and these were selected as processes (gene sets) of interest (Supplementary Table 1).

An answer set was derived from these processes of interest as described in Huttenhower *et al*. (2006). Gene pairs coannotated to any of the 141 terms were considered to be related. A gene pair was unrelated in the gold standard if (1) the two genes were both annotated to some term in the set of 141, (2) the genes were not coannotated to any of these terms and (3) the terms to which the genes were annotated did not overlap with hypergeometric P-value <0.05. All other gene pairs were omitted from the standard (i.e. they were neither related nor unrelated for training and evaluation purposes).

For context-specific learning, this answer set was decomposed into subsets relevant to each process of interest. A gene pair was considered to be relevant to a biological process if either (1) both genes were annotated to the process or (2) one of the two genes was annotated to the process and the pair was unrelated in the standard (i.e. not coannotated to another process).

2.2 Bayesian analysis

2.2.1 Learning Bayesian classifiers One naive Bayesian classifier (Neapolitan, 2004) was learned per biological process of interest; experiments with other network structures were shown to provide negligible performance improvements (Huttenhower and Troyanskaya, 2006). Briefly, a global classifier was learned in which the class to be predicted was gene pair functional relationships (as defined in the gold standard) and each dataset formed one node in the network. One hundred and forty-one function-specific networks were learned with identical structures, each using a subset of the global gold standard as described above. When fewer than 25 gene pairs were available for a particular dataset/relationship combination, the global probability distribution was used for that condition. This defines the predicted probability of functional relationship between genes as a weight:

$$w_F(g_i,g_j)\propto\prod_D P_F[D=d(g_i,g_j)] \tag{3}$$

That is, the weight between genes g_i and g_j in function-specific network F is proportional to (using Bayes rule) the product over all datasets D of D's probability of experimental value $d(g_i,g_j)$ for the two genes.

All Bayes network manipulation was performed with a combination of custom C++ software and the SMILE library from the University of Pittsburgh Decision Systems Laboratory (Druzdzel, 1999).

2.2.2 Predicting functional relationships Each naive Bayesian classifier directly implies a functional relationship network in which nodes represent genes and edge weights consist of the posterior probabilities of functional relationships between gene pairs. The 141 function-specific networks were combined to form a predicted global interaction network by transforming each network's edge weights to z-scores (subtracting the mean predicted probability and dividing by their SD) and averaging each gene pair's weight across all available networks.

2.3 Functional relationship and dataset enrichment predictions

2.3.1 Process/process relationships As described above, for the purposes of this analysis, a biological process was defined as a set of related genes. The strength of a predicted functional relationship between two processes F and G was calculated as the average edge weight in the global interaction network within the edge set:

$$E_{F,G} = \{(g_i, g_j) | g_i \in F, g_j \in G, g_i, g_j \notin F \cap G\} \tag{4}$$

That is, the predicted functional relationship strength between functions F and G is the average weight of all edges in the global interaction network between genes g_i and g_j spanning the two gene sets and not coincident to any gene in their intersection. Note that this specifically excludes process similarity due to overlapping curated annotations and retains only data-driven functional relationships.

Similarly, the functional cohesiveness of a process was measured as the ratio of the average edge weight in the process to the average edge weight incident to the process:

$$cohes(F) = \frac{2|G| \sum_{g_i, g_j \in F} w_F(g_i, g_j)}{|F-1| \sum_{g_i \in F} \sum_{g_j \in G} w_F(g_i, g_j)} \tag{5}$$

where F is the function of interest, G is the genome and $w_F(g_i, g_j)$ is the edge weight between genes g_i and g_j in F's predicted functional relationship network. This normalizes for processes that are inherently more interactive and have uniformly higher probabilities of functional relationship. tRNA genes are omitted for the purposes of these calculations, since they represent a large class of very related genes for which essentially no data is available (thus generating a large number of misleadingly low weights).

2.3.2 Process/dataset relationships The predicted enrichment of each dataset within each biological process was derived from the conditional probability tables learned for that dataset's node within the appropriate function-specific Bayesian classifier. Specifically, the predicted enrichment for process F in dataset D was calculated as the weighted sum of the difference in posterior probability of functional relationship induced in F's classifier by evidence from each possible value of D:

$$rel(F,D) = \sum_{d \in D} P_F[D=d] \, |P_F[FR] - P_F[FR|D=d]| \tag{6}$$

For example, suppose the prior probability of functional relationship in the *ribosome biogenesis* process is 2% ($P_{rb}[\mathrm{FR}]=0.02$). The GRID- and BIND-based *yeast two-hybrid* dataset has two possible values, 0 representing no observed binding and 1 representing binding, thus $D=\{0,1\}$. After learning, the Bayesian classifier for *ribosome biogenesis* indicates that a lack of binding makes little difference ($P_{rb}[\mathrm{FR}|yth=0]=0.025$), but gene pairs that bind are very likely to be functionally related ($P_{rb}[\mathrm{FR}|yth=1]=0.4$). However, there are relatively few such pairs ($P_{rb}[yth=1]=10^{-4}$), since most gene pairs in the genome have not been observed to interact by available two-hybrid data ($P_{rb}[yth=0]=0.9999$). Thus the strength of relationship between the process of *ribosome biogenesis* and the *yeast two-hybrid* dataset is $r=0.9999|0.02-0.025+10^{-4}|0.02-0.4| \approx 0.005$. The exact value may differ due to rounding in this example.

