Molecular Cell, Vol. 12, 1353–1365, December, 2003, Copyright ©2003 by Cell Press

Assigning Function to Yeast Proteins by Integration of Technologies

Tony R. Hazbun,[1,2] Lars Malmström,[3] Scott Anderson,[4] Beth J. Graczyk,[3] Bethany Fox,[3] Michael Riffle,[3] Bryan A. Sundin,[3] J. Derringer Aranda,[2] W. Hayes McDonald,[4] Chun-Hwei Chiu,[3] Brian E. Snydsman,[3] Phillip Bradley,[3] Eric G.D. Muller,[3] Stanley Fields,[1,2] David Baker,[1,3] John R. Yates III,[4] and Trisha N. Davis[3,*]

[1]Howard Hughes Medical Institute
[2]Departments of Genome Sciences and Medicine
[3]Department of Biochemistry
University of Washington
Seattle, Washington 98195
[4]Department of Cell Biology
Scripps Research Institute
La Jolla, California 92037

Summary

Interpreting genome sequences requires the functional analysis of thousands of predicted proteins, many of which are uncharacterized and without obvious homologs. To assess whether the roles of large sets of uncharacterized genes can be assigned by targeted application of a suite of technologies, we used four complementary protein-based methods to analyze a set of 100 uncharacterized but essential open reading frames (ORFs) of the yeast *Saccharomyces cerevisiae*. These proteins were subjected to affinity purification and mass spectrometry analysis to identify copurifying proteins, two-hybrid analysis to identify interacting proteins, fluorescence microscopy to localize the proteins, and structure prediction methodology to predict structural domains or identify remote homologies. Integration of the data assigned function to 48 ORFs using at least two of the Gene Ontology (GO) categories of biological process, molecular function, and cellular component; 77 ORFs were annotated by at least one method. This combination of technologies, coupled with annotation using GO, is a powerful approach to classifying genes.

Introduction

Deciphering the functional roles of a large set of genes and their encoded products is the central challenge in the analysis of an organism once its genome sequence is complete. The yeast *Saccharomyces cerevisiae* has been studied thoroughly by systematic genomic and proteomic technologies. Its ~6000 predicted ORFs have been analyzed for expression under a multitude of conditions (DeRisi et al., 1997; Horak and Snyder, 2002); each ORF has been individually deleted and the resulting strains phenotypically characterized (Giaever et al., 2002; Winzeler et al., 1999); protein interactions have been detected by both biochemical/mass spectrometry (Gavin et al., 2002; Ho et al., 2002) and two-hybrid approaches (Ito et al., 2001; Uetz et al., 2000); and many of the proteins have been localized by indirect immunofluorescence or by fusion to green fluorescent protein (GFP) (Huh et al., 2003; Kumar et al., 2002; Ross-Macdonald et al., 1999). Despite these large-scale studies, as well as numerous small-scale analyses, approximately one-third of the ORFs have not been assigned to a functional category, indicating that large-scale studies yield incomplete data sets and small-scale, focused studies tend to be biased toward specific areas of biology. We focus here on an important subset of these uncharacterized ORFs, those that are essential for yeast viability.

Complete analysis of the yeast proteome requires characterization of proteins refractory to analysis in previous studies. We started with 100 ORFs that were known to be essential for viability but carried out unknown functions. The protein products of these ORFs were subjected to four independent and complementary approaches that assessed protein structure, localization, and interactions, critical properties for determining function. The resulting data were assembled on a web-based informatics platform (http://www.yeastrc.org/unknown_orfs) that allowed their synthesis into a coherent and accessible framework, using the standardized vocabulary of the Gene Ontology Consortium to classify the ORFs (Ashburner et al., 2000). GO terms describe proteins based on three fundamental properties: (1) biological process, the biological objective to which a protein contributes; (2) cellular component, the place in the cell where a protein is active; and (3) molecular function, the biochemical activity. Of the four implemented technologies, two, the identification of copurifying proteins by mass spectrometry and binary protein-protein interactions revealed by two-hybrid, are particularly relevant to assigning biological process by discovering the known proteins that associate with a given unknown. Subcellular localization by fluorescence microscopy can assign cellular component. Sequence similarities or predicted structural similarities to known proteins can yield clues to molecular function. Moreover, the computational prediction of function can be performed on proteins that are difficult to analyze experimentally.

Results

Each technology was optimized before analysis of the uncharacterized ORFs. The tandem affinity purification (TAP) tag, which consists of two IgG binding domains and calmodulin binding peptide (Tasto et al., 2001), was integrated in-frame with a given ORF in a diploid cell, and then haploids were isolated in which the tagged version of the gene was the only copy. This strategy requires that the tagged version of each protein, which is expressed at the normal endogenous level, be able to carry out the essential activity. The proteins copurifying with a given tagged protein were identified by mass spectrometry and multidimensional protein identification technology (MudPIT) (McDonald et al., 2002). For

*Correspondence: tdavis@u.washington.edu

the two-hybrid analysis, ORFs were fused to the Gal4 DNA binding domain and tested against a genome-wide two-hybrid array (Uetz et al., 2000). For fluorescence microscopy, a set of strains with the essential uncharacterized ORFs tagged with the Venus version of YFP (Nagai et al., 2002) was constructed in a fashion that also demanded that the tagged protein provide activity. Like the TAP-tagged proteins, the YFP-tagged proteins must maintain cell viability and thus be functional and properly localized. In sum, one of the four technologies yielded data for 96% of the ORFs, two technologies for over 80%, and three technologies for over 50% (Supplemental Table S1 at http://www.yeastrc.org/unknown_orfs).

To classify the function of each protein, we ascribed Gene Ontology terms. GO biological process terms were determined systematically by first using the GO term finder (http://db.yeastgenome.org/cgi-bin/SGD/GO/goTermFinder) to identify common GO terms for each ORF among the protein purification data set (Table 1, column 4) and among the two-hybrid data set (Table 1, column 5). We did not predict a GO biological process if neither method yielded a common GO term. A single copurifying protein of known function determined the associated GO term for eight uncharacterized ORFs. The cellular component term was assigned based on the fluorescence microscopy (Table 1, column 6). The molecular function term was assigned based on remote homologies to proteins of known function using PSI-BLAST, consensus fold recognition methods, or structure-based matches of de novo structure predictions to proteins of known structures. As proteins with the same fold can have different functions (Todd et al., 2001), assignments were only made if the GO term was consistent with the data generated by other technologies, as was true for 27 out 29 possible annotations (Table 1, column 7). Seventy-seven ORFs were annotated with at least one GO term, 48 ORFs were annotated with at least two GO terms, and 17 ORFs were annotated with all three GO terms (Table 1). During the course of our work, 16 genes were annotated by others (see Experimental Procedures). The newly published information for this set of genes is consistent with our data, hence validating our approach. Moreover, our approach has generated additional information beyond the published data.

Purification and mass spectrometry data allowed the assignment of 32 biological process GO terms and defined 29 complexes (Table 2). Two-hybrid screens identified 271 putative interactions, allowing the annotation of GO process terms for 16 ORFs. The overlap between these two data sets was similar to previously published genomic efforts (von Mering et al., 2002) with 16 interactions identified by both approaches. However, when both methods predicted a GO term, the predictions were uniformly consistent (8/8), although the two-hybrid predictions tended to be more broadly defined GO terms. Localization data allowed the assignment of cellular component terms for 63 ORFs. The cellular component terms were uniformly consistent with the GO process term annotations and thus added confidence to the process term annotations. Remote homology searches and protein structure prediction provided molecular function annotations to 27 ORFs. We describe several examples below where the integration of data collected from complementary technologies, by assignment of GO terms, predicted the cellular role for an uncharacterized protein.

