Supplementary online material

A probabilistic functional network of yeast genes

Insuk Lee, Shailesh V. Date, Alex T. Adai & Edward M. Marcotte

DATA SETS

Saccharomyces cerevisiae genome

This study is based on the verified 5,817 protein encoding open reading frames (ORFs) of yeast genome (downloaded from *Saccharomyces cerevisiae* Genome Database (SGD) (*1*) on August 2003). All linkages and calculations of genome coverage are based on this gene set.

mRNA expression data

Currently, mRNA expression profiling represents one of the most extensive functional genomics data sets, and many such data sets are available in the public domain. Our basic analysis is to measure the co-expression of each pair of genes across multiple experimental conditions, from which we infer functional associations between pairs of the genes. Our approach is described in full below, but in general terms, we measure the Pearson correlation coefficient across a chosen subset of mRNA expression data as an indicator of the strength of functional association in those experiments, and record the strongest such expression linkages found between two genes from across a bank of distinct experiments. All expression data are from the Stanford Microarray

Database (SMD) (*2*), currently the most extensive source of publicly available data. The version of SMD downloaded on June 2003 contains a total of 717 experiments for yeast divided into 27 experimental categories (see Table S1).

Experimental protein-protein interaction data

Physical or genetic protein-protein interactions (PPI) reflect functional association of the corresponding genes in most, if not all, cases. PPI evidence from various small-scale experiments were collected from the Database of Interacting Proteins (DIP) (downloaded on March 2003) (*3*). The other PPI evidence is from large-scale experiments, specifically, mass spectrometry analysis for co-precipitated protein complexes (*4, 5*), high throughput yeast two hybrid (Y2H) assays (*6*-*8*), and high-throughput synthetic lethal (SL) screens (*9*).

Gene context data

With the many genome sequencing projects, genome sequences themselves represent a large source of data from which we can infer functional linkages for pairs of genes. There are several different algorithms that show reasonable performance, and in this study, we use two different genome sequence analyses: the method of phylogenetic profiling (PG) (*10*-*12*), and the Rosetta Stone protein (RS) (or gene-fusion) method (*11, 13*-*15*). Linkages for each method were derived from analysis of a database of 174,901 protein sequences from 57 genomes (41 bacteria, 11 archaea, and 5 eukaryotes). Briefly, each protein sequence was compared to every other sequence using the program BLASTP with default settings (*16*), for a total of $174{,}901^2$, or ~31 billion, sequence comparisons. Phylogenetic profiles were constructed from these comparisons and analyzed as in (*17*), and Rosetta Stone linkages were identified as in (*18*).

Literature mining data

In order to effectively capture known functional relationships in the gene networks, we also identified functional linkages by mining the scientific literature (specifically, Medline abstracts) using the co-citation approach (*19, 20*). This approach is based upon the assumption that pairs of genes mentioned in the same set of Medline abstract in excess of random expectation tend also to be functionally linked. In our study, we analyzed a set of N = 65,807 Medline abstracts that included the word "*Saccharomyces cerevisiae*" in the title, abstract, or MESH terms (*21*) for perfect matches to either the standardized names or common names (or their synonyms) of yeast genes.

For a given pair of genes, we measured the total number of Medline abstract in which each gene name appears (n and m, respectively) and the number of abstracts in which both names appear (k), and calculated the probability of co-citation by random chance from the hypergeometric distribution as

$$p(\text{Number of co-citations} \geq k \mid n, m, N) = 1 - \sum_{i=0}^{k-1} p(i \mid n, m, N),$$

$$\text{where} \quad p(i \mid n, m, N) = \frac{n!(N-n)!\,m!(N-m)!}{(n-i)!\,i!(m-i)!(N-n-m+i)!\,N!}.$$

(Large factorials were calculated with Sterling's approximation, and all probability calculations were performed in logarithmic form to avoid underflow errors.)

Reference and benchmark sets

In order to test the correct assignment of functional linkages in this study, five different benchmark sets were used, based on cellular function, functions of conserved gene families, or sub-cellular localization, and summarized in Table S2.

The Kyoto-based KEGG database (*22*) provides metabolic and regulatory pathway annotation for genes. KEGG maps about 20% of the yeast genes into at least one pathway or cellular system (examples include “glycolysis”, “ribosome”, and “MAPK signaling pathway”). We’ve previously found (*17, 23*) the KEGG database to be an excellent reference set for evaluating functional linkages because of its reasonable coverage of the genome (20%) and its moderately large number of categories (118) and therefore relatively small background probability of matching pathways at random (see Table S2).

As an independently derived set of pathway annotations, we’ve used the Gene Ontology (GO) annotation, provided through the *Saccharomyces cerevisiae* Genome Database (SGD) (*24*). The GO schema lists three hierarchies of function describing “biological process” (i.e., pathways and systems), “molecular function” (i.e., biochemical activities), and “cellular component” (i.e., subcellular localization). For pathway annotations, we have used the GO “biological process” annotation, which contains up to 16 different levels within the hierarchy (assuming the term “biological process” as the first level). Empirical testing (data not shown) revealed the best performance (as assessed by maximizing the coverage of the genome and maximally distinguishing well-curated protein-protein interactions from random gene pairs) at the 8th level of the hierarchy (referred to below as “8pGO”). This level of GO has extensive annotation, covering about 46% of the yeast genome with 494 distinct pathway annotations. The KEGG and GO pathway annotations are therefore reasonably independent annotations, with the extent of GO annotation more than twice that of the KEGG both in terms of genes and pathways.

As the third independent functional linkage benchmark, we opted for the clusters of orthologous group (COG) annotation (*25*), which is based on reconstructing homologous groups of proteins in such a manner as to considerably enrich for orthologous proteins within each group, with the functions of genes assigned within 23 broad categories (such as “Transcription” and “Signal Transduction Mechanisms”) based on the well-annotated proteins with each COG. We use the recently updated COG collection that includes multicellular eukaryotic genomes (named eukaryotic orthologous groups, or KOG) (*26*).

As the fourth benchmark, we chose a completely independent form of data, based not on pathway annotation or homology, but on experimentally-determined subcellular localization. Essentially, we expect that functionally-linked genes should also typically be localized to the same subcellular compartment. For this benchmark, we used the yeast protein localization data generated from genome-wide GFP-tagging and microscopy (*27*). Given the diversity of benchmark sets (especially with regard to the subcellular localization data), we expect that consistent performance across all four of these benchmark sets should indicate legitimate functional linkages between genes.

Finally, in order to effectively summarize broad trends in the data with respect to pathways (for which we desire relatively few categories of pathway annotation, unlike for the case of benchmarking the results), we use a fifth reference set of pathways from the

Munich Information Center for Protein Sequences (MIPS) (*28*). In this latter case, we have used the 11 major pathway categories from the top level MIPS functional category annotation to analyze trends in our resulting modular network.

COMPUTATIONAL METHODS

Overview of the method for integrating functional genomics data

Our working hypothesis is that each set of functional genomics data has an intrinsic error rate and a limited coverage but informs us to some extent about the tendency for genes to operate in the same cellular systems and pathways in the cell. We can therefore construct a more accurate and extensive functional gene network by integrating the information from multiple functional genomics datasets, and in this manner estimate the overall functional coupling between yeast genes across a broad set of experiments. Figure 1 shows our overall strategy of data integration. The prerequisite of this strategy is that we have a unified scoring scheme for testing the many heterogeneous data sets, even when the data sets are accompanied by their own intrinsic scoring schemes (such as for the computational methods). This re-scoring by a single criterion allows us to directly measure the relative merit of each data set, and then to integrate the data sets with weights that reflect this merit.

The integration is performed in four stages. First, different sets of mRNA expression data are analyzed for significant co-expression of pairs of genes, and linkages from the assortment of microarray data sets are integrated to generate a set of functional linkages derived entirely from DNA microarray data. These expression-based linkages are then integrated with other protein-protein interaction experiments, literature mining linkages, and gene context linkages to produce the initial integrated network (referred to as “IntNet” below). An additional set of functional linkages can then be derived from the genes’ linkage patterns in this initial network, in effect using the genes’ contexts in the network to determine whether unlinked genes are more properly linked together (these additional linkages are referred to as “ContextNet” below). Finally, the resulting network context-derived linkages are integrated with the original linkages to generate the final network (“FinalNet”). Therefore, the final network contains both linkages that are derived from direct experimental evidence as well as linkages inferred from the collective weight of indirect evidence implicating pairs of genes into the same cellular system.

A unified scoring method

The scoring method used in this study derives from a Bayesian statistics approach, in which each experiment, computational or physical, adds some degree of evidence that a pair of genes is functionally linked. More specifically, we calculate the odds ratio representing the likelihood that a pair of genes is functionally linked. If *P(L|E)* represents the probability of linkage between a pair of genes conditioned on the given evidence (and *~P(L|E)* represents the probability that these genes are *not* functionally linked), and *P(L)* is the unconditional probability of linkage between a pair of genes, the odds ratio (*OR*) that the given pair of genes is linked is given as:

$$\boldsymbol{OR(L, E) = \frac{P(L|E)/{\sim}P(L|E)}{P(L)/{\sim}P(L)}}$$

In Bayesian terms, the ratio *P(L)/~P(L)* represents the *prior* odds ratio, which is the ratio of the probability of the linkage and its negation before the evidence is seen. This term is estimated by counting the number of gene pairs with any shared functional annotation (using only a single source of functional annotation, for example, the KEGG pathway annotation) and those without any shared functional annotation among all possible gene pairs chosen from the set of annotated yeast genes. The ratio *P(L|E)/~P(L|E)* represents the *posterior* odds ratio, which is the ratio of the probability of the linkage and its negation conditioned on the given evidence. For estimating these conditional odds, we count the number of gene pairs that share or do not share functional annotation and that are also supported by the given evidence. The *OR(L, E)* can therefore be interpreted as the 'likelihood' of the linkage conditioned on the given evidence and corrected for background expectations of linkages. As an example, Figure S1A plots *P(L|E)* and ~*P(L|E)* for genes linked by mRNA co-expression patterns. Regardless of the correlation coefficient between pairs of genes' expression profiles across the microarray experiments, the value of ~*P(L|E)* is consistently low. However, *P(L|E)* shows a positive correlation with the correlation coefficient, especially for the region greater than about 0.3. The *OR(L, E)* is then calculated from these curves and the odds ratio of prior expectations, *P(L)* and ~*P(L)*. In practice, we calculate the natural logarithm of this ratio, the log likelihood ratio, in order to create an additive score, LLS = ln(*OR(L, E)*), plotted in Figure S1B. The ultimate score for a functional linkage is based upon a weighted sum of LLS scores from across the different lines of evidence linking the gene pair. (The weighting scheme is described in more detail below.) Empirically, we find the log likelihood scoring framework to be far more robust than previous benchmarks based on the same pathway data (*23*), providing a ratio of well-measured values rather than measuring more subtle differences in the precise extent of pathway similarities.

“Re-scoring” experiment-specific scoring schemes in the log likelihood framework

For data sets that provide only binary evidence (e.g., “observed to interact” or “not observed to interact”), each linkage derived from the same data set is scored with an identical LLS value calculated as the log of the odds ratio described above. The experimental protein interaction data are scored in this manner (see Table S3).

Other data sets provide functional linkages with associated parametric scores, such as the correlation coefficients indicating the degree of mRNA co-expression, the probability scores of genes being linked by gene fusions or co-citation, and the mutual information score indicating the degree of coinheritance of genes in phylogenetic profiling (as described in (*17*)). For these data sets, we first “re-map” or “re-score” the data in the log likelihood scheme before the linkages derived from this data can be integrated with linkages from other data sets. We rank the gene pairs by the original parametric scores and calculate the log likelihood scores for bins of equal numbers of gene pairs (usually 500 – 2,000 pairs of linked genes per bin). A regression fit is then made from a scatter plot of the log likelihood score versus the parametric score, and the linkages are assigned log likelihood scores as a function of their original parametric scores using the regression relationship between the two scoring schemes. In this manner, all such independent scoring schemes are converted into a single unified scoring scheme. We find log likelihood scores larger than ln(1.5) (i.e., evidence for genes to be linked consisting of odds ratios of larger than 1.5) with bins containing at least 200 pairs annotated by KEGG unlikely represent gene linkages by random chance (Confidence level > 99%, data not shown), thus exclude scores below this threshold for the integration procedure using KEGG. The score threshold varies among different reference/benchmark sets and bin sizes. Data sets that showed qualitatively poor regression fits were excluded from the integration procedure.

A heuristic strategy for integrating links derived from different lines of evidence

Various approaches for integrating information in order to more accurately define physical or functional interactions between proteins have been previously explored in the literature. These approaches range from relatively simple algorithms, such as using the intersection (*23*) or union (*29*) of distinct sets of evidence, to more refined approaches that rely on a variety of scoring schemes (*30*-*33*). Among these more sophisticated approaches, the Bayesian method of integrating information has proved useful in predicting functional associations (*32*) and physical interactions (*33*) between yeast proteins because it captures the relative weights of the various data sets. However, the relative independence of the various datasets can be difficult to estimate in the Bayesian framework. We’ve empirically found that a heuristic modification to the strict Bayesian approach performs extremely well for integrating the diverse functional linkage data sets by incorporating the relative weighting of the data as well as capturing simple aspects of their relative independence.

In this approach, we first collect all available log likelihood scores deriving from the various data sets and lines of evidences, then add the scores with a rank-order dependent weighting scheme. The resulting weighted sum (WS) scoring the functional linkage between a pair of genes is calculated as:

$$WS = \sum_{i=1}^{n} \frac{L_i}{D^{(i-1)}} ,$$

where L represents the log likelihood score for the gene linkage from a single data set, D is a free parameter roughly representing the relative degree of dependence between the various data sets, and i is the rank index in order of descending magnitude of the n log likelihood scores for the given gene pair. The free parameter D ranges from 1 to +∞, and is chosen to optimize overall performance (accuracy and coverage) on the functional benchmark. When $D = 1$, WS represents the simple sum of all log likelihood scores and is equivalent to a *naïve* Bayesian integration. We might expect D to exhibit an optimal value of 1 in the case that all data sets were completely independent. As the optimal value of D increases, WS approaches the single maximum value of the set of log likelihood scores, indicating that the various data sets are strongly redundant (i.e., no new evidence is offered by the additional data sets over what is provided by the first set). Intermediate values of D in effect represent exponentially diminishing belief in the additional evidence. In practice, solving for the optimal value of D provides a simple way to detect and account for strongly dependent data sets (as in the integration of mRNA expression data, Figure S5A). This approach offered a significant improvement in performance over the *naïve* Bayesian integration (for example, see Figures S4A, S5A, and S6C), while requiring optimization of only a single additional parameter.

Inferring additional functional linkages from the network context of genes

Two proteins that interact often tend to share additional interaction partners, or more generally, an interaction neighborhood (*34-37*). We exploited this tendency to discover new functional linkages between genes after the initial integrated gene network was calculated. First, we generated a matrix containing the weighted log likelihood scores (WS) between all pairs of yeast genes, where these scores exceeded a threshold of ln(1.5), or zeroes otherwise. Each row in this matrix was treated as a "context vector" describing the local network context of the associated gene. For each pair of genes, the Pearson correlation coefficient was calculated between the genes' context vectors, including in the calculation only those entries in which at least one of the two genes had a non-zero value. The resulting correlation coefficients indicate the degree of similarity between the overall network neighborhoods of each pair of genes, regardless of whether the genes were previously linked. These correlation coefficients were then treated like any other parametric scoring scheme for discovering linkages and were mapped into the log likelihood framework as described above. The resulting context-derived linkages were then integrated with the initial network to produce the final linkage network (FinalNet, see Figure 1).