The estimated coverage of a process in currently available data was calculated as the average of $rel(F,D)$ over all datasets in our study.

2.3.3 Dataset/dataset relationships This calculation of predicted process/dataset enrichments results in a vector of 141 values in the range [0, 1] for each dataset. To determine the functional similarity between two datasets, each value is first transformed to a log ratio against the average across all datasets:

$$rel'(F,D) = \log(rel(F,D) \cdot |D| / \sum_{d \in D} rel(F,d)) \tag{7}$$

This normalizes against the fact that certain biological processes are inherently more apparent in most high-throughput data (e.g. most microarray datasets have strong signals for processes such as *translation*). The functional similarity between datasets is then the Pearson correlation of the resulting $r\prime$ vectors across all datasets.

2.3.4 Gene/function relationships For the purpose of predicting gene function based on 'guilt by association' with known genes in some process, the connectivity of a gene to a process was assessed as follows. Each gene/process pair was assigned a functional association score equal to the ratio of its average probability of functional relationship to the process over the process's cohesiveness:

$$assoc(g_i, F) = \frac{\sum_{g_j \in F} w_F(g_i, g_j)}{|F| cohes(F)} \tag{8}$$

This calculation was also used to predict each biological process's predicted association enrichment with unknown genes. A list of 1451 genes with no annotation below *biological process* was extracted from the GO. A function's strength of association with unknowns was then the sum of its association scores for these 1451 genes.

2.3.5 Robustness A robustness study was carried out by randomly shuffling data points within each dataset prior to Bayesian learning. The resulting networks had average dataset functional enrichment scores of $4.46 \times 10^{-5} \pm 1.57 \times 10^{-4}$, biological processes cohesiveness of 1.37 ± 1.32, and association between processes of $7.14 \times 10^{-3} \pm 0.0293$, the last due to the greatly reduced differentiation between processes. In contrast, the averages for these values in Supplementary Tables 1–3 are $2.43 \times 10^{-4} \pm 6.02 \times 10^{-4}$, 15.1 ± 35.9, and $1.94 \times 10^{-3} \pm 0.141$, respectively.

2.3.6 Dense subgraphs An implementation of a modified greedy heuristic for discovering heavily weighted subgraphs (Charikar, 2000) was used to mine interaction networks for cohesive modules. Briefly, to discover each module within the network of interest, a node set was initialized with the most cohesive pair in the network. Nodes were added to this set greedily based on edge weight until no node could be added without reducing the average cohesiveness of the node set below the network baseline. The average edge weight of the set was then subtracted from each edge between nodes in the set, and the process was iterated to discover the next module. In pseudocode:

```
(1) $N = \mathrm{argmax}_{\{gi,gj\}}\ cohes(\{g_i, g_j\})$
(2) Loop:
(3)     $g = \mathrm{argmax}_g\ cohes(N \cup \{g\})$
(4)     If $cohes(N \cup \{g\}) < 1$, stop
(5)     $N = N \cup g$
(6) If $|N| > 2$, output $N$
(7) Let $\bar{w}$ be the average edge weight among nodes in $N$
(8) For each $g_i, g_j \in N$
(9)     $w(g_i, g_j) = w(g_i, g_j) - \bar{w}$
(10) Repeat from 1
```

3 RESULTS

By analyzing functional associations among biological processes and functional similarities between high-throughput datasets in a purely data-driven manner, we summarize knowledge from thousands of whole-genome experiments in a biologically informative way. This includes descriptions of the cohesiveness, data coverage and associations of biological processes (Fig. 1), which can guide experimenters towards promising targets for future experimental work (Table 1). Datasets can also be compared based on functional activity, allowing the detection of large-scale functional similarity between the effects of experimental

