Global analysis of protein localization in budding yeast

Won-Ki Huh[1]*, James V. Falvo[1]*, Luke C. Gerke[1], Adam S. Carroll[1], Russell W. Howson[1], Jonathan S. Weissman[1,2] & Erin K. O'Shea[1]

[1]*Howard Hughes Medical Institute, University of California–San Francisco, Department of Biochemistry and Biophysics, and* [2]*Department of Cellular and Molecular Pharmacology, 600 16th Street, San Francisco, California 94143-2240, USA*

* These authors contributed equally to this work

A fundamental goal of cell biology is to define the functions of proteins in the context of compartments that organize them in the cellular environment. Here we describe the construction and analysis of a collection of yeast strains expressing full-length, chromosomally tagged green fluorescent protein fusion proteins. We classify these proteins, representing 75% of the yeast proteome, into 22 distinct subcellular localization categories, and provide localization information for 70% of previously unlocalized proteins. Analysis of this high-resolution, high-coverage localization data set in the context of transcriptional, genetic, and protein–protein interaction data helps reveal the logic of transcriptional co-regulation, and provides a comprehensive view of interactions within and between organelles in eukaryotic cells.

Eukaryotic cells are organized into a complex network of membranes and compartments, which are specialized for various biological functions. Comprehensive knowledge of the location of proteins within these cellular microenvironments is critical for understanding their functions and interactions; this requires assaying the cell's full complement of proteins. The complete genome sequence of the budding yeast *Saccharomyces cerevisiae*[1] coupled with high-throughput experimental techniques has made systematic analyses of a eukaryotic proteome feasible. Recent studies have taken a genome-wide approach to analysing messenger RNA abundance and stability[2,3], biochemical activity[4,5], protein–protein interactions[6–9], transcriptional regulation[10], gene disruption phenotypes[11–14] and protein abundance[15].

Previous large-scale analyses of protein localization in *S. cerevisiae* have depended on transposon-mediated random epitope tagging and plasmid-based overexpression of epitope-tagged proteins[11,16]. However, epitope tagging of partial open reading frames (ORFs) can interrupt important localization signals, and overexpression of proteins may saturate intracellular transport mechanisms, leading to abnormal subcellular localization. To circumvent these potential problems, we generated a yeast strain collection expressing full-length proteins, tagged at the carboxy terminal end with green fluorescent protein (GFP), from their endogenous promoters by inserting the coding sequence of *Aequorea victoria* GFP (S65T)[17] in-frame immediately preceding the stop codon of each ORF. With this strategy, wild-type levels and patterns of protein expression are minimally perturbed. Furthermore, because GFP fluorescence does not require external cofactors, GFP signal can be monitored in living cells without disrupting cellular integrity. We have analysed this strain collection using fluorescence microscopy to comprehensively characterize protein subcellular localization in a simple eukaryotic cell.

Construction and analysis of a GFP-tagged library

We systematically tagged each ORF in its chromosomal location through oligonucleotide-directed homologous recombination (Fig. 1a). For each of the 6,234 annotated ORFs[18] a pair of oligonucleotides was generated that had homology to the desired chromosomal insertion site at the 5′ end of each primer and homology to a vector containing the GFP tag at the 3′ end. These primers were used to amplify the GFP tag and an auxotrophic marker from a plasmid template[19], and the resulting polymerase chain reaction (PCR) products were transformed into a haploid yeast strain. Transformants were assayed by genomic PCR with one primer specific for the GFP tag and a second specific for each ORF, to determine whether the cassette had integrated at the appropriate locus. A total of 6,029 strains with chromosomally GFP-tagged ORFs were grown to mid-logarithmic phase in synthetic medium and analysed by fluorescence microscopy; 4,156 of these showed GFP signals above background levels (Table 1).