Construction of a simplified network of gene modules

As the overall gene network itself is of extremely high complexity, we attempted to simplify analysis of the network by delineating naturally occurring gene modules within the network. From previous examinations of biological networks such as metabolic networks (*38*), we expected a hierarchical organization of gene modules, with genes organized into groupings of ever increasing functional association. Given such an organization, the precise definition of module or cluster boundaries can be difficult (or unreasonable). For this task, we used unsupervised clustering to define coherent groups of genes based upon their linkages. Specifically, we used hierarchical clustering (average linkage), which allows both detection of the hierarchy of gene organization as well as delineation into disjoint groupings of genes, if desired.

First, we took the final set of functional linkages which scored equal or better on the KEGG functional benchmark to our “gold standard” set of small scale experimental data from the DIP database. This set (labeled “ConfidentNet” in Figure 1) consists of 34,000 linkages between 4,686 genes (80.6 % of the genome). A matrix summarizing the links between the 4,686 genes was constructed consisting of the log likelihood scores from linkages in this set and zero otherwise, and hierarchical clustering was performed on this matrix using the program CLUSTER (*39*).

After the hierarchical clustering procedure, the clusters can be separated at different levels of the hierarchy, capturing broader or narrower gene modules in the network. Finding an optimal level of the hierarchy to define the clusters is non-trivial, however. Such functional modules have been previously defined in sparse but highly clustered protein interaction networks (2.7 links / protein, high clustering coefficent (*40*) of 0.6) by breaking clusters according to the sparser edges that connect highly interconnected groups of proteins (*41*), but this approach was less useful for the dense network we present (average number of linkages per protein ~ 7.3, and intermediate clustering coefficient = 0.3). Another approach that has proved useful is constructing a dendrogram of genes by connectivity and delimiting the cluster boundaries by determining the threshold in the dendrogram at which the tree should be cut (*42*) based upon reconstruction of known pathways. We devised a variation of this strategy in which the gene modules are defined by optimizing the "functional coherence" and size of the clusters. The functional coherence is calculated as the fraction of annotated gene pairs that share functional annotation in the given cluster,

$$\text{Coherence} = \frac{\#\text{ of gene pairs with any shared annotation}}{\text{total }\#\text{ of annotated gene pairs}},$$

and tends to be high for small clusters and diminishes as more genes are included. This trend is in turn balanced by a term that maximizes the size of the clusters in the calculation of the "modulation efficiency" at a given depth in the hierarchy,

$$\text{Modulation efficiency} = \sum_{i=1}^{n} \frac{(\text{Coherence}) \times (\#\text{ of genes in the cluster})}{\text{total }\#\text{ of genes in the network}}$$

where *n* is the number of clusters at the given tree depth. The level of the hierarchy with the highest modulation efficiency therefore reflects a compromise between the efficiency of clustering and the degree of functional association between genes in a cluster.

To create a network of gene modules (labeled "ModularNet" in Figure 1), rather than genes, we generated linkages between the modules as a function of the sets of linkages between their component genes. In this module network, nodes and edges represent defined functional clusters and functional association of pairs of clusters, respectively. The degree of association between two functional modules is defined as

$$\text{Distance between module A and B} = \sqrt{\frac{(\#\text{ of genes in A}) + (\#\text{ of genes in B})}{(\#\text{ of genes linked between A and B})}},$$

and was chosen to minimize the effects on the modular network of highly connected single proteins (i.e., single proteins with large numbers of linkages) and more accurately reflect the tendency for proteins in the respective modules to be linked (with the square root taken only to reduce overall network layout size).

Network layout and visualization

All gene networks were visualized in either two-dimensions (2D) or three-dimensions (3D) using the Large Graph Layout (LGL) package (available at http://bioinformatics.icmb.utexas.edu/lgl), treating the network as unweighted, undirected graphs (*43*). The module network was treated as a weighted graph with edge weights proportional to the inter-module distances defined above and with node weights proportional to the number of genes in each module, and visualized in 2D with Graphviz (http://www.research.att.com/~north/graphviz/) (*44*) and in 3D with LGL and Virtual Reality Modeling Language (VRML).

RESULTS

Measuring functional linkage accuracy with the log likelihood scoring scheme

The relative accuracy of each protein-protein interaction data set was measured on the 4 benchmark sets, and the resulting log likelihood scores are summarized in Table S3. In these tests, the small scale protein interaction data from the Database of Interacting Proteins act as a “gold standard” and serve to calibrate the high accuracy end of the measures. Despite the differences in score ranges across the different reference sets, the protein-protein interaction data sets are ranked in an essentially equivalent order regardless of scoring scheme or benchmark set, which implies that the ranking reflects the accuracy of functional linkage discovery in these experiments. As expected, the small scale experimental interaction set from the Database of Interacting Proteins show higher log likelihood scores than any of the large scale experimental interaction assays. Among the large scale interaction data sets, the TAP-tag purification/mass spectrometry analysis of Gavin *et al*. (*4*) exhibited the highest accuracy. The yeast two-hybrid data of Ito *et al.* (*6*) was divided into four subsets of interactions according to their reported reproducibility—log likelihood scores increase accordingly as the number of independent observations of the interaction increases.

The various functional linkage data sets for which independent confidence measures were available were also tested against the four benchmark sets, binning the linkages as described above according to their confidence measures. In each case, the tendency for a pair of genes to be functionally linked increases with the confidence measure associated with the data set. We analyzed mRNA co-expression-derived gene linkages (described in detail below), phylogenetic profile-derived gene linkages (Figure S2A), Rosetta Stone gene linkages (Figure S2B), and the literature mining (co-citation) gene linkages (Figure S2C), each of which shows a significant reconstruction of gene linkages as a function of their confidence measures. Each set of KEGG benchmark scores was fit by an appropriate regression curve (sigmoidal, or rational) and linkages derived from the data sets were assigned log likelihood scores from these KEGG-based regression curves. (Figure S2A-C and Figure S3).

Calculating functional linkages from mRNA expression data sets

Empirical tests indicated that we achieved our best performance from the mRNA co-expression data by first breaking the sets of microarray experiments into selected groups and analyzing for co-expression only within a group, integrating the linkages from these groups second, and only then combining these integrated mRNA expression-based linkages with linkages from other classes of data. Our rationale is as follows: If the complete set of 717 DNA microarray experiments for each gene were analyzed as a single monolithic expression vector, a pair of genes might be strongly and significantly co-expressed in a subset of experiments, but not all, and the resulting signal would be overwhelmed by the associated noise. Therefore, subdividing the microarray experiments into coherent subgroups and testing for co-expression only across a single subgroup allows us to discover the genes co-expressed in one but not all such groups.

We chose groups of experiments designed to perturb a single class of cellular systems (Table S1), following the experimental classes assigned by the Stanford Microarray Database, and then measured co-expression within each subset of experiments. Figure S5A shows that this “divide-test-integrate” approach significantly enhances the data mining performance in terms of both accuracy and coverage. Of 717 available experiments from SMD, 497 experiments (grouped into 12 categories labeled in bold text in Table S1) showed a significant correspondence between the correlation coefficient of co-expression and the log-likelihood scores for the KEGG and GO benchmarks. In each case, the tendency for a pair of genes to be functionally linked increases as the genes tendency to co-express across the set of microarray experiments increases, and genes that strongly co-express (i.e., show high correlation coefficients) exhibit high log likelihood scores, up to an accuracy higher than quality small scale experiment data set.

Linkages between pairs of genes in these 12 sets of experiments were assigned log likelihood scores from the regression fits on the KEGG benchmark data. These linkages were then integrated using the rank-weighted integration scheme described above, optimizing the relative weights between the 12 experiments by choosing the parameter *D* that maximized both accuracy and coverage of the KEGG benchmark set. For comparison, we also tested the performance of predicting functional links from mRNA co-expression across the complete set of 497 experiments (i.e., without subdividing into the 12 subgroups). Figure S4A shows poor performance, especially for the larger size of network, from both this complete set and the *naïve* Bayesian integration (*D* = 1) of the scores from the 12 subgroups, but as the dependency parameter *D* increases, the performance improves, with the best performance seen when *D* is very large (i.e., from taking only the maximum log likelihood score from across the 12 groups of experiments and ignoring the other 11 scores.) This implies that the linkages derived from the 12 experimental groups tend to be highly redundant—filtering the redundancy improves the overall calculation of functional linkages, especially by reducing systematic bias of information retrieval for certain cellular systems (Figure S4C).

To complete the calculation of functional linkages from the DNA microarray data, the integrated scores for the complete set of mRNA co-expression-derived functional links were re-scored in the log likelihood framework (Figure S4B), ensuring their proper weighting relative to links from other data sources. This set of linkages and scores represents the final set of links derived from mRNA co-expression (as analyzed in Figures 2).

Constructing the Initial Integrated Network ("IntNet")

Among the various sets of linkages, the co-expression-derived linkages show the most extensive coverage (> 60% of genome) with an accuracy equivalent to that of the "gold standard" DIP small scale interaction data. However, we expected that both the accuracy and coverage could be improved by adding in the linkages from all of the other data sources. Using the rank-order weighted integration scheme, linkages were integrated from the 11 distinct data sets: co-expression, phylogenetic profiles, Rosetta Stone links, literature mining, and the experimental interaction data listed in Table S3. For the purposes of linkage integration, all scores were derived from the KEGG benchmark tests, with the remaining three benchmark sets held aside as independent test sets. The data dependency parameter *D* was chosen to optimize the accuracy and coverage of the linkages (Figure S5A). Unlike for the integration of the various mRNA expresison data, the *naive* Bayesian integration ($D = 1$) performed relatively well, but not best. Optimal performance was seen with a value of $D = 1.5$, indicating that the data sets are still redundant, but much less than in the case of the mRNA expression-based linkages. To complete the construction of the initial network (referred to as "IntNet"), the integrated scores were re-scored in the log likelihood framework (Figure S5B).

Discovering Additional Linkages from Network Context ("ContextNet")

After the initial network reconstruction, we identified additional functional linkages by analyzing the genes' overall network neighborhoods (*34*-*37*) as described in Methods. Such linkages might be thought of as deriving from the total collection of indirect evidence for the genes' associations, as opposed to the direct evidence linking them.

Empirically, we observed that the quality of linkages identified depended strongly on the initial network's quality. To test this notion formally, we ordered all functional linkages by their initial integration scores and analyzed for network context-derived linkages using sequential subsets of 200,000 linkages to produce the networks. Not surprisingly, networks derived from linkages with lower average scores produced poorer quality context-derived linkages (Figure S6A), suggesting that this approach is best applied to high-quality networks. The final set of context linkages was derived from the IntNet which contains 290,560 linkages, without regard to whether or not the genes were previously linked. The quality of the resulting linkages is plotted in Figure S6B.

Constructing the Final Integrated Network ("FinalNet")

The final integrated network results from the combination of the direct and indirect evidence linking together gene pairs, and was derived by integrating the initial gene network with the context-derived linkages (see Figure 1). Specifically, the scores from IntNet and ContextNet were integrated with the rank-order weighting scheme (optimal $D = 8$; see Figure S6C), then re-scored in the log likelihood framework (see Figure S6D). To more fully evaluate the contribution of the context-derived linkages, we examined the quality of links derived only from network context as well as those derived from both network context and other evidence. Links derived only from network context showed excellent performance, comparable in accuracy to Ito yeast two hybrid data set with minimum 3 hits (Ito3) (Figure S6E). As expected, context-derived linkages that are also supported by other evidence considerably exceed the context-only linkages in quality.

The improvement of the networks throughout the integration process is summarized in Figure S7A-B. Not only did network accuracy and coverage improve, but the degree of clustering in the network (as defined by measuring the average clustering coefficient for each gene in the network (*40*)) increased as well, largely as a result of including the context-derived linkages, which serve to strengthen pre-existing clusters in the initial network. Visualized Networks also show substantially increased clustering of genes in FinalNet (Figure S7C and D). Topological properties of FinalNet and IntNet (each truncated to the top 34,000 linkages) indicate that both are small world networks (*40*) and their connectivity distributions fit a combination of power-law and exponential decay (*45*) rather than single power-law (*46*) (Table S4 and Figure S7E). Although FinalNet and IntNet are roughly comparable in their overall structure and properties, FinalNet exhibits a slightly higher degree of clustering (a local property of the networks) and a slightly longer average shortest pathlength between pairs of genes (a global property of the networks), indicating that the context-derived linkages serve to induce more order in the network (more specifically, the FinalNet resembles a random network's properties less than the IntNet.) The final gene linkages (FinalNet) are listed along with scores for the individual sources of evidence for the linkage, in the accompanying supplemental text file.

Assessing the Overall Quality of the Final Integrated Network ("ConfidentNet")

To evaluate the overall network quality, we compared the performance of the final network to the performance of the various component data sets using the 4 independent benchmark sets: KEGG and GO annotation sets (two independent pathway annotation sets), KOG annotations (providing general functional information for conserved protein families), and the UCSF set of experimentally observed GFP-protein fusion subcellular localizations. These comparisons are summarized in Figure 2 (for KEGG and GO) and S8 (for KOG and UCSF subcellular localizations). Although both KEGG and GO annotations are based on gene functions derived from the literature, they are surprisingly independent, with KEGG sharing only 31% of its linkages with GO (Figure S8C). On all 4 benchmarks, the final integrated network significantly outperforms all individual data sets, even the gold standard set of small scale interaction data. Results are consistent across the 4 benchmarks, suggesting that the tests reflect the true accuracies of the linkages, and that no strong bias was introduced into the networks during the integration procedure. With the small scale interactions to establish the acceptable level of accuracy on the KEGG benchmark, the final integrated network includes 34,000 linkages between 4,681 genes (80.5 % of the proteome; referred to as "ConfidentNet" in the Figure S1). Among the 34,000 linkages, 11,320 linkages are also supported by at least one of the other three benchmarks (GO, KOG, or UCSF subcellular localization) but not by KEGG. This suggests that the improved behavior of the final integrated network does not result from over-training on the KEGG benchmark set.

We tested for systematic bias in the representation of genes in the final network by calculating the gene retrieval rate for genes of different functional classes (as defined by MIPS). Figure S8D shows that proteins involved in protein synthesis were favored among high scoring linkages, probably reflecting these genes' tendency to be highly expressed and easily studied by most functional genomics methods. However, this bias decreased for the lower scores. The final set of 34,000 functional linkages exhibited little overall bias in gene representation, suggesting the bias present in individual data sets (*47*) was successfully reduced in the integration procedure.

In order to compare our functional linkage network against previously constructed protein interaction networks of yeast (*33*, *47*) we tested each network against the 4 pathway benchmark tests, KEGG, GO, KOG, and UCSF-GFP localization, as well as against a set of physical protein interactions derived from protein complexes (*33*), as shown in Figure S9. As expected, FinalNet represented a considerably more accurate and complete set of pathways, while the two protein interaction networks scored better on the physical interaction test. In fact, comparison of FinalNet with an integrated physical protein interaction network derived from at least several of the same data sets (*33*) shows only a small intersection of the linkages (Figure S9F).

Defining Modules of Genes in the Final Integrated Network ("ModularNet")

The resulting final network of genes is highly complex (Figure S7C-D). In order to discover and more conveniently describe the genes' organization, we searched for coherent modules of genes in the network. In short, we've used unsupervised clustering techniques to reveal the higher order organization of the genes. This approach might be considered a "bottom-up" approach to studying the systems of genes (i.e., letting the network connections reveal the genes' intrinsic patterns of organization), in contrast to surveying how known gene functions are distributed across the network (a "top-down" approach), and has the attractive feature of potentially revealing new systems and connections not yet catalogued in existing gene hierarchies.

