Subcellular localization of the yeast proteome

Anuj Kumar,[1] Seema Agarwal,[1] John A. Heyman,[3,5] Sandra Matson,[1] Matthew Heidtman,[1] Stacy Piccirillo,[1] Lara Umansky,[1] Amar Drawid,[2] Ronald Jansen,[2] Yang Liu,[2] Kei-Hoi Cheung,[4] Perry Miller,[4] Mark Gerstein,[2] G. Shirleen Roeder,[1] and Michael Snyder[1,2,6]

[1]Department of Molecular, Cellular, and Developmental Biology, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA; [3]Invitrogen Corporation, Carlsbad, California 92008, USA; [4]Center for Medical Informatics, Department of Anesthesiology, Yale University School of Medicine, New Haven, Connecticut 06510, USA

Protein localization data are a valuable information resource helpful in elucidating eukaryotic protein function. Here, we report the first proteome-scale analysis of protein localization within any eukaryote. Using directed topoisomerase I-mediated cloning strategies and genome-wide transposon mutagenesis, we have epitope-tagged 60% of the *Saccharomyces cerevisiae* proteome. By high-throughput immunolocalization of tagged gene products, we have determined the subcellular localization of 2744 yeast proteins. Extrapolating these data through a computational algorithm employing Bayesian formalism, we define the yeast localizome (the subcellular distribution of all 6100 yeast proteins). We estimate the yeast proteome to encompass ~5100 soluble proteins and >1000 transmembrane proteins. Our results indicate that 47% of yeast proteins are cytoplasmic, 13% mitochondrial, 13% exocytic (including proteins of the endoplasmic reticulum and secretory vesicles), and 27% nuclear/nucleolar. A subset of nuclear proteins was further analyzed by immunolocalization using surface-spread preparations of meiotic chromosomes. Of these proteins, 38% were found associated with chromosomal DNA. As determined from phenotypic analyses of nuclear proteins, 34% are essential for spore viability—a percentage nearly twice as great as that observed for the proteome as a whole. In total, this study presents experimentally derived localization data for 955 proteins of previously unknown function: nearly half of all functionally uncharacterized proteins in yeast. To facilitate access to these data, we provide a searchable database featuring 2900 fluorescent micrographs at http://ygac.med.yale.edu.

[*Key Words*: Protein localization; *S. cerevisiae*; proteomics; machine learning; epitope-tagging]

Received December 18, 2001; revised version accepted February 1, 2002.

A global understanding of the molecular mechanisms underpinning cell biology necessitates an understanding not only of an organism's genome but also of the protein complement encoded within this genome (the proteome). In the past, data regarding an organism's proteome have typically been accumulated piecemeal through studies of a single protein or cell pathway. Genomic methodologies have altered this paradigm: a variety of approaches are now in place by which proteins may be directly analyzed on a proteome-wide scale. Chromatography-coupled mass spectrometry (Gygi et al. 1999; Washburn et al. 2001), large-scale two-hybrid screens (Uetz et al. 2000; Ito et al. 2001; Tong et al. 2002), immunoprecipitation/mass spectrometric analysis of protein complexes (Gavin et al. 2002; Ho et al. 2002), and protein microarray technologies (MacBeath and Schreiber 2000; Zhu et al. 2000, 2001) are yielding unprecedented quantities of protein data. Recent genomic techniques combining microarray technologies with either chromatin immunoprecipitation (Ren et al. 2000; Iyer et al. 2001) or targeted DNA methylation (van Steensel et al. 2001) have been used to globally map binding sites of chromosomal proteins in vivo. Initiatives are even underway to automate and industrialize processes by which protein structures may be solved, potentially providing a library of structural data from which homologous proteins may be modeled (Burley 2000; Montelione 2001).

Although these approaches promise a wealth of information, many fundamental proteomic data sets remain uncataloged. Notably, the subcellular distribution of proteins within any single eukaryotic proteome has never been extensively examined, despite the usefulness and importance of these data. Protein localization is assumed to be a strong indicator of gene function. Localization data are also useful as a means of evaluating pro-

[5]Present address: Active Motif, 104 Avenue Franklin Roosevelt, Box-25, B-1330 Rixensart, Belgium.
[6]Corresponding author.
E-MAIL michael.snyder@yale.edu; FAX (203) 432-6161.
Article and publication are at http://www.genesdev.org/cgi/doi/10.1101/gad.970902.

tein information inferred from genetic data (e.g., supporting or refuting putative protein interactions suggested from two-hybrid analysis; Ito et al. 2001). Furthermore, the subcellular localization of a protein can often reveal its mechanism of action.

To determine the subcellular localization of a protein, its corresponding gene is typically either fused to a reporter or tagged with an epitope. Reporters and epitope tags are fused routinely to either the N or C termini of target genes, a choice that can be critical in obtaining accurate localization data. Organelle-specific targeting signals (e.g., mitochondrial targeting peptides and nuclear localization signals) are often located at the N terminus (Silver 1991); N-terminal reporter fusions may disrupt these sequences, resulting in anomalous protein localizations. In other cases, C-terminal sequences may be important for proper function and regulation, as recently shown from analysis of the yeast γ-tubulin-like protein Tub4p (Vogel et al. 2001). Gene copy number can also have an impact on the accuracy with which a protein is localized; overexpressed protein products may saturate intracellular transport mechanisms, potentially producing an aberrant subcellular protein distribution. In other cases, weakly expressed single-copy genes may not yield sufficient protein to be visualized, particularly by fluorescence microscopy. The effects of copy number and reporter/tag orientation on protein localization, however, have never been studied in a large data set.

To date, few studies have characterized protein localization on a large scale, primarily because few high-throughput methods exist by which reporter fusions or epitope-tagged proteins can be generated and subsequently localized. Typically, systematic approaches have been used to construct a limited number of chimeric reporter fusions applicable to pilot localization studies. For example, >100 human cDNAs have been cloned as N- and C-terminal gene fusions to spectral variants of green fluorescent protein (GFP) as a means of examining the subcellular localization of these proteins in living cells (Simpson et al. 2000). Thus far, the majority of localization studies have been undertaken in yeast, owing primarily to the fidelity of homologous recombination in *Saccharomyces cerevisiae* and the concomitant ease with which integrated reporter gene fusions can be generated. As part of a pilot study in *S. cerevisiae*, Niedenthal et al. (1996) constructed GFP reporter fusions to three unknown open reading frames (ORFs) from yeast Chromosome XIV and subsequently localized these chimeric GFP-fusion proteins by fluorescence microscopy.

In addition to directed cloning methods, strains suitable for localization analysis may be generated through random approaches. Recently, a plasmid-based GFP-fusion library of *Schizosaccharomyces pombe* DNA was constructed by fusing random fragments of genomic DNA upstream of GFP-coding sequence. Fission yeast cells transformed with this library were subsequently screened for GFP fluorescence, and 250 independent gene products were localized (Ding et al. 2000). In *S. cerevisiae*, transposon-based methods have been used to generate random *lacZ* gene fusions (Burns et al. 1994) and epitope-tagged alleles (Ross-MacDonald et al. 1999) for subsequent immunolocalization. Although these transposon-based studies have resulted in the localization of ~300 yeast proteins, the majority of the *S. cerevisiae* proteome has remained uncharacterized in regards to its subcellular distribution.

To address this deficiency, we have undertaken the largest analysis to date of protein localization in yeast. Employing high-throughput methods of epitope-tagging and immunofluorescence analysis, our study defines the subcellular localization of 2744 proteins. By integrating these localization data with those previously published, we identify the subcellular localization of >3300 yeast proteins, 55% of the proteome. Building on these data, we have applied a Bayesian system to estimate the intracellular distribution of all 6100 yeast proteins and have further characterized a subset of nuclear proteins both by immunolocalization on surface spread chromosomal preparations and by phenotypic analysis. In total, our findings provide a wealth of insight into protein function, while formally corroborating an expected link between protein function and localization. Furthermore, this study provides experimentally derived localization data for nearly 1000 proteins of previously unknown function, thereby providing, at minimum, a starting point for informed analysis of this previously uncharacterized segment of the proteome.