The YDR288w complex and the YML023c complex are two new related complexes involved in DNA repair (Figure 1). Both were identified through data collected from all four technologies. Each complex contains the heterodimer Smc5-Rhc18, but the other constituent proteins differ. YDR288w purified in addition with Nse1, Mms21, and the uncharacterized protein Qri2. Mutation of *SMC5*, *NSE1*, (Fujioka et al., 2002), or *MMS21* (Prakash and Prakash, 1977) confers sensitivity to DNA damaging agents, predicting a role in DNA repair for the YDR288w complex. YML023c also purified with the uncharacterized protein Kre29, and the YML023c-Kre29 interaction was observed in the two-hybrid analysis as well. As expected, tagged-Smc5 purified both the YDR288w complex and the YML023c complex. Two-hybrid analysis identified interactions with proteins involved in other biological pathways related to DNA repair, such as sumoylation (Hoege et al., 2002) and chromosome segregation (Figure 2). The localization data were consistent with a prediction of DNA repair because YDR288w and YML023c both localized to the nucleus. We identified structural and sequence homologies for six out of the eight members of these interrelated complexes. All the function predictions were consistent with our process and component annotations (Supplemental Table S2 at http://www.yeastrc.org/unknown_orfs). Hence, the combination of the data from the four technologies yields a strong prediction for the role of these two new complexes.

The YKR079c complex was assigned a role in DNA and RNA catabolism. The protein purification data were consistent with a single stoichiometric complex for YKR079c. However, the localization data suggest that YKR079c forms two different complexes, one in the nucleus with YMR099c and one in the mitochondrion with Nuc1 (Figure 3). Nuc1 is defined as having a role in DNA and RNA catabolism and has both deoxyribonuclease and ribonuclease activity (Dake et al., 1988). The copurification and colocalization of YKR079c with Nuc1, and its protein structure prediction as a metallohydrolase/oxidoreductase, support our annotation of nuclease activity. Consistent with this prediction, YKR079c has a human homolog ELAC2 which is a prostate cancer susceptibility gene that encodes a tRNA 3′processing endoribonuclease activity (Takaku et al., 2003). The role of YKR079c in the nucleus could not be determined because it associates with the uncharacterized protein YMR099c.

Three new mRNA splicing proteins were identified in our analysis. YLR424w and YKR022c purified together and with 18 other spliceosome components, providing a strong prediction for the GO process term. An interaction between YLR424w and YKR022c was also detected by two-hybrid. Two-hybrid analysis of YLR424w was unusual because a large number of interactions were detected including thirty-six other interactions for which nucleobase, nucleoside, nucleotide, and nucleic acid metabolism was the predominant GO process term. These interactions involved proteins mainly associated with RNA processing or transcription. Both YLR424w and YKR022c localized to the nucleus and YLR424w

Table 1. Assignment of GO Terms Based on the Experimental and Computational Evidence from the Four Technologies

	GO Terms Assignment			Experimental and Computational Evidence			
ORF	Biological process	Component	Molecular function	Process predicted by copurification	Process predicted by two-hybrid	Localization	Remote homology search and Protein structure prediction
YJR072C	Aerobic respiration	Cytoplasm	Signal peptide binding	None	Aerobic respiration	Cytoplasm	Signal peptide-binding domain superfamily; GTPase domain of the signal sequence recognition protein Ffh
YJR013W	Amino acid transport	Nuclear envelope-ER	Mannosyl transferase	None	Amino acid transport	Nuclear envelope-ER	Mannosyl transferase (PSI)
YFR003C	Cell cycle	Cytoplasm, nucleus[a]	Protein phosphatase inhibitor	Cell cycle	Morphogenesis*	No signal	Protein phosphatase type 1 inhibitor (PSI)
YIR010W (DSN1)	Chromosome segregation	Kinetochore	Unknown	Chromosome segregation	None*	Kinetochore	Myosin motor domain
YPL233W (NSL1)	Chromosome segregation	Kinetochore	Unknown	Chromosome segregation	None	Kinetochore	Spectrin repeat
YKL088W	Coenzyme A biosynthesis	Cytoplasm	Phosphopanto-thenoylcysteine decarboxylase	Coenzyme A biosynthesis	Cellular process	Cytoplasm	Phosphopantothenoylcysteine decarboxylase; Pantothenate metabolism flavoprotein superfamily
YDR531W	Coenzyme A biosynthesis[b]	Cytoplasm, nucleus	Pantothenate kinase	None	None	Cytoplasm, nucleus	Fumble, similar to pantothenate kinases
YIL083C	Coenzyme A biosynthesis[b]	Cytoplasm, nucleus	Phosphopanto-thenoylcysteine ligase	None	No positives	Cytoplasm, nucleus	Phosphopantothenoylcysteine synthetase/decarboxylase (PSI); Pantothenate metabolism flavoprotein superfamily
YKR079C	DNA catabolism; RNA catabolism	Nucleus, Mitochondrion	Nuclease	DNA or RNA catabolism	No positives	Nucleus, Mitochondrion	tRNA endoribonuclease (PSI) Metallohydrolase/oxidoreductase
YDR489W (SLD5)	DNA repair	Nucleus[a]	Unknown	Unknown complex 1	Unknown complex 1*	No data	Interferon-induced guanylate-binding protein
YOL146W (PSF3)	DNA repair	Nucleus[a]	Unknown	Unknown complex 1	None	No data	Ribosome recycling factor
YDR013W (PSF1)	DNA repair	Nucleus[a]	Unknown	Unknown complex 1	DNA repair*	No signal	None
YJL072C (PSF2)	DNA repair	Nucleus	Unknown	Unknown complex 1	Unknown complex 1*	Nucleus	ARM repeat, Cytokines superfamily
YML023c	DNA repair	Nucleus	Unknown	DNA repair	Cell growth and/or maintenance*	Nucleus	None
YDR288W	DNA repair	Nucleus	DNA binding	DNA repair	Response to DNA damage stimulus*	Nucleus	Adenylation domain of NAD+-dependent DNA ligase; Winged helix DNA-binding domain

(*continued*)

Table 1. Continued

	GO Terms Assignment			Experimental and Computational Evidence			
ORF	Biological process	Component	Molecular function	Process predicted by copurification	Process predicted by two-hybrid	Localization	Remote homology search and Protein structure prediction
YKL082C	Establishment of cell polarity	Nucleolus, nucleus	Structural molecule	No data	Establishment of cell polarity	Nucleolus, nucleus	Nucleolar matrix protein SURF-6 (PSI)
YGR002C	Histone acetylation	Nucleus	DNA binding	Histone acetylation	Nuclear division*	Nucleus	Histone methyltransferase associated binding protein (PSI); Myb-like DNA-binding domain protein family
YGR198W	MAPKKK cascade	Plasma membrane	Transferase	MAPKKK cascade	None	Plasma membrane	N-acetylglucosamine transferase (PSI)
YGR046W	Mitochondrial translocation	Mitochondrion	Unknown	Mitochondrial translocation	None	Mitochondrion	None
YPL063w	Mitochondrial translocation	Mitochondrion	Unknown	Mitochondrial translocation	Transport	Mitochondrion	NLI interacting factor protein family, 4-helical cytokines superfamily
YJR141W	mRNA processing	Unknown	Unknown	No data	mRNA processing	No signal	None
YLR424w	mRNA splicing	Nucleus[a]	RNA binding	mRNA splicing	Nucleobase, nucleoside, nucleotide and nucleic acid metabolism*	No signal	G-patch domain protein family (RNA binding)
YDL209C (CWC2)	mRNA splicing	Nucleus	RNA binding	mRNA splicing	None*	Nucleus	Two RNA-binding domains
YKR022C	mRNA splicing	Nucleus	Unknown	mRNA splicing	None*	Nucleus	None
YNL245c (CWC25)	mRNA splicing	Nucleus	Unknown	mRNA splicing	None	Nucleus	None
YLR132c	mRNA splicing	Nucleus, mitochondrion	Unknown	mRNA splicing	No positives	Nucleus, mitochondrion	None
YNL313C	Nuclear membrane fusion	Cytoplasm, nucleus	Unknown	Nuclear membrane fusion	No positives	Cytoplasm, nucleus	ARM repeat superfamily; TPR-like superfamily
YJL097W	Nuclear membrane fusion	Nuclear envelope-ER, vacuolar membrane, vacuole	Unknown	No data	Nuclear membrane fusion	Nuclear envelope-ER, vacuolar membrane, vacuole	Protein tyrosine phosphatase like (PSI); Membrane all-alpha superfamily
YLL034C	Organelle organization and biogenesis	Nucleolus, nucleus	ATPase	Organelle organization and biogenesis	None	Nucleolus, nucleus	Membrane fusion ATPase
YJR136c	Protein biosynthesis	Unknown	Unknown	Protein biosynthesis	No signal	No data	None
YPR169W	Protein mono-ubiquitination	Nucleolus, Nucleus	Unknown	Protein mono-ubiquitination	No positives	Nucleolus, Nucleus	WD40 repeat superfamily