Hierarchical clustering (average linkage) was applied to the final integrated network (see Methods). Genes were divided into groups according to the hierarchy, essentially by "slicing" the hierarchical tree at a given level and assigning genes still connected in the tree into the same module. In this way, clusters with different numbers of genes and different degrees of functional coherence can be produced (Figures S10A-B), with clusters ordered according to the hierarchical tree in such a way that neighboring clusters tend also to be functionally related (Figure S10C). For clarity, clusters containing only one gene were discarded.

Near the highest level of the tree, 54 modules could be defined, each containing large numbers of genes, that show the broadest level of organization in the networks (Figure S11; genes within each cluster are listed in the accompanying supplementary text file. The interactions among major clusters can be interactively visualized in 3D using a VRML viewer on the associated supplemental VRML file (denoted by the .wrl filename extension.) By raising the threshold at which the hierarchy is cut (i.e., requiring greater similarity between genes in a module), more modules are produced with proportionally fewer genes in each module. For example, at an intermediate level of the hierarchy (where cutoff of similarity between clusters is 0.5), 669 modules are produced, each much more coherent in function. We sought a set of modules that were maximally functionally coherent, yet as large as possible—for this task, we calculated the "modulation efficiency" of each clustering and chose the clustering that maximized this value (Figure S10D). The resulting set (where cutoff of similarity between clusters is 0.39) of 627 gene modules contained 3285 genes (70.2 % of the genes in the network), which are listed in the accompanying supplementary text file (Interactive visualization is also available for this in the accompanying VRML file (indicated by the .wrl filename extension). Labeling the genes according to their major functions in the cell (using MIPS annotations) revealed that the modules were functionally quite coherent (Figure 3B and S12), and also tends to be linked to modules of similar function. Clusters contained 21.3% essential genes (exceeding the 99% confidence interval) as compared to 17.0% for genes which failed to cluster at this threshold.

In practice, we found the modular view of the network to considerably ease the analysis of systems and their connections, as described in more detail in the paper for several specific local regions of the network, such as the systems of DNA repair, vesicle transport, and RNA processing (e.g., see Figure S13). The hierarchical organization of genes in the network is evident in the many occurrences of groupings of functionally-related clusters, such as for energy metabolism, cellular transport, and RNA processing systems. Conversely, modules of cell cycle regulatory genes that are connected to a functionally diverse range of other modules, are not themselves clustered but are distributed throughout the network, including protein fate, cellular transport, signal transduction, metabolism, and transcription. Gene clusters composed primarily of uncharacterized genes or clusters with no dominant function are also numerous.

Testing an Alternate Benchmarking Strategy Based on Cross-Validation

As the integration method we describe relies critically on benchmarks in order to weight data, we also experimented with an alternate method of generating benchmarks: we pooled annotation sets from KEGG, GO, and MIPS, then split the pooled annotation sets into independent training and test benchmark sets. Because data are integrated using weights derived only from the training benchmark, the performances measured on the remaining test benchmark are expected to be free from circular logic and memorization of the annotation set during the training procedure.

Specifically, we generated a pooled annotation set by assembling all pairs of yeast genes that shared the same pathway key at the lowest level of KEGG annotation, that shared the same biological process key at the 8th level of GO process annotation, or that belonged to the same cellular complex key in either the GO components or lowest level MIPS annotation. In total, this represented 121,800 links between 3,390 unique genes. We separated these into disjoint training and test sets by randomly separating the set of 3,390 genes into 2 subsets of 1,695 genes each, retaining all links among genes within the same subset. The resulting training set contained 30,574 links among 1,666 genes; the test set contained 30,261 links among 1,655 genes, with neither links nor genes shared between the sets. The network integration was performed using only the training set for calculating weights and all other steps prior to the final assessment of network accuracy, which was performed on the independent test set. The network derived in this manner is consistent in quality and content with that derived from the KEGG set alone (Figure S14): A comparison of the actual linkages in the top 34,000 linkages derived under the two different training/test regimens shows that 26,736 gene pairs (79%) are shared among the IntNet networks, and 24,599 gene pairs (72%) are shared between the FinalNet networks.

FIGURES

Figure S1

Graphic illustration of unified scoring scheme for a mRNA expression (SMD cellcycle) data set. (A) The P(L|E) and ~P(L|E) can be obtained by measuring frequency of gene pairs operating in the same pathway/process and those in different ones, respectively. All possible gene pairs were sorted by correlation coefficient between expression vectors, and frequencies of genes' sharing or not sharing pathways/processes, based on KEGG or GO process 8th level annotation, were measured for bins of 20,000 gene pairs. The ~P(L|E) is relatively constant across the entire range of correlation coefficients, whereas the P(L|E) is positively correlated with it. In this data set (and many other mRNA expression data sets), P(L|E) is lower that ~P(L|E) for the range of lower correlation coefficient (e.g. < ~0.3 for the SMD cellcycle data set). However, P(L|E) surpasses ~P(L|E) for the range of high correlation coefficient where P(L|E) drastically increases as the correlation coefficient increases. The likelihood, OR(L, E) can be calculated from these probabilities of interaction/non-interaction, conditioned on the given evidence (the given value of correlation coefficient), and from the unconditional probability of interaction/non-interaction. (B) A plot of the log likelihood score, calculated as the natural logarithm of OR(L, E). Note that the log likelihood scores based on two different pathway/process references, KEGG and GO, show very similar distributions.

Figure S1

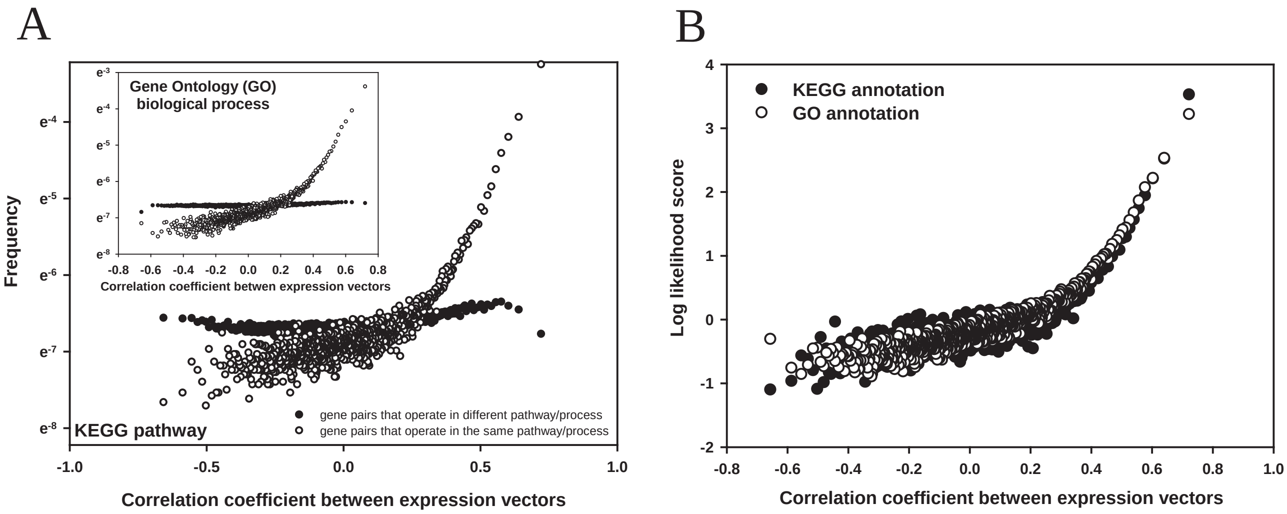

A
Frequency
e-4
e-5
e-6
e-7
e-8
Gene Ontology (GO)
biological process
e-3
e-4
e-5
e-6
e-7
e-8
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
Correlation coefficient betwen expression vectors
● gene pairs that operate in different pathway/process
○ gene pairs that operate in the same pathway/process
KEGG pathway
-1.0
-0.5
0.0
0.5
1.0
Correlation coefficient between expression vectors
B
Log likelihood score
4
3
2
1
0
-1
-2
● KEGG annotation
○ GO annotation
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Correlation coefficient between expression vectors

Figure S2

Performance of the computational methods for reconstructing pathways. Each plot shows regression curve fits of association between the method-specific parameters and the log likelihood scores, reflecting the method's ability to identify biological pathways for three different in-silico methods. The computational methods are (A) Phylogenetic profiling (fit by 4 parameter sigmoid curve). (B) Rosetta stone (or gene fusion) method (fit by 3 parameter rational curve). (C) Co-citation (fit by 3 parameter rational curve). Regression curves were fit based on the KEGG annotation data.

Figure S2

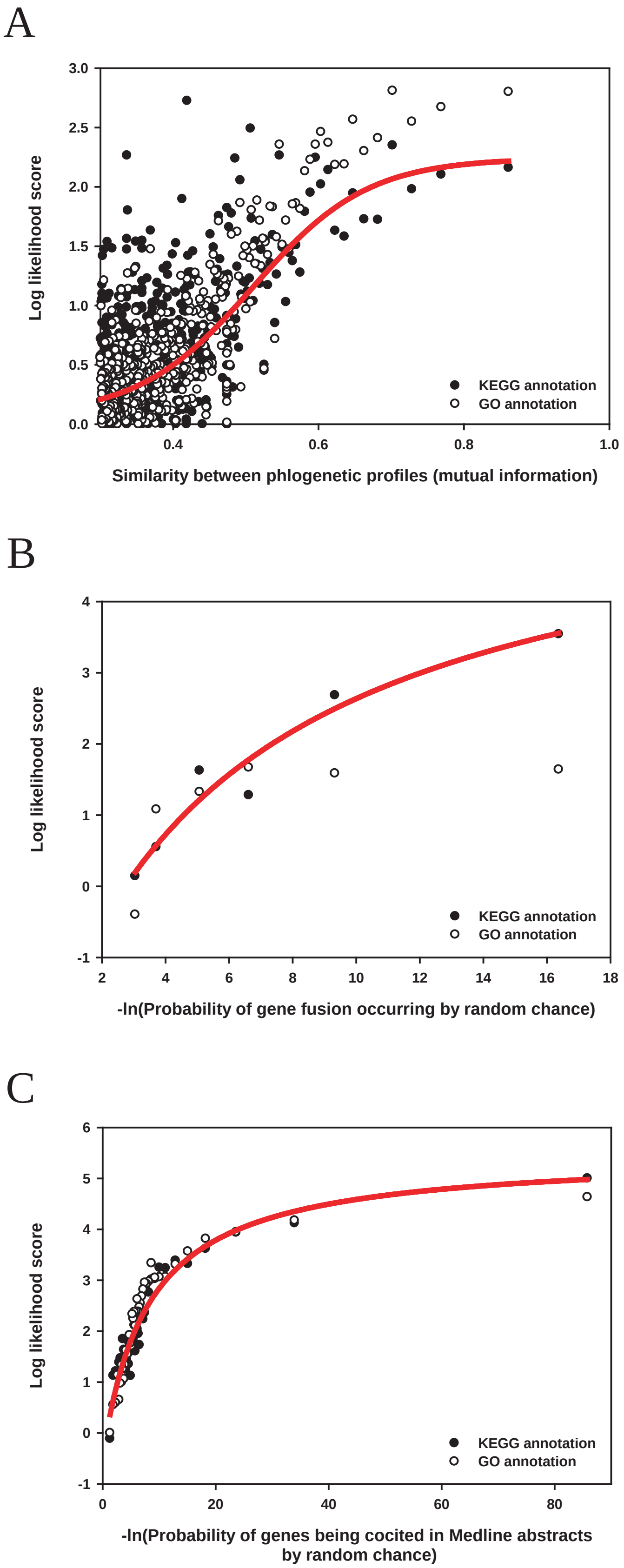

A
Log likelihood score
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.4
0.6
0.8
1.0
Similarity between phlogenetic profiles (mutual information)
● KEGG annotation
○ GO annotation
B
Log likelihood score
4
3
2
1
0
-1
2
4
6
8
10
12
14
16
18
-ln(Probability of gene fusion occurring by random chance)
● KEGG annotation
○ GO annotation
C
Log likelihood score
6
5
4
3
2
1
0
-1
0
20
40
60
80
-ln(Probability of genes being cocited in Medline abstracts
by random chance)
● KEGG annotation
○ GO annotation

Figure S3

Regression curve fits of association between gene pairs' mRNA co-expression (measured by correlation coefficient; CC) and the agreement with biological pathways (measured by the log likelihood score; LLS) for 12 categorized DNA microarray expression sets (*2*) listed in Table S1. In each case, the performances on two independent benchmark reference sets, KEGG pathway (filled circle) and 8pGO (open circle), agree well. Only regression curves for the KEGG pathway data are plotted. All fit by 4 parameter sigmoid curves.

