MCP Papers in Press. Published on January 2, 2007 as Manuscript M600381-MCP200

Towards a Comprehensive Atlas of the Physical Interactome of *Saccharomyces cerevisiae*

Sean R. Collins[1,2,#], Patrick Kemmeren[1,2,3,#], Xue-Chu Zhao[4], Jack F. Greenblatt[5], Forrest Spencer[4], Frank C. P. Holstege[3], Jonathan S. Weissman[1,2] & Nevan J. Krogan[1,2]

1 *Department of Cellular and Molecular Pharmacology and* 2 *The California Institute for Quantitative Biomedical Research, UCSF, San Francisco, California 94158 USA*

3 *Department of Physiological Chemistry, Division of Biomedical Genetics, UMC Utrecht, The Netherlands*

4 *McKusick-Nathans Institute of Genetic Medicine, School of Medicine, Johns Hopkins University Baltimore, Maryland 21205 USA*

5 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S 3E1 Canada

#These authors contributed equally.

- **To whom correspondence should be addressed. E-mail:**
  krogan@cmp.ucsf.edu
  weissman@cmp.ucsf.edu

Tel: 415-476-2980; Fax: 415-514-2073

Mailing Address:
University of California-San Francisco, 1700 4th Street, San Francisco, California, 94143-2540 USA

Running title: A Comprehensive Physical Interaction Map

Abbreviations: PPI, protein-protein interaction; MS, mass spectrometry; TAP, tandem affinity purification; PE, purification enrichment; ROC, receiver operating characteristic

Summary

Defining protein complexes is critical to virtually all aspects of cell biology. Two recent affinity purification/mass spectrometry studies in *Saccharomyces cerevisiae* have vastly increased the available protein interaction data. The practical utility of such high-throughput interaction sets, however, is substantially decreased by the presence of false positives. Here we create a novel probabilistic metric that takes advantage of the high density of these data, including both the presence and absence of individual associations, to provide a measure of the relative confidence of each potential protein-protein interaction. This analysis largely overcomes the noise inherent in high-throughput immunoprecipitation experiments. For example, of the 12,122 binary interactions in the general repository of interaction data (BioGRID) derived from these two studies, we mark 7,504 as being of substantially lower confidence. Additionally, applying our metric and a stringent cutoff identifies a set of 9,074 interactions (including 4,456 which were not among the 12,122 interactions) with accuracy comparable to that of conventional small-scale methodologies. Finally, we organize proteins into coherent multi-subunit complexes using hierarchical clustering. This work thus provides a highly accurate physical interaction map of yeast in a format that is readily accessible to the biological community.

Introduction

Since most cellular functions are mediated by groups of physically associated proteins, or complexes that work in a coherent fashion, it is of great interest to systematically map protein-protein interactions (PPIs). In *Saccharomyces cerevisiae*, these physical connections have been defined in large-scale experiments using the yeast two-hybrid method (1, 2) as well as direct purification of complexes using affinity tags followed by mass spectrometry (MS) analyses. In 2002, two initial studies utilized the latter strategy on subsets of the proteome (3, 4). Ho et al. (4) employed an over-expression strategy combined with a single affinity purification step while Gavin et al. (3) used a tandem affinity purification (TAP) system in which epitope-tagged proteins were expressed under normal physiological conditions. The use of an over-expression system may facilitate detection of weaker or more transitory associations between proteins or protein complexes, but might be less optimal for accurate definition of stoichiometric interactions. Indeed, the purification of proteins expressed under normal physiological conditions followed by mass spectrometry provided the best coverage and accuracy for detection of stable protein complexes (5). Based on these considerations, two separate groups interrogated the physical interactome of *S. cerevisiae* using this strategy (6, 7).

Although a similar approach was used for protein purification and identification, the resulting datasets were subjected to different analytical methods to define PPIs and protein complexes. Gavin et al. exploited a "socio-affinity" scoring system that measures the log-ratio of the number of times two proteins are observed together, relative to what would be expected from their frequency in the dataset. Importantly, this approach takes advantage of not only direct bait-prey connections but also indirect prey-prey

relationships where two proteins are each identified as preys in a purification in which a third protein is used as bait. Krogan et al., on the other hand, used a synthesis of machine learning techniques including Bayesian networks and C4.5-based and Boosted Stump Decision Trees to define confidence scores for potential interactions based on direct bait-prey observations. The two groups also used different clustering algorithms to define protein complexes from their PPI datasets. For example, Krogan et al. used a Markov clustering algorithm (8) for definition of protein complexes while Gavin et al. utilized a different clustering approach to define complexes, each consisting of groups of proteins termed "core", "module" or "attachment". "Modules" were intended to represent subcomplexes that are components of several distinct complexes, and "attachments" were factors less stably associated with stable "core" complexes. Although both of these individual datasets are of high quality, it is not obvious how discrepancies between them should be resolved, and each still contains a substantial number of false positive interactions which can compromise the utility of these data for guiding more focused studies.

In this study, we have merged these two datasets into a single reliable collection of experimentally-based PPIs by analyzing the primary affinity purification data using a novel Purification Enrichment (PE) scoring system. Using a well-defined reference set of manually curated PPIs, we demonstrate that our consolidated dataset is of greater accuracy than the individual sets and is comparable to PPIs defined using more conventional small-scale methodologies. Although algorithms designed to detect multi-protein complexes can be highly effective for extracting additional information from noisy and incomplete datasets, attempting to strictly define protein complexes may not be

the optimal way to analyze such a high-confidence dataset. In particular, any partitioning analysis must either group together distinct complexes which share one or more subunits, or fail to correctly identify all of the components of such complexes. Additionally, weak interactions between proteins or protein complexes may be lost. In this work, we have subjected the entire high confidence PPI dataset to a relatively unbiased hierarchical clustering from which one can more easily identify shared components of distinct complexes as well as weak associations between complexes. We argue that this representation provides a convenient tool for biologists to gather information about a protein of interest rapidly. Finally, this depiction potentially mimics the *in vivo* environment: a continuum of weak associations between stable protein complexes.