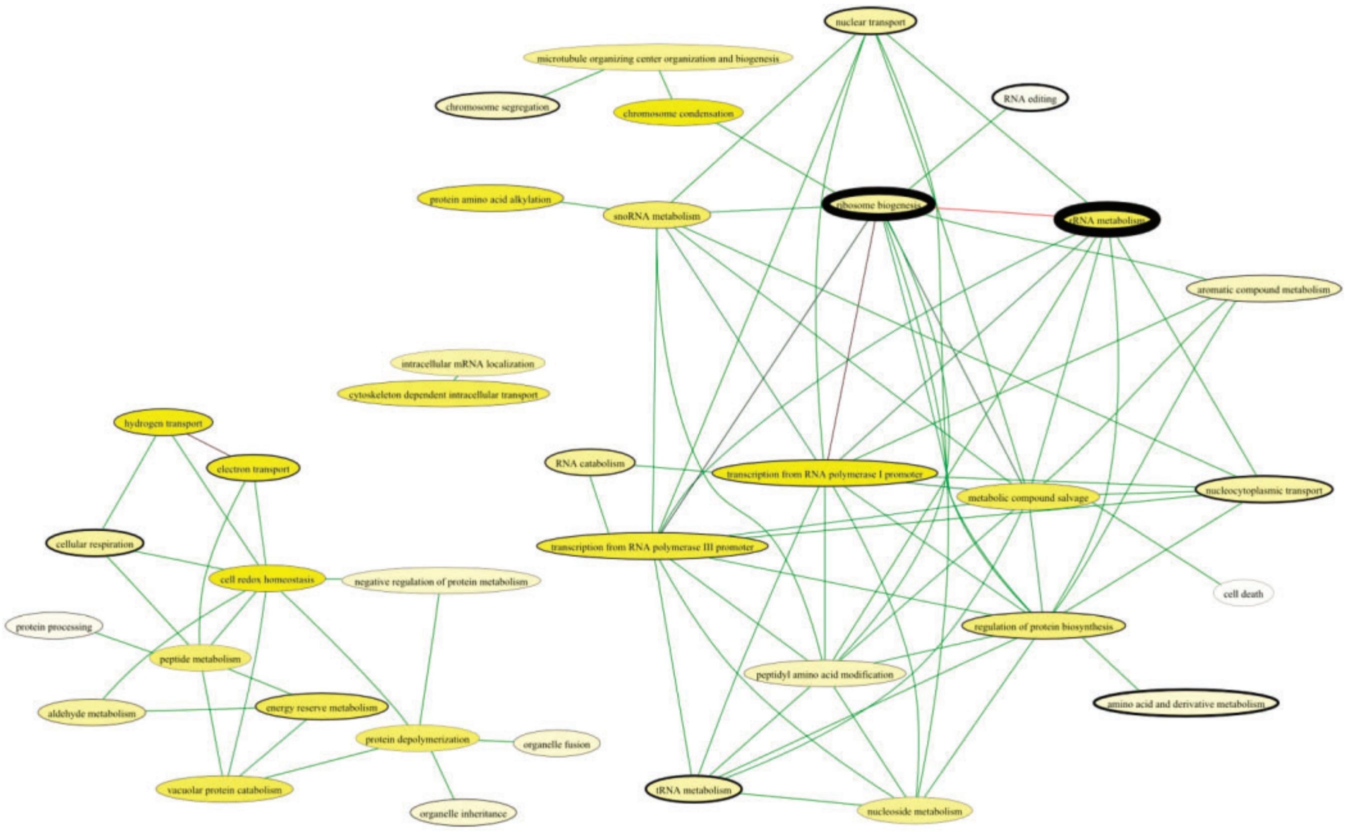
nuclear transport
microtubule organizing center organization and biogenesis
RNA editing
chromosome segregation
chromosome condensation
protein amino acid alkylation
snoRNA metabolism
ribosome biogenesis
rRNA metabolism
aromatic compound metabolism
intracellular mRNA localization
cytoskeleton dependent intracellular transport
hydrogen transport
electron transport
RNA catabolism
transcription from RNA polymerase I promoter
metabolic compound salvage
nucleocytoplasmic transport
cellular respiration
transcription from RNA polymerase III promoter
cell redox homeostasis
negative regulation of protein metabolism
cell death
protein processing
peptide metabolism
regulation of protein biosynthesis
peptidyl amino acid modification
aldehyde metabolism
energy reserve metabolism
amino acid and derivative metabolism
protein depolymerization
organelle fusion
vacuolar protein catabolism
tRNA metabolism
organelle inheritance
nucleoside metabolism

Fig. 1. High-confidence associations between biological processes predicted from large-scale data integration. Each node represents a biological process extracted from the GO; edges represent predicted functional associations between these terms based on their constituent genes' behavior in a compendium of >180 *S.cerevisiae* datasets. Node color intensity represents cohesiveness of the process, a measure of predicted relationship density within the process's gene set (white indicates background cohesiveness, yellow maximum cohesiveness). Border thickness summarizes estimated coverage of the biological process by available data. These edges represent only the strongest associations in the complete network, so coloration is relative, ranging from green (least strong) through black to red (strongest). Biological processes with high cohesiveness but low data coverage represent particularly promising targets for future experimental screens.

perturbations (Figs 2 and 3). These analyses provide an important global summary of interplay between pathways, and they identify processes, process associations and dataset similarities likely to benefit from experimental investigation.

3.1 Discovering data-driven functional associations between biological processes

Two or more biological processes can interact and work together to perform cellular functions in a manner analogous to a relationship between individual genes. A pair of genes might be functionally related if they operate in the same complex, pathway or transcriptional module. Our focus is at a higher level, where two processes might be functionally associated if they interact to achieve the same cellular goals; for example, nutrient sensing and the translation of new proteins at the ribosomes are distinct processes, but they interact to allow controlled cellular growth. These process–process associations are thus an extension of gene functional relationships: processes are functionally associated if they achieve related cellular goals, and we predict such an association if their constituent genes behave similarly in datasets determined to be good functional indicators. A small segment of our predicted process association network appears in Figure 1, made up of only the most confidently associated biological processes (see Supplementary Table 1 for complete results).