Figure S3

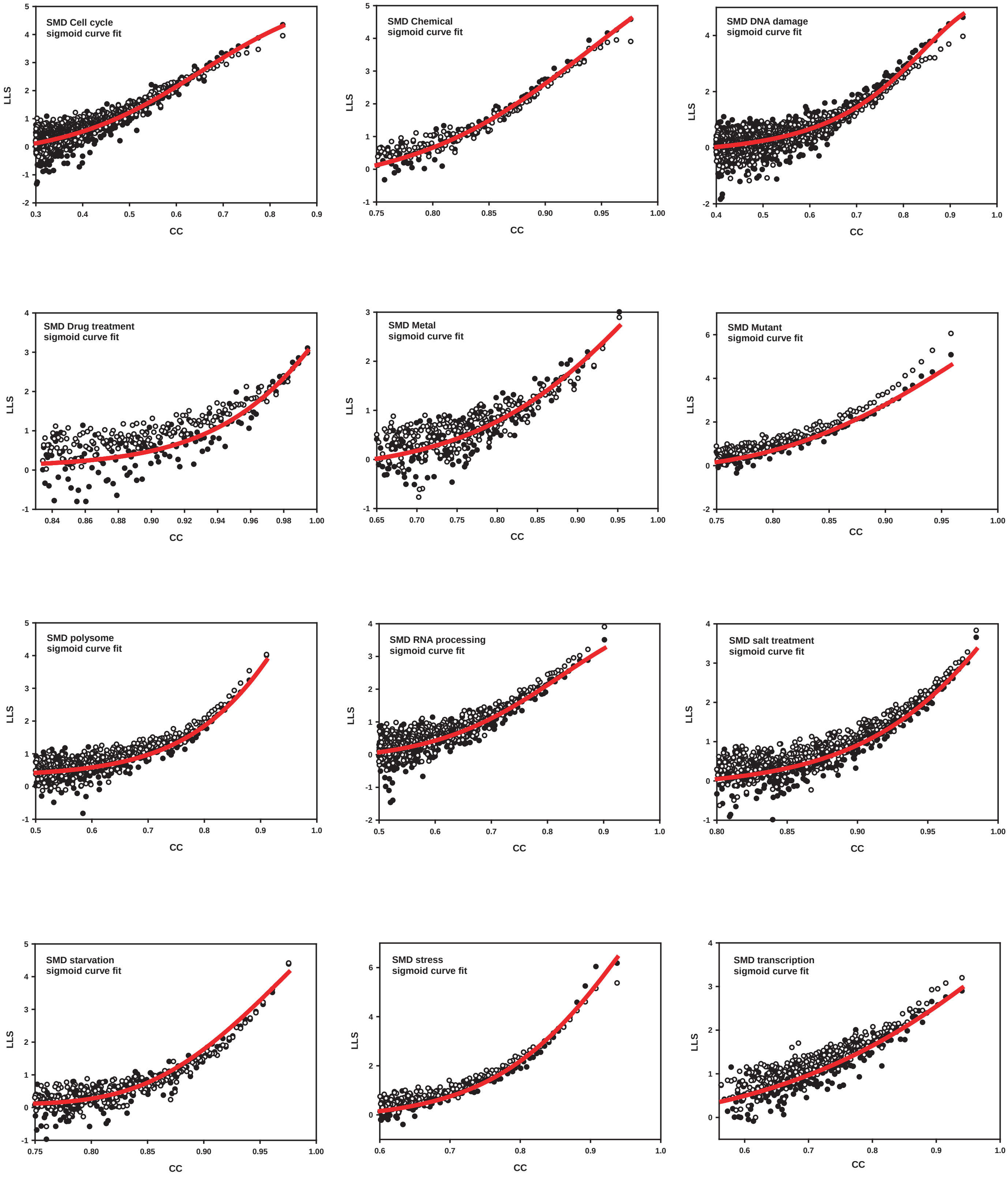
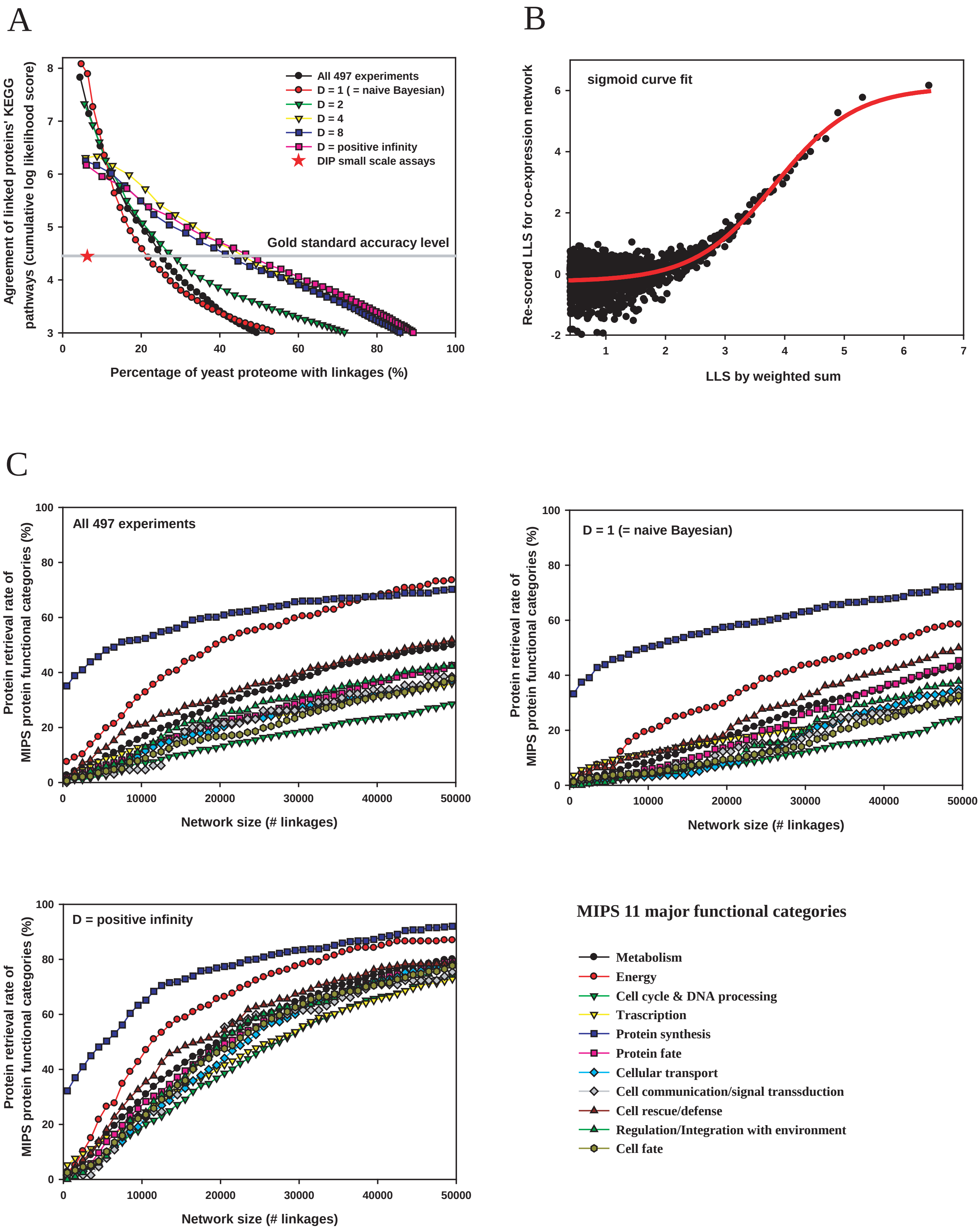

SMD Cell cycle
sigmoid curve fit
LLS
CC
5
4
3
2
1
0
-1
-2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
SMD Chemical
sigmoid curve fit
LLS
CC
5
4
3
2
1
0
-1
0.75
0.80
0.85
0.90
0.95
1.00
SMD DNA damage
sigmoid curve fit
LLS
CC
4
2
0
-2
0.4
0.5
0.6
0.7
0.8
0.9
1.0
SMD Drug treatment
sigmoid curve fit
LLS
CC
4
3
2
1
0
-1
0.84
0.86
0.88
0.90
0.92
0.94
0.96
0.98
1.00
SMD Metal
sigmoid curve fit
LLS
CC
3
2
1
0
-1
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
SMD Mutant
sigmoid curve fit
LLS
CC
6
4
2
0
-2
0.75
0.80
0.85
0.90
0.95
1.00
SMD polysome
sigmoid curve fit
LLS
CC
5
4
3
2
1
0
-1
0.5
0.6
0.7
0.8
0.9
1.0
SMD RNA processing
sigmoid curve fit
LLS
CC
4
3
2
1
0
-1
-2
0.5
0.6
0.7
0.8
0.9
1.0
SMD salt treatment
sigmoid curve fit
LLS
CC
4
3
2
1
0
-1
0.80
0.85
0.90
0.95
1.00
SMD starvation
sigmoid curve fit
LLS
CC
5
4
3
2
1
0
-1
0.75
0.80
0.85
0.90
0.95
1.00
SMD stress
sigmoid curve fit
LLS
CC
6
4
2
0
0.6
0.7
0.8
0.9
1.0
SMD transcription
sigmoid curve fit
LLS
CC
4
3
2
1
0
0.6
0.7
0.8
0.9
1.0

Figure S4

Integration of the gene functional linkages from the12 different DNA microarray data sets. (A) Benchmarking, by measuring the agreement of linked proteins' KEGG pathways (accuracy) and the percentage of the yeast proteome with linkages (coverage), of the resulting functional networks created by integrating mRNA co-expression data sets with different assumptions of independence. Treating the entire set of mRNA expression data as a single set of 497 element vectors, the scoring co-expression based on correlation of these vectors is indicated by 'All 497 experiments'. The remaining curves show different methods of integrating linkages from the 12 separate data sets, as described in the text: Integration with free parameter D = 1 is mathematically equivalent to naïve Bayesian approach. Integration with D = positive infinity can be performed, in practice, by taking only the largest log likelihood scores for each gene pairs. Here, integration shows the best performance at the level of our 'Gold standard' (the DIP small scale assays data set) where D = positive infinity. (B) For incorporation of these integrated linkages with other linkages, they are re-scored. Here the top-scoring integrated log likelihood scores (achieved with D = positive infinity) re-scored on the log likelihood score test. (C) The top-scoring method of integration (D = positive infinity) also shows the least functional bias, as measured by counting the rate at which linkages are generated for the annotated yeast genes in each of 11 major functional categories by MIPS (*28*). The naïve Bayesian integration (D = 1) is strongly biased to include links from protein synthesis, but the top-scoring method not only achieves improved accuracy and coverage but also reduced systematic bias in the functional annotations.

Figure S4



A
Agreement of linked proteins' KEGG pathways (cumulative log likelihood score)
Percentage of yeast proteome with linkages (%)
All 497 experiments
D = 1 (= naive Bayesian)
D = 2
D = 4
D = 8
D = positive infinity
DIP small scale assays
Gold standard accuracy level
B
sigmoid curve fit
Re-scored LLS for co-expression network
LLS by weighted sum
C
All 497 experiments
D = 1 (= naive Bayesian)
D = positive infinity
Protein retrieval rate of MIPS protein functional categories (%)
Network size (# linkages)
MIPS 11 major functional categories
Metabolism
Energy
Cell cycle & DNA processing
Trascription
Protein synthesis
Protein fate
Cellular transport
Cell communication/signal transsduction
Cell rescue/defense
Regulation/Integration with environment
Cell fate

Figure S5

Construction of IntNet (Figure 1) by integration of gene functional linkage information of all available functional genomics data sets. The integration mRNA expression information from the previous step was used in this second step integration. (A) Benchmarking, as described in Figure S4A, of gene functional network by integration of all available functional genomics data with different degree of dependence. The best performance is observed where D = 1.5, indicating that the different functional genomics data sets are reasonably independent in terms of the linkages represented. (B) Re-scoring of original log likelihood scores to new ones after the integration of all functional genomics data.

Figure S5

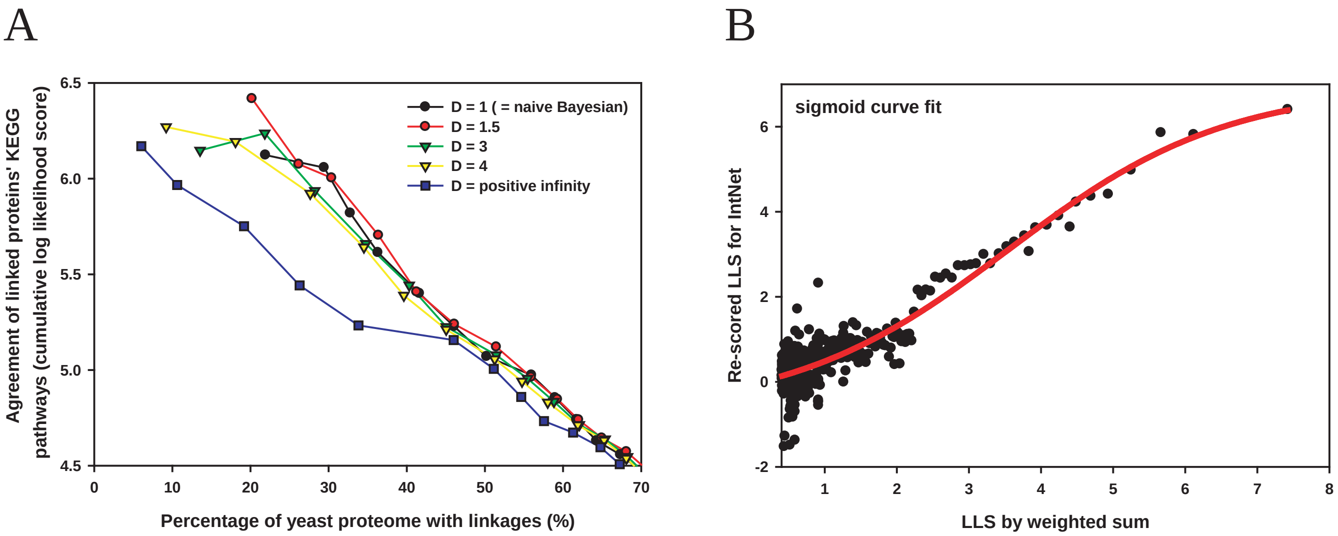

A
B
6.5
6.0
5.5
5.0
4.5
0
10
20
30
40
50
60
70
Agreement of linked proteins' KEGG pathways (cumulative log likelihood score)
Percentage of yeast proteome with linkages (%)
D = 1 (= naive Bayesian)
D = 1.5
D = 3
D = 4
D = positive infinity
sigmoid curve fit
6
4
2
0
-2
1
2
3
4
5
6
7
8
Re-scored LLS for IntNet
LLS by weighted sum

Figure S6

Construction of ContextNet and FinalNet (Figure 1). (A) Testing the importance of the quality of the initial network (IntNet) on the quality of the resulting network created by introducing context-derived linkages into the initial network (ContextNet). Three different initial networks with identical size but different quality were made by replacing high quality linkages with low ones (essentially, ranking linkages by their LLS scores and choosing 200,000 successive linkages to form the network, selecting linkages 1-200,000 as the most accurate set, 5,000-205,000 as the second most accurate, and so on). A poorer quality initial network results in much lower cumulative log likelihood score ranges in ContextNet. (B) Re-scoring of the network neighborhood similarity (measured by correlation coefficient) in the log likelihood scoring scheme (LLS). ContextNet is defined using these re-scored log likelihood scores. (C) Choosing the optimal integration of the initial linkages with the context-derived linkages by benchmarking, as described in Figure S4A, the functional gene networks resulting from different choices of the dependency parameter D. A high degree of dependence, where D is 8, shows optimized performance, and naïve Bayesian performs poorly. (D) Re-scoring of the original log likelihood scores after the (weighted) integration of initial linkages with context-derived linkages. (E) Assessment, as described in Figure S4A, of the quality of linkages derived by analysis of network context. Linkages derived from only network context (i.e. without supporting evidence from other functional genomics data) show significant accuracy, while linkages supported by both direct functional genomics evidence and network context have higher accuracy for smaller networks but lower for larger ones. The total set of linkages derived by this context-based approach (ContextNet) shows improved accuracy and coverage over the initial network (IntNet).

