Analyzing yeast protein–protein interaction data obtained from different sources

Gary D. Bader and Christopher W.V. Hogue*

High-throughput methods for detecting protein interactions, such as mass spectrometry and yeast two-hybrid assays, continue to produce vast amounts of data that may be exploited to infer protein function and regulation. As this article went to press, the pool of all published interaction information on *Saccharomyces cerevisiae* was 15,143 interactions among 4,825 proteins, and power-law scaling supports an estimate of 20,000 specific protein interactions. To investigate the biases, overlaps, and complementarities among these data, we have carried out an analysis of two high-throughput mass spectrometry (HMS)–based protein interaction data sets from budding yeast, comparing them to each other and to other interaction data sets. Our analysis reveals 198 interactions among 222 proteins common to both data sets, many of which reflect large multiprotein complexes. It also indicates that a "spoke" model that directly pairs bait proteins with associated proteins is roughly threefold more accurate than a "matrix" model that connects all proteins. In addition, we identify a large, previously unsuspected nucleolar complex of 148 proteins, including 39 proteins of unknown function. Our results indicate that existing large-scale protein interaction data sets are nonsaturating and that integrating many different experimental data sets yields a clearer biological view than any single method alone.

Proteomics technologies, such as mass spectrometry (MS) and yeast two-hybrid assays, are currently providing a wealth of data on gene function through molecular interactions and post-translational protein modifications[1]. Protein–protein interactions mediate many aspects of cellular behavior[2] and are the basis for assemblies of molecular machines, such as RNA polymerase II. Estimates of the number of protein interactions range from two to ten per protein[3].

Two recent high-throughput analyses of protein complex composition in *S. cerevisiae* by Gavin *et al.*[4] and Ho *et al.*[5] have generated an unprecedented amount of protein interaction information. Both methods use tagged proteins as baits for high-affinity capture of complexes whose protein components are subsequently identified using MS[6]. Ho *et al.* use overexpressed bait proteins in a mild, single-step purification protocol based on the FLAG (DYKDDDDK) epitope tag and ultrasensitive liquid chromatography (LC)–tandem MS for protein identification (HMS-PCI; high-throughput mass-spectrometric protein complex identification). Gavin *et al.* use a more stringent two-step purification based on the tandem-affinity purification (TAP) tag using native bait protein expression and less precise peptide mass fingerprinting by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS for identification.

There are clear advantages and disadvantages to each approach, as noted by a recent study by von Mering *et al*[7]. Because each study detected interactions covering ~25% of the predicted yeast proteome, each represents a partial analysis of protein–protein interaction space. Together, the two HMS data sets provide functional information for 2,283 yeast proteins.

In this article, we first measure the biases and accuracy of these data sets using the Biomolecular Interaction Network Database (BIND)[8] and its associated bioinformatics infrastructure. BIND was designed to collect diverse experimental data on molecular interactions, complexes, and pathways in a machine-readable format. After comparing the HMS-PCI and TAP data sets to each other, we then undertake a global analysis of all current electronically accessible knowledge of experimentally determined yeast protein interaction data sets, including large-scale two-hybrid screens[9–13]. We then apply gene ontology (GO)[14]–derived annotation for *S. cerevisiae* proteins to examine functional connections in the genome-scale experiments and apply a recently described method based on *k*-cores[11] to find and visualize molecular complexes. (For a detailed explanation of methods and the protein interaction data sets used in this study, see Supplementary Experimental Protocol and Supplementary Table 1 online, respectively.)

Comparison of data sets

The overall networks of the two HMS data sets are remarkably different in connectivity, despite being similar in size. The HMS-PCI data set appears much more interconnected, whereas the TAP data set comprises more clusters of protein complexes that are sparsely connected (Fig. 1). An increased number of regulatory network proteins may create a higher level of connectivity between well-known protein complexes.

To assess whether the HMS-PCI and TAP data sets are different in this respect, we computed a high-level ratio of regulatory to housekeeping protein GO annotation. The regulatory category contains processes that include the words "response" (e.g., stress response), "control" (e.g., cell shape and cell size control), and "cycle" (e.g., cell cycle), processes (e.g., mating, budding) that are not involved in typical housekeeping roles, and any process having to do mainly with protein level regulation and cell signaling (e.g., protein degradation, phosphorylation and dephosphorylation)

*Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON, Canada M5G 1X5, and Department of Biochemistry, University of Toronto, Toronto, ON, Canada M5S 1A8. *Corresponding author (hogue@mshri.on.ca).*

(see Supplementary Table 2 online). For the yeast proteome data set, this ratio was 0.45, whereas for TAP it was 0.43 and for HMS-PCI it was 0.77. Thus, there are a greater number of regulatory proteins in the HMS-PCI data set than in the proteome and TAP data sets. This may partially explain the higher level of connectivity in the HMS-PCI data set. However, there are still large fractions of unknown and unannotated proteins, and we cannot determine what the true fraction is for any of these data sets.