(*continued*)

Table 1. Continued

	GO Terms Assignment			Experimental and Computational Evidence			
ORF	Biological process	Component	Molecular function	Process predicted by copurification	Process predicted by two-hybrid	Localization	Remote homology search and Protein structure prediction
YKL195w	Protein targeting	Mitochondrion	Unknown	Protein targeting	No positives	Mitochondrion	None
YKR038c	Response to dessication	Nucleus	ATPase	Response to dessication	No data	Nucleus	Glycoprotein endopeptidase (PSI); Actin-like ATPase domain
YNR054C	rRNA processing	Nucleolus[a]	RNA binding	No data	rRNA processing	No data	Three RNA binding domains
YGR145W	rRNA processing	Nucleolus	Unknown	rRNA processing	No data	Nucleolus	WD40 repeat superfamily
YDR365C	rRNA processing	Nucleolus	Unknown	rRNA processing	None	Nucleolus	Unknown
YJL010C	rRNA processing	Nucleolus, nucleus	RNA binding	rRNA processing	None	Nucleolus, nucleus	Pumilio family RNA binding
YNL124w (NAF1)	rRNA processing	Nucleolus, nucleus	Unknown	rRNA processing	None	Nucleolus, nucleus	Beta and beta-prime subunits of DNA dependent RNA-polymerase superfamily
YJL091C (GWT1)	Secretory pathway	Nuclear envelope-ER	Unknown	Secretory pathway	No data	Nuclear envelope-ER	4-helical cytokines superfamily
YOR287C	Sporulation	Unknown	Unknown	No data	Sporulation	No data	Tropomyosin superfamily
YHR122W	Transcription	Cytoplasm, nucleus	Unknown	No data	Transcription	Cytoplasm, nucleus	None
YJR012C	Transport	Unknown	Unknown	Transport	No positives	No signal	None
YLR145w	TRNA processing	Nucleolus, nucleus	Unknown	tRNA processing	No positives	Nucleolus, nucleus	None
YNL207W (RIO2)	Unknown	Cytoplasm	Protein kinase	None	None	Cytoplasm	Chk2 kinase; Cyclin-dependent PK
YDR527W	Unknown	Cytoplasm	Unknown	None	No positives	Cytoplasm	ARM repeat superfamily
YLR022C	Unknown	Cytoplasm	Unknown	None	None	Cytoplasm	None
YOR262W	Unknown	Cytoplasm	Unknown	No data	None	Cytoplasm	Domain of the SRP/SRP receptor G-proteins
YGR277C	Unknown	Cytoplasm, nucleus	Nucleotidyltrans-ferase	None	None	Cytoplasm, nucleus	Nucleotide-diphospho-sugar transferases superfamily; Cytidylyltransferase family
YDR267C	Unknown	Cytoplasm, nucleus	Transcription regulator[c]	None	None	Cytoplasm, nucleus	Tup1, C-terminal domain; WD40 repeat superfamily
YGL047W	Unknown	Cytoplasm, nucleus	Transferase, transferring hexosyl groups	None	No positives	Cytoplasm, nucleus	UDP-Glycosyltransferase/ glycogen phosphorylase superfamily
YNL260C	Unknown	Cytoplasm, nucleus	Unknown	Unknown complex 3	No positives	Cytoplasm, nucleus	None

(*continued*)

Table 1. Continued

ORF	GO Terms Assignment			Experimental and Computational Evidence			
	Biological process	Component	Molecular function	Process predicted by copurification	Process predicted by two-hybrid	Localization	Remote homology search and Protein structure prediction
YOR060C	Unknown	Cytoplasm, nucleus, Nuclear envelope-ER	Unknown	No data	No data	Cytoplasm, nucleus, Nuclear envelope-ER	None
YNR046W	Unknown	Cytoplasm, nucleus, nucleolus	Unknown	No data	None	Cytoplasm, nucleus, nucleolus	None
YDR367W	Unknown	Golgi apparatus	Unknown	No data	None	Golgi apparatus	Membrane all-alpha superfamily
YHR083W	Unknown	Mitochondrion	Unknown	No data	None	Mitochondrion	None
YNL026W	Unknown	Mitochondrion	Unknown	No data	No data	Mitochondrion	None
YMR211W (DML1)	Unknown	Mitochondrion, cytoplasm	Structural molecule	No data	None	Mitochondrion, cytoplasm	Tubulin: stathmin-like domain complex
YDR196C	Unknown	Nuclear envelope-ER	Dephospho-CoA kinase	None	None	Nuclear envelope-ER	Dephospho-CoA kinase
YNL181W	Unknown	Nuclear envelope-ER	Oxidoreductase	None	No data	Nuclear envelope-ER	Tyrosine-dependent oxidoreductases
YNL158W	Unknown	Nuclear envelope	Unknown	None	None	Nuclear envelope	None
YLR440C	Unknown	Nuclear envelope-ER	Unknown	No data	No data	Nuclear envelope-ER	ARM repeat superfamily
YMR134W	Unknown	Nuclear envelope-ER	Unknown	No data	No data	Nuclear envelope-ER	None
YMR298W	Unknown	Nuclear envelope-ER	Unknown	No data	No data	Nuclear envelope-ER	None
YNL149C	Unknown	Nuclear envelope-ER	Unknown	No data	No data	Nuclear envelope-ER	Winged helix DNA-binding domain superfamily
YDR437W	Unknown	Nuclear envelope-ER	Unknown	No data	None	Nuclear envelope-ER	Bacterial enterotoxin and exotoxin superfamily
YDL193W	Unknown	Nuclear membrane, lipid particle	Prenyl-transferase	No data	No data	Nuclear membrane, lipid particle	Undecaprenyl diphosphate synthase
YOR004W	Unknown	Nucleolus	Unknown	None	No data	Nucleolus	None
YIL091C	Unknown	Nucleolus, nucleus	RNA helicase	None	None	Nucleolus, nucleus	DEAD box RNA helicase
YLR051C	Unknown	Nucleolus, nucleus	Unknown	No data	No positives	Nucleolus, nucleus	None
YGR251W	Unknown	Nucleolus, nucleus	Unknown	No data	No data	Nucleolus, nucleus	None
YHR040W	Unknown	Nucleus	Unknown	No data	None	Nucleus	None
YHR197W (RIX1)	Unknown	Nucleus[a]	Unknown	Unknown complex 2	Unknown complex 2*	No signal	None
YHR085W	Unknown	Nucleus	Unknown	Unknown complex 2	None	Nucleus	None

(*continued*)

Table 1. Continued

ORF	GO Terms Assignment			Experimental and Computational Evidence			
	Biological process	Component	Molecular function	Process predicted by copurification	Process predicted by two-hybrid	Localization	Remote homology search and Protein structure prediction
YNL182C	Unknown	Nucleus	Unknown	Unknown complex 2	Unknown complex 2*	Nucleus	Tup1, C-terminal domain; WD40 repeat superfamily
YLR008C	Unknown	Unknown	Chaperone	No data	None	No data	Chaperone J-domain
YNL152W	Unknown	Unknown	Phospholipid binding	No data	No data	No data	Synaptogamin C2 domain; NEDD4 WWIII domain
YLR243W	Unknown	Unknown	Signal peptide binding	No data	No data	No data	Signal peptide-binding domain superfamily; GTPase domain of the signal sequence recognition protein Ffh

The GO term finder was used to predict biological processes from the copurification or two-hybrid data sets. Localization was a direct assay that determined the component term. Protein structure prediction and remote homology searches were used to annotate the molecular function term. (PSI indicates a remote homology search; ER, endoplasmic reticulum network; *, overlap between the copurification and two-hybrid data sets).