Figure S6

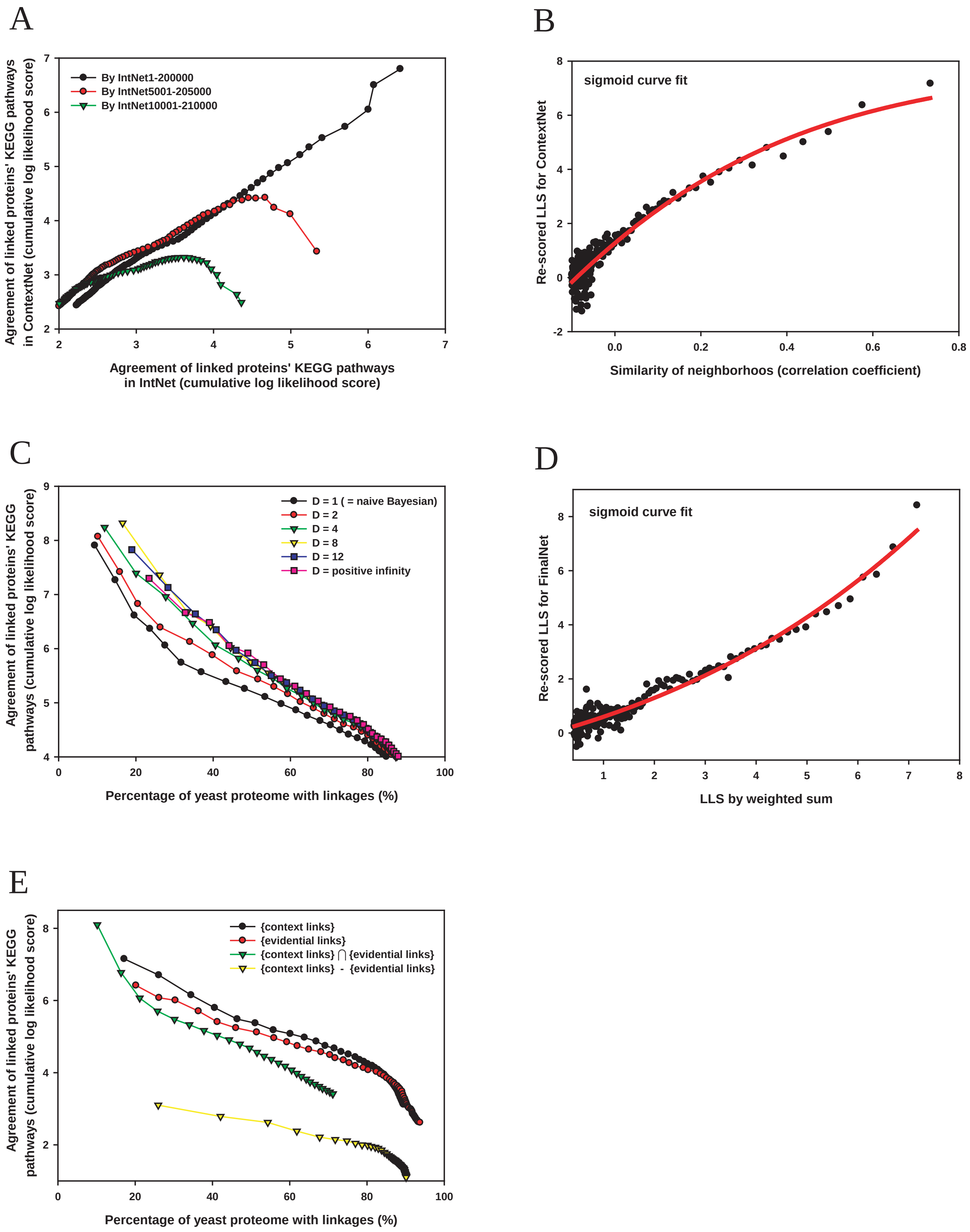

A
Agreement of linked proteins' KEGG pathways in ContextNet (cumulative log likelihood score)
Agreement of linked proteins' KEGG pathways in IntNet (cumulative log likelihood score)
By IntNet1-200000
By IntNet5001-205000
By IntNet10001-210000
2
3
4
5
6
7
B
sigmoid curve fit
Re-scored LLS for ContextNet
Similarity of neighborhoos (correlation coefficient)
-2
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
C
Agreement of linked proteins' KEGG pathways (cumulative log likelihood score)
Percentage of yeast proteome with linkages (%)
D = 1 (= naive Bayesian)
D = 2
D = 4
D = 8
D = 12
D = positive infinity
4
5
6
7
8
9
0
20
40
60
80
100
D
sigmoid curve fit
Re-scored LLS for FinalNet
LLS by weighted sum
0
2
4
6
8
1
2
3
4
5
6
7
8
E
Agreement of linked proteins' KEGG pathways (cumulative log likelihood score)
Percentage of yeast proteome with linkages (%)
{context links}
{evidential links}
{context links} ⋂ {evidential links}
{context links} - {evidential links}
2
4
6
8
0
20
40
60
80
100

Figure S7

(A) A summary of the improvement in quality of the functional network with each step of the linkage integration procedure. Data integration noticeably improves the network. (B) A summary of the degree of clustering in the network after each integration step. Note that data integration significantly increases clustering of genes in the network, representing improved definition of functional modules/systems. The addition of context-derived linkages also significantly increases clustering, but the effect is moderated somewhat in the final network. (C-D) A comparison of the initial integrated functional network (C) (plotting the top-scoring 24,000 linkages of IntNet) with the final integrated network (D) (plotting the top-scoring 24,000 linkages of FinalNet). While both networks show local clustering, the inclusion of context-derived linkages in FinalNet results in more extensive clustering of genes into modules, evident in the "clumping" in (D). Each network is visualized with LGL (*43*). (E) The connectivity distribution of IntNet and FinalNet was assessed using the top ranked 34,000 linkages from each (i.e. IntNet34000 and FinalNet34000 (= ConfidentNet)). Both networks' connectivity distributions were fit by a combined power-law and exponential decay function (*45*), $f(x) = a(1+x)^{-b}exp^{-cx}$ with r^2 for the fit greater than 0.99 (IntNet34000: a = 1487 ± 18, b = 0.87 ± 0.02, c = 0.045 ± 0.002; FinalNet34000: a = 1018 ± 18, b = 0.39 ± 0.02, c = 0.104 ± 0.004).

Figure S7

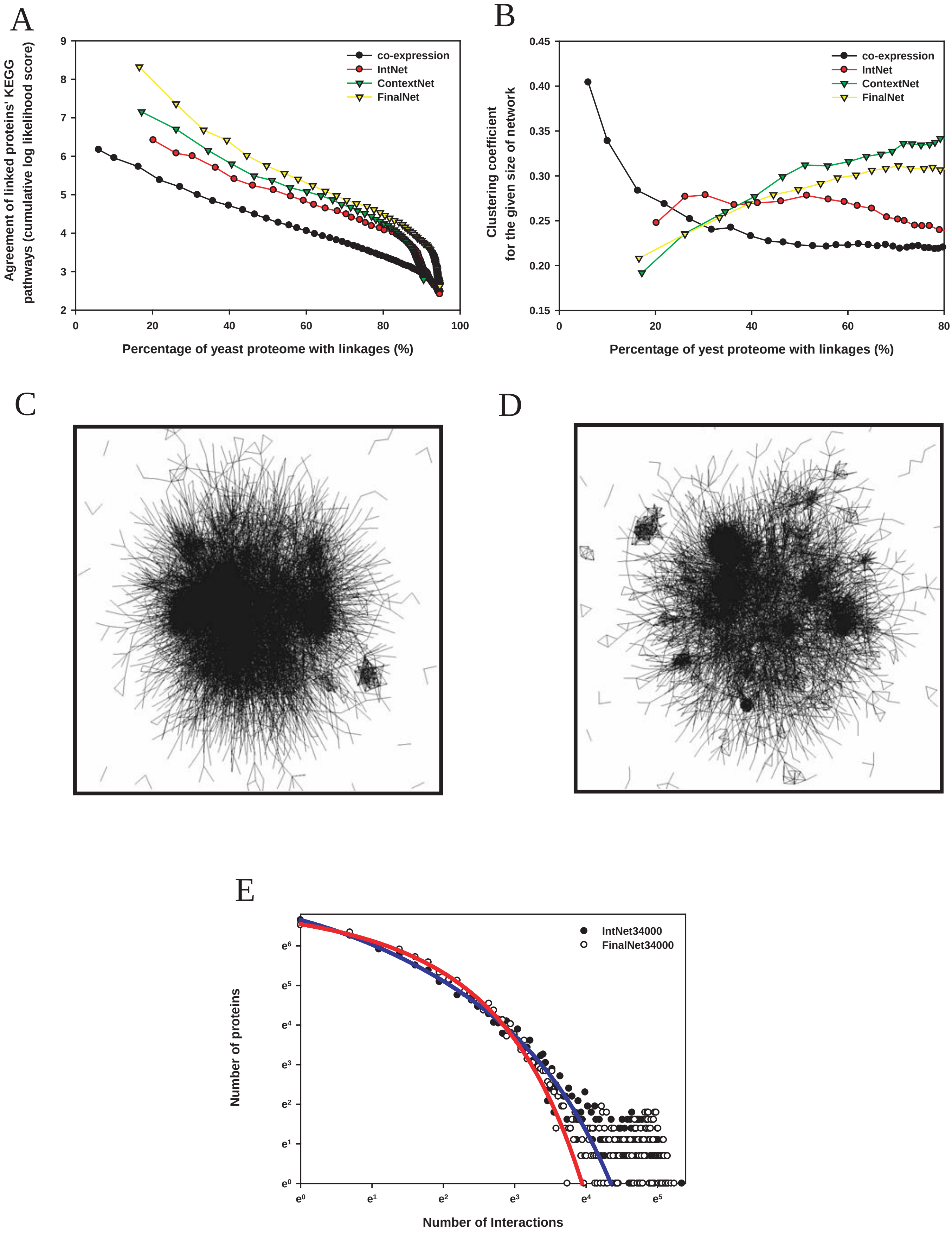

A
Agreement of linked proteins' KEGG pathways (cumulative log likelihood score)
9
8
7
6
5
4
3
2
0
20
40
60
80
100
Percentage of yeast proteome with linkages (%)
co-expression
IntNet
ContextNet
FinalNet
B
Clustering coefficient for the given size of network
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0
20
40
60
80
Percentage of yest proteome with linkages (%)
co-expression
IntNet
ContextNet
FinalNet
C
D
E
Number of proteins
e^6
e^5
e^4
e^3
e^2
e^1
e^0
e^0
e^1
e^2
e^3
e^4
e^5
Number of Interactions
● IntNet34000
○ FinalNet34000

Figure S8

In addition to the benchmarks of Figure 2, the accuracy and extent of each functional genomics data set and the resulting integrated networks were measured on two additional benchmark annotation sets. A critical point is the comparable performance of the networks on all four independent benchmarks. Besides the tests of Figure 2, we assessed the tendencies for linked genes to share (A) GO process annotations (*24*) and (B) KOG annotation (*26*) Each x-axis indicates the percentage of protein encoding yeast genes provided with linkages by the plotted data sets; each y-axis indicates the measured agreement of the linked genes' annotations on one of the four benchmark sets. The "gold standards" of accuracy, used to calibrate the benchmarks and indicated by a red star, are small scale protein-protein interaction data from the Database of Interaction Proteins (DIP) (*3*). Experimental data sets are indicated by colored markers, computational by gray markers. The initial integrated network (lower black line), although trained only using the KEGG benchmark, has measurably higher accuracy than any of the individual data sets on each of the four benchmarks; adding context-inferred linkages to create the final integrated network (upper black line) further improves the accuracy and extent of the network. (C) A Venn diagram showing annotation overlap among ConfidentNet, KEGG pathway, and GO biological process 8th level. Notice that the redundancy of annotation among them is minimal, with only 31% of the linkages derived from the KEGG set also present in the GO set. (D) The final network shows little representational bias for different gene functions, as measured by the percent of genes in each major functional category (defined by MIPS (*28*)) incorporated in the network as a function of network size. Approx. 90% of the genes in each MIPS functional category are included in the network at the gold standard confidence level (~34,000 linkages).

Figure S8

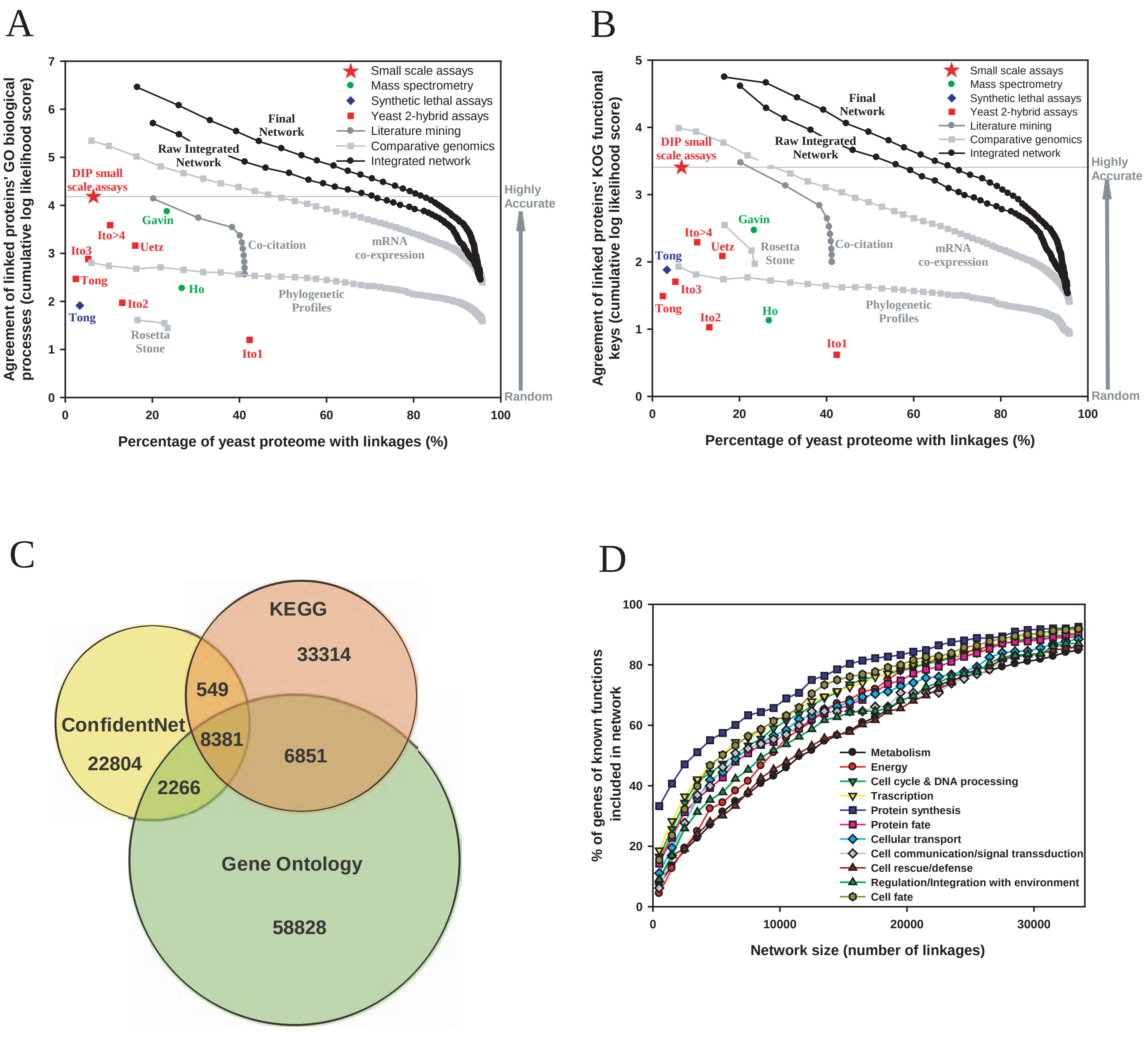

A
Agreement of linked proteins' GO biological processes (cumulative log likelihood score)
0
1
2
3
4
5
6
7
Percentage of yeast proteome with linkages (%)
0
20
40
60
80
100
Small scale assays
Mass spectrometry
Synthetic lethal assays
Yeast 2-hybrid assays
Literature mining
Comparative genomics
Integrated network
Final Network
Raw Integrated Network
DIP small scale assays
Gavin
Ito>4
Uetz
Ito3
Tong
Ho
Ito2
Tong
Rosetta Stone
Ito1
Co-citation
mRNA co-expression
Phylogenetic Profiles
Highly Accurate
Random
B
Agreement of linked proteins' KOG functional keys (cumulative log likelihood score)
0
1
2
3
4
5
Percentage of yeast proteome with linkages (%)
0
20
40
60
80
100
Small scale assays
Mass spectrometry
Synthetic lethal assays
Yeast 2-hybrid assays
Literature mining
Comparative genomics
Integrated network
Final Network
Raw Integrated Network
DIP small scale assays
Gavin
Ito>4
Uetz
Tong
Rosetta Stone
Co-citation
mRNA co-expression
Ito3
Tong
Ito2
Ho
Phylogenetic Profiles
Ito1
Highly Accurate
Random
C
KEGG
33314
549
ConfidentNet
8381
6851
22804
2266
Gene Ontology
58828
D
% of genes of known functions included in network
0
20
40
60
80
100
Network size (number of linkages)
0
10000
20000
30000
Metabolism
Energy
Cell cycle & DNA processing
Trascription
Protein synthesis
Protein fate
Cellular transport
Cell communication/signal transsduction
Cell rescue/defense
Regulation/Integration with environment
Cell fate

Figure S9

Comparison of FinalNet to previously re-constructed physical interaction networks of yeast proteins (*33*, *47*). FinalNet, two other yeast protein networks, and the "gold standard" DIP small scale interaction data, were assessed as described in Figure S4A with five different reference sets; (A) KEGG, (B) GO 8th level process annotation, (C) KOG, (D) UCSF-GFP localization, and (E) Jansen's "gold standard" composed of known protein complexes based on small scale protein interaction experiments (*33*). FinalNet represents a more accurate and extensive pathway set, based on functional (A, B), protein co-evolutionary (C), and sub-cellular location (D), but the highest scoring physical interaction data sets score better than FinalNet on the physical interaction benchmark set (E) (although FinalNet achieves better accuracies for large networks), presumably because FinalNet has been "trained" by pathway data. (F) This interpretation is supported by the relatively small intersection between ConfidentNet (top 34,000 linkages of FinalNet) and the Jansen *et al.* physical protein complex network (with L > 600) (*33*), in which only 16 % of ConfidentNet linkages are found.