The edges in this process association network summarize information regarding the interactions between biological processes. A single biological process is internally *cohesive* in the currently available experimental results if its constituent genes also show strong individual functional relationships. If most gene pairs within a process are confidently functionally related, that process is reflected well by the available data: its annotations are in agreement with measured cellular behavior. If gene pairs within a process are related with low confidence, it often indicates an area of biology where further experimentation or annotation efforts may be most beneficial. The cohesiveness of biological processes in Figure 1 is represented by node color, where more cohesive processes appear in brighter yellow.

Finally, we also determined the degree to which each biological process is *covered* by available data. Our integration method provides a statistical measure of how active each biological function is within each dataset; we can thus sum over all datasets to estimate a biological process' total representation within the data. This coverage measure is summarized by border width in Figure 1, with thicker borders indicating well-covered processes. Cohesive biological processes (yellow nodes) not covered well by available data thus represent promising candidates for future investigation: they show evidence of strong functional similarity, but they may not yet have been specifically targeted by high-throughput studies.

This interplay between functional associations, cohesiveness and data coverage is evident in several of the example processes in Figure 1. *Ribosome biogenesis* and *rRNA metabolism*, for example, are processes strongly evident in most microarray data (Myers *et al*., 2006), and this ubiquity is demonstrated by their extremely strong coverage and association. They are not as cohesive as many other processes, however, due to the large number of snRNAs and rRNAs annotated to these processes for which little or no high-throughput data is available. This analysis thus highlights an area for future exploration, even in an area as thoroughly studied as the ribosome. Other processes with relatively low coverage for their size (data not shown in Fig. 1) include *protein complex assembly*, *ion homeostasis and transport* and *mitochondrion organization*, all representing opportunities for future directed screens. Processes with low cohesiveness can either be particularly diverse (e.g. *amino acid and derivative metabolism*, *protein processing*) or not yet fully characterized, representing further opportunities for future experimental investigations.

Table 1. Biological processes highly associated with yeast genes currently uncharacterized in the GO

Process	Size (Genes)	Cohes.	Rel. Data Coverage	Assoc. wt. Unch.
Carbohydrate metabolism	233	2.09	3.75	972.1
Phosphorus metabolism	201	1.95	2.35	895.3
Reproductive physiological process	308	1.87	1.95	863.5
Establishment of protein localization	279	1.82	1.77	862.0
Sporulation	120	2.48	1.68	832.7
Autophagy	40	3.69	1.22	797.6
One-carbon compound metabolism	94	1.94	2.57	794.9
Cell wall organization and biogenesis	196	2.11	1.40	788.2
Chromosome organization and biogenesis	557	1.96	4.53	773.1
Cofactor metabolism	169	2.60	2.52	743.8

Association with uncharacterized genes is measured as the sum of predicted functional relationships between genes in a process and uncharacterized genes, normalized by the cohesiveness (and thus size) of the process. The cohesiveness of a process indicates the ratio of average in-process relationship weight to the average out-of-process relationship weight (with 1.0 thus the genomic background). Relative data coverage is a scaled sum of all datasets' predicted association weight with the given biological process. Because of their likely association with uncharacterized genes, these processes represent good candidates for future genomic screens.

3.1.1 Processes predicted to be enriched for uncharacterized genes

Networks of functional associations between processes represent a richly structured summarization of high-throughput data; they implicitly encode predicted details regarding pathway structure, association between gene sets and the functional diversity of currently available data. In addition to associating known processes and pathways, though, similar relationships can also be inferred to find areas of biology enriched for uncharacterized genes. These represent specific processes for which targeted genomic screens might uncover substantial new information.