[a] Component term was assigned from copurification or two-hybrid data.

[b] Process term was assigned from remote homology searches or protein structure prediction.

[c] Function term was assigned from two-hybrid.

was predicted to be a member of a G-patch domain protein family, which is involved in RNA binding. The third novel mRNA splicing protein, YLR132c, appears to be a bifunctional protein. It copurified with Prp19 and Snt309 and localized to the nucleus, indicating a role in mRNA splicing. It also copurified with Cor1 and localized in the mitochondrion, suggesting an additional role in aerobic respiration.

Discussion

Despite the completion of the *S. cerevisiae* genome sequence seven years ago, numerous genome-wide functional genomics analyses, and thousands of more focused studies, many ORFs remain uncharacterized in this organism. We demonstrate a targeted approach involving the integration of multiple protein-based technologies that are specifically relevant to describing a protein in the GO format. These technologies provided information for nearly all 100 uncharacterized and essential genes, allowing annotation of ~50% of this set in at least two of the three GO categories, resulting in a large reduction in the number of uncharacterized and essential genes in yeast. Our work provides a model for other studies, including those focused on more complex organisms, in which multiple data sets are synthesized into the coherent framework of GO terms.

Originally, GO terms were defined to provide a standardized vocabulary to permit software-driven comparisons between organisms. Here they united assignments made by technologies that may not otherwise share a common vocabulary. Our GO term assignments for biological process and molecular function provide a set of predictions ready to be tested by other researchers. To facilitate the transfer of pertinent information to the research community, we have provided a website that provides extensive supporting data and search features (http://www.yeastrc.org/unknown_orfs).

Comparison to previously published large-scale studies of protein function reveals several features of our methods. Our targeted applications of protein-based technologies have an advantage for elucidating function in that they are direct assays of the properties of a selected group of proteins. We detected 1246 interactions of the proteins encoded by the uncharacterized ORFs by mass spectrometry. The recently published Bayesian networks approach (Jansen et al., 2003) predicted 25 of these interactions. The large-scale protein purification and mass spectrometry analyses identified 79 (Gavin et al., 2002) and 10 (Ho et al., 2002) of the interactions. The low level of overlap is largely due to the low representation of the uncharacterized ORFs in the other data sets, suggesting the value of targeted characterizations. We, along with Gavin et al. (2002), further analyzed our respective mass spectrometric data sets to identify protein complexes. Twenty-six of our complexes show almost no overlap with the Gavin et al. complexes. However, three of our complexes agree well (YDL209C, YLR424w, and YKR079c). A comparison to previously published two-hybrid interactions reveals 10 out of the 271 interactions were identified in previous high-throughput studies (Ito et al., 2001; Uetz et al., 2000), indicating the lack of saturation of protein-protein interaction data even in a well-studied organism such

Table 2. Copurifying Sets of Proteins Identified by Mass Spectrometry as Described in Experimental Procedures Except as Noted

ORF that Defined the Complex	Biological Process	Copurifying Proteins
1. YFR003C	Cell cycle	Glc7, Sds22, *YFR003C
2. YIR010W (DSN1) and YPL233W (NSL1)	Chromosome segregation	Ame1, Chl4, Ctf3, Ctf19, *Dsn1, Mcm22, Mtw1, Nkp1, Nnf1, *Nsl1, Okp1
3. YKL088W[a]	Coenzyme A biosynthesis	Sis2, Vhs3, *YKL088W
4. YKR079C[b]	DNA or RNA catabolism	Nuc1, *YKR079C, YMR099C
5. YJL072C (PSF2)	DNA repair (Unknown complex 1)	Psf1, *Psf2, Psf3, Sld5
6. YDR288W	DNA repair	Mms21, Nse1, Qri2, Rhc18, *Smc5, *YDR288W
7. YML023C	DNA repair	Kre29, Mms21, Qri2, Rhc18, *Smc5, *YML023C
8. YGR002C	Histone acetylation	Arp4, Epl1, Esa1, Rvb1, Rvb2, Tra1, Vid21, Yaf9, Yng2, YDR334W, YEL018W, *YGR002C
9. YGR198W	MAPKKK cascade	Stt4, *YGR198W
10. YPL063W (TIM50)[a]	Mitochondrial translocation	Tim50, Tom40
11. YGR046W[a]	Mitochondrial translocation	Hsp60, *YGR046W
12. YDL209C (CWC2)	mRNA splicing	Brr2, Cdc40, Cef1, Clf1, *Cwc2, Cwc22, Cwc23, Ecm2, Isy1, Lea1, Prp8, Prp19, Prp43, Prp45, Prp46, Smb1, Smd1, Smd3, Snt309, Snu114, Syf1, Syf2, Yju2, YKR022C, *YLR424W
13. YLR424W and YKR022C	mRNA splicing	Brr2, Cdc40, Cef1, Clf1, *Cwc2, Cwc23, Ecm2, Prp8, Prp19, Prp43, Prp45, Prp46, Smb1, Smd3, Snt309, Smx3, Snu114, Syf1, YKR022C, *YLR424W
14. YLR132C[b]	mRNA splicing	Cor1, Prp19, *YLR132C
15. YNL245C[a] (CWC25)	mRNA splicing	Clf1, *Cwc25, Prp8, Prp19, Snu114
16. YNL313c	Nuclear membrane fusion	Grs1, Kar2, Tub3, *YNL313C
17. YLL034C	Organelle organization and biogenesis[c]	Amn1, Axl2, Imh1, Pex19, Pfk27, Rgt2, Ric1, Rpa190, Sst2, Trx1, Trx2, *YLL034C, YLR035C-A, YLR084C
18. YJR136C	Protein biosynthesis	Rsm23, *YJR136C
19. YPR169W	Protein monoubiquitination	Bre1, *YPR169W
20. YKL195W	Protein targeting	Adh1, Kap123, Tom40, *YKL195W
21. YKR038C	Response to dessication	Gon7, *YKR038C
22. YNL124W (NAF1)	rRNA processing	Cbf5, *Naf1
23. YJL010C	rRNA processing	Gcd6, Nsr1, Snu13, *YJL010C
24. YDR365C and YGR145W	rRNA processing	Bfr2, Hca4, Lcp5, Nop58, Utp9, *YDR365C, YGR145W
25. YJL091C (GWT1)	Secretory pathway	Ded81, *Gwt1, Mrc1, Sec7, Sec63, YJR100C
26. YJR012C	Transport	Hol1, Mmp1, Pex7, Plb1, *YJR012C
27. POP1[d]	tRNA processing	Rpp1, Pop1, Pop3, Pop4, Pop5, Pop6, Pop8, Snm1, *YLR145W
28. YHR085W, YHR197W (RIX1), and YNL182C (IPI3)	Unknown complex 2	Ipi3, Rix1, *YHR085W
29. YNL260C	Unknown complex 3	Yae1, *YNL260C

Proteins used as a TAP-tagged bait are denoted by an asterisk.

[a] For these ORFs, the copurifying set of proteins were the proteins with high coverage in the mass spectrometric analysis.

[b] The localization and mass spectrometry data suggest that the asterisked protein forms two complexes, one in the mitochondrion and one in the nucleus (see text).

[c] The GO annotation was chosen as the significant annotation involving more than two of the ORFs.