Experimental Procedures

Calculation of Purification Enrichment Scores

Purification Enrichment (PE) scores were modeled after a discriminant function for a Bayes classifier (9) as a measure of the likelihood of observed experimental results given the hypothesis that an interaction is genuine relative to the likelihood of the same results if the interaction is not real. These scores incorporate ideas from the socio-affinity scoring system reported by Gavin et al. (6), but differ in several significant ways. First, these scores take into account not only positive evidence for an interaction contained in the identification of two proteins in the same purification, but also negative evidence against interactions wherein one protein fails to be identified as a prey when another is used as a bait. This negative evidence has typically not been used in previous interaction scoring techniques, and it can be particularly useful for distinguishing non-interacting pairs of proteins that share many interaction partners from pairs that do exist in stable complexes. Second, PE scores more powerfully exploit situations in which a particular bait protein was used in multiple separate purifications. Third, the PE scoring strategy uses a different model for the likelihood of observing a pair of proteins in the same purification if these proteins do not interact.

PE scores were motivated by the probabilistic framework of a (Naïve) Bayes classifier. In a Bayes classifier, an estimate of the probability one hypothesis (here that an interaction is real) relative to the probability of a second hypothesis (here that the interaction is not real), given a set of observations, is calculated to determine which hypothesis is more likely. Both of these probabilities are calculated using Bayes' Theorem, and a

discriminant function *f* is calculated as the log-ratio of these probabilities. An interaction is classified as real if *f*>0 and false if *f*<0 (9). The function *f* is defined as:

$$f(all\_observations) = \log_{10} \frac{P(all\_observations \mid true\_PPI) \times P(true\_PPI)}{P(all\_observations \mid false\_PPI) \times P(false\_PPI)}$$

where *P(true_PPI)* and *P(false_PPI)* represent prior expectations for the fraction of all protein pairs which do and do not interact physically. The above equation can be rewritten as:

$$f(all\_observations) = \log_{10} \frac{P(true\_PPI)}{P(false\_PPI)} + \sum_{i=1}^{num\_observations} \frac{P(observation_i \mid true\_PPI)}{P(observation \mid false\_PPI)}$$

While the accuracy of a Bayes classifier will rely on an appriopriate value for *P(true_PPI)* and the correct value is not obvious, an incorrect choice of this value will not affect the ordering of scores for putative interactions. We therefore computed PE scores as a sum of the evidence supporting or disaffirming each potential interaction over all relevant purifications in the dataset. For a particular observation, this evidence was computed as an estimate of the corresponding term in the above sum:

$$Evidence_{observation} = \log_{10} \frac{P(observation \mid true\_PPI)}{P(observation \mid false\_PPI)} \qquad (1)$$

A PE score of zero then indicates that no evidence for or against the validity of a particular interaction was collected (and in theory the probability that such an interaction is true should be equal to the prior estimate of *P(true_PPI)*). In particular, we considered two types of observations in the construction of PE scores: bait-prey observations when one of the proteins of interest was used as a bait and prey-prey observations when the two proteins of interest both appeared as preys in the purification of a third protein. As a result, similar to socio-affinity scores (6), PE scores can be written as a sum of direct

bait-prey components (S) and an indirect prey-prey component (M). Thus, for a potential interaction between proteins i and j:

$$PE_{ij} = S_{ij} + S_{ji} + M_{ij}$$

Here S_{ij} measures evidence from purifications where protein i was used as bait, S_{ji} measures evidence from purifications where protein j was used as bait, and M_{ij} measures indirect evidence due to co-occurrence of proteins i and j as preys in the same purifications. Below we give detailed equations used to compute the S and M components:

$$S_{ij} = \sum_k s_{ijk}$$

where each value of k indicates a distinct purification in which protein i was used as bait and s_{ijk} represents the corresponding evidence computed using equation (1). The probabilities *P(observation | true_PPI)* and *P(observation | true_PPI)* used to define s_{ijk} were calculated based on estimates of two underlying probabilities: *r* representing the probability that a true association will be preserved and detected in a purification experiment and p_{ijk} representing the probability that a bait-prey pair will be observed for nonspecific reasons. Using these quantities, we calculate:

$$s_{ijk} = \log_{10} \frac{r + (1-r) \times p_{ijk}}{p_{ijk}}$$

if protein j appeared as a prey in purification k using bait i, and

$$s_{ijk} = \log_{10} \frac{(1-r) \times (1-p_{ijk})}{(1-p_{ijk})} = \log_{10}(1-r)$$

otherwise. Values for *r* and p_{ijk} could in principle be estimated in a number of ways. Here, we estimated *r* using the observed frequency of successful purification over a very

high confidence set of interactions (the intersection of MIPS complexes and MIPS small-scale experiments). For the Krogan et al., Gavin et al., and Ho et al. data, this gave values of 0.51, 0.62, and 0.265, respectively. For p_{ijk} we used an estimate of the probability that a given bait-prey pair would be observed for nonspecific reasons at least once in the dataset, calculated using the Poisson distribution as:

$$p_{ijk} = 1 - \exp(-f_j n_{ik}^{prey} n_i^{bait})$$

where n_{ik}^{prey} is the number of preys identified in purification k with bait i, n_i^{bait} is the number of times protein i was used as bait, and f_j is an estimate of the nonspecific frequency of occurrence of prey j in the dataset. The relative values of the f_j are estimates of relative rates at which different preys occur nonspecifically (and can be considered measures of relative promiscuity), and the sum of the f_j can be considered to be the fraction of all prey identifications which are nonspecific. Although alternate strategies could be used, for simplicity we allowed the sum of the f_j to be one, and we computed f_j as Bayesian posterior estimates based on the observed frequency of occurrence of preys in the dataset and the prior hypothesis that all preys occur nonspecifically with equal frequency:

$$f_j = \frac{n_j^{prey\_obs} + n_{psuedo}}{n_{tot}^{prey\_obs} + (n_{distinct\_preys} \times n_{psuedo})}$$

Here, $n_j^{prey\_obs}$ is the total number of observations of protein j as a prey, $n_{tot}^{prey\_obs}$ is the total number of observations of all preys, $n_{distinct\_preys}$ is the number of distinct preys observed, and n_{pseudo} is a number of pseudocounts added for each prey which determines the weight given to the prior hypothesis. Values of 20, 10, and 5 were used for n_{pseudo} for the Krogan et al., Gavin et al., and Ho et al. datasets, respectively. The value of n_{pseudo}

was the only parameter adjusted to optimize the PE scoring system. Adjustments were done using the MIPS complexes as a reference, and for this reason results of all comparisons made using a reference set based on the MIPS complexes were duplicated using an independent reference set generated from the SGD complexes.