Common baits

Of ~6,300 proteins ostensibly encoded by the yeast genome, Ho *et al.* selected 725 baits and Gavin *et al.* chose 1,739 baits; of these, 68% (493/725) and 26% (454/1739) yield detectably associated proteins (see Supplementary Table 3 online). These may be considered method efficiency ratios and may reflect differences in the bait expression systems selected. Only 115 baits are common to both studies, and of these 81 are associated with identifiable proteins in both data sets. Seven common baits do not associate with any proteins in either experiment, and 27 have partners in one method but not in the other (see Supplementary Tables 4 and 5 online).

To evaluate the biological relevance of these two methods, we compare the 115 common purifications from each method to a literature benchmark consisting of 1,762 proteins with 3,310 published interactions (obtained by low-throughput methods), which are presumed to be real, garnered from the Munich Information for Protein Sequences (MIPS) Yeast Genome Database[15], Yeast Proteome Database (YPD)[16], and PreBIND data set of BIND.

Modeling biochemical complexes as binary interactions

The purification processes used in the FLAG and TAP tag–based experiments isolate complexes of proteins that are sufficiently self-assembled around the tagged bait protein to withstand the purification protocol. Not all proteins in any given complex will interact directly with the bait protein, because interactions may be bridged by other molecules in the mixture (e.g., RNA or proteins) or interact with the bait at the same time (e.g., if the bait protein is involved in multiple physiologically relevant complexes). Consequently, in a computational analysis, the bait and associated proteins must be considered a population of biomolecular complexes of unknown topology.

While it is relatively straightforward to compare this information to known complexes in databases, most protein association information has been recorded as pairwise protein interactions resulting from experimental methods ranging from yeast two-hybrid screens to biochemical purification protocols, such as co-immunoprecipitation. Two models that represent complexes of unknown topology as collections of hypothetical pairwise interactions can be used to compare multiprotein complexes to previously determined protein interaction data sets: the "spoke" model and the "matrix" model.

The "spoke" model. This model assumes that the protein bait interacts directly with each one of the proteins in the population of complexes, like spokes of a wheel, as shown here for a single purification:

Population of complexes:

$C = \{b, c, d, e\}$ (b = bait)

Spoke model hypothetical interactions:

$i_S = \{b–c, b–d, b–e\}$

The spoke model excludes consideration of any homodimer formation or higher-ordered self-oligomerization of any protein in the set. It also yields fewer interactions than may actually be present and may misrepresent indirect interactions. Both Gavin *et al.*[4] and Ho *et al.*[5] implicitly used the spoke model when determining criteria for filtering promiscuously binding proteins based on frequency of occurrence. Spoke model representation is useful to reduce complexity in data visualization.

The "matrix" model. This approach assumes that any two proteins within the population of complexes have a pairwise interaction, as shown below:

Population of complexes;

$C = \{b, c, d, e\}$

Matrix model hypothetical interactions;

$$i_M = \{b–b, b–c, b–d, b–e, c–c, c–d, c–e, d–d, d–e, e–e\}$$

The matrix model contains all possible true interactions within the experimental data, but necessarily has a large number of false interactions as well, a problem that grows quadratically with the number of subunits in the complex. Furthermore, matrix topologies are physically implausible for larger multiple-subunit complexes because of probable steric clash. Both Gavin *et al.*[4] and Ho *et al.*[5] used a matrix model to determine their maximum data set overlap with previous large-scale yeast two-hybrid data sets.

A recent analysis of large-scale protein interaction data sets[7] used the matrix model to represent and compare HMS-PCI and TAP data and to derive measures of accuracy. The matrix model amplifies the effect of nonspecific interacting proteins by connecting them to all other associated proteins in the complex. The functional distribution of interactions for the spoke-modeled HMS-PCI and TAP data sets more closely resembles that for literature and large-scale yeast two-hybrid interactions than matrix-modeled data do (Fig. 3). The spoke-modeled HMS-PCI and TAP data sets have similar interaction density patterns along the diagonal of the function interaction matrix; however, TAP has less interfunctional group interaction density (below the diagonal), possibly signifying less nonspecific interactions between proteins in this set. While information is discarded in the spoke model, this may be an appropriate trade-off because spoke data are roughly threefold more accurate to our literature benchmark than matrix representation (see Table 2; accuracy here equals size of literature benchmark overlap/size of data set). If a matrix representation is used, it may be useful to weight the direct bait protein to associated protein (spoke) interactions with a higher significance score than other matrix interactions.

We urge caution when interpreting these diagrams as assessments of interaction data set reliability, as many modular proteins have multiple annotations. A set of functional annotation terms can be chosen to maximize or minimize interaction density along the diagonal of the functional matrix graphs. Thus, while interesting, these graphs cannot provide a complete view of most large-scale data sets, and conclusions drawn from methodological comparisons will be questionable until the *Saccharomyces cerevisiae* proteome is fully mapped and annotated using multiple methods.