[d] The POP1 complex was purified using TAP tagged-YLR145W, an uncharacterized essential protein. Although used as the bait, YLR145W was detected in only 1 of 3 mass spectrometric analyses. Pop3, Pop6, and Pop8 were also only detected in one of the three analyses.

as yeast. On the other hand, cellular localization of GFP-tagged proteins by fluorescence microscopy is remarkably reproducible. Of the 58 proteins that were localized in both this study and in Huh et al. (2003), 60% of the localization assignments were in exact agreement, and greater than 90% were in partial agreement. In only two cases were our images significantly different (YMR298w, YOR004w). The greater clarity in our images allowed us to observe finer details such as distinguishing kinetochores from spindle pole bodies and detecting lipid particles in our DIC images. A likely explanation for the difference in image quality is that in the large-scale study of Huh et al., greater than 4000 proteins were localized, required mounting cells in glass bottom 96-well plates, whereas we mounted cells under more ideal optical conditions.

Membrane proteins present a particular challenge to biochemical analyses. Localization appears to be particularly suited for annotating these proteins although the tag could interfere with targeting sequences. Eleven proteins localize to the endoplasmic reticulum and only one (YJL091C) could be purified and analyzed by mass spectrometry. For three of the ER proteins, the TAP tag was toxic or lethal even though the corresponding and otherwise isogenic YFP tagged strain was healthy (see Status List, http://www.yeastrc.org/unknown_orfs). Two of the proteins from the ER (YJR013W and YJL097W) interacted with numerous other membrane proteins by two-hybrid, and these interactions provided a biological process annotation.

The modification of our current technologies or addition of alternate technologies could enhance the predictions for some classes of proteins such as membrane proteins. Furthermore, the integration of additional approaches including synthetic genetic interactions (Tong et al., 2002, 2001), induced proteolysis (Kanemaki et al.,

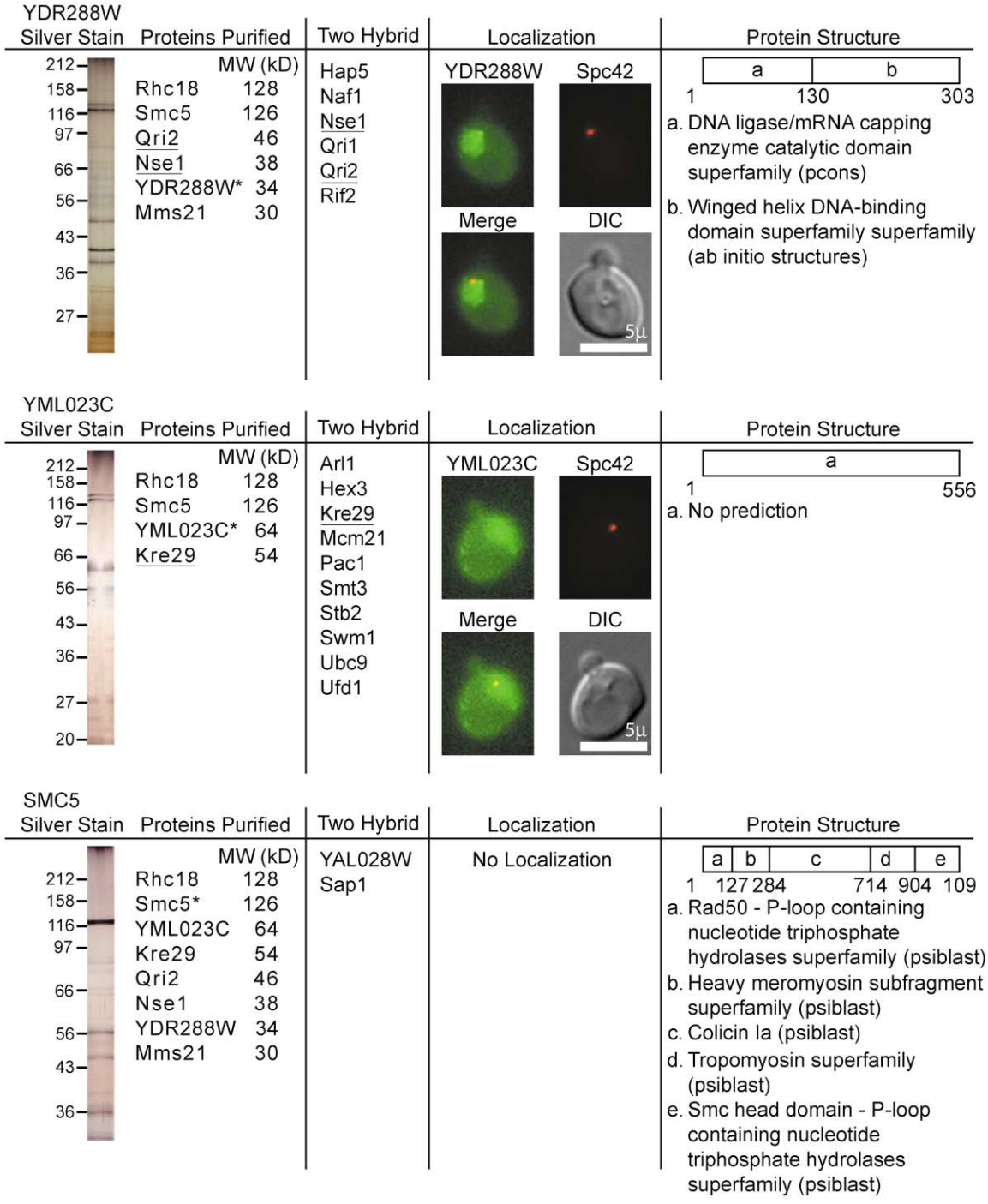

Figure 1. Parallel Analysis of Three Proteins Involved in Two Related DNA Repair Complexes

Four technologies were applied to YDR288w, YML023c, and SMC5 and the results displayed from left to right are purification and mass spectrometry, two-hybrid analysis, localization, and protein structure prediction. The TAP-tagged protein in each purification is asterisked. The eluate from the purification was subjected to SDS-PAGE and the proteins visualized by silver staining. Mass spectrometry analysis of the eluate identified copurifying proteins that are listed adjacent to the gel with their respective molecular weights. Proteins identified by two-hybrid analysis are listed alphabetically. The proteins identified by both mass spectrometry and two-hybrid analyses are underlined. Each ORF was tagged with Venus and the fusion protein was localized by fluorescence microscopy as described in the supplemental data. Spc42 fused to CFP was used as a marker for the nucleus and spindle pole body. For protein structure prediction, the protein sequence was computationally parsed into domains, and the structure of each domain was predicted using a sequential hierarchy of methods as described in the supplemental data.

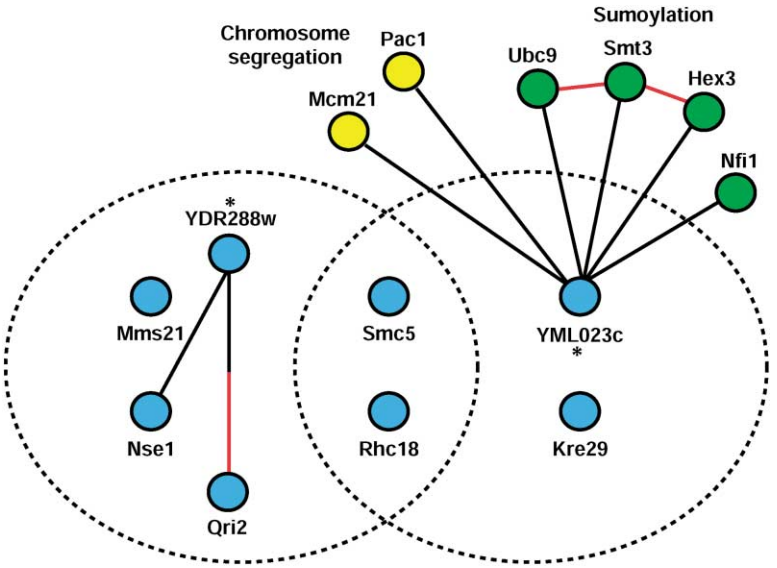

Figure 2. Schematic of DNA Repair Complexes and Their Interaction Networks Identified by Copurification and Two-Hybrid Analysis

DNA repair complexes that were identified by copurification of TAP-tagged versions of the uncharacterized ORFs YDR288w and YML023c (asterisk denotes TAP-tagged proteins). Proteins copurifying with each TAP-tagged uncharacterized ORF are blue and are encircled with a dashed black line. TAP-tagged Smc5 purified all members of both complexes but is not depicted here. All the members of the YDR288w and YML023c complexes are essential. Two-hybrid interactions identified in this report are represented as black lines and previous two-hybrid interactions as red lines. Proteins identified by two-hybrid interactions that have a role in sumoylation are represented in green and those involved in chromosome segregation are in yellow. A total of 11 other interactions identified by two-hybrid analysis are not represented in this diagram. Asterisk denotes protein used as two-hybrid bait.