The M component was calculated as:

$$M_{ij} = \sum_{k} m_{ijk}$$

where each value of k indicates one purification in which proteins i and j were simultaneously observed as preys. In this case, our approach differs slightly from the full Bayesian classifier approach, which would either sum over all purifications or sum over all purifications in which at least one of the two proteins was identified as a prey. We did not use a sum over all purifications because it would require an enormous number of calculations and because estimation of all of the relevant probabilities is itself a very difficult problem. We instead created an approximate implementation of equation (1) for m_{ijk} calculated only for observations where both preys were observed in the same purification. Significantly, we did not include a negative term for the case in which only one of the two proteins was observed as a prey in a purification. This was because two proteins can interact, yet also be components of alternate complexes. Our implementation was again based on estimates for two underlying probabilities. Here, we use r to represent the probability that a true association between proteins i and j will be preserved and detected during a purification experiment and p_{ijk} to represent the probability that proteins i and j will appear as preys in the same purification for nonspecific reasons.

$$m_{ijk} = \log_{10} \frac{r + (1 - r) \times p_{ijk}}{p_{ijk}}$$

We used the same estimate for r as calculated above, and for p_{ijk} we use an estimate of probability that proteins i and j will occur nonspecifically as preys in the same purification at least once in the dataset. This value for p_{ijk} is calculated using the Poisson distribution as:

$$p_{ijk} = 1 - \exp(-f_i f_j n_{tot}^{prey-prey})$$

where f_i and f_j are computed as described above, and $n_{tot}^{prey-prey}$ is the total number of prey-prey pairs observed in the dataset.

The Krogan et al. and Gavin et al. data were combined by computing a score for each putative interaction independently over each dataset and adding them as:

$$PE_{ij}^{Combined} = 0.5 \times PE_{ij}^{Krogan} + PE_{ij}^{Gavin}$$

This weighted sum was used instead of a straight sum because empirically it was a more effective predictor of PPIs, and in practice this may be due to redundancy of the Krogan LCMS-MS and MALDI-TOF data.

Clustering of PPI data

First, scaled PE scores were computed for use in hierarchical clustering to minimize variation in scores which does not correspond to variation in the reliability of the represented interactions. For example, PE scores of 10 and 20 may both correspond to extremely reliable interactions, but a score of zero likely indicates a non-interaction. The scaled scores range from zero to one and were intended to approximate confidence values (i.e., a scaled score of 0.8 would correspond to 80% likelihood of a true interaction). However, these values were not carefully trained and should not be taken as reliable confidence values. Equations used for calculating these values are detailed below. A

vector of scaled PE scores was then created for each protein which had at least one scaled score of 0.2 or higher (corresponding to a PE score threshold of 1.85). A value of 1 was assigned for the diagonal elements (representing self-interaction) so that interacting proteins would tend to cluster together. This data was then hierarchically clustered using the uncentered correlation metric and the average linkage method with the Cluster 3.0 program (10). Results were visualized and figure images were created using the JavaTreeview program (http://jtreeview.sourceforge.net/).

Scaled scores represent a monotonic mapping of PE scores onto the interval zero to one. They would represent confidence values given the approximations that: 1) binary interactions in MIPS complexes represent an unbiased subset of the set of all true binary protein-protein interactions, 2) MIPS small-scale experiments are approximately 95% accurate, and 3) the set of MIPS complexes is independent of the results contained in MIPS small-scale experiments. They were computed using the slope of a "coverage curve" of the cumulative number of interactions detected which were annotated in MIPS complexes versus the total number of interactions identified (see Suppplemental Fig. 3). For each PE score, a corresponding slope in the coverage curve was computed by local linear regression. The resulting slopes were made monotonic (as a function of PE score) and smoothed using the Pool Adjacent Violators Algorithm (11) and Loess regression (12). To convert these slopes to scaled scores, they were divided by the fraction of interactions included in the MIPS small-scale experiments (excluding two-hybrid studies) which were also contained in the MIPS complexes (461/1081). The resulting values were multiplied by 0.95, and an upper bound of 0.99 was applied. Scaled scores below 0.05 were set zero for computational expediency.

GO & GOslim Annotations

GeneOntology (GO) (13) and GOslim annotations were obtained from SGD (14) on March 7, 2006. Any feature annotated as ORF, pseudo_gene or transposable_element_gene in SGD was used to calculate the total number of proteins in each GOslim category.

MIPS & SGD Complexes

MIPS complexes were obtained from the MIPS database on March 7, 2006 using the funcat scheme version 2.0 (15). SGD complexes were extracted from the SGD database using the GO cellular component annotations. GO annotations containing the words "complex", "subunit", "ribosome", "proteasome", "nucleosome", "repairosome", "degradosome", "apoptosome", "replisome", "holoenzyme" or "snRNP" were used to assign proteins with the same GO annotation to a complex.

MIPS small-scale experiments

A collection of 1081 putative protein-protein interactions identified in small-scale experiments was obtained from the MIPS database on March 7, 2006 (15). Two-hybrid experiments were excluded from this set because they appeared to be of lower accuracy. The collection from MIPS was used rather than the larger collection contained in the BioGRID database (16), because the collection in MIPS appeared to be of greater accuracy by each of the metrics we considered.

True Positive & True Negative Calculation

True positives (TP) were calculated for PPIs within complexes (for MIPS and SGD). True negatives (TN) were taken to be connections between proteins in different complexes if the proteins have a different subcellular localization according to Huh et al. (17) and Kumar et al. (18) or show significant mRNA expression anti-correlation (calculated using a standard correlation coefficient, distance > 1.108328 (corresponds to $R < -0.108328$ or a $P < 0.001$) over a set of 1000 microarray experiments (19)).