Micrographs of each GFP-tagged strain (Fig. 1b; see also Supplementary Fig. S1), lacking ORF identifiers, were independently evaluated by two scorers and initially classified into one or more of 12 subcellular localization categories (Table 2). We then refined these categories by performing a series of co-localization experiments. Haploid reference strains expressing monomeric red fluorescent protein (mRFP)[20] fusions to proteins whose localization had been characterized previously (Table 2) were mated to approximately 700 GFP strains that were not assigned definitive localizations by GFP microscopy alone, and the resulting diploid cells were analysed by fluorescence microscopy (Fig. 1c). On the basis of this analysis, proteins were assigned to an additional 11 localization categories (Table 2). All information was captured into a database (http://yeastgfp.ucsf.edu).

Subcellular localization of yeast proteins

The 4,156 proteins for which we defined subcellular localizations in the GFP library represent 75% of the yeast proteome[15,21,22]. Our results provide localization data for about 70% of previously unlocalized yeast proteins, constituting about 30% of the proteome (Fig. 2a). Over 90% of the proteins visible in the GFP collection were also detected by western blot analysis of a collection of TAP (tandem affinity purification)-tagged strains[15], suggesting that the false-positive rate in this study is extremely low.

The distribution of protein subcellular localization reveals that, as expected, many proteins are found in the nucleus or cytoplasm, whereas 1,839 proteins, 44% of the total observed, localize to other specific subcellular regions (Fig. 2b). Notably, over 40% of the proteins that we assigned to the cytoplasm, late Golgi/clathrin and lipid particle represent new localization assignments. There are limitations to the subcellular localizations in yeast discernible by fluorescence microscopy; for example, we cannot distinguish kinetochore versus spindle pole body, or membrane versus lumen for mitochondria or the endoplasmic reticulum. However, use of the GFP tag and co-localization with RFP-tagged reference proteins allowed us to resolve many related subcellular compartments with

confidence. For example, the nucleus, nuclear periphery and the endoplasmic reticulum are distinct (Fig. 1b, top row), as are the vacuole and vacuolar membrane, and multiple compartments of the secretory pathway. This level of precision greatly facilitates our assignment of protein localization as well as integration with other genome-wide data sets.

Previously published localization data from the *Saccharomyces* Genome Database (SGD)[18], including data from earlier large-scale studies[11,16], were available for a total of 2,526 proteins visible in the GFP library—we found that there was 80% agreement between our data and those of the SGD. We also found that our localization assignments generally agree with those of the pioneering studies of the Snyder laboratory[11,16]. However, for those assignments that differ, our results show closer agreement with the SGD (Supplementary Fig. S2). Direct comparison between our data and the results of a mass spectrometric analysis of the nuclear pore complex[23] (NPC) revealed that, of 29 identified NPC components, 25 were visible in our study: 23 proteins (92%) were localized to the nuclear periphery and one each was localized to the nucleus/cytoplasm and endoplasmic reticulum. Furthermore, of 16 spindle-pole-body components identified by mass spectrometry[24], all 14 of the proteins visible in this study were localized to the spindle pole. We found an additional 20 proteins localized to the nuclear periphery and 14 to the spindle pole; of these, 11 had not been detected previously in the nuclear periphery and 7 had not been detected in the spindle pole (Supplementary Table S1). The strong correlation between the data we obtained by fluorescence microscopy and localization data obtained by other methods supports the reliability of this study in defining new protein localizations.

A potential source of discrepancy between our data and those from other studies is that the C-terminal fusion of the GFP protein (approximately 27 kDa) may cause mislocalization through steric hindrance or interruption of critical C-terminal localization/retention sequences (Supplementary Table S2). For example, the small GTP-binding protein Ras2 was localized to the nucleus and the cytoplasm in this study, but it is known to be localized to the plasma membrane due to modification of its C terminus with palmitoyl and farnesyl groups[25]. Proteins localized to the cell wall[26] and subsets of proteins localized to the peroxisome[27] and endoplasmic reticulum[28] also contain C-terminal targeting signals, and these were often mislocalized in this study.