Figure S9

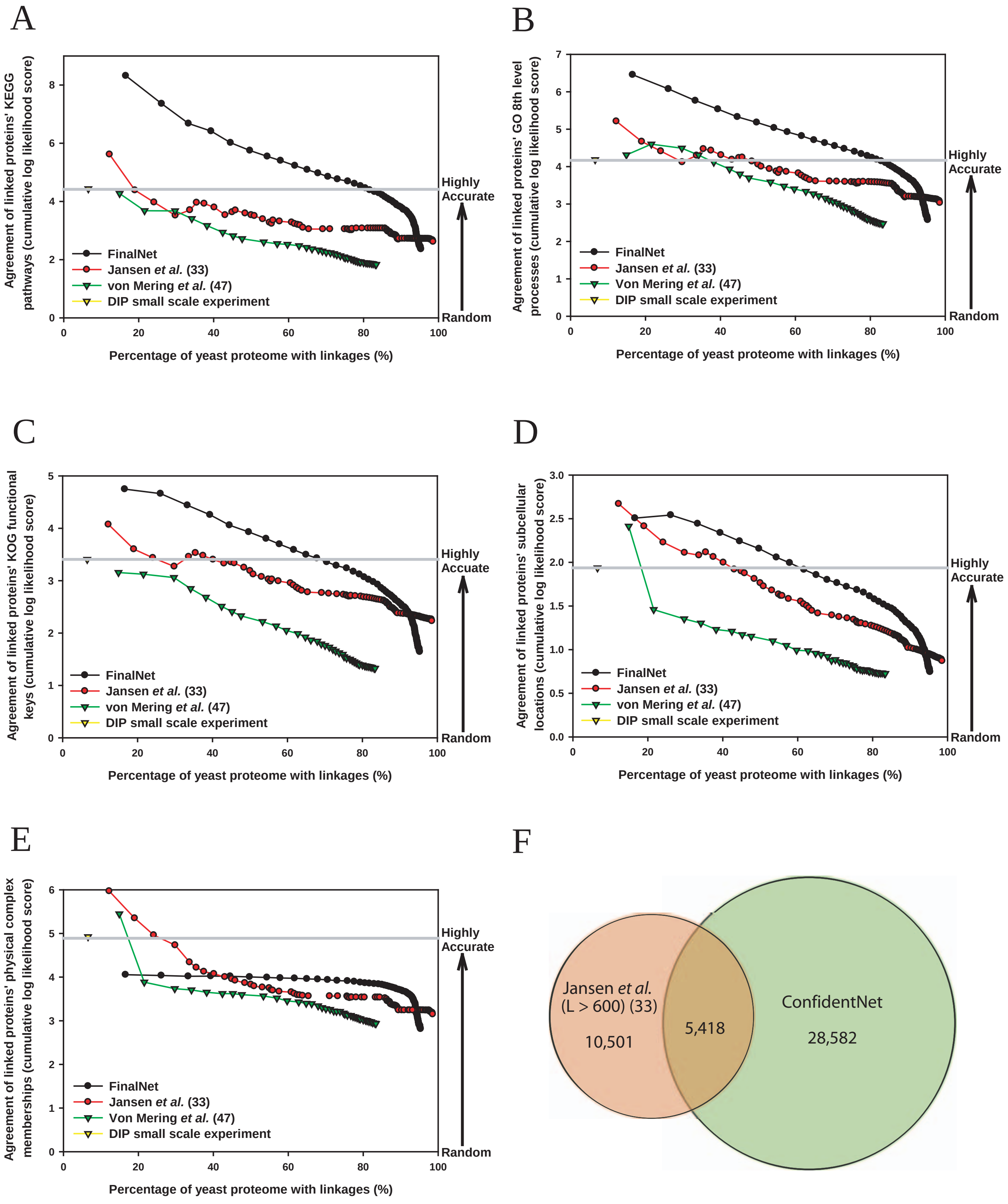

A
Agreement of linked proteins' KEGG pathways (cumulative log likelihood score)
0
2
4
6
8
0
20
40
60
80
100
Percentage of yeast proteome with linkages (%)
Highly Accurate
Random
FinalNet
Jansen et al. (33)
von Mering et al. (47)
DIP small scale experiment
B
Agreement of linked proteins' GO 8th level processes (cumulative log likelihood score)
0
1
2
3
4
5
6
7
0
20
40
60
80
100
Percentage of yeast proteome with linkages (%)
Highly Accurate
Random
FinalNet
Jansen et al. (33)
Von Mering et al. (47)
DIP small scale experiment
C
Agreement of linked proteins' KOG functional keys (cumulative log likelihood score)
0
1
2
3
4
5
0
20
40
60
80
100
Percentage of yeast proteome with linkages (%)
Highly Accurate
Random
FinalNet
Jansen et al. (33)
von Mering et al. (47)
DIP small scale experiment
D
Agreement of linked proteins' subcellular locations (cumulative log likelihood score)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
20
40
60
80
100
Percentage of yeast proteome with linkages (%)
Highly Accurate
Random
FinalNet
Jansen et al. (33)
von Mering et al. (47)
DIP small scale experiment
E
Agreement of linked proteins' physical complex memberships (cumulative log likelihood score)
0
1
2
3
4
5
6
0
20
40
60
80
100
Percentage of yeast proteome with linkages (%)
Highly Accurate
Random
FinalNet
Jansen et al. (33)
Von Mering et al. (47)
DIP small scale experiment
F
Jansen et al. (L > 600) (33)
10,501
5,418
ConfidentNet
28,582

Figure S10

Finding the optimized similarity cutoff for defining hierarchical clusters. We tested the dependence of desirable properties of the resulting clusters and chose a similarity cutoff that maximized these properties; including (A) the differences in the number of clusters (B) in the number of proteins belong to any cluster and (C) in the mean degree of functional coherence of the clusters. (D) We defined function, the modulation efficiency, based on these 3 parameters, then chose the similarity cutoff that maximized the modulation efficiency. For the hierarchically clustered ConfidentNet, we found that a cutoff of similarity of 0.39 achieved the maximized modulation efficiency, producing 627 gene clusters.

Figure S10

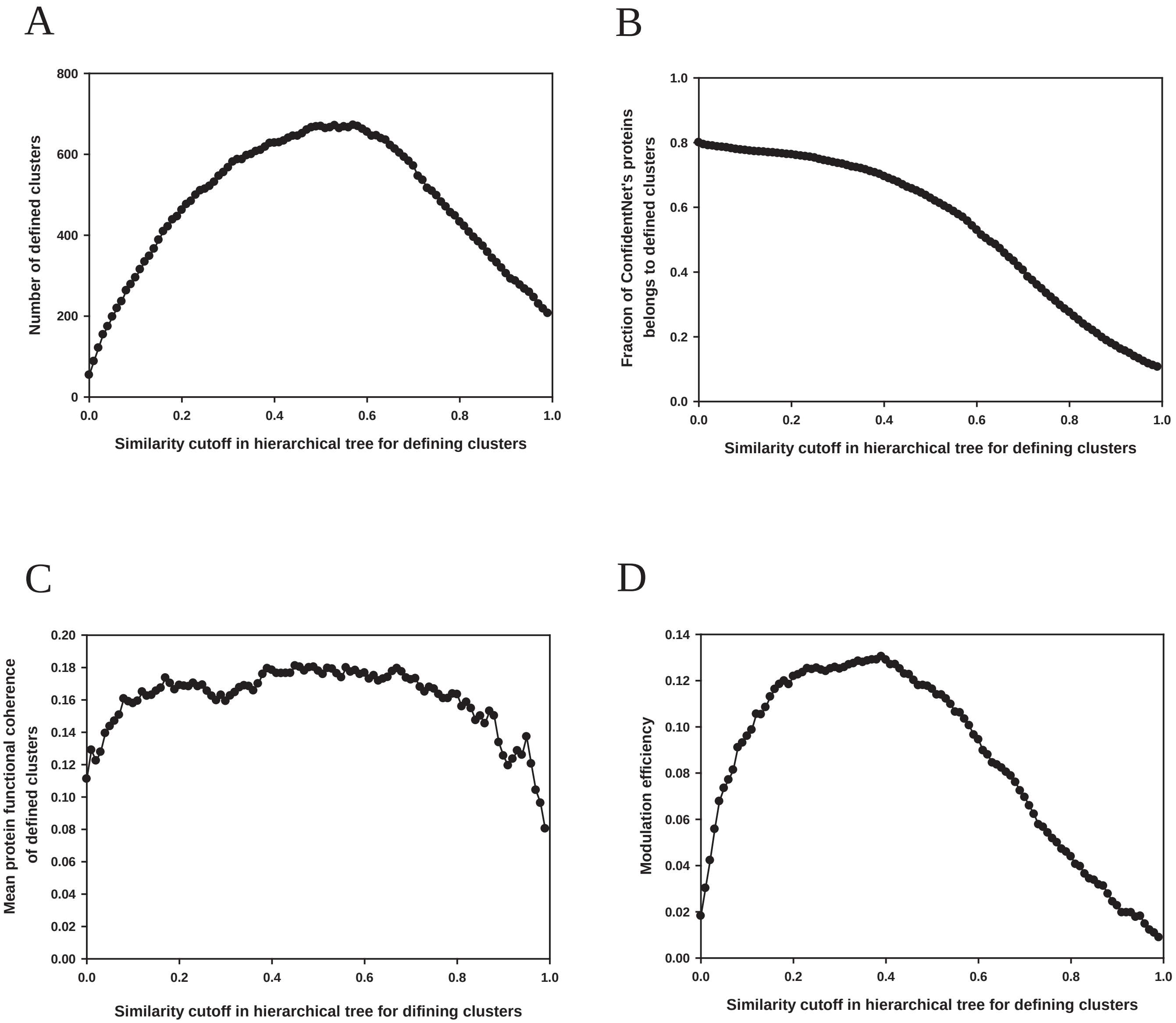
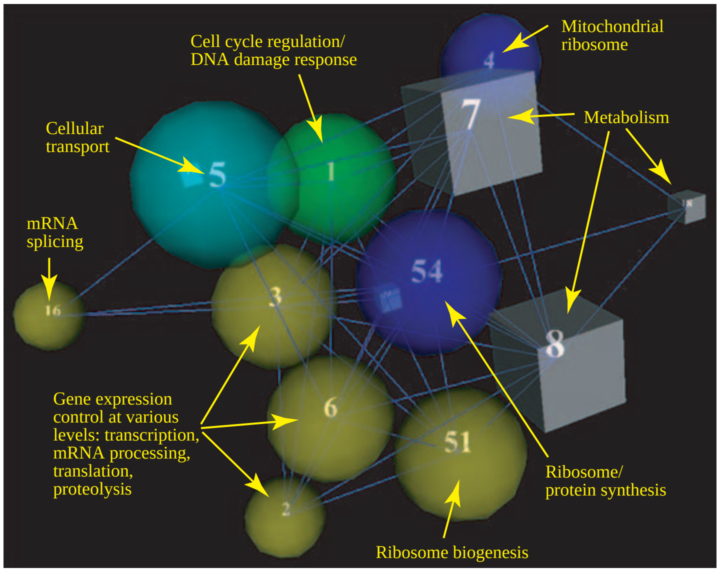

A
800
600
400
200
0
Number of defined clusters
0.0
0.2
0.4
0.6
0.8
1.0
Similarity cutoff in hierarchical tree for defining clusters
B
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of ConfidentNet's proteins
belongs to defined clusters
0.0
0.2
0.4
0.6
0.8
1.0
Similarity cutoff in hierarchical tree for defining clusters
C
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Mean protein functional coherence
of defined clusters
0.0
0.2
0.4
0.6
0.8
1.0
Similarity cutoff in hierarchical tree for difining clusters
D
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Modulation efficiency
0.0
0.2
0.4
0.6
0.8
1.0
Similarity cutoff in hierarchical tree for defining clusters

Figure S11

(A) Views of gene modules at the most inclusive level of hierarchical clustering that defines 54 gene modules encompassing 3,744 of the 4,681 genes (~80%) of ConfidentNet—we plot the 18 major modules connected by the 50 strongest inter-module linkages. These broadly inclusive clusters and their connections provide an overview of the yeast gene network in which the main features of the network are easily visualized—dominant features include the ribosome, ribosomal biogenesis system, cellular transport, cell cycle regulation, and general categories of metabolism. To simplify visualization of the networks of modules, we indicate gene modules as shapes whose sizes are proportional to the number of member genes, connected by edges whose lengths are inversely proportional to the fraction of genes directly linked between clusters. The color and shape of each module indicates the major function of the associated genes, as defined by MIPS (*28*) and listed at the bottom right. (B) We analyzed the gene functions represented in each cluster, using the 11 MIPS protein functional categories. The number of annotated proteins in a cluster in each functional category was counted. Only 12 clusters showed significant grouping of genes with reasonably homogeneous functions, representing very broadly defined functional modules in the network, even at this gross level generalization.

Figure S11

A



MIPS Functional category	Vertex color	Vertex shape
Metabolism	white	box
Energy	red	sphere
Cell cycle & DNA processing	green	sphere
Transcription/RNA processing/RNA transport	yellow	sphere
Protein synthesis	blue	sphere
Protein fate	blue	box
Cellular transport	cyan	sphere
Cell communication/signal transduction	cyan	box
Cell rescue/defense	Red	box
Regulation/integration with environment	green	box
Cell fate	yellow	box
Not clear	white	sphere