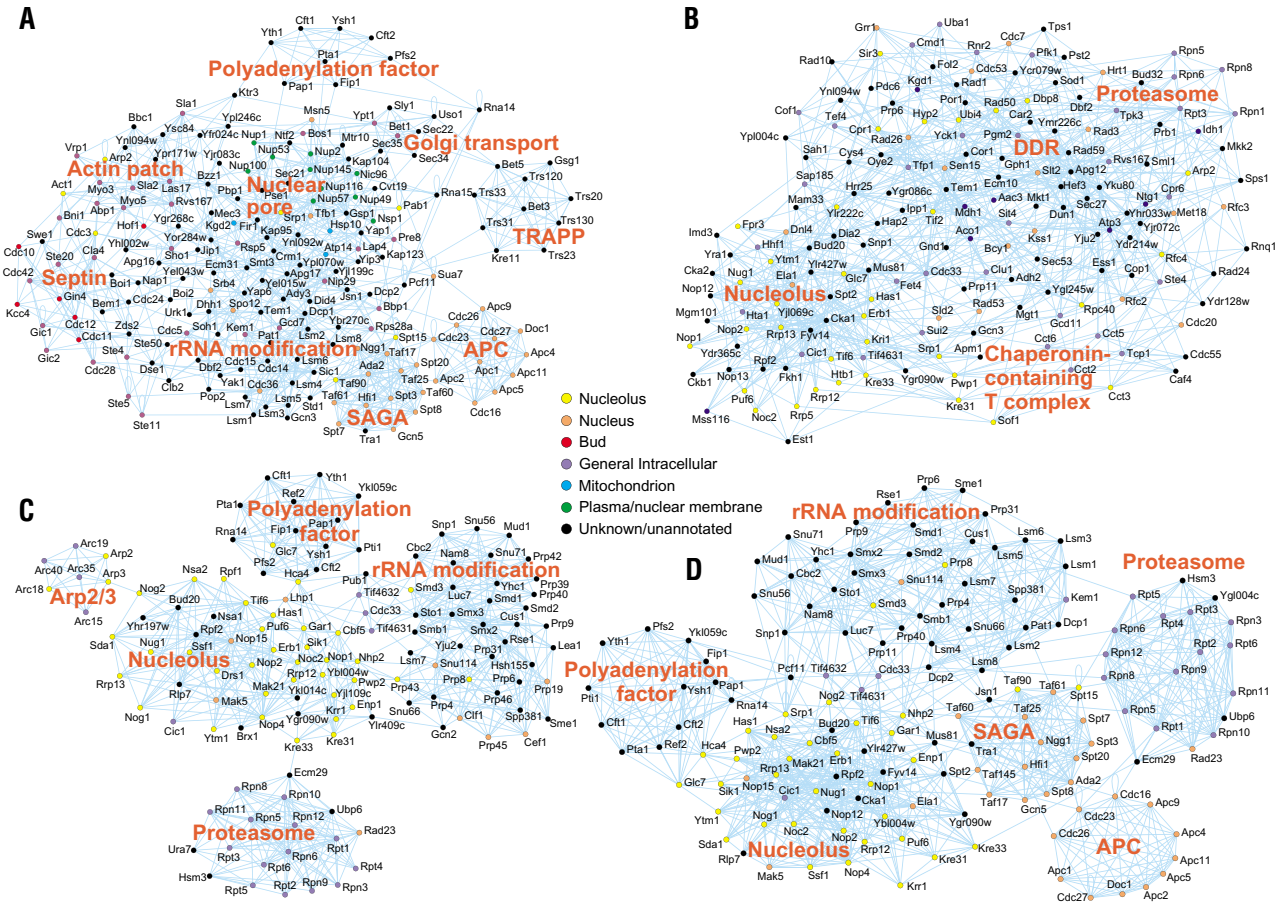

Figure 1. Visual representation of molecular complexes in protein interaction networks found using the *k*-core method. Although there are higher *k*-cores in these sets, a *k*-core level was chosen that represents as many nucleolar annotated proteins as possible without becoming too large. (A) Six-core of the integrated yeast protein interaction network before addition of HMS data. (B) Six-core of the HMS-PCI data set. (C) Six-core of the TAP data set. (D) Nine-core of the integrated yeast data set after addition of HMS data. The complex connectivity surrounding the nucleolus is clearer and more complete in the fully integrated data set (D), indicating that data integration is necessary for better understanding of a biological system. APC, Anaphase-promoting complex; SAGA, Spt-Ada-Gcn5-acetyltransferase transcriptional activator–histone acetyltransferase complex; DDR, DNA damage response; TRAPP, transport protein particle complex; 19S regulatory subunit of the proteasome labeled "proteasome". Proteins are colored according to GO cellular component, although nucleolar-localized annotation was supplemented with yeast orthologs of human proteins recently found to be in the human nucleolus[17]. In 1,000 randomly permuted networks from (A), (B), (C), and (D), the mean highest *k*-core was 5 (s.d. = 0), 5.85 (s.d. = 0.36), 5 (s.d. = 0), and 7 (s.d. = 0), respectively. Thus, the high *k*-core numbers in (A), (C), and (D) are highly unlikely to occur by chance.