ROC Curve Calculations

ROC curves were calculated using PE (and in some cases socio-affinity) scores calculated for all pairs of proteins in the full reference set. Thus a sensitivity value of 1 indicates detection of all true positive examples in the reference set, and a 1-specificity value of 1 indicates detection of all true negative examples in the reference set. For all ROC curves plotted on the same graph, an identical reference set was used to calculate the curves.

Supporting Website and Database

A searchable website, which contains all the PE scores and PPI clustering, has been created at http://interactome-cmp.ucsf.edu using perl, php and a PostgreSQL relational database.

Diploid Bimater Assay

To compare *yol054wΔ/yol054wΔ* cells to wild-type, 1cm^2 patches of each were made from independent single colonies, replica plated to a lawn of tester cells, cultured for 6 hours at 30°C, and again replicated to medium selective for rare matings (20). The number of colonies on each patch was counted manually with the median number of colonies on each patch being used to calculate fold-change (mutant/wild-type ratio). Selection was based on histidine prototrophy since experimental genotypes were: MATa/MATα, *his3Δ/his3Δ* (control) or MATa/MATα, *his3Δ/his3Δ*, *yol054wΔ/yol054wΔ* (experiment) and the mating testers were: MATa *his1* or MATα *his1*.

a-Like Faker Assay

To compare MATα *yol054wΔ his3* to MATα *his3,* 1cm^2 patches from independent single colonies were replica plated to medium selective for rare matings, based on histidine prototrophy as above (21).

Results

A Metric for Defining Protein-Protein Interactions

The recently completed high-throughput affinity purification experiments provided hundreds of thousands of putative PPIs. The challenge is then to convert this array of affinity purification data into a set of high confidence PPIs. Due to the high-density nature of these studies, there are often many separate observations that provide evidence supporting or disaffirming a potential interaction, as well as a significant amount of experimental noise intrinsic to high-throughput affinity purification approaches. Clearly, and as appreciated in the original studies (6, 7), a simple cataloguing of observed associations does not adequately exploit these data. Instead, one would like to integrate all of the data in a uniform manner to fully exploit direct evidence for interactions where one protein was used as a bait and another was identified as a prey, indirect evidence due to the co-occurrence of a pair of preys in identical purifications, as well as evidence against the validity of an interaction when one protein was used as bait and the other was not identified as a prey. This problem can naturally be cast in terms of Bayesian statistics where one can quantify the evidence that each relevant observation provides for or against the validity of an interaction in terms of the probabilities of making such an observation if the interaction is true and the probability if the interaction is not true:

$$Evidence_{observation} = \log_{10} \frac{P(observation \mid true\_PPI)}{P(observation \mid false\_PPI)}$$

Motivated by this framework, we have created a novel metric, which we term the Purification Enrichment (PE) score. For each putative interaction, this score is a sum of

the evidence calculated for each relevant observation in a dataset (detailed equations are provided in Experimental Procedures).

By several independent metrics including the ability to predict membership in previously annotated complexes, the PE scores appear to identify interactions of higher confidence than the socio-affinity scores of Gavin et al. (6) (Supplemental Fig. 1). PE scores also performed better than scores that only took advantage of the direct bait-prey data from purification experiments (Fig. 1A "Krogan PPI" point and data not shown). The use of indirect prey-prey information was also a component of the socio-affinity score, and it is conceptually related to a computational approach taken to predict PPIs based on shared interaction partners (22). While it is clear from those studies (and our own) that there is a wealth of information contained in inferences from indirect prey-prey associations, some care should be taken with interactions inferred solely in this way, as it appears that incorrect linkages may occasionally be inferred between proteins sharing a large number of common interaction partners. For this reason, we have preserved annotations indicating which interactions were and were not observed directly (see below). We also note that, given a set of purification results, a PE score can be computed for any pair of proteins including, but not limited to, pairs of proteins for which direct or indirect evidence for an interaction was observed. Pairs that never co-purified will either be assigned scores of zero (if neither protein was used as a bait), or negative scores, indicating that evidence against the potential interactions was collected. Finally, it is important to be aware that the negative interaction data may exhibit some bias with respect to tagging artifacts, protein abundance, and mass spectrometry issues, however

we have found that including this information in the analysis increases the quality of the final dataset.

Assessing Confidence of Binary Interactions

A standard method to evaluate the accuracy of a scored interaction dataset is to measure it against a high-confidence reference set which is taken to be correct (22). For the calculated PE scores as well as previous mass spectrometry-based datasets, we evaluated accuracy and coverage using a reference set of true positive and true negative interactions generated from manually curated complexes obtained from either MIPS (15) or SGD (14) (Supplemental Fig. 2, and see Experimental Procedures for a more detailed description). True positive interactions were taken to be connections between proteins that were annotated as belonging to the same complex in the database (MIPS or SGD). While such a reference set will contain some false positives, this contamination is unlikely to be biased in favor of a particular dataset. Generating an unbiased set of non-interacting pairs of proteins, or true negative interactions, is challenging. Nevertheless, our results did not seem to be particularly sensitive to the method used to define this set. We defined our set of true negative interactions to be connections between pairs of proteins which were annotated only to distinct complexes, and which either had non-overlapping cellular localizations as determined by GFP-fusion studies (17, 18) or had significantly anti-correlated mRNA expression patterns. While the localization and co-expression criteria we applied probably each have their own biases, they both largely deplete known interactions from the true negative set (17, 23). With reference sets constructed, we could measure the relative accuracy and coverage of different datasets by

creating Receiver Operating Characteristic (ROC) curves which measure the tradeoff between accuracy and completeness as a function of a score threshold (12). We find that when the PE metric is applied to either the new Gavin or the Krogan primary coprecipitation results it is possible to identify a substantial (although non-identical) fraction of known protein complexes while excluding the vast majority of the true negative set (Fig. 1A and Supplemental Fig. 2). Application of the PE score to the co-precipitation data in Ho et al. (4) was significantly less successful at identifying known PPIs (Supplemental Fig. 2), although the difference may be largely due to this dataset's smaller quantity of raw data. In each of the ROC curves, there is a significant portion of the curve that is linear and has a slope similar to that of the random background. This trend is due to interactions in the reference set which were neither supported nor disaffirmed by the dataset and received scores of zero.