B

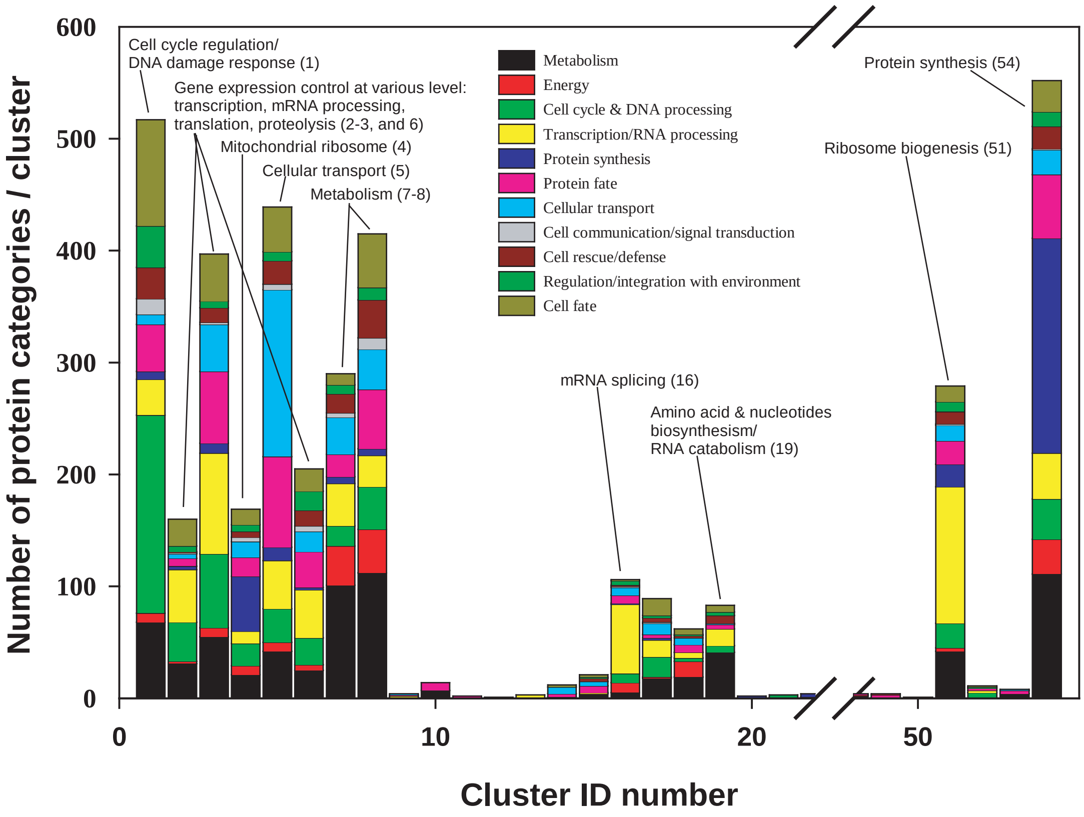
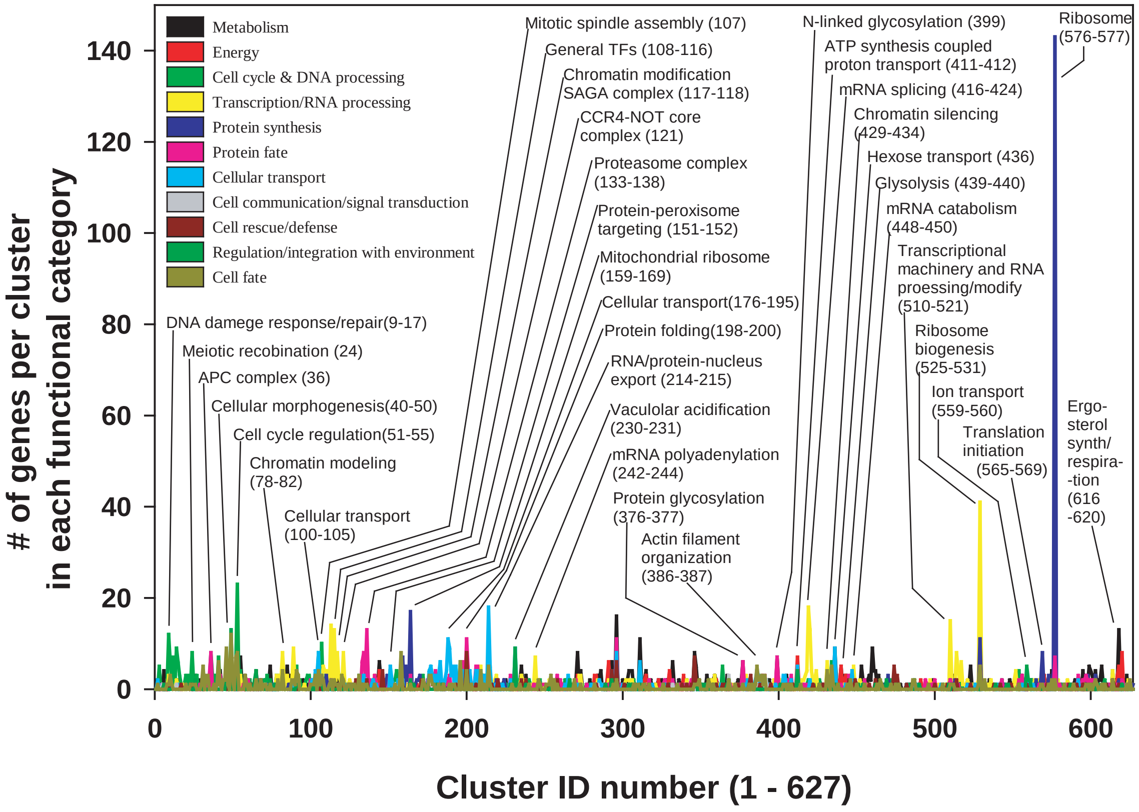


Figure S12

For the network of Figure 3B, we summarize the functions of genes in each cluster by plotting the distribution of MIPS functional categories among the 627 modules, ordered according to the hierarchical clustering tree. The height of each plot indicates the number of genes per cluster in a given functional category, indicated by color. The functional coherence of genes in each cluster is apparent; adjacent modules (indicated by sequential numberings) are often functionally related.

Figure S12



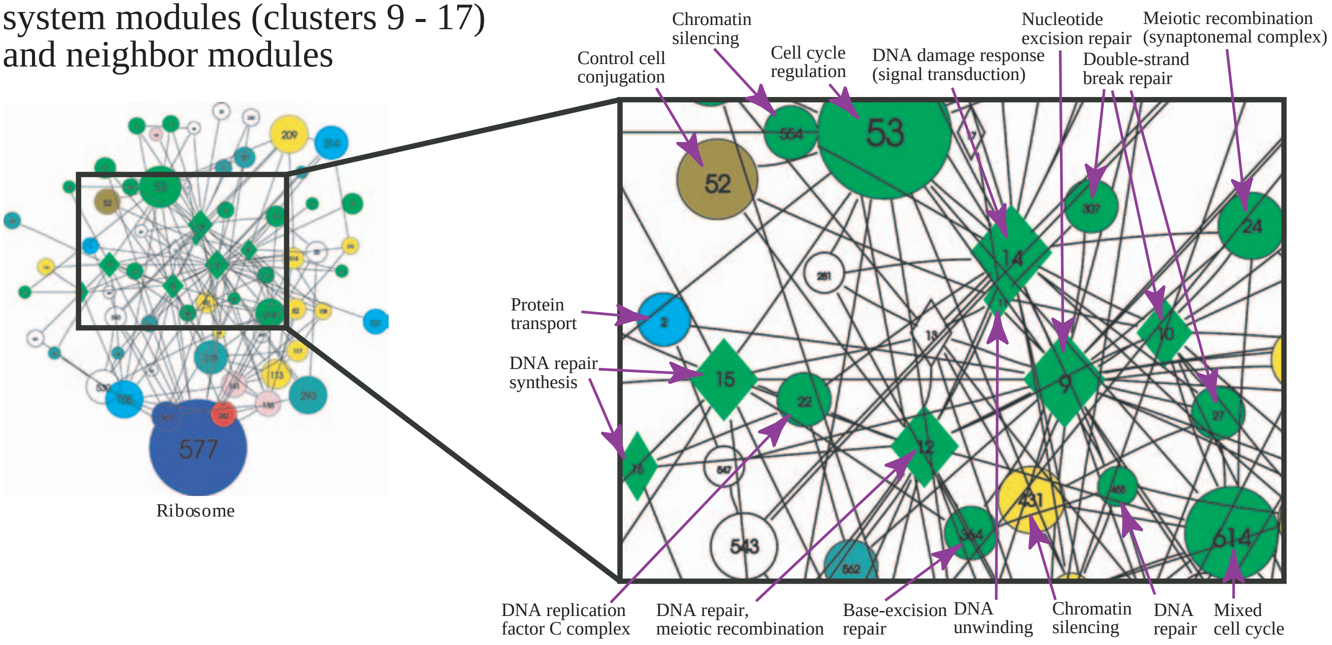
Metabolism
Energy
Cell cycle & DNA processing
Transcription/RNA processing
Protein synthesis
Protein fate
Cellular transport
Cell communication/signal transduction
Cell rescue/defense
Regulation/integration with environment
Cell fate
# of genes per cluster in each functional category
0
20
40
60
80
100
120
140
0
100
200
300
400
500
600
Cluster ID number (1 - 627)
DNA damege response/repair(9-17)
Meiotic recobination (24)
APC complex (36)
Cellular morphogenesis(40-50)
Cell cycle regulation(51-55)
Chromatin modeling (78-82)
Cellular transport (100-105)
Mitotic spindle assembly (107)
General TFs (108-116)
Chromatin modification SAGA complex (117-118)
CCR4-NOT core complex (121)
Proteasome complex (133-138)
Protein-peroxisome targeting (151-152)
Mitochondrial ribosome (159-169)
Cellular transport(176-195)
Protein folding(198-200)
RNA/protein-nucleus export (214-215)
Vaculolar acidification (230-231)
mRNA polyadenylation (242-244)
Protein glycosylation (376-377)
Actin filament organization (386-387)
N-linked glycosylation (399)
ATP synthesis coupled proton transport (411-412)
mRNA splicing (416-424)
Chromatin silencing (429-434)
Hexose transport (436)
Glysolysis (439-440)
mRNA catabolism (448-450)
Transcriptional machinery and RNA proessing/modify (510-521)
Ribosome biogenesis (525-531)
Ion transport (559-560)
Translation initiation (565-569)
Ribosome (576-577)
Ergo-sterol synth/ respira- -tion (616 -620)

Figure S13

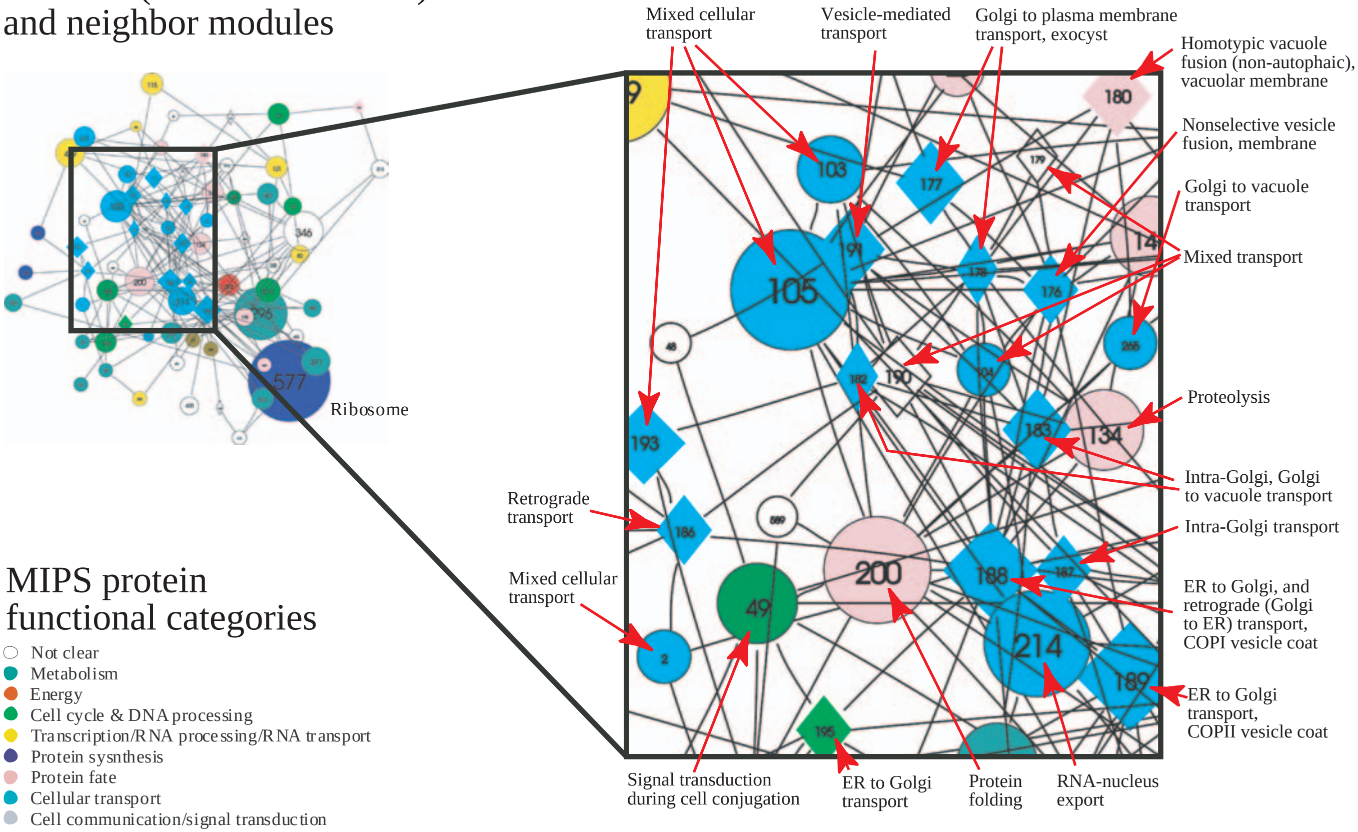
Detailed views of the reconstructed gene modules participating in DNA damage response/repair (top) and intracellular transport (bottom). In both examples, each shape represents one of the 627 modules defined by clustering genes in the network, with colors indicating module functions (see legend, bottom left) and sizes proportional to the number of genes in a module. We chose several modules of interest (diamonds) and plot all modules directly linked to these. The hierarchical organization of genes in the network is obvious: genes within the same module have precisely related functions, as labeled, while connected modules have more generally related functions, still within the same broad cellular category. For example, we find the genes participating in base-excision repair within a single module (#364). In the same local region of the network, we find modules devoted to double-strand break repair (#10, #27, #307), DNA unwinding (#11), and signal transduction (#14). Moving beyond these modules are more generally associated systems, such as cell cycle control (#53, #614). Similarly for the bottom example, modules dedicated to intracellular transport, such as the Cop I system (#188) are connected in the network to functionally related modules, such as the Cop II system (#189), vacuolar transport (#182, #183), and nuclear export (#214). These clusters (especially those related to protein transport) are in turn connected to those for protein fate, whose systems are strongly functionally coupled to the transport systems. Networks were visualized using NEATO (*44*)

Figure S13

DNA damage response/repair system modules (clusters 9 - 17) and neighbor modules



Cellular transport system modules (clusters 176 - 195) and neighbor modules


Ribosome
577
346
Mixed cellular transport
Vesicle-mediated transport
Golgi to plasma membrane transport, exocyst
Homotypic vacuole fusion (non-autophaic), vacuolar membrane
Nonselective vesicle fusion, membrane
Golgi to vacuole transport
Mixed transport
Proteolysis
Intra-Golgi, Golgi to vacuole transport
Intra-Golgi transport
ER to Golgi, and retrograde (Golgi to ER) transport, COPI vesicle coat
ER to Golgi transport, COPII vesicle coat
Retrograde transport
Mixed cellular transport
Signal transduction during cell conjugation
ER to Golgi transport
Protein folding
RNA-nucleus export
180
103
177
179
105
191
178
176
190
193
134
186
200
188
49
2
214
189
195