The PreBIND set encompasses the known PubMed literature concerning all HMS-PCI baits, thus it can be considered comprehensive for the limited common bait subset. The TAP (628 interactions, 522 proteins) and HMS-PCI (875 interactions, 651 proteins) spoke model data sets from common baits contained 87 and 66 benchmark interactions involving 116 and 94 proteins, respectively. In contrast, the TAP (4,916 interactions, 522 proteins) and HMS-PCI (7,618 interactions, 651 proteins) matrix model sets from common baits had 264 and 193 benchmark interactions, involving 216 and 118 proteins, respectively. Thus, the TAP method is ~30% better at finding previously published interactions, at least for the limited intersection set. Interestingly, the HMS-PCI method finds 32% more unknown or unannotated proteins than TAP for the set of proteins associated with common baits (see Supplementary Table 6 online).

Table 1. Properties of large yeast interaction data sets

Data set	Proteins	Interactions	Homodimers
Ho "spoke"	1,578	3,618	0
Ho "matrix"	1,578	28,252	1,578
Gavin "spoke"	1,363	3,225	0
Gavin "matrix"	1,363	18,677	1,363
Uetz	1,001	946	43
Ito "full"	3,274	4,468	82
Ito "core"	796	805	52
PreBIND	859	1,196	0
MIPS	964	1,353	51
YPD	1,538	2,205	283
MIPS + PB + YPD	1,762	3,310	303

Ho, ref. 5; Gavin, ref. 4; Uetz, ref. 10; Ito, ref. 9.

Comparing common hits

Given that each data set encompasses 25% of the yeast proteome, the two data sets show little overall overlap, despite ~70% internal reproducibility within each data set[4,5]. In part, this minimal overlap reflects bait selection by different functional criteria and differing expression systems effects. The intersection of the two data sets using the spoke data representation model contains only 198 associations among 222 proteins (Fig. 2). This subset is probably the most reliable data in the two experimental sets, as it was independently found by both methods.