A High Confidence Consolidated Dataset

Subjecting the Gavin et al. and Krogan et al. datasets to the same log-likelihood scoring function allowed us to directly combine them into a single comprehensive set that encompasses all of the high-throughput TAP purification experiments completed to date. We computed combined scores from both the Krogan et al. and Gavin et al. datasets (see Experimental Procedures for detailed equations) and, not surprisingly, this consolidated dataset provided greater coverage and accuracy than either of the individual datasets (Fig. 1A and Supplemental Fig. 3). In particular, it is possible to capture approximately 50% of the previously reported interactions within protein complexes, although the true coverage may be substantially higher since this reference set likely still contains false positives.

We chose not to include the Ho et al. data in our consolidated dataset because it was created using a different experimental method and its inclusion resulted in negligible changes to the resulting ROC curves (data not shown).

Using the true positive and true negative sets of protein pairs described above not only allowed us to compare the processed results of this consolidated dataset to previous high-throughput datasets, but it also provided an opportunity to compare our new results to those obtained in small-scale experiments which are often taken as a standard for high accuracy (24, 25). Consistent with earlier analyses, we find that previous high-throughput efforts do not reach the level of accuracy obtained in small-scale studies (25). However, using the consolidated dataset, it is possible to define a large set of PPIs with the same calculated true positive to true negative rate as the collection of 1081 pairwise interactions obtained from small-scale experiments (excluding two-hybrid studies) in the MIPS database (Fig. 1A and Supplemental Fig. 4). This true positive to true negative rate suggests a score threshold (of 3.19) that defines a set of 9074 high confidence interactions among 1622 distinct proteins. Consistent with an earlier analysis based on smaller protein-protein interaction networks (26), we find that this network, which is probably enriched for stable interactions relative to more transient ones, is not scale-free (i.e. although the network contains a substantial number of nodes with high degree, the node degree distribution is not described by a power law) (Supplemental Fig. 5).

The suggestion that this subset of 9074 interactions from the consolidated dataset is of comparable confidence to that of a manually curated set of interactions identified in small-scale experiments was tested by three additional independent measures: subcellular co-localization, GO annotation and mRNA co-expression. First, since proteins that

interact physically tend to have the same subcellular localizations (17, 18, 25), we compared the published experimentally determined localizations of the putatively interacting protein pairs. Unlike pairs identified in previous high-throughput studies, we found that pairs in this high confidence set were more likely to have matching localizations than pairs identified in small-scale experiments (Fig. 1B). Next, we found that three different classes of GO annotations (cellular component, biological process, and molecular function) were either equally or more likely to match for pairs of interacting proteins in our new set compared to pairs derived from small-scale experiments (Fig. 1B). Finally, it is known that genes encoding physically interacting proteins are more likely to have similar expression profiles (10, 23, 27, 28), and so we examined the distribution of Pearson correlation coefficients between expression patterns of interacting pairs over a set of 1000 previously published microarray experiments (19). Relative to the pairs identified in small-scale experiments, our new high confidence set is significantly enriched for gene pairs with highly similar expression patterns (Fig. 1C and Supplemental Fig. 4). While this enrichment may reflect better coverage of the ribosome and proteins involved in ribosome biogenesis, the new high confidence set also shows an almost identical lack of anti-correlated gene pairs when compared to the small-scale set (Fig. 1C and Supplemental Fig. 4), providing further evidence that the consolidated set of PPIs has a very low false positive rate which compares favorably to that of the MIPS small-scale dataset.

Comparison of the PPIs generated in this study to ones deposited into BioGRID (16) (which is a primary source for SGD (14)) from the original studies clearly demonstrates that we have defined a more reliable dataset (Fig. 1A). In particular, the

4456 PPIs unique to our set appear to be of confidence comparable to that of the small-scale experiments, whereas those unique to either the Gavin (2963) or Krogan (4512) sets deposited in the databases appear to be of markedly lower confidence as judged by cellular localization and GO annotation (Fig. 2). It should be noted that using the socio-affinity scoring system described by Gavin et al. (6) provides a dataset that, although of lower coverage and accuracy than the new datasets we define here, is of higher confidence than the set deposited in the major databases (Supplemental Fig. 1). We also note that although in general they should be considered of lower confidence, the interactions unique to the Gavin or Krogan sets are still likely to contain a number of physiologically relevant associations. The high confidence set of interactions defined here, similar to other PPI datasets derived from high-throughput studies (5-7), shows some apparent bias towards high-abundance proteins and against proteins from certain cellular compartments (such as the cell wall and the plasma membrane) (Supplemental Fig. 6). These biases probably reflect experimental limitations, but may also to some extent reflect real features of the distribution of protein complexes in yeast.

A Portrait of the Physical Interactome Map

Several methods to accurately define protein complexes were explored using the high confidence consolidated PPI dataset. Using such analyses as a final representation, however, often results in unwanted consequences such as the merging of several clearly distinct complexes that share one or more subunits. Also, information regarding weak associations between protein complexes can be lost. To overcome these difficulties and in an attempt to visualize the physical interactome as it exists *in vivo*, we subjected the