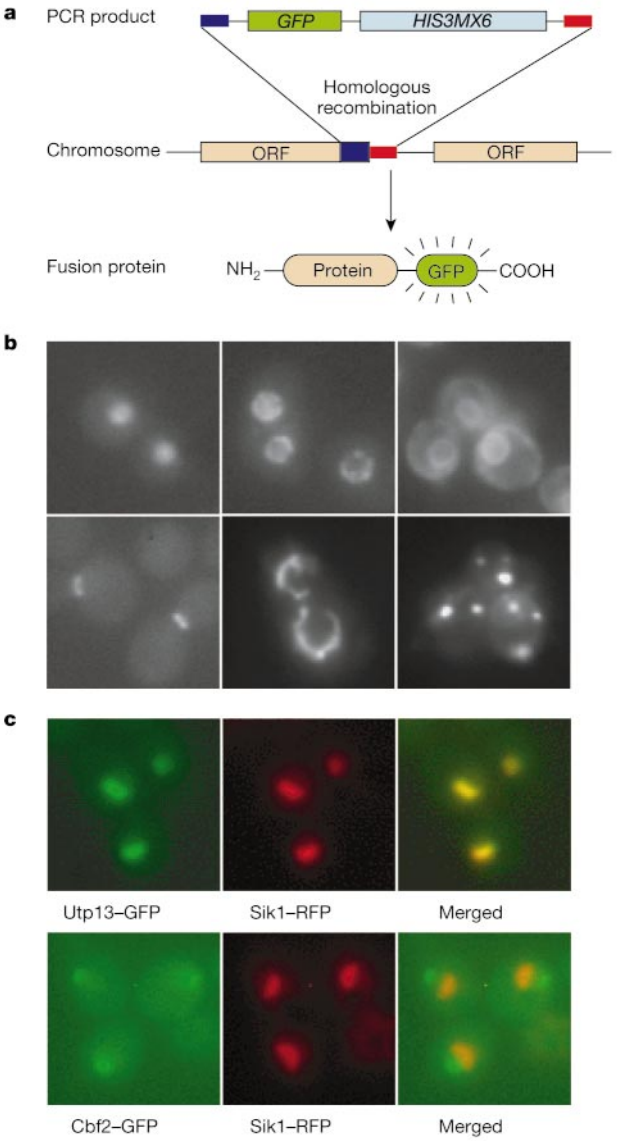

Figure 1 Microscopic analysis of yeast strains expressing GFP-tagged proteins. Data and images are accessible at http://yeastgfp.ucsf.edu. **a**, Strategy for library construction. PCR products containing the GFP tag and a selectable marker gene were inserted at the C terminus of each ORF through homologous recombination, yielding a C-terminally GFP-tagged protein. **b**, Representative GFP images of Rox3–GFP (nucleus; top left), Nic96–GFP (nuclear periphery; top middle), Pho86–GFP (endoplasmic reticulum; top right), Hof1–GFP (bud neck; bottom left), Ilv6–GFP (mitochondrion; bottom middle) and Erg6–GFP (lipid particle; bottom right). **c**, Representative co-localization experiment. A Utp13–GFP or Cbf2–GFP yeast strain was mated with a strain containing Sik1–RFP as a nucleolar marker, then fluorescence images for GFP (left) and RFP (middle) were taken and merged (right).

Organellar proteomics of the nucleolus

The identification of subsets of proteins in various organelles is an initial step towards the understanding of biological processes at the cellular level. 'Organellar proteomics' studies[29] would benefit especially from the comprehensive localization data for yeast proteins provided by this study. For example, we detected 164 proteins in the nucleolus in this study; 82 of these overlap the 127 nucleolar proteins catalogued in the SGD, but another 82 are newly defined (Fig. 2c). Of the remaining 45 nucleolar proteins from the SGD, 28 were not visualized in our study, whereas the others were localized to the nucleus (7 proteins), nucleus/cytoplasm (7 proteins), nuclear periphery (2 proteins) and cytoplasm (1 protein). These proteins may occupy the nucleolus in a transient fashion, at levels not detectable by our methods, or under conditions distinct from those of our study—mislocalization may also result from the GFP tag. A number of the nucleolar proteins found in this study are involved in ribosomal RNA transcription and processing and in ribosome biogenesis, in accordance with the classical role of the nucleolus; for some of these proteins, we provide the first direct demonstration that they reside or are enriched in the nucleolus (Fig. 2d). Given that some nucleolar proteins are involved in cell cycle control and gene regulation[30–33], it will be very interesting to investigate the functional roles of nucleolar proteins newly defined in this study.