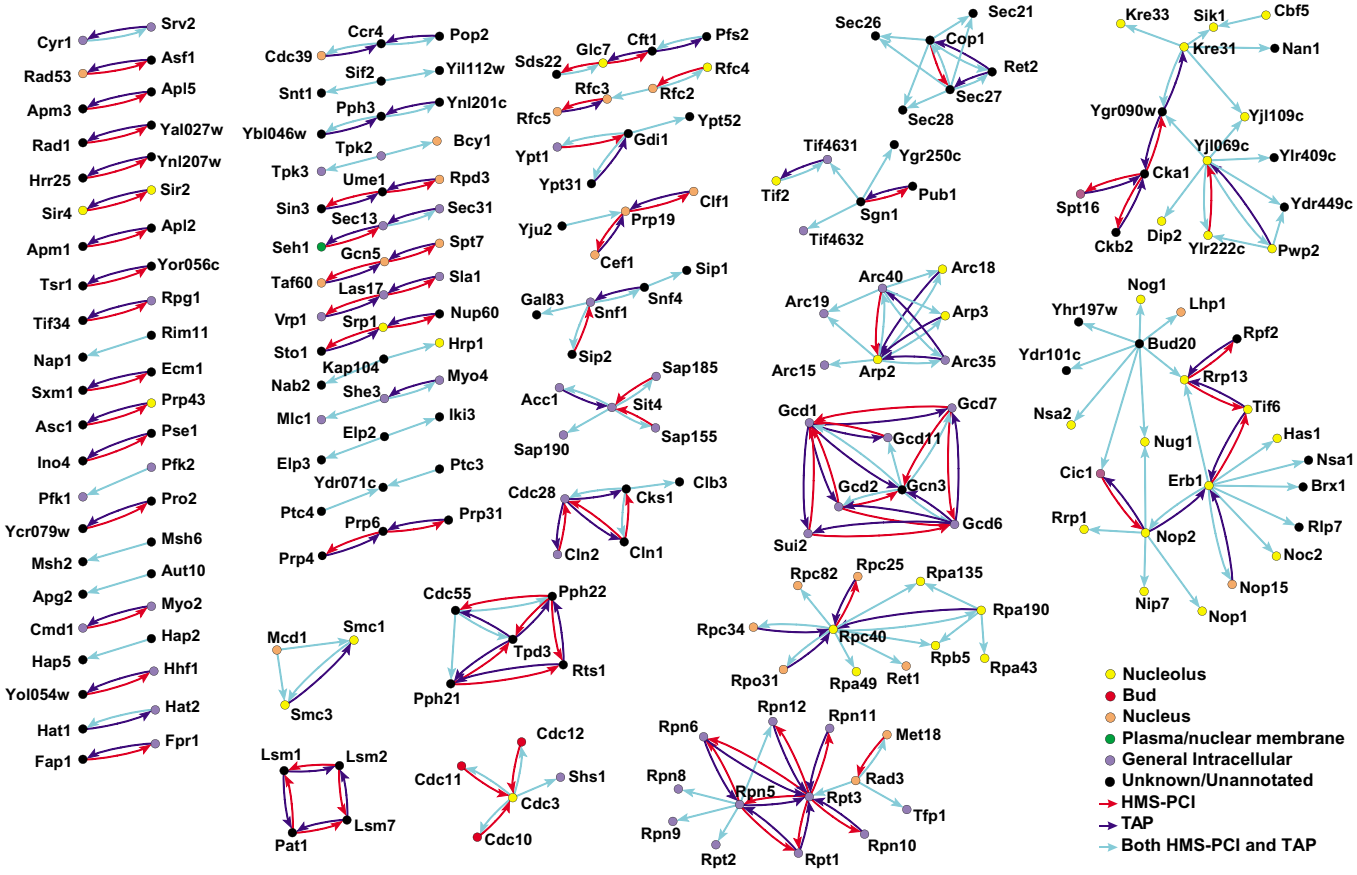

Figure 2. Overlap of the spoke models of TAP and HMS-PCI. There are 222 proteins and 310 arrows representing 198 protein associations. Arrows represent spoke interactions and point from bait to associated protein. Arrows are colored according to which study found the interaction: red, HMS-PCI; blue, TAP; cyan, both HMS-PCI and TAP. Proteins are represented as nodes, labeled with the common *S. cerevisiae* gene name and are colored by GO-derived cellular localization annotation: yellow, nucleolus; red, bud; orange, nucleus; green, membrane; purple, intracellular; black, unknown or unannotated.

The two largest common networks in the intersection comprise nucleolar proteins, including yeast orthologs of novel proteins recently detected in purified human nucleolar preparations[17,18]. One nucleolar network in the intersection set contains six essential proteins of unknown function: Ydr449c, Yjl069c, Yjl109c, Ygr090w, Ylr222c, and Ylr409c (Fig. 2). Several other smaller complexes are observed, many with known function. These include components of the proteasome regulatory particle, polyadenylation and elongation factors, chromosomal segregation, mitotic exit complexes and proteins involved in mRNA splicing, vesicle trafficking, glucose repression, and cytoskeleton rearrangement (see Fig. 2).

Functional bias

We examined various subsets of the experimental results to see if they were enriched in proteins of specific biological function

Table 2. Large yeast interaction data set cross-comparison

	Proteins\interactions\homodimers shared by datasets									
Data set	**MIPS+PB+YPD**	**YPD**	**MIPS**	**PreBIND**	**Ito core**	**Ito full**	**Uetz**	**Gavin matrix**	**Gavin spoke**	**Ho matrix**
Ho "spoke"	265\210\0	230\168\0	161\119\0	169\113\0	71\41\0	109\64\0	88\55\0	333\366\0	222\198\0	1578\3618\0
Ho "matrix"	448\480\135	385\357\126	226\202\21	246\192\0	101\69\13	162\117\22	120\86\12	658\2230\658	362\549\0	
Gavin "spoke"	361\333\0	276\198\0	249\230\0	163\117\0	71\40\0	97\55\0	78\47\0	1363\3225\0		
Gavin "matrix"	537\691\121	452\418\111	319\412\23	227\188\0	118\73\5	182\122\15	134\91\9			
Uetz	168\106\3	142\86\3	117\70\1	77\47\0	201\133\10	276\187\15				
Ito "full"	205\135\10	175\112\10	114\69\1	94\54\0	796\804\52					
Ito "core"	127\82\7	109\68\7	76\46\1	61\35\0						
PreBIND	859\1196\0	579\554\0	442\402\0							
MIPS	964\1353\51	803\834\31								
YPD	1538\2205\283									

Ho, ref. 5; Gavin, ref. 4; Uetz, ref. 10; Ito, ref. 9.

(functional bias), according to yeast functional annotation terms derived from the GO[14]. A full GO annotation of the TAP data set that corresponds to the published GO annotations of Ho *et al.* is provided in Supplementary Table 7 online. In general, Ho *et al.* focus on regulatory pathways in cell cycle control, DNA damage response and repair, signal transduction, and protein phosphorylation/dephosphorylation. In contrast, the baits and associated proteins expressed by Gavin *et al.* are enriched in general metabolism, nucleolar and ribosome biogenesis, protein metabolism, and transcription. The HMS-PCI data set also has more membrane-localized proteins, but otherwise subcellular compartments are evenly represented in both bait selection sets (see Supplementary Table 8 online).