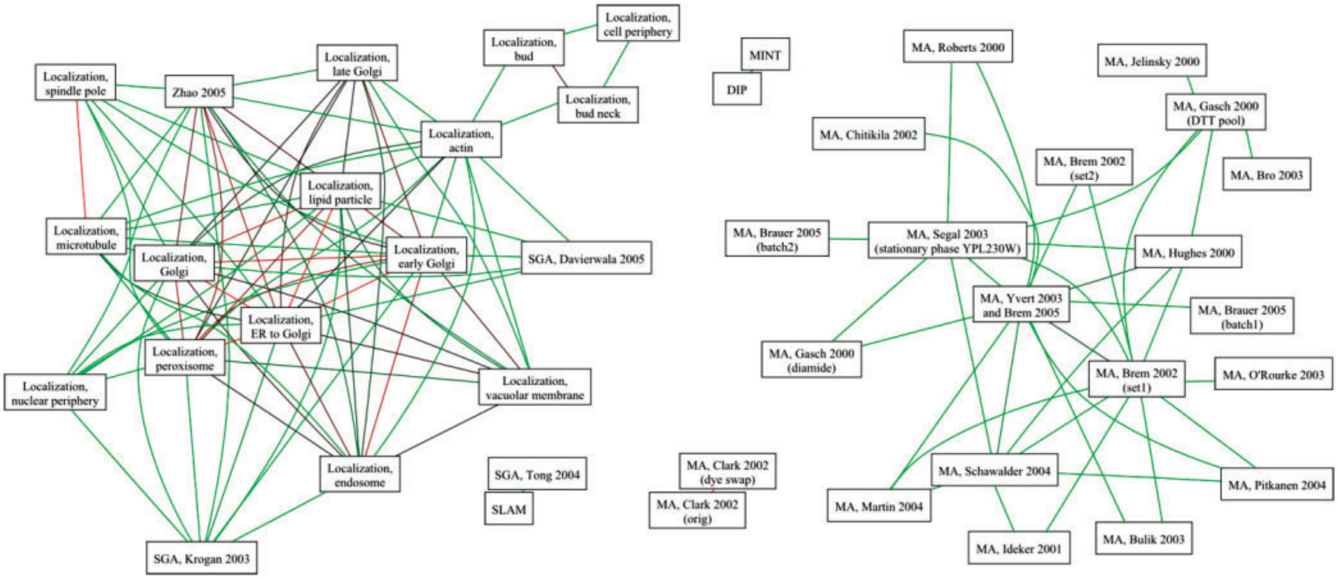

Fig. 2. Similarities in functional activity between high-throughput datasets. Each node represents a dataset, each edge the correlation between two datasets' functional activity profiles. These edges represent only the strongest correlations (by Kendall's τ), so coloration is relative from green (least strong) to red (strongest). This associates collections of datasets that explore related areas of biology, either by specific experimental design (e.g. protein localization) or by provoking similar biological responses (e.g. the diauxic shift and stationary phase growth). This also confirms that multiple genetic (SLAM and Tong *et. al.* 2004) and physical (DIP and MINT) interaction collections offer similar functional coverage.

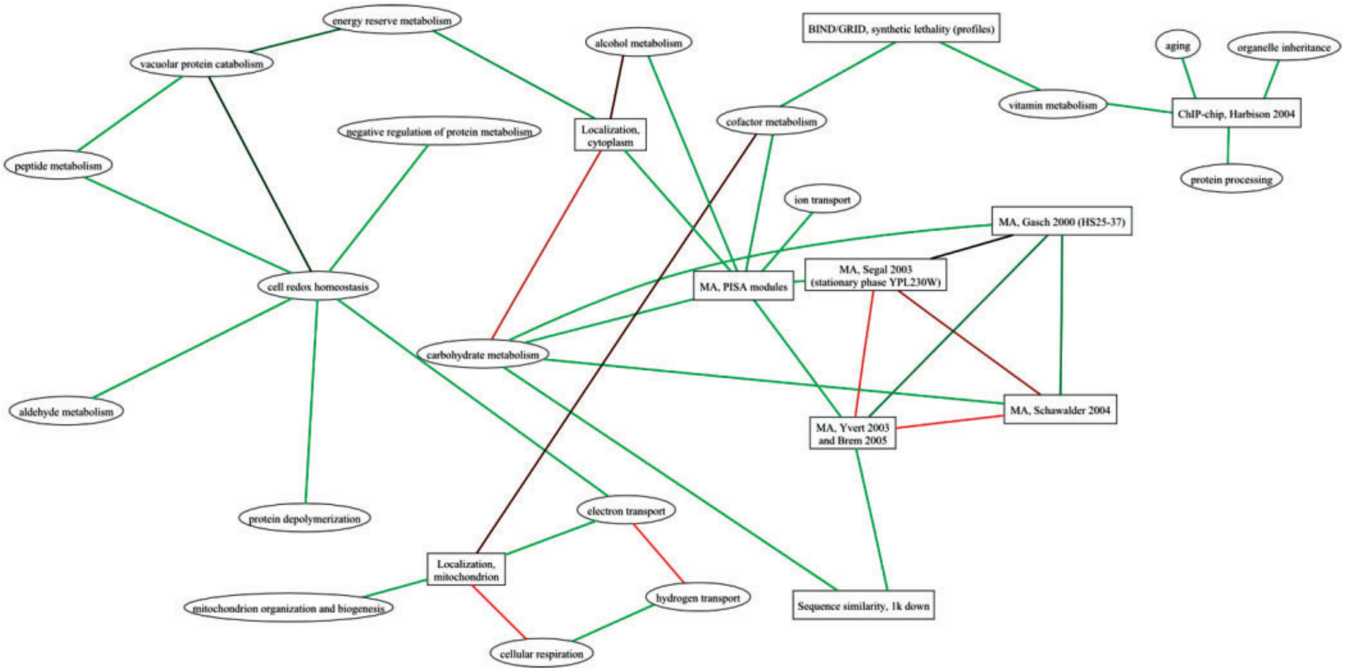

Fig. 3. Coclustering datasets and biological processes in an area of dense functional associations. By mining associations between biological processes for dense subgraphs, we recover a collection of processes (rectangular nodes) predicted to be highly related based solely on experimental data. We then extract the datasets (oval nodes) most informative for those processes and display the most confident process/process, dataset/dataset and dataset/process associations among these nodes. Each edge type is individually weighted, and only the strongest edges are shown, ranging in weight from green (least strong) to red (strongest). This network thus represents a snapshot of one area of yeast biology, the interconnections among its constituent processes and datasets exploring these processes.