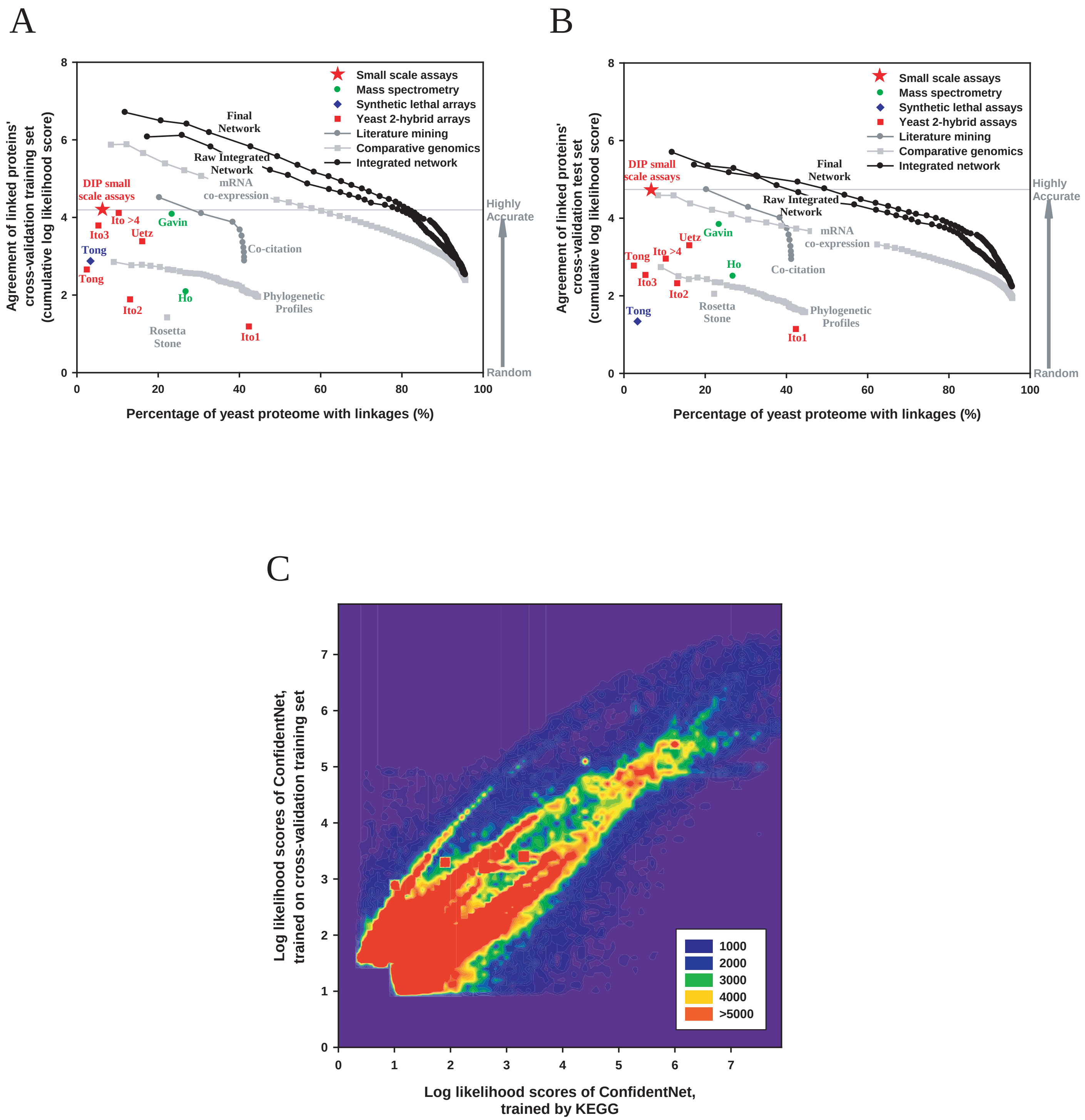
MIPS protein functional categories

- ○ Not clear
- ● Metabolism
- ● Energy
- ● Cell cycle & DNA processing
- ● Transcription/RNA processing/RNA transport
- ● Protein sysnthesis
- ● Protein fate
- ● Cellular transport
- ● Cell communication/signal transduction
- ● Cell rescue/defence
- ● Regulation/integration with environment
- ● Cell fate

Figure S14

An alternate method of data integration produces a network consistent with that described. Rather than “training” versus the KEGG database and testing versus other annotation sets, such as GO, MIPS, KOGs, and subcellular localization, we assembled a pooled annotation set from KEGG, GO, and MIPS, then split this set into two disjoint sets of linkages. One served as a training set for calculating weights for data set integration; the other annotation set was held aside for a final, independent test of integration accuracy. (A) shows the performance of the various data sets and integrated networks on the annotation set they were trained upon, while (B) shows their accuracies as measured using the independent test set. The similar performance between (A) and (B) indicates the effectiveness of data integration in successfully combining experimental and computational data to reconstruct pathways. (C) shows a histogram comparing the FinalNet network calculated in this fashion with that described previously (i.e., trained on KEGG). Scores for the top 100,000 linkages calculated under each training regime are binned in bins of size 0.1 x 0.1, and the resulting distribution plotted with color indicating height of the histogram—the preponderance of scores near the diagonal shows consistency between the networks. Multiple trends visible in the histogram reflect the slightly differing weights given to the major sources of data under the two training regimes. The consistent scores shown by linkages in the two networks argue that the integration method correctly predicts pathways and is reasonably robust to the precise choice of training set.

Figure S14


A
Agreement of linked proteins' cross-validation training set (cumulative log likelihood score)
Percentage of yeast proteome with linkages (%)
0
2
4
6
8
0
20
40
60
80
100
★ Small scale assays
● Mass spectrometry
◆ Synthetic lethal arrays
■ Yeast 2-hybrid assays
Literature mining
Comparative genomics
Integrated network
Final Network
Raw Integrated Network
mRNA co-expression
DIP small scale assays
Ito >4
Ito3
Gavin
Uetz
Tong
Tong
Co-citation
Ito2
Ho
Phylogenetic Profiles
Rosetta Stone
Ito1
Highly Accurate
Random
B
Agreement of linked proteins' cross-validation test set (cumulative log likelihood score)
Percentage of yeast proteome with linkages (%)
0
2
4
6
8
0
20
40
60
80
100
★ Small scale assays
● Mass spectrometry
◆ Synthetic lethal assays
■ Yeast 2-hybrid assays
Literature mining
Comparative genomics
Integrated network
DIP small scale assays
Final Network
Raw Integrated Network
mRNA co-expression
Uetz
Gavin
Tong
Ito >4
Ito3
Ito2
Ho
Co-citation
Tong
Rosetta Stone
Phylogenetic Profiles
Ito1
Highly Accurate
Random
C
Log likelihood scores of ConfidentNet, trained on cross-validation training set
Log likelihood scores of ConfidentNet, trained by KEGG
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
7
1000
2000
3000
4000
>5000

TABLES

Table S1

Categories of DNA microarray mRNA expression data for yeast, downloaded from the Stanford Microarray Database. A total of 717 experiments in 27 categories were tested, and, among them, 497 experiments in 12 categories showed significant regression between the mRNA co-expression patterns and the log likelihood scores (i.e., see Figure S3). These 12 categories are indicated by bold letters in the Table.

TABLE S1

SMD Category	# Exp	Description
Calcium	24	Time course of calcineurin/Crz1p-dependent gene expression following addition of 200 mM Ca^{2+}
Cell cycle	**87**	Time course of cell cycle dependent gene expression in cultures synchronized by alpha factor arrest, elutriation, cdc15-ts mutant
Chemical	**15**	Time course after treatment with dithiothreitol
Chemostat	5	Differential gene expression between parent strain and evolved strain after many generations in chemostat
ChIP	44	Finding TF binding sites using ChIP and intergenic microarray
DNA damage	**42**	Comparison of wild-type cells with mutants defective in Mec1 signaling, including mec1, dun1, and crt1 mutants, under normal growth conditions and in response to the methylating-agent methylmethane sulfonate (MMS) and ionizing radiation.
Drug treatment	**8**	Response to the sulfhydryl-oxidizing agent diamide
Evolution	45	Changes in DNA copy number were assessed after 100-500 generations of growth in glucose-limited chemostats
GCAT	31	Variety of experiments by many different groups
Genomic DNA	2	Genomic DNA analyses
Over expression	1	GAL vector over-expression
Martel	4	Response to different iron conditions
Mating	6	Differential expression of MCM1 or MCM7 (DNA replication init)
Metal	**15**	Response to different metal treatments
Mutant	**36**	Stationary phase, osmotic treatment, heat shock
ORF-IN enrichment	34	Genomic binding distributions of promoter specific transcr. factors
Osmotic	3	Transcriptional response to different osmotic conditions
Phosphate	8	Transcriptional response to low extracellular P_i concentrations
Polysome	**43**	Analyses of RNA bound to different polysomal fractions
RNA processing	**58**	Measured decay rates of yeast mRNAs after thermal inactivation of temperature-sensitive RNA polymerase II
Salt treatment	**18**	Time course of calcineurin/Crz1p-dependent gene expression following addition of 0.8 M Na^{+}
Sporulation	9	Time course during sporulation
Starvation	**36**	Time course in various conditions of nutritional starvation
Stress	**117**	Transcriptional responses to different environmental stresses
Transcription	**22**	Response to deletion of general TFs
Unfolded protein response	2	
Yeast expression	2	

Table S2

Five independent reference sets were used in this study. The 11 major top level MIPS protein functional category (MIPS major 1st level funcat to provide benchmarks of linkage accuracy) was used primarily in functional profiling of the final network or its derived clusters, while the other four references were used for assessing the quality of linkages between proteins.

TABLE S2

Reference	Number of terms	Number of annotated proteins (% of proteome)	O $_{prior}$	Download date
KEGG pathway	118	1166 (20.0)	0.078	08/07/2003
GO process 8th level	494	2675 (46.0)	0.022	03/05/2003
KOG	23	3025 (52.0)	0.093	11/26/2003
UCSF-GFP localization	22	3965 (68.2)	0.475	10/08/2003
MIPS major 1st level funcat	11	3753 (64.5)	0.302	06/25/2003

O $_{prior}$ = P(share any annotation unconditionally) / P(share no annotation unconditionally)

Table S3

Assessment the accuracies of protein-protein interaction data sets in the log likelihood scoring framework. Ten different interaction data sets are ranked by descending order of quality (indicated by log likelihood score; LLS) based on the KEGG pathway reference. It is notable that each independent reference set provides similar rankings, indicating that the scores provided by the log likelihood scheme correctly reflect data accuracy are reasonably independent of the precise reference set used, whether derived from pathways (KEGG), biological processes (GO process 8th), sequence homology (KOG), of protein sub-cellular localization (UCSF-GFP localization). Protein-protein interaction data set derived from many individual small scale experiments (collected from the Database of Interacting Proteins, DIP) outperformed all data sets derived from large scale experiments, and were adopted as the gold standard for interaction quality.

TABLE S3

PPI data set	# unique protein	# interaction	LLS by KEGG pathway	LLS by GO process 8th	LLS by KOG	LLS by UCSF-GFP localization
DIP small scale	382	2822	4.43	4.18	3.41	1.94
Gavin MS	1361	3221	3.83	3.87	2.47	1.80
Ito Y2H 4 above hit	598	521	3.25	3.59	2.29	0.97
Tong2002 Y2H	141	211	3.06	2.47	1.49	0.43
Uetz Y2H	934	854	3.03	3.16	2.09	1.28
Ito Y2H 3 hit	307	212	2.33	2.88	1.71	0.95
Ho MS	1560	3589	1.97	2.27	1.13	0.95
Tong2001 SL	195	275	1.71	1.91	1.89	0.28
Ito Y2H 2 hit	761	625	1.62	1.97	1.03	0.41
Ito Y2H 1 hit	2462	2628	0.68	1.20	0.62	0.14

Table S4

The integrated network shows a "small world" topology (*40*). Here, we compare network topology between IntNet and FinalNet (using only the top 34,000 linkages of each). Although the topology of FinalNet appears more ordered than IntNet (i.e., has a higher mean shortest path length between proteins and a higher clustering coefficient), both are "small world" networks. The higher coverage and accuracy of the FinalNet is evident in the fraction of proteome represented and the log likelihood scores.

TABLE S4

Parameters	IntNet34000	FinalNet34000
# of links	34000	34000
# of unique proteins	4478	4681
Proteome coverage (%)	77.0	80.5
Log likelihood score by KEGG	4.19	4.46
Average connectivity (links / gene)	7.60	7.26
C (clustering coefficient)	0.244	0.308
C of random counterpart	0.004	0.003
C of regular lattice counterpart	0.697	0.694
N.C. (normalized C by regular)	0.351	0.444
N.C. of random counterpart	0.006	0.004
L (length of shortest path)	4.28	5.03
L of random counterpart	3.40	3.46
L of regular lattice counterpart	148.34	162.02
N.L. (normalized L by regular)	0.029	0.031
N.L. of random counterpart	0.023	0.021

REFERENCES

1. J. M. Cherry *et al.*, *Nucleic Acids Res* **26**, 73 (1998).
2. J. Gollub *et al.*, *Nucleic Acids Res* **31**, 94 (2003).
3. I. Xenarios *et al.*, *Nucleic Acids Res* **30**, 303 (2002).
4. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
5. Y. Ho *et al.*, *Nature* **415**, 180 (2002).
6. T. Ito *et al.*, *Proc Natl Acad Sci U S A* **98**, 4569 (2001).
7. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
8. A. H. Tong *et al.*, *Science* **295**, 321 (2002).
9. A. H. Tong *et al.*, *Science* **294**, 2364 (2001).
10. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc Natl Acad Sci U S A* **96**, 4285 (1999).
11. M. Huynen, B. Snel, W. Lathe, 3rd, P. Bork, *Genome Res* **10**, 1204 (2000).
12. Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, E. V. Koonin, *Genome Res* **11**, 356 (2001).
13. E. M. Marcotte *et al.*, *Science* **285**, 751 (1999).
14. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, C. A. Ouzounis, *Nature* **402**, 86 (1999).
15. I. Yanai, A. Derti, C. DeLisi, *Proc Natl Acad Sci U S A* **98**, 7940 (2001).
16. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
17. S. V. Date, E. M. Marcotte, *Nat Biotechnol* **21**, 1055 (2003).
18. C. J. Verjovsky Marcotte, E. M. Marcotte, *Applied Bioinformatics* **2**, 93 (2002).
19. B. J. Stapley, G. Benoit, *Pac Symp Biocomput*, 529 (2000).
20. T. K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, *Nat Genet* **28**, 21 (2001).
21. E. M. Marcotte, I. Xenarios, D. Eisenberg, *Bioinformatics* **17**, 359 (2001).
22. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, *Nucleic Acids Res* **30**, 42 (2002).
23. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, D. Eisenberg, *Nature* **402**, 83 (1999).
24. S. S. Dwight *et al.*, *Nucleic Acids Res* **30**, 69 (2002).
25. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
26. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
27. W. K. Huh *et al.*, *Nature* **425**, 686 (2003).
28. H. W. Mewes *et al.*, *Nucleic Acids Res* **30**, 31 (2002).
29. I. Yanai, C. DeLisi, *Genome Biol* **3**, research0064 (2002).
30. R. Jansen, N. Lan, J. Qian, M. Gerstein, *J Struct Funct Genomics* **2**, 71 (2002).
31. C. von Mering *et al.*, *Nucleic Acids Res* **31**, 258 (2003).
32. O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, D. Botstein, *Proc Natl Acad Sci U S A* **100**, 8348 (2003).
33. R. Jansen *et al.*, *Science* **302**, 449 (2003).
34. M. Thompson, E. Marcotte, M. Pellegrini, T. Yeates, D. Eisenberg, in *Currents in Computational Molecular Biology* S. Miyano, R. Shamir, T. Takagi, Eds. (Universal Academy Press, Inc., 2000).
35. D. S. Goldberg, F. P. Roth, *Proc Natl Acad Sci U S A* **100**, 4372 (2003).
36. M. P. Samanta, S. Liang, *Proc Natl Acad Sci U S A* **100**, 12579 (2003).

37. T. Schlitt *et al.*, *Genome Res* **13**, 2568 (2003).
38. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. L. Barabasi, *Science* **297**, 1551 (2002).
39. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (1998).
40. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
41. B. Snel, P. Bork, M. A. Huynen, *Proc Natl Acad Sci U S A* **99**, 5890 (2002).
42. A. W. Rives, T. Galitski, *Proc Natl Acad Sci U S A* **100**, 1128 (2003).
43. A. T. Adai, S. V. Date, S. Wieland, E. M. Marcotte, *J. Mol. Biol.* **340**, 179 (2004).
44. E. R. Gansner, S. C. North, *Softw. Pract. Exper* **00(S1)**, 1 (1999).
45. L. Giot *et al.*, *Science* **302**, 1727 (2003).
46. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabasi, *Nature* **407**, 651 (2000).
47. C. von Mering *et al.*, *Nature* **417**, 399 (2002).