Results

Genome-wide epitope-tagging and large-scale immunolocalization

Yeast proteins immunolocalized in this study were epitope-tagged using two approaches: directed cloning of PCR-amplified ORFs into a yeast tagging/expression vector, and random tagging by transposon mutagenesis. By the former approach, 2085 annotated *S. cerevisiae* ORFs were cloned into the yeast high-copy expression vector pYES2/GS through topoisomerase I-mediated ligation (Fig. 1A). PCR-amplified yeast ORFs were inserted immediately upstream of sequence encoding the V5 epitope (from the P and V proteins of paramyxovirus SV5; Heyman et al. 1999) and downstream of the galactose-inducible *GAL1* promoter, such that galactose induction in yeast could be used to drive expression of each gene as a fusion protein carrying the V5 epitope at its C terminus. For purposes of this study, sequence-verified plasmids bearing yeast genes were transformed into an appropriate strain of *S. cerevisiae* in a 96-well format (see Materials and Methods). Cloned genes were expressed in yeast by galactose induction; the induction period was kept as brief as possible to minimize potential artifacts associated with gene overexpression. Protein products were subsequently localized by indirect immunofluorescence using monoclonal antibodies directed against the V5 epitope. To accommodate higher throughput, yeast cells were prepared for immunofluorescence analysis in a 96-well format as described (Kumar et al. 2000b).

Yeast genes were also epitope-tagged by means of in-

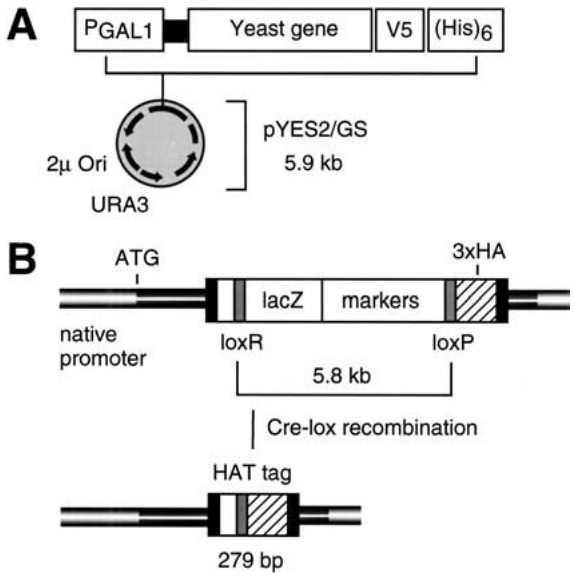

Figure 1. Genome-wide epitope-tagging strategies. (*A*) Yeast ORFs were amplified by PCR and cloned by topoisomerase I-mediated ligation into the yeast expression vector pYES2/GS. The pYES2/GS vector carries the yeast 2μ origin of replication for maintenance of high copy number. Yeast genes were inserted into pYES2 such that they are under transcriptional control of the *GAL1* promoter and fused at their 3′ ends to sequence encoding the V5-epitope and polyhistidine tag $(HIS)_6$. By galactose induction in yeast, cloned genes were overexpressed as V5-tagged proteins for subsequent immunolocalization with α-V5 antibodies (in 96-well formats). (*B*) Modified bacterial transposons were used to randomly tag yeast genes at their native genomic loci with sequence encoding three copies of the viral haemagglutinin epitope (3×HA epitope). The transposon carries a promoterless and 5′-truncated *lacZ* reporter enabling selection of in-frame insertions by β-galactosidase assay. In-frame insertions were subsequently modified in yeast by Cre-*lox* recombination, such that the majority of the transposon sequence was excised. The remaining HA-epitope insertion element (HAT tag) encodes no stop codons in the specified reading frame. The indicated 279-bp HAT-tag insertion includes a 5-bp duplication in target site sequence associated with Tn*3* transposition. HAT-tagged proteins were immunolocalized with monoclonal α-HA antibodies in a 96-well format.

sertional mutagenesis using a series of bacterial transposons, each modified to carry sequence encoding a reporter gene, bacterial and yeast selectable markers, a pair of internal *lox* sites, and three copies of the HA epitope (Fig. 1B; Ross-Macdonald et al. 1997). By shuttle mutagenesis (Seifert et al. 1986), transposon-mutagenized fragments of genomic DNA were introduced into a diploid strain of yeast; insertion alleles integrated at their corresponding genomic loci by homologous recombination. Insertions in-frame with gene-coding sequences were selected and subsequently modified in vivo by Cre-*lox* recombination such that all transposon-encoded reporters, stop codons, and selectable markers were excised. The remaining transposon insertion element encodes 93 amino acids, primarily consisting of the triplicate HA epitope. Proteins carrying this transposon-encoded HA tag (HAT tag) were localized by indirect immunofluorescence with α-HA monoclonal antibodies (see Materials and Methods). By this approach, 11,417 HAT-tagged strains were generated, encompassing 2958 different proteins suitable for subsequent immunolocalization.

In total, we have examined by indirect immunofluorescence >13,000 strains harboring epitope-tagged alleles of 3565 different genes (~60% of the yeast proteome). Of 2085 genes cloned into V5-tagging/expression vectors, 2022 gene products showed a staining pattern above background upon immunofluorescence analysis. Of 2958 HAT-tagged proteins similarly examined, 1083 proteins yielded staining patterns appreciably distinct from background. As these two data sets partially overlap, we define here the subcellular localization of 2744 different proteins. The subcellular compartmentalization of these proteins is indicated in Table 1. Example staining patterns resulting from indirect immunofluorescence analysis of HAT-tagged proteins and V5-tagged proteins are presented in Figure 2.

Subcellular localization of 2744 yeast proteins

Tagged proteins were localized in yeast to a wide variety of organelles and intracellular structures including the nucleus, mitochondria, endoplasmic reticulum, plasma membrane, vacuole, cytoplasm, and cell neck (Fig. 2). The majority (48%) of proteins tested in this study were found localized throughout the cytoplasm, typically showing a finely punctate pattern of staining. In addition, 68 proteins (2.5% of those tested) localized predominantly in clusters within the cytoplasm, visualized as intense areas of staining or patches occasionally overlaid on a background of general cytoplasmic staining. This patchy staining was often evident in strains carrying tagged alleles of known cytoskeletal or cytoskeleton-associated proteins. For example, immunofluorescence analysis of tagged Hsp42p revealed a patchy pattern of cytoplasmic staining; Hsp42p is a small heat shock protein functioning in reorganization of the actin cytoskeleton following thermal stress (Gu et al. 1997). In total, 18 known cytoskeletal proteins showed this staining pattern upon immunolocalization. Patches of cytoplasmic staining were also observed in cells carrying tagged proteins identified previously as components of the Golgi apparatus or other membrane-bound vesicles of the yeast secretory pathway. Van1p, a mannosyltransferase residing in the early Golgi compartment (Cho et al. 2000), showed this patchy staining pattern upon HAT-tagging and subsequent immunolocalization.