patterns of PPIs for all proteins having at least one interaction with a scaled PE score (see Experimental Procedures) above 0.20, a criterion encompassing almost 2400 proteins, to hierarchical clustering (Fig. 3A) (see Experimental Procedures for a more detailed description). The threshold used here, which is lower than the one used above to define 9074 high confidence interactions, was used to allow a more complete interaction map. Stable, stoichiometric protein complexes are, for the most part, accurately recapitulated as distinct blocks along the diagonal while PPIs that reside off the diagonal either represent shared subunits of complexes or weak associations between complexes (Fig. 3B and C). A clear example of the former emerges from four complexes that are involved in chromatin function: NuA4 (29-31), SWR-C (31-33), INO80C (34) and the helicase chaperone complex, Tah1/Pih1 (35) (Fig. 3D). Visual inspection of the off-diagonal connections demonstrates that the DNA helicases, Rvb1 and Rvb2, are components of the INO80C, SWR-C and Tah1/Pih1, but not NuA4, protein complexes (Fig. 3D). Similarly, Swc4 and Yaf9 are shared components of SWR-C and NuA4, while the actin-related proteins, Act1 and Arp4, are part of SWR-C, NuA4 and INO80, but not the Tah1/Pih1, complexes. Further inspection of the off-diagonal connections (Fig. 3B) reveals that Tra1 is a shared subunit of SAGA (36) and NuA4 (29), Taf14 resides both in TFIIF and INO80C (37) and actin (Act1) physically associates with several factors involved in cytoskeleton formation in addition to being a subunit of multiple chromatin remodeling complexes (38). A different region of the clustergram nicely demonstrates that Sec13 is part of both the Nup84 nucleoporin (39) and the coatomer COPII complexes (40) (Fig. 3E). Further inspection reveals that Sec23, a component of the COPII complex, seems to be independently associated with the three members of the Sec24 family, Sec24, Sfb2

and Sfb3, a phenomenon that has been previously characterized (41) (see http://interactome-cmp.ucsf.edu for more comprehensive views of the clustergrams).

Using a two-color scheme overlaid on the clustering analysis (Fig. 3A), we highlighted interactions that were observed directly as bait-prey pairs (yellow) from those that were solely inferred on the basis of co-purification as preys in the same experiments (blue). Strikingly, the physical composition of the ribosome is primarily inferred from indirect (prey-prey) interactions. Mainly due to the purification protocols used, neither the Krogan et al. nor Gavin et al. studies successfully purified tagged subunits of the ribosome, although both works often obtained ribosomal proteins as preys. Krogan et al. filtered these promiscuous proteins from their dataset, and although Gavin et al. retained the ribosomal protein data, it resulted in the inference of many complexes containing various subsets of the ribosome. Instead, in this unbiased representation, the ribosome remarkably appears as a single complex along the diagonal, largely free of non-specific off-diagonal connections.

As a further demonstration that hierarchical clustering of the consolidated data is potentially more informative than the lists of complexes presented in the original studies, we used a different two-color scheme (yellow and red) to highlight interactions that were not present in either the inferred protein complexes from Gavin et al. or Krogan et al. (Fig. 4A-D). These new interactions may have been identified due to the improved scoring system, the simultaneous consideration of both raw datasets, or a combination of these factors. Consistent with the trends observed in the ROC curves (Fig. 1A), a number of previously characterized PPIs were only seen with the new analyses. For example, six subunits of the transcriptional elongation complex Elongator have been previously

characterized (42-44), but only in this new representation is the smallest subunit, Elp6, actually confirmed (Fig. 4A). Similarly, in our new merged PPI dataset is it clear that Sec20 is a component of the Dsl1 complex, required for stability of the Q/t-SNARE complex at the endoplasmic reticulum (45) (Fig. 4B), and that Dad2 and Dad3 are components of the DASH microtubule ring complex (46) (Fig. 4C).

An example of a weak association between two distinct sets of proteins revealed by the hierarchical clustering is represented in Fig. 4D. The MIND (Mtw1 Including Nnf1-Nsl1-Dsn1) complex (47) is seemingly associated with the kinetochore complex through one of it subunits, Ame1. A relatively weak association also exists between several subunits of the inner and outer kinetochore (Ame1, Mcm22, Okp1, Chl4, Nkp2) (48) and an uncharacterized protein, Yol054w, a connection not present in the complexes derived in the Krogan et al. and Gavin et al. studies (Fig. 4D). Consistent with this hypothesis, deletion of *YOL054w* results in genomic instability as measured by bimater (20, 49), and "a-like faker" (21) assays (Fig. 4E,F).

Discussion

With the two largest high-throughput studies of protein-protein interactions in yeast (or any other organism) recently completed, two questions arise: how completely have interactions been identified, and how accurately have they been determined? With respect to coverage, a rough calculation based on the degree of overlap between the two recent studies suggests that they cover approximately 80% of interactions accessible to the TAP approach under the conditions used.

In terms of accuracy, we demonstrate here that high-throughput identification of protein-protein interactions has reached a new landmark. For the first time, this consolidated dataset can match the reliability of small-scale experiments. By simultaneously analyzing the two recent studies with one scoring system and creating a single merged dataset, we were able to generate a large set of PPIs ordered according to a score that indicates the strength of experimental evidence supporting their validity. In particular, we are able to identify a large subset of approximately 9000 of these interactions, which by several independent metrics appear to be of equal or greater accuracy than that attained in a collection of small-scale experiments. More valuable than these high accuracy binary interactions, however, may be the portrait of the yeast physical interactome that emerges from them through hierarchical clustering. The weak, but reproducible interactions that appear between well-defined complexes or between the individual components within these complexes and other proteins can be used to generate a number of hypotheses for future research.

Even though identification of stable protein complexes that survive TAP purification may be nearing saturation for *Saccharomyces cerveisiae*, much work remains

in characterizing PPIs. For example, since a precise estimate of the false positive rates for the PPI datasets presented here remains elusive, a systematic re-analysis of a subset of these putative interactions using small-scale methods may be very valuable. Also, further identification of transient associations between well-defined complexes, perhaps by further exploiting the yeast two-hybrid system, will prove insightful. An understanding of the dynamics of protein-protein interactions in response to changes in the environment has yet to be systematically explored. Obtaining low-resolution structural analyses of the defined complexes using electron microscopy and determining which protein post-transcriptional modifications are involved in mediating PPIs are also of immediate interest. Furthermore, efforts should be made to more quantitatively characterize protein-protein interactions perhaps by using technologies amenable to detecting PPIs in vivo. Finally, considering that such significant biological information was extracted from yeast using this approach, a similar comprehensive strategy for defining the physical interactome in more complex organisms must be endeavored.