A selection of processes that we find to be highly associated with uncharacterized genes is shown in Table 1, in addition to statistics describing the processes (see Supplementary Table 3 for complete results). The *autophagy* term, despite being the smallest and most cohesive process in this subset, still maintains a very strong association with uncharacterized genes. It is moderately well covered by available data, falling roughly in the middle of our 141 coverage estimates; it is thus possible that further information regarding autophagy could be gleaned from existing data, even though few experiments have specifically investigated the process in yeast. However, this predicted association with uncharacterized genes also suggests that substantial new functional assignments could be made by targeted screens for involvement in autophagy.

3.2 Similar functional activity in high-throughput datasets

While most high-throughput experiments are designed with fairly specific goals in mind, almost every dataset contains information about a variety of biological processes, and our analysis provides several ways of exploring these data. Our Bayesian learning process results in a probabilistic score indicating the activity of each biological process within each dataset. Collecting all such scores for a single dataset results in a functional profile for the dataset, and these numerical vectors can be compared between datasets to evaluate functional similarity. The network in Figure 2 contains a selection of datasets with similar functional activities (see Supplementary Table 2 for complete results).

Even in this small subset of analyzed datasets, several patterns are apparent. On the left, the first of the two main clusters contains primarily localization data from (Huh *et al*., 2003). Within the localization subsets, dataset similarity is correlated with cellular localization: the periphery and the bud are associated with the main body of data by way of actin, the Golgi stages are associated with each other, the endosome and peroxisome are related, and so forth. Three synthetic genetic array screens are also similar to the localization data. Davierwala *et al*. (2005) is associated primarily with the Golgi and ER, and one of the primary findings of this study was the characterization of PGA1, a gene essential for ER activity. Krogan *et al*. (2003) and Zhao *et al*. (2005) show similar functional activity to a variety of localization subsets (including several not shown in Fig. 2) and to Krogan *et al*. (2004), all of which are enriched for nuclear functions (*DNA packaging, chromosome organization, transcription, RNA elongation*, etc.) These functional similarities were generated solely by automatic data mining and call out important biological associations between disparate experimental results.

On the right, the cluster of microarray data is centered around a core of large datasets exploring very diverse conditions and thus enriched for many different biological processes (Brem and Kruglyak, 2005; Brem *et al*., 2002; Hughes *et al*., 2000; Yvert *et al*., 2003). The other main components of the cluster are stationary-phase growth and carbon metabolism (Brauer *et al*., 2005;

Ideker *et al*., 2001; Martin *et al*., 2004; Pitkanen *et al*., 2004; Segal *et al*., 2003) and various stresses (Bro *et al*., 2003; Gasch *et al*., 2000; Jelinsky *et al*., 2000; O'Rourke and Herskowitz, 2004). Interestingly, (Bulik *et al*., 2003; Chitikila *et al*., 2002), and (Schawalder *et al*., 2004) are all likely included due to their use of galactose-inducible promoters while investigating other diverse processes; these datasets all share a *carbohydrate metabolism* enrichment in addition to their more specific targets [e.g. *biopolymer biosynthesis*, a parent of *chitin biosynthesis*, in Bulik *et al*. (2003)]. This demonstrates the power of associative functional analysis to uncover both primary and secondary enrichments, a consideration essential to getting the most out of any experimental result.

3.3 Simultaneous association of datasets and biological processes

Because our method assesses functional activity within datasets, functional similarities between datasets and associations between biological functions, it provides a means of coclustering datasets and processes in a biologically meaningful way. This raises the possibility of exploring complex data, potentially summarizing millions of individual measurements, in an intuitive manner. Each predicted weight between two datasets, two processes or a dataset and a process represents a measure of similar biological function, and thus an investigation of heavily weighted subgraphs in this space provides a way of exploring groups of related data and processes.

An example of such a cluster appears in Figure 3, which highlights one of the densest functional areas and the datasets in which these functions are most active. This consists of metabolic processes including *alcohol, aldehyde* and *carbohydrate metabolism, cellular respiration, hydrogen* and *electron transport*, and *mitochondrion biogenesis*; while they have been removed for visual clarity, several other related processes are also members of this cluster, including *cofactor metabolism, autophagy* and *aging*. The group of associated microarrays again represent a combination of broad genomic response (Brem and Kruglyak, 2005; Yvert *et al*., 2003), carbon metabolism (Schawalder *et al*., 2004; Segal *et al*., 2003) and stresses (Gasch *et al*., 2000), the latter likely included due to the relationship between stress response and growth rate (Brauer *et al*., 2008). These are linked into the cluster of biological processes primarily through *carbohydrate metabolism*, but also through the biclustering modules (PISA). These biclustering results incorporate all of the available microarray conditions, in contrast to the normalized correlation scores used to analyze individual datasets. Biclustering thus represents a view of expression data orthogonal to pairwise correlations and tends to be more sensitive to metabolic functions in general (*phosphorus, amino acid* and *nitrogen compound metabolism* in addition to those appearing in Fig. 3).