Approximately 1200 of all 2744 localized proteins were compartmentalized to discrete subcellular organelles such as the nucleus, mitochondria, or endoplasmic reticulum. Of these proteins, a significant fraction (25.2%) showed a mixed compartmentalization, localizing predominantly to a single organelle but also showing appreciable cytoplasmic staining upon immunofluorescence analysis. For example, 82 proteins were localized to the endoplasmic reticulum and cytoplasm, including

Table 1. *Summary of localized proteins*

Overexpressed, V5-tagged proteins		HAT-tagged proteins		Cumulative	
Staining pattern	# proteins	Staining pattern	# proteins	Staining pattern	# proteins (w/known functions)
Cell periphery	30	Cell periphery	43	Cell periphery	64 (51)
Cuto. (patches)	265	Cyto. (patches)	51	Cyto. (patches)	68 (53)
Cytoplasmic	928	Cytoplasmic	573	Cytoplasmic	1314 (760)
Nuclear rim/ER	94	Nuclear rim/ER	30	Nuclear rim/ER	115 (83)
Nucleus	371	Nucleus	130	Nucleus	451 (292)
Mitochondria	284	Mitochondria	74	Mitochondria	332 (233)
Mixed:	219	Mixed:	116	Mixed:	302 (235)
Cyto./ER	55	Cyto./ER	41	Cyto./ER	82 (61)
Cyto./Nucleus	153	Cyto./Nucleus	73	Cyto./Nucleus	207 (165)
Cell neck	1	Cell neck	4	Cell neck	5 (5)
Spindle pole body	4	Spindle pole body	3	Spindle pole body	5 (4)
Vacuole	5	Vacuole	6	Vacuole	11 (9)
Other:	60	Other:	53	Other:	77 (64)
Total	2022	Total	1083	Total	2744 (1789)

A subset of yeast proteins is represented within both the V5- and HAT-tagged data sets; therefore, the cumulative totals correspond to the number of distinct proteins within the union of these two data sets. The number of functionally characterized proteins (as extracted from the MIPS CYGD) showing each respective staining pattern is indicated in parentheses beside the cumulative totals (see Fig. 6). Major subcategories within the mixed and other categories are indicated. Specific protein localization data and corresponding immunofluorescence images may be accessed at http: //ygac.med.yale.edu (Protein Localization in Yeast link).

the vesicular transport protein Sec17p. Previous studies have suggested a role for Sec17p in vesicle-mediated endoplasmic reticulum-to-Golgi transport (Waters et al. 1991); the observed cytoplasmic/endoplasmic reticulum staining pattern resulting from immunolocalization of tagged Sec17p is typical of secretory vesicle proteins (see Discussion).

Interestingly, an even greater number of proteins (207) colocalized to the cytoplasm and nucleus. The majority of these proteins (for which functions had been assigned

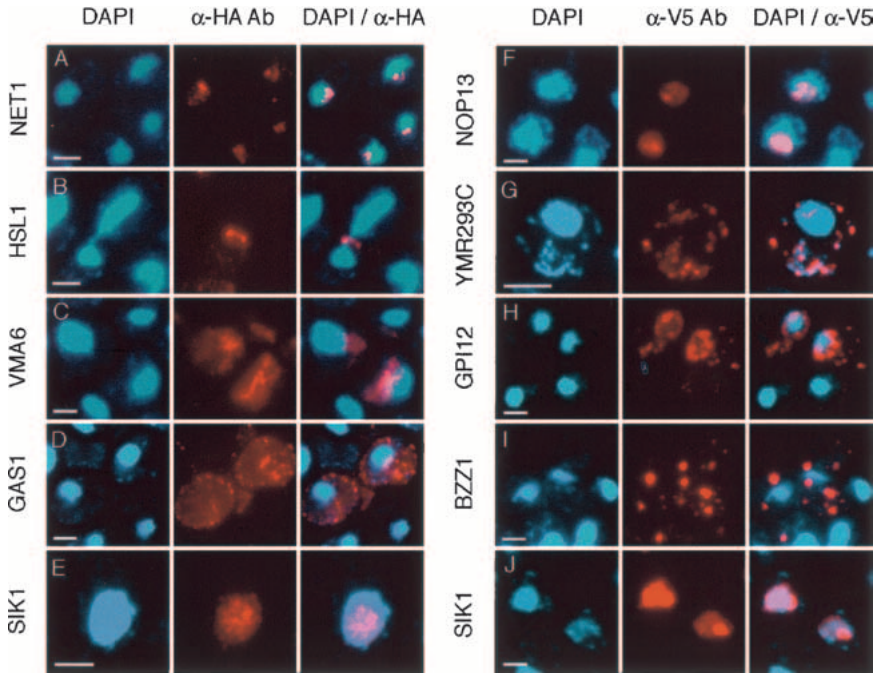

Figure 2. Immunolocalization of epitope-tagged proteins. (*A–E*) Vegetative cells containing HAT-tagged proteins were stained with the DNA-binding dye 4′,6-diamidino-2-phenylindole (DAPI; *left* image) and monoclonal antibody against HA (*center*). Per row, the DAPI-stained and α-HA-stained images are shown merged in the rightmost panel. Typical nucleolar staining patterns can be seen in strains containing HAT-tagged alleles of the rRNA-binding proteins Net1p (*A*) and Sik1p (*E*). Staining of the cell neck is evident in cells containing HAT-tagged Hsl1p (*B*). HAT-tagging of the vacuolar ATPase Vma6p is shown in row *C*. Staining of the cell periphery can be seen upon HAT-tagging of the cell surface glycoprotein Gas1p (*D*). (*F–J*) Vegetative cells carrying V5-tagged proteins were stained with monoclonal antibody directed against the V5 epitope (*center*). Corresponding DAPI-stained images and merged images are shown to the left and right, respectively. Nucleolar staining is apparent in cells carrying V5-tagged Nop13p (*F*). Note, however, that V5-tagging and mild overexpression of *SIK1* (*J*) results in a nuclear staining pattern, as opposed to the nucleolar pattern evident upon HAT-tagging of this same gene (*E*). Mitochondrial staining (*G*) can be seen in cells carrying a tagged allele of *YMR293C*; overlap between DAPI- and α-V5 staining is shown in the merged image. V5-tagged Gpi12p localizes to the endoplasmic reticulum (*H*), visible as an area of strong staining around the nuclear rim. A patchy pattern of cytoplasmic staining can be seen in cells carrying V5-tagged Bzz1p (*I*). Bar, 2 μm.

previously) are involved either in processes of transcription or cytoskeletal organization. In our study, many transcription factors were localized, at least in part, to the cytoplasm. For example, we found the transcriptional activator Pho4p localized predominantly to the cytoplasm, and only slightly in the nucleus, under conditions of vegetative growth on standard media. This finding agrees with published work in which Pho4p was found concentrated in the nucleus only under conditions of phosphate starvation (O'Neill et al. 1996). In some cases, however, cytoplasmic staining may be an artifact resulting from a disrupted nuclear localization signal or saturated nuclear transporters.

To estimate the frequency with which such artifacts are present within our data, we have compared all localizations from this study with previously published localization data extracted from the Yeast Protein Database (YPD; Costanzo et al. 2001), the SwissProt Database (Bairoch and Apweiler 2000), and the Munich Information Center for Protein Sequences Comprehensive Yeast Genome Database (MIPS CYGD; Mewes et al. 2000). Comparison of 694 protein localizations indicated >85% agreement with data from existing literature. In particular, our findings are in agreement with previously published results in 93% of cases in which we localize a protein to the mitochondria (134 comparisons total) and 90% of cases in which we localize a protein to the nucleus (230 comparisons total). We do recognize biases in our method, as certain classes of proteins (e.g., spindle pole proteins) are underrepresented in our results. A more detailed analysis of the accuracy and efficiency of our methods is provided in the Discussion.

Mapping protein sorting signals by transposon-tagging

As transposition occurs nearly at random, our genome-wide methods of transposon mutagenesis often generate multiple insertions within a single gene (see Discussion). The availability of these multiple insertion alleles can be advantageous, providing a means by which intragenic sequences important for proper localization and function may be mapped. For example, from immunolocalization of several HAT-tagged variants of the yeast peroxisomal membrane protein Pex22p, we have identified a putative peroxisomal membrane-targeting signal at the N terminus of this protein: HAT-tag insertion at residue 10 of Pex22p disrupts peroxisomal localization, whereas an insertion 55 residues C-terminal of this site does not. Interestingly, a functional homolog of Pex22p in *Pichia pastoris* contains a known 25-amino-acid membrane-targeting signal at its extreme N terminus (Koller et al. 1999).