No significant functional bias is found in baits that yield associated proteins versus those that do not (see Supplementary Tables 9 and 10 online). However, examination of the set of all identified proteins (1,579 from HMS-PCI and 1,363 from TAP; see Supplementary Table 11 online) as well as the set of only associated proteins (1,317 from HMS-PCI and 1,179 from TAP; see Supplementary Table 12 online) reveals that the functional bias mirrors the choice of baits, as might be expected from previous results showing that proteins of like function in yeast associate[19]. The only exception to this correlation is that metabolic proteins are overrepresented in the HMS-PCI interaction set compared with the bait set (see Supplementary Tables 6 and 8 online). This may reflect the propensity of the more sensitive LC–tandem MS method to detect low levels of nonspecifically associated background proteins. It may be that contaminant frequency filter cutoffs need be adjusted after examining the comparison of these two data sets (see Supplementary Table 3 online). Interestingly, proteins of unknown and/or unannotated GO biological process make up 41% and 35% of HMS-PCI and TAP data sets, respectively. Thus, HMS methods may help to provide functional connections for the large unannotated portion of the yeast proteome (see Supplementary Table 9 online).

Assuming that baits should generally pull down proteins of like function, it is expected that the distribution of function in the set of proteins associated with the 115 common baits will be similar in each experiment. Cell cycle and unknown proteins are heavily represented in the set of 115 common baits. In the set of proteins interacting with the common baits, the HMS-PCI data set contains more proteins involved in general metabolism, transport, signal transduction, and of unknown function, whereas the TAP data set contains more proteins involved in DNA damage response and repair, nucleolar and ribosome biogenesis, transcription, RNA localization and processing, or that are localized in the nucleus/nucleolus (see Supplementary Tables 6 and 13 online). Functional bias of the protein exclusion list does not explain this bias (see Supplementary Table 14 online), thus it most likely relates to biological sample handling, such as cell disruption techniques.

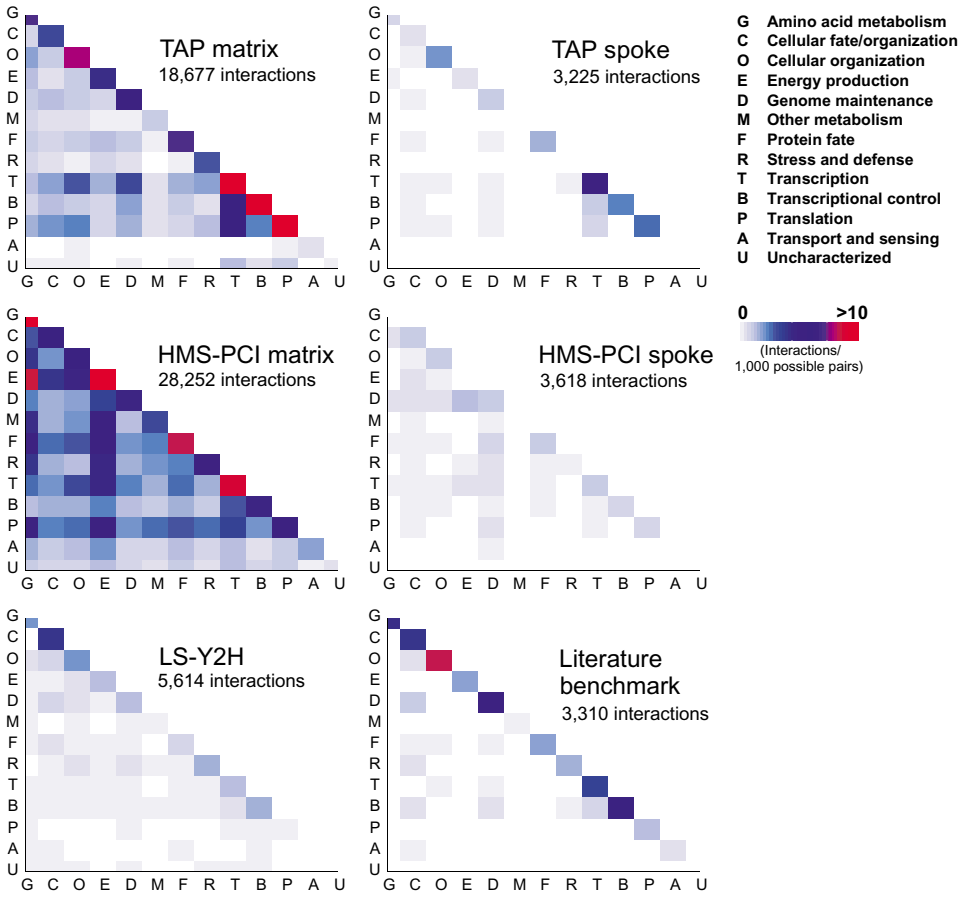

Figure 3. Functional annotation matrices[26] showing the distribution of interactions of six data sets. Annotation is as in von Mering *et al.*[7] to aid comparison. The HMS-PCI matrix interaction set is corrected compared to the von Mering version, as it was derived from original immunoprecipitation (IP) data (see Supplementary Table 1 online), whereas the published HMS-PCI data collapsed multiple IPs into one protein set.