2003), conditional expression of essential genes (Peng et al., 2003), and correlated mRNA expression (Hughes et al., 2000) should enable a greater success rate or even more robust predictions for all classes of proteins. The targeted application of multiple orthogonal approaches should propel the systematic analysis of other complements of uncharacterized proteins.

Experimental Procedures

Selection Criteria for Essential Uncharacterized ORFs

We used the following criteria to define our list of 100 uncharacterized ORFs based on information from the *Saccharomyces* Genome Database. (1) The deletion of the gene was lethal. (2) The gene was annotated as biological process unknown. (3) The gene did not have a name. Information about several of these genes were published during our research but we did not use this information in our analysis: PSF1, PSF2, PSF3, and SLD5 (Takayama et al., 2003); GWT1 (Tsukahara et al., 2003); NAF1 (Fatica et al., 2002); YIL083c, YKL088w, and YDR196c (Daugherty et al., 2002); CWC2 and CWC25 (Ohi and Gould, 2002); DML1 (Gurvitz et al., 2002); RIO2 (Vanrobays et al., 2003); NSL1 and DSN1 (Euskirchen, 2002; Nekrasov et al., 2003), TIM50 (Geissler et al., 2002; Yamamoto et al., 2002).

The yeast genome was recently revised and SGD has labeled 9 of the 100 ORFs as dubious based on comparative genomics data (Cliften et al., 2003; Kellis et al., 2003) and these are noted on our website. We were able to tag only one of these ORFs, YJR012c, which we detected by Western blot analysis. Therefore, we propose that YJR012c encodes a protein. For two of the dubious ORFs, YDR196w and YDR413c, we detected significant homology with membrane proteins but we were unable to tag or characterize them. Perhaps these ORFs are part of a nearby ORF. For the other 6 dubious ORFs, our inability to tag or characterize them is consistent with the idea that they do not encode proteins.

Localization

Strain BSY9 has the genotype: MATa/MATα *ade2-1^{oc}/ade2-1^{oc} ADE3/ade3Δ can1-100/can1-100 CYH2^{s}/cyh2^{r} his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 trp1-1/trp1-1 ura3-1/ura3-1*. Strain BSY110

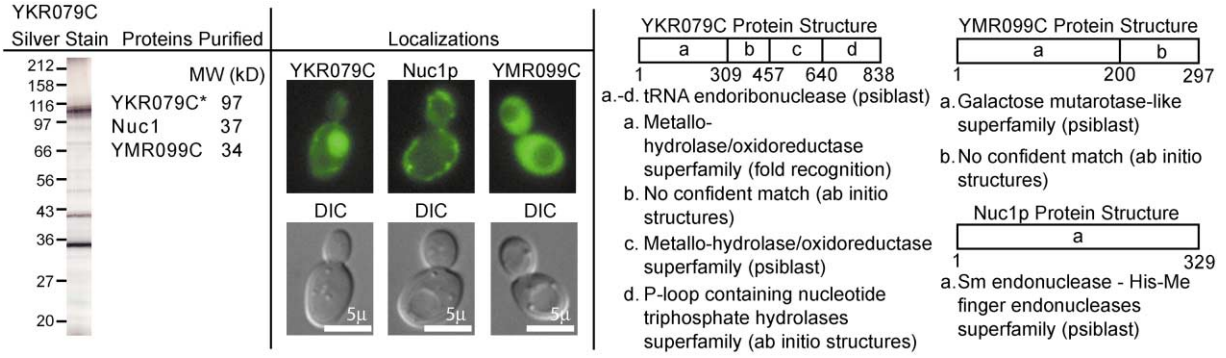

Figure 3. YKR079c Copurifies and Colocalizes with Nuc1 and YMR099c

Left panel, purification of TAP-tagged YKR079c (asterisked) and subsequent identification of copurifying proteins. The silver-stained gel of the eluate from the copurification displays three dominant bands and mass spectrometry identifies YKR079c, Nuc1, and YMR099c. Central panel, localization of YKR079c and Nuc1 fused to Venus and YMR099c fused to YFP. YKR079c localizes in the mitochondria and nucleus. Nuc1 localizes in the mitochondria, and YMR099c localizes in the nucleus and cytoplasm. These localizations suggest that YKR079c forms a complex in the mitochondria with Nuc1 and forms a complex in the nucleus with YMR099c. Right panel, the protein structure predictions for domains identified in each protein as described in the legend for Figure 1. The PSI-BLAST matches with the full-length ORF are also indicated.

has the same genotype as BSY9 except it also has *CFP-SPC42/SPC42*. Strain BSY110 was constructed by integrating a CFP tag on the 5′ end of *SPC42* as described using plasmid pBS5 as the template (Prein et al., 2000). Plasmid pBS5 was made by changing the GFP in plasmid pyGFP (Prein et al., 2000) to CFP by Quikchange site-directed mutagenesis (Stratagene). Plasmid pBS7 was made in two steps. First, the YFP in pDH6 was converted to citrine to make plasmid pDH27. Then the citrine in pDH27 was converted to Venus (Nagai et al., 2002) by site-directed mutagenesis. Note that Venus in pBS7 has two additional mutations Q69M and Q80R not found in the original Venus.

Each uncharacterized ORF was tagged at the 3′ end with the Venus version of YFP (Nagai et al., 2002) as described (Wach et al., 1997) using plasmid pBS7 as the template. The Venus tag was integrated in a diploid strain heterozygous for *CFP-SPC42* (strain BSY110) to provide a marker for the SPB and the nucleus. The N-terminal CFP tag is adjusted for yeast preferred codons and contains little sequence homology with the C-terminal Venus cassette in plasmid pBS7. Thus, homologous recombination strongly favors integration at the 3′ end of the uncharacterized ORF. The diploid was then subjected to random spore analysis. Haploids were selected by resistance to cycloheximide and tested for resistance to G418, which marks the tagged gene. If the tagged copy of the gene did not appear in half of the progeny, tetrads were dissected to determine if the tag was lethal or toxic (deleterious) to the strain. Viable haploids were analyzed by PCR to demonstrate that the only copy of the gene was tagged with Venus. Haploids containing both the Venus-tagged ORF and CFP-Spc42 were analyzed by fluorescence microscopy on the DELTAVISION system, which incorporates an Olympus IL-70 microscope, a CoolSnap HQ digital camera from Roper Scientific (Tucson, AZ), and optical filter sets from Omega Optical (Brattleboro, VT). If the signal was very low, the tagged strain was mixed with an untagged strain, so that the experimental and a control strain could be imaged on the same slide. If the tagged and untagged strains could not be distinguished in the YFP channel, then the tagged ORF was labeled as having no signal. The tagged and untagged strains could be distinguished in the CFP channel by the presence of CFP-Spc42 in one but not the other.