It has been reported that essential proteins and orthologues are

Table 1 **Summary of GFP-tagged library construction**

ORF category	Number of ORFs	Success rate (%)
ORFs processed for PCR of the GFP cassette	6,234* (1,100)	100 (100)
ORFs with successful transformation	6,151 (1,018)	99 (93)
ORFs with positive homologous recombination confirmed by genomic PCR	6,029 (953)	97 (87)
ORFs with positive GFP signal	4,156 (827)	67 (75)

Values for essential ORFs are indicated in parentheses.
*ORFs annotated in SGD, 17 April 2001.

enriched in related protein complexes isolated from yeast and humans[8]. Of the proteins localized to the nucleolus in this study, 99 proteins (60%) are known to be essential, substantially more than the 20% required for viability in the proteome as a whole[12,14]. Recently, mass spectrometric analysis of the human nucleolus identified 271 proteins, 166 of which have homologues in yeast[34]; 52 of these proteins are classified as nucleolar in this study (Supplementary Table S3). Of the 112 proteins remaining from the 164 proteins that we have detected in the yeast nucleolus, 73 have human homologues and 33 of these are localized to the nucleolus or have biological functions related to transcription and processing of rRNAs and ribosome biogenesis (Supplementary Table S4) according to the Human Proteome Survey Database[35]. Given the enrichment of essential proteins in the yeast nucleolus and the enrichment of essential proteins and orthologues in related protein complexes from yeast and humans[8], we expect that many of the remaining human homologues of yeast proteins detected in the nucleolus in this study will also be nucleolar proteins.

Protein localization and mRNA co-expression

Many genome-wide analyses have demonstrated that mRNA transcript expression patterns are similar for groups of functionally related genes[2,36–38]; mRNA abundance is also similar within certain cellular compartments[39]. However, transcriptional co-regulation has not been directly compared to subcellular protein localization on a proteome-wide scale. To assess the extent of this correlation, we made use of a study that identified 33 transcriptional 'modules' of genes with marked co-regulation based on analysis of over 1,000 microarray data sets reflecting the results of different mutant strain backgrounds or environmental perturbations[38,40]. For each module, the fraction of proteins with a given subcellular localization was calculated and divided by that fraction in the whole proteome to generate fold enrichment in each subcellular localization category (Fig. 3a). We obtained statistically significant enrichments (one-sided binomial test with $P < 0.05$) for 19 of the 22 most highly expressed modules, indicating that co-localization is strongly correlated with transcriptional co-expression and, by extension, with biological function.

The combination of protein localization and transcriptional co-expression can be used to corroborate or predict the function of unnamed ORFs in a specific module. For example, YGL068W and YNL122C, both of which belong to the mitochondrial ribosomal protein transcriptional module, localize to the mitochondrion in our study, as do 13 other members of this module, strongly supporting the function predicted by the module (Fig. 3b). Indeed, the sequence of YGL068W shows 49% similarity to that of the human mitochondrial ribosomal protein L12 (ref. 41).