Integration and analysis

To assess the proteome coverage provided by all HMS and yeast two-hybrid studies to date, the spoke and matrix models of the HMS-PCI and TAP data sets (see "Modeling biochemical complexes as binary interactions") were combined and compared with a compiled data set of interactions from multiple large-scale yeast two-hybrid experiments[9–13]. We find 173 interactions between 265 proteins common to yeast two-hybrid assays (5,614 interactions, 3,652 proteins) and spoke MS (6,645 interactions, 2,283 proteins), and 304 interactions between 388 proteins common to the yeast two-hybrid assays and matrix HMS (44,680 interactions, 2,283 proteins). We collected and integrated all machine-readable data from various data sets[4,5,9–13,15,16] to form a nonredundant set of 15,143 experimentally determined yeast protein interactions encompassing 4,825 proteins, or ~76% of the proteome.

The largest component of this integrated network contains 15,059 interactions among 4689 proteins, leaving only 136 proteins not part of the main group. A full $N \times N$ comparison among selected large-scale individual data sets is shown in Table 1. The combined HMS matrix data set overlaps with only 33% of interactions in the the MIPS + PreBIND + YPD literature benchmark, leaving 67% of previously found protein interactions involving proteins in the combined HMS data set undetected. We conclude from this analysis that even with the advent of recent HMS studies, the detectable protein interaction space in the yeast system is far from saturated.

As described earlier by Barabasi *et al.*[20,21], the integrated network follows a power-law node connectivity distribution. Within this distribution, essential proteins show a higher level of connectivity (10.7 average connections) than nonessential proteins (5.0 average connections). Furthermore, by scaling the power-law connectivity distribution of the integrated data set (4,825 proteins), defined above, to the yeast proteome (6,334 proteins[22,23]), we estimate on the order of 20,000 protein interactions in yeast, a lower estimate than that provided by von Mering *et al.*[7]

The large integrated data set contains a higher percentage of proteins of unknown function and localization than the proteome (compare Supplementary Tables 11 and 15 online). Of the ~1,500 predicted open reading frames (ORFs) not identified by any protein interaction method, 75% are of unknown biological process and 80% have no localization GO annotation (see Supplementary Table 16 online). These ORFs may be present in extremely low abundance in the cell or may only be expressed during specific developmental stages (e.g., spore formation).

Conclusions

Large-scale experiments have the potential to discover previously unknown functional connections among components of the cell (see "A novel nucleolar network"), and thus promise to expand rapidly our knowledge of biology. However, data quality is of paramount importance in this knowledge expansion. Thus far, large-scale techniques do not show enough internal consistency to warrant complete acceptance of the resulting data. This indicates that each screen will have to be carried out multiple times before achieving a high enough data quality for a particular method. While it is relatively straightforward to accomplish systematic identification of stable multiprotein complexes, or "cellular machines", detecting transient regulatory interactions, often involved in signaling pathways, metabolons, and hyperstructures[24], is still difficult. Considering these constraints, it is important concurrently to develop computational systems, such as BIND[8,25], that can integrate, visualize, and mine available molecular interaction data sets to speed the emergence of a clear view of protein complexes and associated regulatory interactions.

Note: Supplementary information is available on the Nature Biotechnology website.

A novel nucleolar network

Using a method of complex detection in interaction networks based on finding *k*-cores, as earlier described[11], we have determined that both high-throughput mass spectrometry data sets contain a dense, previously unsuspected nucleolar network. A *k*-core of a network, or graph, is a subgraph in which all proteins are connected to at least *k* other proteins in the subgraph, where *k* is 0, 1, 2, 3 ... The *k*-core method was applied to the integrated yeast interaction network without HMS data (Fig. 1A), to the HMS-PCI (Fig. 1B) and TAP (Fig. 1C) HMS data sets alone, and to the fully integrated network including all HMS data (Fig. 1D).

The nucleolar network emerges as the data set size is increased. Notably, only a few nucleolar proteins are present in the highly connected regions of the network before HMS data inclusion (Fig. 1A). In contrast, both the individual HMS-PCI and TAP data sets contain highly connected networks involving nucleolar proteins. Many of the proteins in the nucleolar network are orthologs of human proteins recently found in highly purified human nucleoli[17,18].