Copurification and Mass Spectrometry

The purification protocol of Rigaut et al. (1999) was optimized. The stringency of the washes of the first affinity purification was increased from 150 mM NaCl to 300 mM NaCl and the washes of the final affinity purification were decreased from 200 to 20 column volumes. With these modifications, the Tub4p complex, used as a test, purified to near homogeneity with stoichiometric amounts of each of the three components and minor contamination from ribosomal proteins. (The detailed optimized protocol can be found at http://depts.washington.edu/yeastrc/ms_tap1.htm.)

The TAP tag was integrated into diploid strain BSY9 using plasmid pFA6a-CTAP-MX6-2XPA (Tasto et al., 2001) as the template. Haploids were isolated as above and tested by PCR to ensure that the TAP-tagged copy of the gene was the only copy of the gene. Proteins that were successfully tagged were subjected to the optimized purification protocol. The purified eluates were analyzed by SDS-polyacrylamide gel electrophoresis and proteins detected by silver staining according to the directions of the manufacturer (Bio-Rad). Mass spectrometry was performed to identify the copurifying proteins using MudPIT analysis as described previously (McDonald et al., 2002). The silver-stained gels and the detailed results from the mass spectrometry analysis can be viewed and downloaded at our website (http://yeastrc.org/unknown_orfs).

The copurifying proteins that specifically associate with a given essential uncharacterized protein were determined in two steps (Table 2). First, proteins that occurred in nine or more purifications and ribosomal proteins were excluded from consideration. Second, we ranked the relative statistical significance of the presence of each of the remaining proteins that copurified with the uncharacterized protein. A probability model was derived based on a hypergeometric distribution. The formula applied to each copurifying protein was:

$$P(I) = \frac{\binom{A}{I} * \binom{T-A}{B-I}}{\binom{T}{B}}$$

where, A is the number of mass spectrometry runs containing the uncharacterized protein. A includes all instances where the protein appeared, including when it was not the targeted purified protein. B is the number of runs containing the copurifying protein. T is the total number of mass spectrometry runs in our dataset ($T = 83$). I is the number of runs containing both proteins. P(I) is the probability of I runs containing both proteins by random chance, given only the number of runs containing protein A, the number of runs containing protein B and the total number of runs (T).

A P score was then assigned to the copurifying protein:

$$\text{P score} = \sum_{i=I}^{\min(A,B)} P(i),$$

where min*(A,B)* is the minimum of the two values A and B.

The P score represents the likelihood that the two proteins, the uncharacterized protein and the copurifying protein, would appear together by random chance I or more times. We established our significance threshold empirically such that if the uncharacterized protein appeared only once, proteins that only appeared in that run were considered significant. Given our total number of runs, the exact cutoff for a significant P score was 0.01205.

Two-Hybrid

Genome-wide two-hybrid screens were performed in a high-throughput manner using robotics as described previously (Drees et al., 2001; Uetz et al., 2000). In brief, the essential uncharacterized ORFs were fused to the Gal4 DNA binding domain and screened in duplicate against an array of ~6000 yeast strains containing each of the ~6000 *S. cerevisiae* ORFs expressed as fusions to the Gal4 activation domain. The array was generated by recombination cloning into the activation domain vector pOAD as previously described, except that instead of selecting two colonies from each ORF transformation plate we pooled all the colonies from the transformation plates that were 3 times higher than the vector only control. The essential uncharacterized ORFs were cloned by recombination into the DNA binding domain vector pOBD2 and individual clones were sequenced through the whole ORF. Putative interacting partners were identified as reproducible two-hybrid positives that were observed twice out of the duplicate high-throughput screens. Positives that were identified only once were presumably a result of a false positive colony that is not reproducible or due to inefficient pinning by the robot. In some cases, a confirmation screen of these positives was performed by re-arraying the activation domain strains corresponding to the positives that appeared singly or doubly from the duplicate high-throughput screens into 96-well microtiter plates. Subsequent screening of these re-arrayed strains by the DNA binding domain hybrid enabled the identification of single positives that were the result of inefficient pinning and allowed them to be classified as double positives. The results of the two-hybrid analysis can be viewed and downloaded at our website (http://yeastrc.org/unknown_orfs).

Protein Structure Prediction and Sequence Homology Detection

Domain parsing and structure prediction were performed as follows. An iterative procedure called Ginzu (Chivian et al., 2003) was used to parse each sequence into domains and to predict the structure of each domain. The basic concept behind Ginzu is to start with a sequence search using reliable database search methods, mask out any matched portions of the sequence which are taken to be independent domains, and subject the unmatched regions to searches using less reliable but more sensitive methods. First, PSI-BLAST (Altschul et al., 1997) searches for homologous sequences in the nonredundant NCBI database were used to generate multiple sequence alignments for each ORF (5 iterations and an e-value cutoff of 0.001). Scoring matrices (PSSMs) generated from the multiple sequence alignments were then used to search the PDB database for sequences with known structures. The homologous regions of the query sequence were annotated with the Protein Data Bank accession number (PDB id) (Berman et al., 2002) from the match and the homologous regions were masked. Unmasked regions were submitted first to ORFeus (Ginalski et al., 2003), a fold-recognition server, and then Pcons2 (Lundstrom et al., 2001), a consensus fold recognition server. Significant matches were again masked and annotated with the PDB id. Still unmatched regions were then searched

against the PFAM database with Hmmer (Bateman et al., 2002), and domains matching a protein family were annotated.

After this procedure, ROSETTA de novo structure prediction (Bonneau et al., 2001; Simons et al., 1997, 1999) was carried out for remaining domains shorter than 200 amino acids with no structure annotation. The ensemble of protein structures produced by ROSETTA for each sequence was clustered to identify broad energy minimum, and one representative (the cluster center) was selected from each of the 20 largest clusters. The 20 cluster centers were then compared to a comprehensive set of known structures using the Mammoth structure-structure comparison method (Ortiz et al., 2002). For each significant structure-structure match, an alignment was forced between sequence profiles generated using PSI-BLAST for the sequence of the predicted structure and the sequence of the PDB match using the MVP protocol (P.B., unpublished data) to probe for residual sequence similarity. The score for the predicted structure-PDB structure match was based on the similarity of the structures and the agreement between the forced sequence alignment and the Mammoth structure alignment. Additionally, matches of ROSETTA predictions to PDB structures were highlighted if the previous PSIBLAST, ORFEUS, or PCONS searches using the sequence of the protein had identified low to moderate confidence matches to the same SCOP superfamily (Murzin et al., 1995).

We also sought information based purely on similarities in primary sequence with other proteins. The scoring matrices (PSSMs) generated in the PSI-BLAST searches described above were used to search the NCBI nonredundant database with an e-value cutoff of 0.01 and no sequence number cutoff. This final less-restrictive search allows greater sensitivity with less risk of contamination of the scoring matrices with information from unrelated sequences.

GO Term Assignment

To enable the objective assignment of GO terms, we used the GO Term Finder on SGD to categorize our data (http://db.yeastgenome.org/cgi-bin/SGD/GO/goTermFinder). The process term that had the highest p value was used to describe the ORF unless the highest p value was >0.01; then the ORF was not annotated. In some cases, a process term could not be derived because proteins associating with the uncharacterized protein were not annotated. Annotation by physical association is equivalent to the GO evidence code of IPI (inferred by physical association). The GO definition of the IPI code allows the assignment of the two other categories, cellular component and molecular function. The few times we relied on this method are noted in Table 1 of the manuscript.

The observed localization of an ORF fused to a fluorescent protein was used to determine the component term and is equivalent to the GO code of IDA (inferred by direct assay). Assignment of the molecular function derived from the remote homology detection or protein structure prediction is equivalent to the GO code of ISS (inferred by sequence or structural similarity). We assigned molecular function terms if the domain or superfamily identified was associated with a GO molecular function. There were several examples of notable structural homologies and PSI-BLAST similarities that were not assigned a molecular function term because they were partial matches to a domain or they did not contain the key elements of a domain. We also did not annotate domains that are associated with diverse cellular processes and are mainly involved in protein interactions.