The non-microarray datasets associated with this functional cluster are diverse, including mitochondrial localization (in association with several mitochondrial and respiratory functions), cytoplasmic localization (in association with more general metabolism), two sequence-based analyses [downstream sequence similarity and shared transcription factor binding sites from Harbison *et al*. (2004)] and synthetic lethality interaction profiles from GRID (Stark *et al*., 2006) and BIND (Alfarano *et al*., 2005). Synthetic lethality profiles and shared binding sites both provide good coverage of many biological processes and are included largely due to moderate association with many of the functions within the cluster (most edges are not shown in Fig. 3); this is reflected in their relative isolation in the network. Broad downstream (and upstream) sequence similarity tends to capture structural features of the genome, in this case the close positional association of the GAL genes.

3.4 A case study: detecting a specific biological response in diverse data

At a more specific level, these interprocess associations and functional descriptions of datasets can be used to uncover detailed biological responses in high-throughput data. We were struck by the correlation in functional activities between three seemingly diverse datasets: Chitikila *et al*. (2002), an investigation of TBP inhibitors, Martin *et al*. (2004), an analysis of *tor2* mutants described in Helliwell *et al*. (1998), and Pitkanen *et al*. (2004), a *pmi40* deletion assayed over varying mannose concentrations. These three microarray collections share functional enrichments with other datasets assaying similar conditions [e.g. the nutritional cluster discussed above including Martin *et al*. (2004) and Pitkanen *et al*. (2004)], and no one pair of the three correlations is unusually high. They also represent two different experimental platforms: Martin *et al*. (2004) and Pitkanen *et al*. (2004) both employ single channel microarrays, while Chitikila *et al*. (2002) uses a two-color array. However, the average functional correlation between the three datasets is highly significant ($\overline{rel'} = 0.316$, $P < 10^{-3}$) for arrays under such apparently diverse conditions.

All three datasets are enriched for activity in distinct biological processes, and all three present unique biological conclusions that are in no way undermined by this unexpected similarity. Upon inspection of the three datasets' experimental protocols, however, the common factor appears to be the use of a specific plasmid shuffle transformation employing a strain background of the form *ura3 trp1 leu2 his3* or *his4*. We have confirmed this similarity in a fourth dataset we are currently developing investigating temperature-sensitive *dbf4* mutants (Myers *et al*., 2005). Although the overarching biological conditions of our dataset share little in common with Chitikila *et al*. (2002), Martin *et al*. (2004) and Pitkanen *et al*. (2004), our mutants were also constructed using a similar plasmid transformation, and the resulting microarrays produce highly correlated functional profiles. Even when strain background and reference channels (when applicable) are all properly controlled, the plasmid shuffle process and associated auxotrophies result in subtle changes in global transcription detectable by large-scale functional analysis.

This effect is quite subtle, a fact which we stress for two reasons. First, it is a secondary effect within the more prominent biological features assayed by these three datasets, and it is only by large-scale analysis of their functional content in the context of many other datasets that the similarity was discovered. Second, we emphasize that it in no way diminishes these datasets' primary results, and instead provides additional functional insight into their coexpression measurements. Most previous computational data integration has focused on associating genes with functions or genes with genes. As more high-throughput data becomes available, it opens up opportunities for associating entire datasets with broad functional activity and with other datasets, allowing the detection of biological signals and similarities that would remain undetectable at smaller scales.

4 DISCUSSION

We present a high-level functional analysis of very large compendia of genomic data and apply it to *S.cerevisiae*. By computationally summarizing thousands of whole-genome experimental conditions, we elucidate the current data coverage of *S.cerevisiae* biological processes, the cohesiveness of its functional annotations, and associations among these processes based on high-throughput experimental results. We also determine the functional activity in high-throughput datasets, allowing us to discover subtle relationships such as shared strain backgrounds in otherwise diverse microarray conditions. This analysis begins with specific functional relationships between individual genes predicted from large-scale data integration, and it extends into high-level information including functional associations between datasets, uncharacterized genes and biological processes.

A primary application of this system lies in directing future experimental efforts. In particular, high-throughput screens of any sort can be costly to implement and assay fairly general conditions; for example, if two proteins bind only during fermentation, their interaction will not be observed in a genomic screen during respiratory growth. A high-level functional analysis serves to call out underrepresented biological processes and those with increased likelihoods of novel discovery, which can in turn provide focus for experimental screens. This is analogous to candidate gene selection at a whole-genome level, a form of 'candidate process' selection, just as our predicted associations between biological processes represent functional relationships at a larger scale.

High-level functional analysis also provides very specific information on individual experimental results, in addition to its larger scale applications. This is exemplified by the functional signature of the plasmid shuffle strain discussed above; given any new high-throughput dataset, microarray or otherwise, we provide a means for establishing its functional activity in the context of existing data. Both this *post hoc* analysis and the *a priori* predictions of underrepresented functions are of particular use in less well-studied organisms. By designing experiments to explore processes shown to lack functional coverage and by leveraging all available data to interpret new results, laboratory work can be quickly guided to areas of biological interest and potential.

Finally, the functional information summarized by our system can also be employed in the continuous process of functional cataloging. While we have used examples from the GO, any sets of functionally related genes could drive analyses such as this, and the results can guide annotators in cataloging existing data much as they can guide experimenters in generating new data. By providing a means of directing annotators to potentially under-annotated functions and the datasets associated with them, our analysis simplifies a curation and cataloging task that grows with each new publication. By analyzing and presenting the large-scale functional structure of genome-scale data, we hope to guide annotators and experimenters alike in exploring the potential of the ongoing genomic revolution.