Acknowledgments

We thank C. J. Ingles and N. Friedman for critically reading the manuscript and A. Roguev, S. Wodak and S. Pu for stimulating discussion. S.R.C. was supported by a pre-doctoral fellowship from the Burroughs Wellcome Fund, P.K. by a Netherlands Genomics Initiative (NGI) fellowship, nr. 050-72-417, J.S.W by the Howard Hughes Medical Institute and N.J.K. by a Sandler Family Fellowship.

References

1. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98, 4569-4574.
2. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403, 623-627.
3. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-147.
4. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415, 180-183.
5. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399-403.
6. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631-636.
7. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S.

J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440, 637-643.
8. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30, 1575-1584.
9. Duda, R. O., Hart, P. E., and Stork, D. G. (2001) *Pattern Classification*, Second Ed., John Wiley & Sons, Inc., New York, NY.
10. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863-14868.
11. Robertson, T., Wright, F. T., and Dykstra, R. L. (2005) *Order Restricted Statistical Inference*, New York, John Wiley and Sons Ltd.
12. Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Springer, New York.
13. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32, D258-261.
14. Hong, E. L., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Livestone, M. S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C. L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Hitz, B., Miyasato, S., Schroeder, M., Sethuraman, A., Weng, S., Dolinski, K., Botstein, D., and Cherry, J. M. Saccharomyces Genome Database.
15. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., and Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* 32, D41-44.
16. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34, D535-539.
17. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* 425, 686-691.
18. Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K. H., Miller, P., Gerstein, M., Roeder, G. S., and Snyder, M. (2002) Subcellular localization of the yeast proteome. *Genes Dev* 16, 707-719.
19. Ihmels, J., Bergmann, S., and Barkai, N. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20, 1993-2003.

20. Spencer, F., Gerring, S. L., Connelly, C., and Hieter, P. (1990) Mitotic chromosome transmission fidelity mutants in Saccharomyces cerevisiae. *Genetics* 124, 237-249.
21. Warren, C. D., Eckley, D. M., Lee, M. S., Hanna, J. S., Hughes, A., Peyser, B., Jie, C., Irizarry, R., and Spencer, F. A. (2004) S-phase checkpoint genes safeguard high-fidelity sister chromatid cohesion. *Mol Biol Cell* 15, 1724-1735.
22. Goldberg, D. S., and Roth, F. P. (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* 100, 4372-4376.
23. Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A., and Holstege, F. C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* 9, 1133-1143.
24. Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88-93.
25. Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004) A probabilistic functional network of yeast genes. *Science* 306, 1555-1558.
26. Tanaka, R., Yi, T. M., and Doyle, J. (2005) Some protein interaction data do not exhibit power law statistics. *FEBS Lett* 579, 5140-5144.
27. Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res* 12, 37-46.
28. Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* 29, 482-486.
29. Allard, S., Utley, R. T., Savard, J., Clarke, A., Grant, P., Brandl, C. J., Pillus, L., Workman, J. L., and Cote, J. (1999) NuA4, an essential transcription adaptor/histone H4 acetyltransferase complex containing Esa1p and the ATM-related cofactor Tra1p. *Embo J* 18, 5108-5119.
30. Krogan, N. J., Baetz, K., Keogh, M. C., Datta, N., Sawa, C., Kwok, T. C., Thompson, N. J., Davey, M. G., Pootoolal, J., Hughes, T. R., Emili, A., Buratowski, S., Hieter, P., and Greenblatt, J. F. (2004) Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4. *Proc Natl Acad Sci U S A* 101, 13513-13518.
31. Kobor, M. S., Venkatasubrahmanyam, S., Meneghini, M. D., Gin, J. W., Jennings, J. L., Link, A. J., Madhani, H. D., and Rine, J. (2004) A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol* 2, E131.
32. Krogan, N. J., Keogh, M. C., Datta, N., Sawa, C., Ryan, O. W., Ding, H., Haw, R. A., Pootoolal, J., Tong, A., Canadien, V., Richards, D. P., Wu, X., Emili, A., Hughes, T. R., Buratowski, S., and Greenblatt, J. F. (2003) A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol Cell* 12, 1565-1576.
33. Mizuguchi, G., Shen, X., Landry, J., Wu, W. H., Sen, S., and Wu, C. (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* 303, 343-348.

34. Shen, X., Mizuguchi, G., Hamiche, A., and Wu, C. (2000) A chromatin remodelling complex involved in transcription and DNA processing. *Nature* 406, 541-544.
35. Zhao, R., Davey, M., Hsu, Y. C., Kaplanek, P., Tong, A., Parsons, A. B., Krogan, N., Cagney, G., Mai, D., Greenblatt, J., Boone, C., Emili, A., and Houry, W. A. (2005) Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. *Cell* 120, 715-727.
36. Grant, P. A., Schieltz, D., Pray-Grant, M. G., Yates, J. R., 3rd, and Workman, J. L. (1998) The ATM-related cofactor Tra1 is a component of the purified SAGA complex. *Mol Cell* 2, 863-867.
37. Kabani, M., Michot, K., Boschiero, C., and Werner, M. (2005) Anc1 interacts with the catalytic subunits of the general transcription factors TFIID and TFIIF, the chromatin remodeling complexes RSC and INO80, and the histone acetyltransferase complex NuA3. *Biochem Biophys Res Commun* 332, 398-403.
38. Grummt, I. (2006) Actin and myosin as transcription factors. *Curr Opin Genet Dev* 16, 191-196.
39. Siniossoglou, S., Wimmer, C., Rieger, M., Doye, V., Tekotte, H., Weise, C., Emig, S., Segref, A., and Hurt, E. C. (1996) A novel complex of nucleoporins, which includes Sec13p and a Sec13p homolog, is essential for normal nuclear pores. *Cell* 84, 265-275.
40. Salama, N. R., Chuang, J. S., and Schekman, R. W. (1997) Sec31 encodes an essential component of the COPII coat required for transport vesicle budding from the endoplasmic reticulum. *Mol Biol Cell* 8, 205-217.
41. Peng, R., De Antoni, A., and Gallwitz, D. (2000) Evidence for overlapping and distinct functions in protein transport of coat protein Sec24p family members. *J Biol Chem* 275, 11521-11528.
42. Krogan, N. J., and Greenblatt, J. F. (2001) Characterization of a six-subunit holo-elongator complex required for the regulated expression of a group of genes in Saccharomyces cerevisiae. *Mol Cell Biol* 21, 8203-8212.
43. Li, Y., Takagi, Y., Jiang, Y., Tokunaga, M., Erdjument-Bromage, H., Tempst, P., and Kornberg, R. D. (2001) A multiprotein complex that interacts with RNA polymerase II elongator. *J Biol Chem* 276, 29628-29631.
44. Winkler, G. S., Petrakis, T. G., Ethelberg, S., Tokunaga, M., Erdjument-Bromage, H., Tempst, P., and Svejstrup, J. Q. (2001) RNA polymerase II elongator holoenzyme is composed of two discrete subcomplexes. *J Biol Chem* 276, 32743-32749.
45. Reilly, B. A., Kraynack, B. A., VanRheenen, S. M., and Waters, M. G. (2001) Golgi-to-endoplasmic reticulum (ER) retrograde traffic in yeast requires Dsl1p, a component of the ER target site that interacts with a COPI coat subunit. *Mol Biol Cell* 12, 3783-3796.
46. Westermann, S., Avila-Sakar, A., Wang, H. W., Niederstrasser, H., Wong, J., Drubin, D. G., Nogales, E., and Barnes, G. (2005) Formation of a dynamic kinetochore- microtubule interface through assembly of the Dam1 ring complex. *Mol Cell* 17, 277-290.