Localization and co-regulation data can also be used to gain insight into biological function when proteins in a given transcription module are enriched in more than one localization category. This allows us to subdivide sets of co-expressed proteins, providing a level of information that cannot be gleaned solely from their classification in the same module based on their expression profiles. For example, proteins in the G1 module (representing processes coordinated at the G1/S transition) localize to three basic categories: nucleus, bud/bud neck and spindle pole (Fig. 3c). The basic functions of proteins in the G1 module, where known, can be divided by localization; proteins localized to the bud/bud neck are involved in bud formation, whereas nuclear proteins from this module are involved mainly in chromosome cohesion, transcription, and DNA replication, repair and recombination. Thus, given

Table 2 **Subcellular localization categories used in this study**

GFP localization	GFP and RFP co-localizaton
Cell periphery	Cytoskeleton
Bud	Actin cytoskeleton
Bud neck	Spindle pole
Cytoskeleton	Nucleolus
Microtubule	Nuclear periphery
Cytoplasm	Golgi apparatus
Nucleus*	Transport vesicle
Mitochondrion*	Early Golgi/COPI
Endoplasmic reticulum	Late Golgi/clathrin
Vacuole	Endoplasmic reticulum to Golgi
Vacuolar membrane	Endosome
Punctate	Peroxisome
Ambiguous	Lipid particle

Co-localization experiments used the following RFP-tagged markers: Sac6 (actin cytoskeleton), Spc42 (spindle pole), Sik1 (nucleolus), Nic96 (nuclear periphery), Anp1 (Golgi apparatus), Cop1 (early Golgi/COPI), Chc1 (late Golgi/clathrin), Sec13 (endoplasmic reticulum to Golgi), Snf7 (endosome), Pex3 (peroxisome) and Erg6 (lipid particle).
*Initially assigned by DAPI staining; some punctate proteins were subsequently confirmed as mitochondrial using MitoTracker red.