Subcellular compartmentalization of the yeast proteome using an integrated Bayesian system

By integrating our results with those publicly available in YPD, SwissProt, and MIPS CYGD, we can definitively assign subcellular localizations to 3343 yeast proteins. In complement, we have employed a hydrophobicity-based predictive algorithm (Krogh et al. 2001) to identify all yeast proteins possessing two or more transmembrane domains—an approach estimated to identify integral membrane proteins with 99% accuracy (Krogh et al. 2001). In total, 1029 integral membrane proteins were identified in the yeast proteome; 387 of these predicted membrane proteins were already assigned a subcellular compartment from our immunolocalization data and/or previously published data. We have estimated the relative subcellular distribution of the remaining 642 previously unstudied membrane proteins by extrapolating from the relative compartmentalization of membrane proteins observed in our experimentally derived localization data set. We, therefore, define a molecular environment (transmembrane or soluble) and subcellular localization for 3985 yeast proteins.

To intelligently predict the subcellular distribution of the remaining 2147 soluble yeast proteins for which no localization data are available, we have used a Bayesian system (Drawid and Gerstein 2000) that extrapolates from our findings and also integrates additional types of data potentially correlative to protein localization (see Materials and Methods). For purposes of this analysis, we have used all available data describing experimentally determined protein localizations in yeast to calculate a default localization probability. This initial probability is sequentially updated for each previously uncharacterized protein using Bayes's rules and a diverse set of 30 features (including motif analysis, surface composition, isoelectric point, and mRNA expression, among others), thereby generating a final localization probability for each protein. Localization probabilities were subsequently summed, providing an estimate as to the overall population of each subcellular compartment. The estimated compartment populations were added to those experimentally determined to arrive at the total subcellular compartmentalization of the yeast proteome (Fig. 3).

By this approach, we estimate 47% of all yeast proteins to be cytoplasmic and an additional 27% to be nuclear. Approximately equal fractions of the yeast proteome (13%) compartmentalize to the mitochondria and exocytic network. As expected, we find the majority of yeast integral membrane proteins localized to the endoplasmic reticulum or other secretory vesicles. In the categorization scheme employed here, plasma membrane proteins have been incorporated into the cytoplasmic compartment; therefore, the membrane fraction of cytoplasmic proteins is higher than otherwise would be expected. In total, the yeast proteome consists of 1029 transmembrane proteins and 5103 soluble proteins. Comprehensive results from this Bayesian analysis may be accessed at http://genecensus.org/localize.

Chromosomal association and phenotypic analysis of nuclear-localized proteins

By our analysis (Fig. 3), the yeast proteome may encompass in excess of 1600 nuclear proteins. The presence of this surprisingly large nuclear complement raises an in-

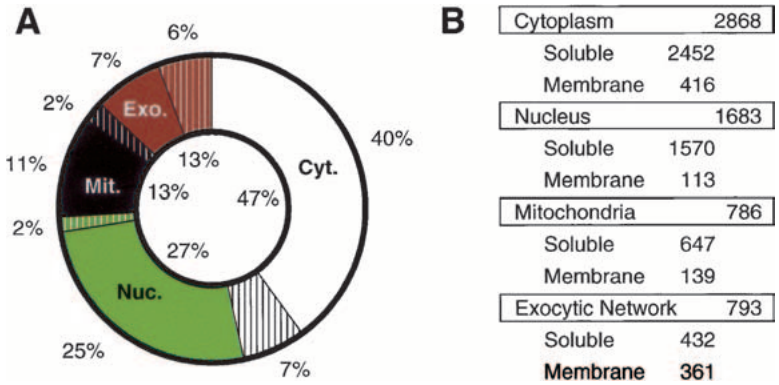

Figure 3. Subcellular compartmentalization of the yeast proteome. (*A*) Cellular compartments are as follows: cytoplasmic (Cyt.), nuclear (Nuc.), mitochondrial (Mit.), and exocytic (Exo.). The membrane fraction of each compartment is indicated in stripes. The percentage of the yeast proteome contained within the respective membrane and soluble fractions of each compartment is indicated outside the chart; the total percentage of the proteome contained within each of the four main compartments is indicated inside the chart. Plasma membrane proteins are included in the cytoplasmic compartment for purposes of this analysis. (*B*) The corresponding protein population of each cellular compartment and membrane/soluble subfraction is indicated.

teresting question: what fraction of these proteins associates with chromosomes? Furthermore, how many of these nuclear proteins are essential for cell viability? To address these questions, we have analyzed the chromosomal localization and disruption phenotypes associated with a subset of yeast nuclear proteins identified in this study. Transposon-tagged strains were chosen for this analysis, as a single transposon insertion can be used to generate both a gene disruption as well as an epitope-tagged allele (Ross-Macdonald et al. 1997), facilitating phenotypic study and immunofluorescence analysis, respectively. To assess the ability of these proteins to associate with chromosomal DNA, 56 HAT-tagged nuclear proteins were immunolocalized on surface-spread preparations of meiotic chromosomes isolated from late-zygotene-to-pachytene nuclei. A sampling of observed staining patterns is presented in Figure 4; complete results are indicated in Figure 5. In addition, corresponding alleles of each gene carrying full-length transposon insertions were assayed for their effect on spore viability (see Materials and Methods). Results from this phenotypic analysis are also presented in Figure 5.

In total, 21 nuclear proteins of 56 tested (38%) were found localized to meiotic chromosomes. Specifically, 16 proteins (including six of previously unknown function) showed staining patterns indicative of general chromosomal binding, typically with 40 or more chromosomal foci per nucleus. Two proteins, Orc4p (Fig. 4S–U) and Rap1p, bound telomeric DNA. Orc4p is a component of the origin recognition complex (ORC) and is involved in transcriptional silencing at telomeres (Bell et al. 1995); Rap1p is a transcription factor also involved in telomeric silencing as well as telomere maintenance (Ray and Runge 1998). Two additional proteins, the DNA replication factor component Rfc3p (Fig. 4A–C; Li and Burgers 1994) and the chromatin remodeling protein Rsc6p (Cairns et al. 1996), bound telomeric sequence while also recognizing more than 20 other chromosomal sites each. As expected, the centromere-binding factor Cbf1p (Baker and Masison 1990) bound centromeric sequence, visualized as a single staining spot in the center of each chromosome. Nine gene products (16% of those tested) localized predominantly, if not exclusively, to the nucleolus, including two previously uncharacterized proteins encoded by *YGR090W* (Fig. 4J–L) and *YHR196W* (Fig. 4P–R). The majority of chromosomal and nucleolar proteins, such as these, likely bind DNA, either chromosomal DNA or nucleolar rDNA; however, a significant fraction may only associate with chromosomes through interactions with other chromosomal proteins (e.g., histone modification proteins).

Phenotypic analysis of the 56 nuclear proteins tested here revealed 19 genes (34%) indispensable for spore viability (Fig. 5)—a fraction approximately twice as great as that found for the genome as a whole (Winzeler et al. 1999). Interestingly, 6 of these 19 essential genes encode nucleolar proteins, including the aforementioned *YGR090W* and *YHR196W* gene products. An additional 13 genes produced a slow-growth phenotype upon disruption. Five of these genes encode chromosomal-associated proteins; three encode nucleolar proteins. In total, all nine nucleolar proteins conferred observable phenotypes (spore inviability or slow-growth) upon disruption; 52% of chromosomal proteins conferred these same phenotypes, underscoring the fundamental importance of the nucleus/nucleolus and its protein complement.

Protein localization correlates strongly with protein function

The studies presented here provide a unique opportunity to examine more rigorously the assumption that protein function can be inferred from protein localization, an assumption best tested by correlating proteome-wide data sets of protein localization with corresponding data sets of protein function. Accordingly, we have tallied all molecular functions (extracted from MIPS CYGD) associated with the 2744 yeast proteins immunolocalized in this study. The most frequently observed functions associated with each of the eight most populous compartments of the yeast proteome are indicated in Figure 6.