ACKNOWLEDGEMENTS

The authors would like to thank Chad Myers, Matthew Hibbs, Florian Markowetz and David Hess for insightful comments and conversations and Camelia Chiriac for experimental assistance.

Funding: This research is partially supported by NSF CAREER award DBI-0546275, NIH grant R01 GM071966, NIH grant T32 HG003284 and NIGMS Center of Excellence grant P50 GM071508. O.G.T. is an Alfred P. Sloan Research Fellow.

Conflict of Interest: none declared.

REFERENCES

Alfarano,C. *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Res.*, **33**, D418–D424.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Brauer,M.J. *et al*. (2008) Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol. Biol. Cell*, **19**, 352–367.

Brauer,M.J. *et al*. (2005) Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Mol. Biol. Cell*, **16**, 2503–2517.

Brem,R.B. and Kruglyak,L. (2005) The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 1572–1577.

Brem,R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Bro,C. *et al*. (2003) Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells, *J. Biol. Chem.*, **278**, 32141–32149.

Bulik,D.A. *et al*. (2003) Chitin synthesis in Saccharomyces cerevisiae in response to supplementation of growth medium with glucosamine and cell wall stress, *Eukaryot. Cell*, **2**, 886–900.

Charikar,M. (2000) Greedy approximation algorithms for finding dense components in a graph. *Third International Workshop on Approximation Algorithms for Combinatorial Optimization*. Springer, Saarbrücken, Germany.

Chitikila,C. *et al*. (2002) Interplay of TBP inhibitors in global transcriptional control, *Mol. Cell*, **10**, 871–882.

David,F.N. (1949) The moments of the Z and F distributions. *Biometrika*, **36**, 394–403.

Davierwala,A.P. *et al*. (2005) The synthetic genetic interaction spectrum of essential genes. *Nat. Genet.*, **37**, 1147–1152.

Druzdzel,M.J. (1999) {SMILE}: Structural Modeling, Inference, and Learning Engine and {GeNIe}: a development environment for graphical decision-theoretic models. *Sixteenth National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, Orlando, FL.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Franke,L. *et al*. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, *Am. J. Hum. Genet.*, **78**, 1011–1025.

Gansner,E.R. and North,S.C. (2000) An open graph visualization system and its applications to software engineering. *Software Pract. Exper.*, **30**, 1203–1233.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415**, 141–147.

Giaever,G. *et al*. (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, **418**, 387–391.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Helliwell,S.B. *et al*. (1998) TOR2 is part of two related signaling pathways coordinating cell growth in Saccharomyces cerevisiae. *Genetics*, **148**, 99–112.

Hibbs,M.A. *et al*. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Huh,W.K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Huttenhower,C. *et al*. (2006) A scalable method for integration and functional analysis of multiple microarray datasets, *Bioinformatics*, **22**, 2890–2897.

Huttenhower,C. and Troyanskaya,O.G. (2006) Bayesian data integration: a functional perspective, *Computational Syst. Bioinform. / Life Sci. Soc.*, **5**, 341–351.

Ideker,T. *et al*. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934

Jansen,R. *et al*. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Jelinsky,S.A. *et al*. (2000) Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes, *Mol. Cell Biol.*, **20**, 8157–8167.

Karaoz,U. *et al*. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks, *Proc. Natl Acad. Sci. USA*, **101**, 2888–2893.

Kloster,M. *et al*. (2005) Finding regulatory modules through large-scale gene-expression data analysis. *Bioinformatics*, **21**, 1172–1179.

Krogan,N.J. *et al*. (2003) Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Mol. Cell Biol.*, **23**, 4207–4218.

Krogan,N.J. *et al*. (2004) High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell*, **13**, 225–239.

Lee,I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Martin,D.E. *et al*. (2004) Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics*, **5**, 148.

Myers,C.L. *et al*. (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

Myers,C.L. *et al*. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol.*, **6**, R114.

Myers,C.L. and Troyanskaya,O.G. (2007) Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, **23**, 2322–2330.

Neapolitan,R.E. (2004) *Learning Bayesian Networks*. Prentice Hall, Chicago, IL.

O'Rourke,S.M. and Herskowitz,I. (2004) Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol. Biol. Cell*, **15**, 532–542.

Pitkanen,J.P. *et al*. (2004) Excess mannose limits the growth of phosphomannose isomerase PMI40 deletion strain of Saccharomyces cerevisiae, *J. Biol. Chem.*, **279**, 55737–55743.

Schawalder,S.B. *et al*. (2004) Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1. *Nature*, **432**, 1058–1061.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

SGD (2006) Saccharomyces Genome Database. Available at http://www.yeastgenome.org

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Tong,A.H. *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Troyanskaya,O.G. *et al*. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.

Yvert,G. *et al*. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.*, **35**, 57–64.

Zhao,R. *et al*. (2005) Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. *Cell*, **120**, 715–727.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, **15**, 607–611.