Acknowledgments

We thank M. Johnston (Washington University, St. Louis) and C. Boone (University of Toronto) for critical reading of the manuscript; R. Aebersold (Institute for Systems Biology) for insightful comments on the project; and A. Zelter (University of Washington) for comments on the website. This work was a collaborative effort of the Yeast Resource Center and was funded by the National Center for Research Resources of the National Institutes of Health (PHS # P41 RR11823). S.F. and D.B. are investigators of the Howard Hughes Medical Institute.

Received: September 15, 2003
Revised: November 10, 2003
Accepted: November 12, 2003
Published: December 12, 2003

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic Acids Res. *30*, 276–280.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. (2002). The Protein Data Bank. Acta Crystallogr. D Biol. Crystallogr. *58*, 899–907.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E., and Baker, D. (2001). Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins (*Suppl 5*), 119–126.

Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E., Bonneau, R., Rohl, C.A., and Baker, D. (2003). Automated prediction of CASP-5 structures using the Robetta server. Proteins *53* (*Suppl 6*), 524–533.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science *301*, 71–76.

Dake, E., Hofmann, T.J., McIntire, S., Hudson, A., and Zassenhaus, H.P. (1988). Purification and properties of the major nuclease from mitochondria of *Saccharomyces cerevisiae*. J. Biol. Chem. *263*, 7691–7702.

Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crecy-Lagard, V., and Osterman, A. (2002). Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. J. Biol. Chem. *277*, 21431–21439.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680–686.

Drees, B.L., Sundin, B., Brazeau, E., Caviston, J.P., Chen, G.C., Guo, W., Kozminski, K.G., Lau, M.W., Moskow, J.J., Tong, A., et al. (2001). A protein interaction map for cell polarity development. J. Cell Biol. *154*, 549–571.

Euskirchen, G.M. (2002). Nnf1p, Dsn1p, Mtw1p, and Nsl1p: a new group of proteins important for chromosome segregation in *Saccharomyces cerevisiae*. Eukaryot. Cell *1*, 229–240.

Fatica, A., Dlakic, M., and Tollervey, D. (2002). Naf1 p is a box H/ACA snoRNP assembly factor. Rna *8*, 1502–1514.

Fujioka, Y., Kimata, Y., Nomaguchi, K., Watanabe, K., and Kohno, K. (2002). Identification of a novel non-structural maintenance of chromosomes (SMC) component of the SMC5–SMC6 complex involved in DNA repair. J. Biol. Chem. *277*, 21585–21591.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Geissler, A., Chacinska, A., Truscott, K.N., Wiedemann, N., Brandner, K., Sickmann, A., Meyer, H.E., Meisinger, C., Pfanner, N., and Rehling, P. (2002). The mitochondrial presequence translocase: an essential role of Tim50 in directing preproteins to the import channel. Cell *111*, 507–518.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. Nature *418*, 387–391.

Ginalski, K., Pas, J., Wyrwicz, L.S., von Grotthuss, M., Bujnicki, J.M., and Rychlewski, L. (2003). ORFeus: detection of distant homology

using sequence profiles and predicted secondary structure. Nucleic Acids Res. *31*, 3804–3807.

Gurvitz, A., Hartig, A., Ruis, H., Hamilton, B., and de Couet, H.G. (2002). Preliminary characterisation of DML1, an essential *Saccharomyces cerevisiae* gene related to misato of Drosophila melanogaster. FEM. Yeast Res. *2*, 123–135.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature *415*, 180–183.

Hoege, C., Pfander, B., Moldovan, G.L., Pyrowolakis, G., and Jentsch, S. (2002). RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. Nature *419*, 135–141.

Horak, C.E., and Snyder, M. (2002). Global analysis of gene expression in yeast. Funct. Integr. Genomics *2*, 171–180.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109–126.

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.

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. USA *98*, 4569–4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449–453.

Kanemaki, M., Sanchez-Diaz, A., Gambus, A., and Labib, K. (2003). Functional proteomic identification of DNA replication proteins by induced proteolysis in vivo. Nature *423*, 720–725.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., et al. (2002). Subcellular localization of the yeast proteome. Genes Dev. *16*, 707–719.

Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci. *10*, 2354–2362.

McDonald, W.H., Ohi, R., Miyamoto, D.T., Mitchison, T.J., Yates, I., and John, R. (2002). Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. Int. J. Mass Spectrom. *219*, 245–251.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nat. Biotechnol. *20*, 87–90.

Nekrasov, V.S., Smith, M.A., Peak-Chew, S., and Kilmartin, J.V. (2003). Interactions between centromere complexes in *Saccharomyces cerevisiae*. Mol. Biol. Cell. *14*, 4931–4946.

Ohi, M.D., and Gould, K.L. (2002). Characterization of interactions among the Cef1p-Prp19p-associated splicing complex. RNA *8*, 798–815.

Ortiz, A.R., Strauss, C.E., and Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. *11*, 2606–2621.

Peng, W.T., Robinson, M.D., Mnaimneh, S., Krogan, N.J., Cagney, G., Morris, Q., Davierwala, A.P., Grigull, J., Yang, X., Zhang, W., et al. (2003). A panoramic view of yeast noncoding RNA processing. Cell *113*, 919–933.

Prakash, S., and Prakash, L. (1977). Increased spontaneous mitotic segregation in MMS-sensitive mutants of *Saccharomyces cerevisiae*. Genetics *87*, 229–236.

Prein, B., Natter, K., and Kohlwein, S.D. (2000). A novel strategy for constructing N-terminal chromosomal fusions to green fluorescent protein in the yeast *Saccharomyces cerevisiae*. FEBS Lett. *485*, 29–34.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030–1032.

Ross-Macdonald, P., Coelho, P.S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L., et al. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature *402*, 413–418.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins *34*, 82–95.

Takaku, H., Minagawa, A., Takagi, M., and Nashimoto, M. (2003). A candidate prostate cancer susceptibility gene encodes tRNA 3′ processing endoribonuclease. Nucleic Acids Res. *31*, 2272–2278.

Takayama, Y., Kamimura, Y., Okawa, M., Muramatsu, S., Sugino, A., and Araki, H. (2003). GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. Genes Dev. *17*, 1153–1165.

Tasto, J.J., Carnahan, R.H., McDonald, W.H., and Gould, K.L. (2001). Vectors and gene targeting modules for tandem affinity purification in *Schizosaccharomyces pombe*. Yeast *18*, 657–662.

Todd, A.E., Orengo, C.A., and Thornton, J.M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. *307*, 1113–1143.

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science *294*, 2364–2368.

Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science *295*, 321–324.

Tsukahara, K., Hata, K., Nakamoto, K., Sagane, K., Watanabe, N.A., Kuromitsu, J., Kai, J., Tsuchiya, M., Ohba, F., Jigami, Y., et al. (2003). Medicinal genetics approach towards identifying the molecular target of a novel inhibitor of fungal cell wall assembly. Mol. Microbiol. *48*, 1029–1042.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature *403*, 623–627.

Vanrobays, E., Gelugne, J.P., Gleizes, P.E., and Caizergues-Ferrer, M. (2003). Late cytoplasmic maturation of the small ribosomal subunit requires RIO proteins in *Saccharomyces cerevisiae*. Mol. Cell. Biol. *23*, 2083–2095.

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.

Wach, A., Brachat, A., Alberti-Segui, C., Rebischung, C., and Philippsen, P. (1997). Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in *Saccharomyces cerevisiae*. Yeast *13*, 1065–1075.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901–906.

Yamamoto, H., Esaki, M., Kanamori, T., Tamura, Y., Nishikawa, S., and Endo, T. (2002). Tim50 is a subunit of the TIM23 complex that links protein translocation across the outer and inner mitochondrial membranes. Cell *111*, 519–528.