Interestingly, three of the subcomplexes that are visually apparent in Figure 1D correspond to the known substructure of the nucleolus as determined by electron microscopy[27]. The fibrillar component (FC) involved in pre-rRNA transcription corresponds to a subcomplex of proteins with likely transcriptional functions, labeled "SAGA" (Fig. 1D). All 14 known components of the SAGA complex are visible in Figure 1D, although two other proteins are also highly connected to SAGA: Taf145 and Spt15. Taf145 and Spt15 are known to participate in the RNA polymerase II general transcription factor complex with other SAGA components.

The dense fibrillar component (DFC) is the site of rRNA processing and corresponds to the complex of proteins labeled "rRNA modification". Known nucleolar links with snRNA-associated proteins are visible in the many links between the nucleolar complex and RNA modification complexes (e.g., U4/U6 snRNP, U4/U6.U5 tri-snRNP complex, U2 snRNP, and U1 snRNP complexes). All nine known components of polyadenylation factor I (PFI) are clustered in Figure 1D along with Rna14 and Ref2, known to be associated with PFI, and Pti1, a protein of unknown function that seems to be a previously unknown component of PFI. The granular component (GC) involved in assembling preribosomal proteins, corresponds to the protein cluster labeled "nucleolus". Consistent with recent findings of nucleolar functional links to cell cycle control[28], the anaphase-promoting complex (APC) is seen connecting to the nucleolus, SAGA, and the proteasome (Cdc23 interacts with Spt2, Ada2, and Rpt1; Cdc16 interacts with Mus81 and Rpt1). All 11 known components of APC are visible in Figure 1D. Of the 18 known 19S proteasome regulatory particle (PRP) components, the nine-core in Figure 1D misses Rpn1, Rpn2, Rpn4, and Rpn7. These are connected to the 19S PRP in the underlying data set, but not by nine interactions, and so do not appear in the nine-core. Interestingly, Ecm29, Hsm3, Rad23, Ubp6, and Ygl004c appear highly connected with the 19S PRP. Ubp6 and Rad23 are known to be associated with elements of the proteasome, but Ecm29, Hsm3, and Ygl004c, a WD40 repeat–containing protein, are not, although their high connectivity suggests that they may be components of PRP. While Jsn1 is not known to be part of any complex, it has been shown to interact with >160 proteins almost exclusively in high-throughput yeast two-hybrid screens. Jsn1 has been shown to bind to SAGA, APC, protein components of the proteasome, nucleolus, and the region on Figure 1D labeled "rRNA modification," although these interactions may be mediated by at least one RNA-bridging molecule, because Jsn1 has been predicted to bind RNA. Thus, as illustrated by identification of a large nucleolar complex, sufficient nondirected coverage of protein interactions can reveal large-scale functional domains, without a priori knowledge of the functional annotation in the integrated data set.

Acknowledgments

We thank Mike Tyers, Charlie Boone, and Tony Pawson for helpful discussions. This work was supported in part from grants from the Canadian Institutes of Health Research (CIHR), the Ontario Research and Development Challenge Fund and MDS-Sciex to C.H. G.D.B. is supported by an Ontario Graduate Scholarship (OGS).

Received 20 February 2002; accepted 18 August 2002

1. Fields, S. Proteomics. Proteomics in genomeland. *Science* **291**, 1221–1224 (2001).
2. Pawson, T., Gish, G.D. & Nash, P. SH2 domains, interaction modules and cellular wiring. *Trends Cell Biol.* **11**, 504–511 (2001).
3. Marcotte, E.M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
4. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
5. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
6. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
7. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
8. Bader, G.D. *et al.* BIND—The biomolecular interaction network database. *Nucleic Acids Res.* **29**, 242–245 (2001).
9. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
10. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
11. Tong, A.H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324 (2002).
12. Drees, B.L. *et al.* A protein interaction map for cell polarity development. *J. Cell Biol.* **154**, 549–571 (2001).
13. Fromont-Racine, M. *et al.* Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17**, 95–110 (2000).
14. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
15. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
16. Costanzo, M.C. *et al.* YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29**, 75–79 (2001).
17. Andersen, J.S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
18. Harnpicharnchai, P. *et al.* Composition and functional characterization of yeast 66S ribosome assembly intermediates. *Mol. Cell* **8**, 505–515 (2001).
19. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nat.Biotechnol.* **18**, 1257–1261 (2000).
20. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
21. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
22. Pruitt, K.D. & Maglott, D.R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
23. Chervitz, S.A. *et al.* Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res.* **27**, 74–78 (1999).
24. Norris, V. *et al.* Hypothesis: hyperstructures regulate bacterial structure and the cell cycle. *Biochimie* **81**, 915–920 (1999).
25. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
26. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29**, 482–486 (2001).
27. Olson, M.O., Dundr, M. & Szebeni, A. The nucleolus: an old factory with unexpected capabilities. *Trends Cell Biol.* **10**, 189–196 (2000).
28. Visintin, R. & Amon, A. The nucleolus: the magician's hat for cell cycle tricks. *Curr. Opin. Cell. Biol.* **12**, 752 (2000).