Within each organelle or compartment, a plurality of proteins participate, at least partially, in maintaining structural integrity. Secondary functions also correlate well with major organelle-specific processes: for example, 34% of all nuclear-localized proteins are involved in the process of transcription, and 26% of all mitochondrial proteins function directly in cellular respiration.

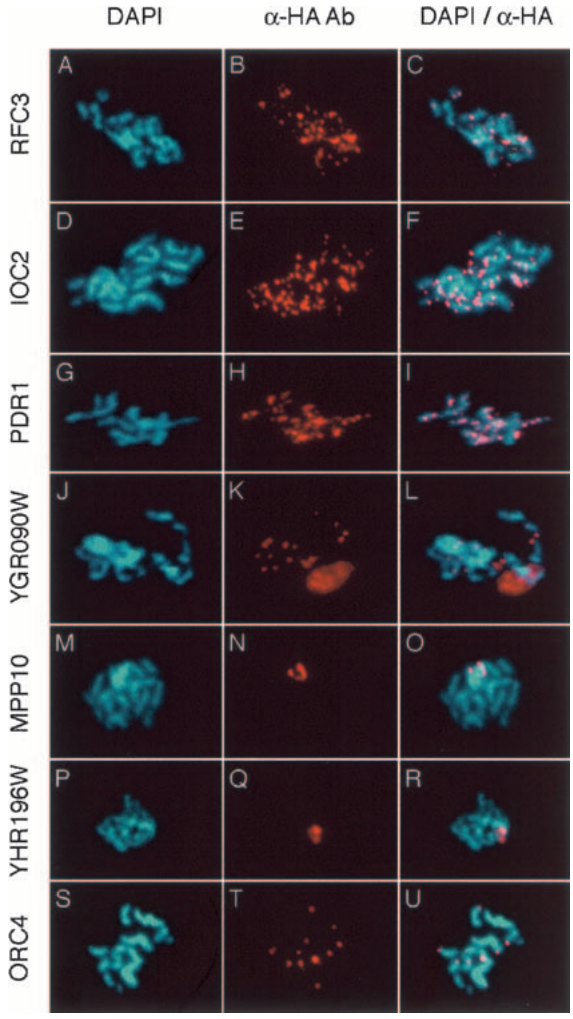

Figure 4. Immunolocalization of nuclear proteins on surface-spread meiotic chromosomes. Meiotic chromosomes were surface spread and stained with the DNA-binding dye DAPI (*left*) and monoclonal anti-HA antibodies (*center*). Corresponding merged images are shown to the right. A general pattern of chromosomal binding can be seen from immunofluorescence analysis of cells containing HAT-tagged alleles of *RFC3* (*A–C*), *IOC2* (*D–F*), and *PDR1* (*G–I*). Nine proteins localized predominantly to the nucleolus; typical nucleolar staining patterns are shown here in cells containing HAT-tagged alleles of *YGR090W* (*J–L*), *MPP10* (*M–O*), and *YHR196W* (*P–R*). Specific binding to telomeric DNA can be seen upon HAT-tagging and immunolocalization of the origin recognition complex subunit Orc4p (*S–U*). Bar, 1 μm.

Furthermore, specific functions can be correlated with subtly distinct localization patterns. For example, 17% of the proteins that colocalized to the nucleus and cytoplasm are cytoskeletal, whereas cytoskeletal functions are not as strongly associated with proteins that localize only to the nucleus or only to the cytoplasm. Similarly, cytoskeletal proteins and Golgi proteins constitute the bulk of those proteins showing patchy patterns of cytoplasmic staining; however, identical functions are not significantly represented among proteins yielding fine, granular, or punctate cytoplasmic staining patterns. This strong correlation between function and localization suggests that broad functional categories can now be ascribed to the 955 proteins of previously unknown function localized in this study.

An on-line database and visual library of protein localization in yeast

To catalog the data presented here, we have developed an on-line database of yeast protein localization accessible from our lab homepage at http://ygac.med.yale.edu (Protein Localization in Yeast link). For this site, we have developed a new user interface specifically accommodating our V5-tagged data set; new HAT-tagged data may now be accessed from our TRIPLES web site (Kumar et al. 2000a). In both cases, we supply search options by which users can access data for any gene of interest. Alternatively, complete data sets for all proteins localizing to a given site may be downloaded as tab-delimited text. Tabular data sets are supplemented with fluorescent micrographs of staining patterns observed upon immunofluorescence analysis of each indicated protein. In total, this new site houses >2893 micrographs, establishing it as the largest visual library of eukaryotic protein localization to date.

Discussion

Constituting the first proteome-wide analysis of protein localization, this study is uniquely suited to address a number of issues regarding both the methods by which such a project may be undertaken as well as the utility and applications of the end data. Here, we have used two common approaches by which epitope-tagged alleles may be generated on a genomic scale: directed cloning methods and random transposon-based approaches. By comparing the localization data generated from each respective set of tagged alleles, we can rigorously assess the efficiency and accuracy of each approach. In particular, our results may be used to consider the accuracy with which overexpressed proteins can be localized as compared with the localization accuracy associated with endogenously expressed proteins. The resulting localization data sets correlate strongly with protein function, providing further means by which proteins may be ascribed functions on a proteome-wide scale. This analysis also offers specific insight into the relative distribution of functions and phenotypes associated with nuclear proteins, while providing data regarding nearly 1000 proteins of unknown function.

Two genomic epitope-tagging approaches: respective efficiencies

Directed cloning and random transposon-tagging each possess advantages and disadvantages as approaches for genome-wide epitope-tagging. For large-scale directed cloning, a significant investment in labor and reagents is

	Chromosomal	Nucleolar	Non-chromosomal
Inviable	RSC6 - Telomere / Chrom.	NAB2	ROX3
	TAF61	YGR090W	BCP1
	RFC3 - Telomere / Chrom.	YHR196W	RRP4
	RAP1 - Telomere	MPP10	TRM5
	GCR1	RRP5	PRP8
	ORC4 - Telomere	RPB8	SRP1 - Nuclear pores / rim
			DIS3
Slow-Growth	VID21	FYV13	TUP1
	NSR1	SRP40	YDR101C
	RPA34	DOT4 - Nucleolus / Chrom.	HMO1
	SPT8		POL32
	BDF1		DBP2
Viable	MBP1		FUN30
	WTM1		HHT1
	PLO2		LYS20
	YER049W		UME6
	PDR1		NPL3
	ARP4		DOT6
	CBF1 - Centromere		HAP2
	YKL160W		YJL122W
	YKL005C		RGT1
	IOC2		YLR374C
			RGM1
			TOP1
			YOR227W
			SGV1

Figure 5. Chromosomal localization and phenotypic analysis of nuclear proteins. Chromosomal localization indicates a general pattern of chromosomal binding, typically with >40 staining foci per nucleus. Strains disrupted for each gene were assayed for spore viability or growth defects; observed disruption mutants are categorized as viable, inviable, or slow-growth, accordingly.

initially required; however, the final collection possesses little or no redundancy in gene representation. Furthermore, for purposes of immunolocalization, directed approaches are efficient: in this study, 93% of genes cloned into a tagging/expression vector subsequently yielded staining patterns above background upon immunofluorescence analysis. Transposon-based methods, in contrast, are economical but inefficient. Only 30% of transposon-tagged proteins showed a staining pattern distinct from background. In addition, owing both to the stochastic nature of transposition and to the insertional biases associated with bacterial transposons (e.g., Tn*3*), transposon-based methods can prove problematic as a means of saturating a given genome. Small genes are less likely to be mutagenized by transposon mutagenesis than are large genes. Also, insertional collections possess greater redundancy in gene representation than do collections generated by directed methods: the collection of HAT-tagged genes generated in this study shows approximately fourfold redundancy in gene representation on average (11,417 HAT-tagged alleles representing 2958 different genes). As shown, however, this redundancy can be beneficial in mapping domains within a given protein.