47. De Wulf, P., McAinsh, A. D., and Sorger, P. K. (2003) Hierarchical assembly of the budding yeast kinetochore from multiple subcomplexes. *Genes Dev* 17, 2902-2921.

48. Kitagawa, K., and Hieter, P. (2001) Evolutionary conservation between budding yeast and human kinetochores. *Nat Rev Mol Cell Biol* 2, 678-687.

49. Haber, J. E. (1974) Bisexual mating behavior in a diploid of Saccharomyces cerevisiae: evidence for genetically controlled non-random chromosome loss during vegetative growth. *Genetics* 78, 843-858.

Figure Legends

Figure 1

Generation of a High Confidence PPI Dataset

(**A**) ROC curve comparing PE analysis based on the primary data from Gavin et al. (6) (green), Krogan et al. (7) (red) and the consolidated (blue) dataset. MIPS complexes were used to generate true positive and true negative reference sets (see Experimental Procedures). The black diamond represents MIPS small-scale experiments (excluding two-hybrid experiments), and the dotted line indicates the set of points with the same true positive to true negative rate as MIPS small-scale experiments. The gold square represents earlier results from Gavin et al. (3). The binary interactions deposited in the SGD and BioGRID databases from the original Krogan et al. (7) and Gavin et al. (6) studies are represented by red and green circles, respectively. Inset, the same ROC curves are shown with expanded axis limits. (**B**) A subset of 9074 interactions from the consolidated dataset was obtained by applying a score threshold based on the true positive to true negative rate of the MIPS small-scale experiments (see Experimental Procedures). Shown is the fraction of interacting protein pairs with identical annotations from a random background set (gray), the initial Gavin et al. data (3) (gold), the MIPS small-scale experiments (black), and the above described subset of the consolidated data (blue) for the indicated categories. (**C**) The Pearson's correlation between the expression patterns (over a collection of approximately 1000 microarray experiments (19)) of pairs of genes encoding putatively interacting proteins were computed for the same sets as in B. The same color scheme is used. The curves have been normalized according to the

frequency at a correlation of 0.16, which corresponds to the maximum of the distribution for the small-scale experiments.

Figure 2

The High Confidence PPI subset of the Consolidated Dataset is of Higher Confidence than the original Deposited Datasets.

(**A**) Venn diagram showing the overlap between the high confidence PPI subset of the consolidated dataset, PPIs from Gavin et al. as deposited in BioGRID and PPIs from Krogan et al. as deposited in BioGRID. The number of PPIs defined in each dataset is indicated. (**B**) Shown is the fraction of interacting protein pairs with identical annotations for the indicated categories. The bars in the different categories represent a random background set (grey), PPIs found within the deposited Gavin et al. dataset only (green), the deposited Krogan et al. dataset only (red), the consolidated subset only (blue), in all three datasets (brown) and from the MIPS small-scale experiments (black).

Figure 3

The Unified, Physical Interactome Map.

(**A**) Hierarchical clustering of PPIs with a scaled PE score above 0.20. Directly observed interactions are labeled black (0) to yellow (1) and interactions purely inferred from indirect observations are labeled black (0) to blue (1). (**B**) and (**C**) Details from A showing an enlarged view of some of the complexes defined around the diagonal. The white rectangles in (**B**) indicate off-diagonal interactions between different complexes

and/or shared subunits. (**D**) Hierarchical clustering and corresponding venn-diagram of the NuA4 (blue), SWR-C (red), INO80C (green) and Tah1/Pih1 complexes. (**E**) Representation of the Nup84 nucleoporin and the coatomer COPII complexes.

Figure 4

PPIs and Protein Complexes Evident from the Consolidated Dataset

Hierarchical clustering of the PPI dataset accurately reveals the Elongator (**A**), Dsl1 and Q/t-SNARE (**B**), DASH microtubule ring (**C**), the inner and outer kinetochore and MIND (**D**) complexes. Interactions reported in either of the original datasets and the consolidated dataset are labeled black (0) to yellow (0.5) and interactions only identified by the consolidated dataset with our new analysis are labeled black (0) to red (0.5). Homozygous diploid *yol054w*Δ cells were tested for elevated 'bimater' (**E**) and 'a-like faker' phenotypes (**F**). Representative patches of homozygous wild-type and *yol054w*Δ strains after mating with MATa and MATa testers are shown. Histograms show the median number of colonies (n>5, error bar = 1 s.d.).