Respective accuracy of each approach

To estimate the accuracy of data generated by each tagging strategy, we have compared all protein localizations determined experimentally in this study with previously published localization data. Both approaches (i.e., mild overexpression of C-terminal, V5-tagged alleles vs. endogenously expressed random HAT-tagged alleles) yielded data sets of similar accuracy (~85%) when compared with published localization results. This internal comparison, however, is complicated by the fact that different proteins are represented in the V5- and HAT-tagged data sets, respectively. To estimate the relative

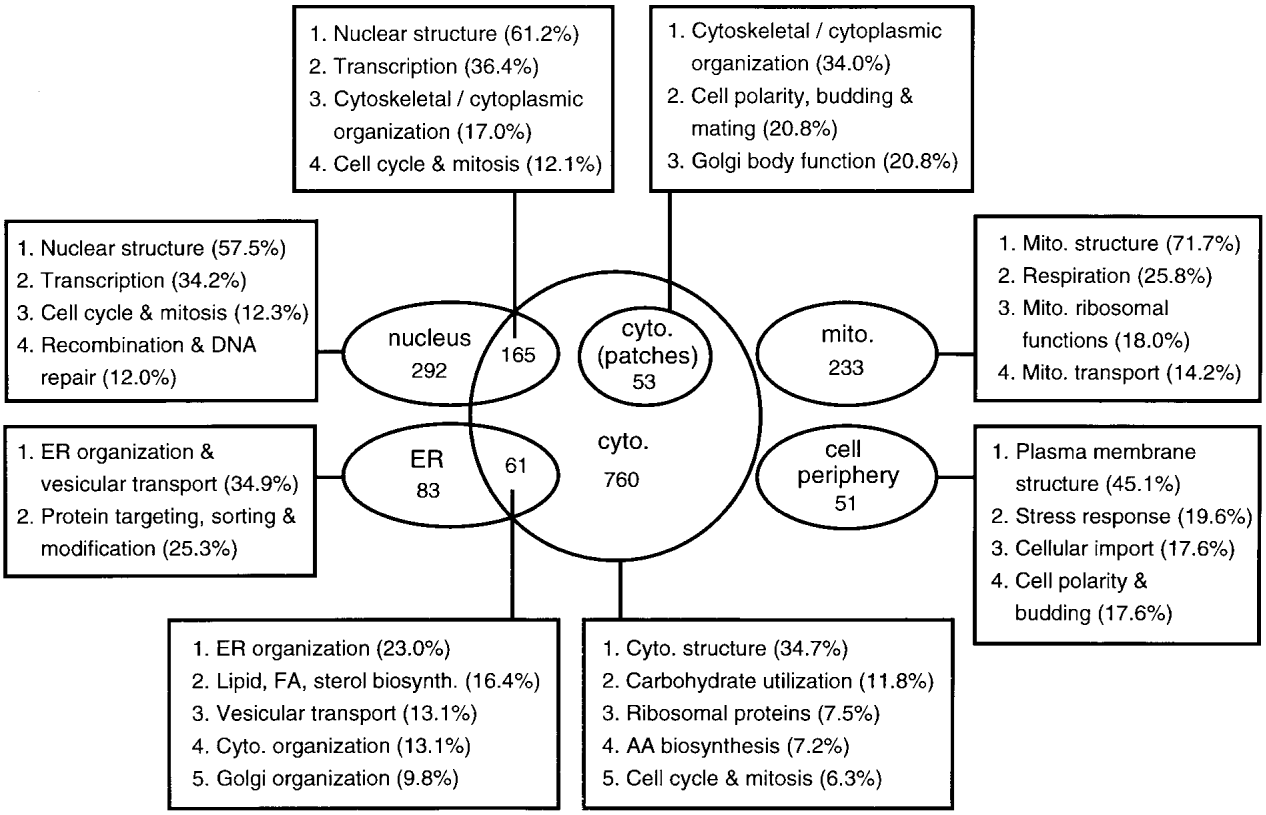

Figure 6. Prevalent functions associated with cellular compartments in yeast. Functional categorizations (compiled from published literature) were extracted from the MIPS CYG database for all proteins experimentally localized in this study. In total, functions were available for 1789 proteins; the number of functionally categorized proteins localized to each of the indicated cellular compartments is shown. Mixed localizations are also represented: 165 functionally characterized proteins were colocalized to the cytoplasm and nucleus; similarly, 61 such proteins were colocalized to the cytoplasm and endoplasmic reticulum (ER). Functions were tallied for all proteins within a given cellular compartment. The most frequently occurring functions per compartment are shown boxed. Multiple functions may be associated with a single protein. Therefore, the listed percentage following each function refers to the fraction of proteins within each compartment associated with that particular cellular process, and the sum total of these percentages within a given compartment will not equal 100%.

accuracy of each approach more rigorously, we have limited the comparison to only those 361 proteins common to both data sets. Of these proteins, 295 yielded identical or very similar staining patterns upon immunofluorescence analysis regardless of the tagging approach used; the remaining 66 proteins yielded differing results. For 29 of these proteins, no previously published localization data are available. Of the remaining 37 proteins analyzed, localization data derived from V5-tagged proteins agreed more closely with published results in 20 cases; in 17 cases, HAT-tagged data proved more accurate (e.g., Sik1p shown in Fig. 2E,J). Therefore, both approaches may be used to generate epitope-tagged alleles for subsequent immunolocalization with comparable degrees of accuracy, suggesting, furthermore, that the effects of tag size, placement, and expression may be less severe than generally thought.

The yeast proteome: subcellular distribution and functional implications

Using the tagging/immunofluorescence strategies discussed above, we have determined the subcellular localization of 2744 different yeast proteins (Table 1). As expected, large sets of proteins were localized to the cytoplasm and nucleus (~50% and ~25% of the yeast proteome, respectively); however, a surprisingly large number of proteins showed a mixed staining pattern, localizing to more than one subcellular compartment. In total, mixed localization patterns were evident in 11% of all samples tested. Transcription factors and cytoskeletal proteins were frequently found distributed within both the nucleus and cytoplasm as discussed. Also, vesicular proteins (e.g., Sec17p) often colocalized to the endoplasmic reticulum and cytoplasm. Secretory polypeptides that have been either epitope-tagged or overproduced may be processed less efficiently for export, and therefore accumulate in the endoplasmic reticulum: randomly placed tags may disrupt signal peptide sequence, and overexpressed proteins may saturate mechanisms responsible for membrane protein traffic. Therefore, colocalization of secretory vesicle proteins to the endoplasmic reticulum is expected.

The fact that a given protein may be distributed among more than one cellular compartment is a relevant con-

sideration in developing a computational system by which our data may be extrapolated over the yeast proteome. For this purpose, we have used a Bayesian system by which the relative protein population of each yeast cellular compartment may be estimated without requiring a definitive localization for every constituent protein. The accuracy of this method is dependent largely on the availability of a large and unbiased localization data set from which a default series of localization probabilities can be calculated. In previous applications of this approach (Drawid and Gerstein 2000), an initial data set was constructed from all known yeast protein localizations cataloged in public databases (1342 localizations in total). This data set, however, is biased toward nuclear proteins, as they have been traditionally studied in greater detail than other protein classes. Our study provides a less biased data set: the transposon-based approaches used here are near-random, and the population of yeast genes successfully cloned into pYES2 may reflect only a marginal enrichment for short ORFs that tend to be easily amplified by PCR. By merging our experimentally determined localizations with published yeast localization data and predicted transmembrane classifications (Krogh et al. 2001), we define a subcellular localization for 3985 yeast proteins. Applying our probabilistic system to the remaining yeast protein complement, we arrive at the proteome-wide protein compartmentalization indicated in Figure 3. This distribution agrees well with previous theoretical estimates of protein localization in yeast (Drawid and Gerstein 2000).

Because protein localization and function are tightly correlated (Fig. 5), our global localization analysis provides a means by which gene function in yeast may be inferred on a genome-wide scale. Extrapolating from the functional categorizations maintained in the MIPS CYGD and our localization data, ~45% of all yeast proteins function at least partly in maintaining cytoplasmic and organelle-specific organization and integrity. From our localization analysis, we estimate that the yeast proteome contains nearly 800 mitochondrial proteins—the majority of which function, as expected, in processes of cellular respiration (Fig. 6).

Similar predictions can be made regarding the yeast nuclear protein complement. In this study, we have identified 457 nuclear-localized proteins for which functional data are currently available (including 165 proteins that colocalize to the nucleus and cytoplasm). Of these nuclear proteins, 34.8% function in transcription. Extrapolating this fraction to the total nuclear protein compartment (1683 proteins; Fig. 3), we estimate that nearly 10% of the yeast proteome is dedicated to processes of mRNA transcription. Consistent with this prediction, we have found that ~38% of all nuclear proteins (or 10% of the yeast proteome) are associated with chromosomal DNA as determined by immunofluorescence analysis of tagged proteins on surface-spread meiotic chromosomes (Fig. 5). Although caution must be exercised in extrapolating from a limited population of 56 nuclear proteins, our phenotypic studies suggest that roughly 34% of all nuclear proteins (>570 proteins in total) are essential for spore viability. In contrast, over the genome as a whole, 18% of yeast genes (~1100) are thought to be essential as indicated from systematic analyses of yeast deletion mutants (Winzeler et al. 1999). Therefore, slightly more than half of all essential genes in yeast are likely to be nuclear.

Integrating localizome data

Large-scale localization data sets provide a fundamental complement to other existing varieties of proteomic data. For example, our localization data may be used to screen sets of putative protein–protein interactions, enriching for genuine protein associations by virtue of the expectation that two interacting proteins will share a common cellular compartment and show similar localization patterns. At present, large catalogs of protein interactions in yeast have been generated through genome-wide applications of the two-hybrid method (Uetz et al. 2000; Ito et al. 2001) and systematic, high-throughput approaches using mass spectrometric analysis of immunoprecipitated protein complexes (Gavin et al. 2002; Ho et al. 2002). We have correlated our localization results with a sampling of interaction data drawn from each of these studies. Of 155 randomly selected two-hybrid interactions identified either by Uetz et al. (2000) or Ito et al. (2001), only 73 (47%) contain a protein pair localized to the same cellular compartment. In contrast, however, within a set of 105 two-hybrid interactions independently identified by both groups, 87 protein pairs (83%) show a shared localization pattern. Analysis of data generated by Gavin et al. (2002) and Ho et al. (2002) yields similar results. Of 100 sampled protein associations (encompassing 10 different bait proteins), 67 interactions consist of two proteins from the same cellular compartment. Of these 100 protein associations, 23 were identified by both groups: all 23 of these protein pairs show compatible localization patterns. These correlations suggest that confidence can be placed preferentially in protein interactions independently identified within more than one study, while simultaneously demonstrating the usefulness of localization data in distinguishing spurious protein–protein interactions is likely to be spurious.

As illustrated by these comparisons, results from independent studies may be effectively integrated to provide more accurate and complete genomic findings. The accuracy of our own localization results may be improved through comparison with a set of known and established protein–protein interactions, a corollary of the analysis above. A more comprehensive representation of yeast protein function may be achieved by integrating multiple proteomic data sets, because all such individual data sets are presently incomplete (i.e., encompass <6000 yeast proteins). Collectively, this union of diverse proteomic and genomic approaches will prove mutually complementary and necessary as a means of understanding global processes of eukaryotic cellular function.

Materials and methods

Epitope-tagging and immunolocalization

Yeast genes were amplified by polymerase chain reaction (PCR) and cloned into the yeast expression vector pYES2/GS by topoisomerase I-mediated ligation as described previously (Heyman et al. 1999). Vector constructs carrying cloned yeast genes were introduced into haploid strain YNN218 [*ura3-52 lys2-801 ade2-101 his3Δ200*] by DNA transformation (Ito et al. 1983). To induce gene expression, yeast transformants were first grown to saturation in synthetic medium lacking uracil (SC-Ura) with raffinose as its carbon source; cultures were then washed in sterile water prior to resuspension in SC-Ura with galactose as a carbon source. Transformant cultures were incubated in galactose for 1 h. Multiple incubation periods were tested to determine the optimum time for galactose induction such that artifacts resulting from gene overexpression are minimized. Following galactose induction, cells were prepared for immunofluorescence analysis in a 96-well format as described (Kumar et al. 2000b). V5-tagged proteins were immunolocalized by indirect immunofluorescence using anti-V5 mouse monoclonal IgG2a antibody (Invitrogen) and Cy3-conjugated goat anti-mouse IgG (Jackson Labs).

Yeast genes were HAT-tagged using transposon-based methods presented previously (Ross-MacDonald et al. 1999; Kumar et al. 2000b). Tagged genes were generated in a Y800 background [*MAT***a** *leu2-Δ98cry1*^{R}/*MATα leu2-Δ98CRY1 ade2-101 HIS3/ade2-101 his3-Δ200 ura3-52 cani*^{R}/*ura3-52CAN1 lys2-801/lys2-801 CYH2/cyh2*^{R} *trp1-1/TRP1 Cir*^{0}] carrying pGAL-*cre* (*amp*, *ori*, *CEN*, *LEU2*) (Burns et al. 1994). Asynchronous cultures of HAT-tagged yeast strains were grown and prepared for immunofluorescence analysis in 96-well microtiter plates (Kumar et al. 2000b). Transposon-tagged proteins were immunolocalized as above, except that mouse monoclonal anti-HA 16B12 (MMS101R, BAbCO) was used as the primary antibody.

Computational methods

For purposes of this analysis, all yeast proteins were divided into four localization categories: cytoplasm (Cyt), nucleus (Nuc), mitochondria (Mit), and exocytic network (Exo; endoplasmic reticulum, Golgi apparatus, vacuoles, vesicles, peroxisome, and extracellular proteins). A different localization prediction procedure was applied for soluble and membrane proteins of all categories.

We identified 1029 yeast proteins as integral membrane proteins. These proteins were predicted to possess two or more transmembrane helices using TMHMM (Krogh et al. 2001); many also had a verifying match to a Pfam family of known membrane proteins (L. Yang and M. Gerstein, in prep.). Of these putative membrane proteins, ~380 had been localized to one of the four categories described above by transposon-tagging and subsequent immunolocalization as described here. We believe the distribution of these proteins among the four categories to be random and accurate. Consequently, we applied this distribution to the ~650 membrane proteins of unknown localization.

The remaining 5101 proteins in the yeast genome were considered to be soluble proteins. In total, experimentally derived localization data are available for ~2950 of these soluble proteins both from this study as well as from data previously deposited in the MIPS, YPD, and SwissProt databases. This served as the training set for our Bayesian method (Drawid and Gerstein 2000). The Bayesian system integrates a large number of different features related to yeast proteins, including sequence patterns, such as the nuclear localization signal or signal sequence, expression information, and many varieties of phenotypic data (e.g., viability of corresponding null mutants). The incorporation of expression information is particularly unique and is derived from the observation that cytoplasmic proteins possess much higher levels of expression than those in other compartments (Drawid et al. 2000).

By transposon-tagging/immunolocalization, we have defined the localization of ~2500 soluble proteins. Because we believe these proteins represent a random sample from the yeast genome, we have used their localization proportions as our priors (Cyt, 52%; Nuc, 27%; Mit, 14%; Exo, 7%). The subcellular localization of the remaining 2150 soluble proteins (for which no localization data are available) was predicted using our Bayesian method and the above prior and training data. We directly integrated the proportions of these 2150 proteins to yield an overall prediction for protein compartmentalization within the yeast proteome. We added together the number of soluble and membrane proteins to obtain the pie chart presented in Figure 3A.

Immunocytology and phenotypic analysis of nuclear proteins

Meiotic chromosomes from HAT-tagged strains were surface spread and stained as described using mouse anti-HA (Covance) at 1:400 dilution and DAPI (Agarwal and Roeder 2000). To examine phenotypes of strains (Y800 background) containing disruptive, full-length transposon insertions within genes encoding nuclear proteins, diploids heterozygous for the insertion were sporulated; tetrads were subsequently dissected and assayed for spore viability and transposon-encoded β-galactosidase activity as described previously (Burns et al. 1994).

Acknowledgments

We thank James R. Chambers, Shannon Hattier, and Jon Rowland of Invitrogen Corporation for strain organization and DNA preparation. This work was supported by NIH Grant R01-CA77808 to M.S. A.K. is supported by a postdoctoral fellowship from the American Cancer Society.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Agarwal, S. and Roeder, G.S. 2000. Zip3 provides a link between recombination enzymes and synaptonemal complex proteins. *Cell* **102:** 245–255.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Baker, R.E. and Masison, D.C. 1990. Isolation of the gene encoding the *Saccharomyces cerevisiae* centromere-binding protein CP1. *Mol. Cell. Biol.* **10:** 2458–2467.

Bell, S.P., Mitchell, J., Leber, J., Kobayashi, R., and Stillman, B. 1995. The multidomain structure of Orc1p reveals similarity to regulators of DNA replication and transcriptional silencing. *Cell* **83:** 563–568.

Burley, S.K. 2000. An overview of structural genomics. *Nat. Struct. Biol.* **Suppl:** 932–934.

Burns, N., Grimwade, B., Ross-Macdonald, P.B., Choi, E.-Y., Finberg, K., Roeder, G.S., and Snyder, M. 1994. Large-scale characterization of gene expression, protein localization and gene disruption in *Saccharomyces cerevisiae*. *Genes & Dev.*

8: 1087–1105.
Cairns, B.R., Lorch, Y., Li, Y., Zhang, M., Lacomis, L., Erdjument-Bromage, H., Tempst, P., Du, J., Laurent, B., and Kornberg, R.D. 1996. RSC, an essential, abundant chromatin-remodeling complex. *Cell* **87:** 1249–1260.
Cho, J.H., Noda, Y., and Yoda, K. 2000. Proteins in the early Golgi compartment of *Saccharomyces cerevisiae* immunoisolated by Sed5p. *FEBS Lett.* **469:** 151–154.
Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Lennon-Hopkins, K., Kondu, P., et al. 2001. YPD™, PombePD™ and WormPD™: Model organism volumes of the BioKnowledge™ Library, an integrated resource for protein information. *Nucleic Acids Res.* **29:** 75–79.
Ding, D.Q., Tomita, Y., Yamamoto, A., Chikashige, Y., Haraguchi, T., and Hiraoka, Y. 2000. Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library. *Genes Cells* **5:** 169–190.
Drawid, A. and Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *J. Mol. Biol.* **301:** 1059–1075.
Drawid, A., Jansen, R., and Gerstein, M. 2000. Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* **16:** 426–430.
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.
Gu, J., Emerman, M., and Sandmeyer, S. 1997. Small heat shock protein suppression of Vpr-induced cytoskeletal defects in budding yeast. *Mol. Cell. Biol.* **17:** 4033–4042.
Gygi, S.P., Rist, B., Gerber, S.A., Tureck, F., Gelb, M.H., and Aebersold, R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotech.* **17:** 994–999.
Heyman, J.A., Cornthwaite, J., Foncerrada, L., Gilmore, J.R., Gontag, E., Hartman, K.J., Hernandez, C.L., Hood, R., Hull, H.M., Lee, W.-Y., et al. 1999. Genome-scale cloning and expression of individual open reading frames using topoisomerase I-mediated ligation. *Genome Res.* **9:** 383–392.
Ho, Y., Gruhler, A., Heilbut, A., Bader, G., 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.
Ito, H., Fukuda, Y., Murata, K., and Kimura, A. 1983. Transformation of intact yeast cells treated with alkali cations. *J. Bacteriol.* **153:** 163–168.
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.* **98:** 4569–4574.
Iyer, V.R., Horak, C.A., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533–538.
Koller, A., Snyder, W.B., Faber, K.N., Wenzel, T.J., Rangell, L., Keller, G.A., and Subramani, S. 1999. Pex22p of *Pichia pastoris*, essential for peroxisomal matrix protein import, anchors the ubiquitin-conjugating enzyme, Pex4p, on the peroxisomal membrane. *J. Cell Biol.* **146:** 99–112.
Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305:** 567–580.
Kumar, A., Cheung, K.-H., Ross-Macdonald, P., Coelho, P.S.R., Miller, P., and Snyder, M. 2000a. TRIPLES: A database of gene function in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **28:** 81–84.
Kumar, A., des Etages, S.A., Coelho, P.S.R., Roeder, G.S., and Snyder, M. 2000b. High-throughput methods for the large-scale analysis of gene function by transposon tagging. *Methods Enzymol.* **328:** 550–574.
Li, X. and Burgers, P.M. 1994. Molecular Cloning and expression of the *Saccharomyces cerevisiae* RFC3 gene, an essential component of replication factor C. *Proc. Natl. Acad. Sci.* **91:** 868–872.
MacBeath, G. and Schreiber, S.L. 2000. Printing proteins as microarrays for high-throughput function determination. *Science* **289:** 1760–1763.
Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., et al. 2000. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **28:** 37–40.
Montelione, G.T. 2001. Structural genomics: An approach to the protein folding problem. *Proc. Natl. Acad. Sci.* **98:** 13488–13489.
Niedenthal, R.K., Riles, L., Johnston, M., and Hegemann, J.H. 1996. Green fluorescent protein as a marker for gene expression and subcellular localization in budding yeast. *Yeast* **12:** 773–786.
O'Neill, E.M., Kaffman, A., Jolly, E.R., and O'Shea, E.K. 1996. Regulation of PHO4 nuclear localization by the PHO80–PHO85 cyclin–CDK complex. *Science* **271:** 209–212.
Ray, A. and Runge, K.W. 1998. The C terminus of the major yeast telomere binding protein Rap1p enhances telomere formation. *Mol. Cell. Biol.* **18:** 1284–1295.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Ross-Macdonald, P., Sheehan, A., Roeder, G.S., and Snyder, M. 1997. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* **94:** 190–195.
Ross-MacDonald, P., Coelho, P.S.R., 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.
Seifert, H.S., Chen, E.Y., So, M., and Heffron, F. 1986. Shuttle mutagenesis: A method of transposon mutagenesis for *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* **83:** 735–739.
Silver, P.A. 1991. How proteins enter the nucleus. *Cell* **64:** 489–497.
Simpson, J.C., Wellenreuther, R., Poustka, A., Pepperkok, R., and Wiemann, S. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Reports* **1:** 287–292.
Tong, A.H.Y., 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.
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.

van Steensel, B., Delrow, J., and Henikoff, S. 2001. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27:** 304–308.

Vogel, J., Drapkin, B., Oomen, J., Beach, D., Bloom, K., and Snyder, M. 2001. Phosphorylation of γ-tubulin regulates microtubule organization in budding yeast. *Dev. Cell* **1:** 621–631.

Washburn, M.P., Wolters, D., and Yates III, J.R. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19:** 242–247.

Waters, M.G., Griff, I.C., and Rothman, J.E. 1991. Proteins involved in vesicular transport and membrane fusion. *Curr. Opin. Cell. Biol.* **3:** 615–620.

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.

Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A., Klemic, K.G., Smith, D., Gerstein, M., Reed, M.A., and Snyder, M. 2000. Analysis of yeast protein kinases using protein chips. *Nat. Genet.* **26:** 283–289.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. 2001. Global analysis of protein activities using proteome chips. *Science* **293:** 2101–2105.