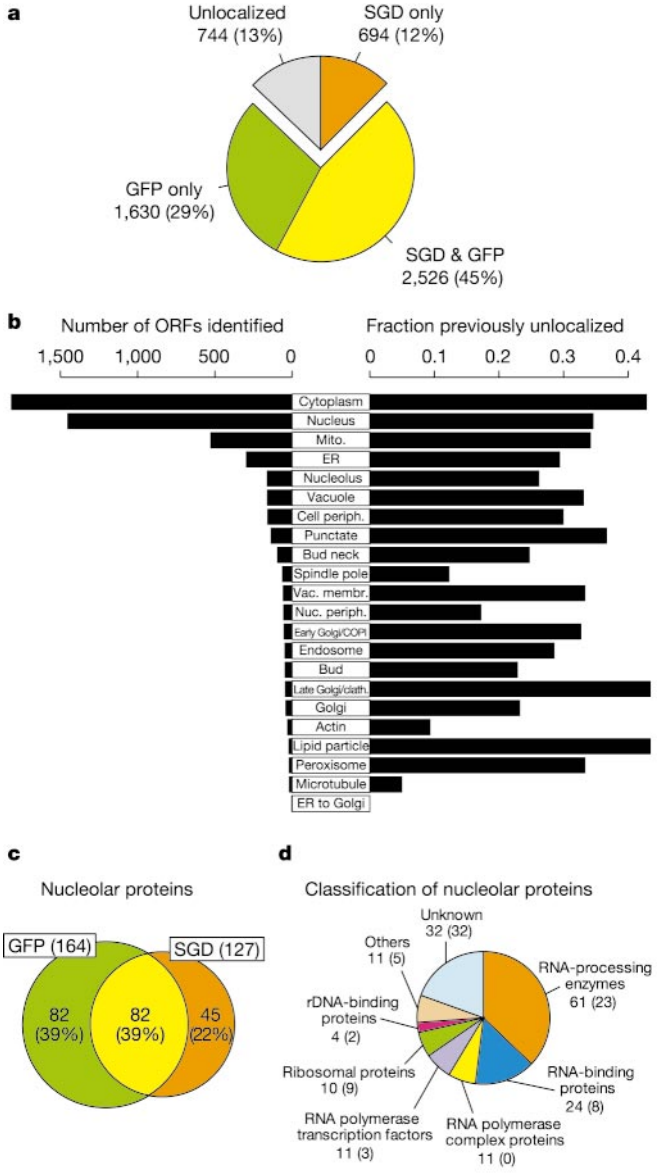

Figure 2 Subcellular localization of yeast proteins. **a**, Contribution of this study to expanding localization data for the yeast proteome. Six hundred and forty ORFs thought to be spurious[15] were excluded from the pie chart. **b**, Distribution of the subcellular localizations for proteins visualized in this study. Also shown is the fraction of previously unlocalized proteins in each category. ER, endoplasmic reticulum. **c**, Comparison of the nucleolar proteins identified from this study with those from SGD. **d**, Functional classification of 164 nucleolar proteins defined in this study[18]. The numbers of ORFs with previously unknown subcellular localizations are shown in parentheses.

that the G1 module proteins Hif1, Hsn1, YGR151C, YKR077W and YMR144W are localized to the nucleus, it is likely that they share the functions of nuclear proteins from this module.

Comparison with genetic and physical interactions

Recent genome-wide studies have sought to enumerate all protein–protein interactions that occur in *S. cerevisiae*[6–9]. Despite the large scale of these efforts, the agreement between studies[42] suggests that total coverage is poor and false-positive rates remain high. To interact physically proteins must exist in close proximity, at least transiently, suggesting that co-localization may be an effective means for evaluating hypothetical interactions. To assess the relationship between co-localization and interaction, we chose as a reference set the sum of all genetic and protein–protein interactions reported in the GRID database[43]. Although this set is certain to contain a considerable fraction of false-positive interactions, it was chosen to minimize systematic bias in individual screens that inevitably results from alternative interaction detection methods. We determined the subcellular localizations of each interacting protein pair from this reference set and the fraction of the total number of interactions occurring for each localization pair. A set of randomized protein pairs was also generated from the whole proteome, and localization pair statistics were collected on this set in the same way. We calculated the fold enrichment observed for each localization pair in our reference data set as compared with the randomized data set to generate an interaction matrix (Fig. 4a).

This analysis supports and extends interaction data from other studies. As expected, interactions are strongly enriched between proteins that co-localize (one-sided binomial test with $P < 0.001$), but the degree of enrichment varies widely by compartment. For example, interactions between cytoplasmic proteins are 1.3-fold enriched above chance, whereas interactions between microtubule proteins are 56-fold enriched above chance, implying that co-localization of two putative interacting proteins to the microtubule cytoskeleton provides better evidence of physical and functional interaction than the simple fact that they do co-localize.

Of particular interest is that enrichment in interactions was observed between distinct localization categories in our study, shown as red off-diagonal circles in the matrix (Fig. 4a). Such off-diagonal circles are indicative of functional relationships between subcellular localizations: they are neither the result of systematic errors from individual interaction data sets (Supplementary Fig. S3) nor of proteins being assigned multiple localizations (Supplementary Fig. S4). An extensive network of interactions occurs between proteins localized to the actin cytoskeleton and those localized to the bud neck (Fig. 4b). Notably, some proteins not previously known to localize to these regions have known or predicted functions consistent with their assigned localizations. The Rho-GTPase activator protein Bem2, for example, has been shown through genetic studies to be involved in bud growth, establishment of cell polarity, and organization and biogenesis of the actin and microtubule cytoskeleton and of the cell wall[44]. These functions are consistent both with localization to the bud neck and functional interaction with the actin cytoskeleton. Similarly, Chs7 is involved in cell wall chitin biosynthesis[45], consistent with the role of certain bud neck proteins[46], and Akl1 has a predicted role in the organization of the actin cytoskeleton[47].

The biological importance of statistically significant interactions between localization categories in the GFP library is fully revealed when these interactions are considered in the context of a eukaryotic cell (Fig. 4c). A network of interactions connects subcellular regions that are functionally and physically related. Strongly interconnected localizations can reflect dynamic interchange of proteins between compartments; for example, compartments of the secretory path-

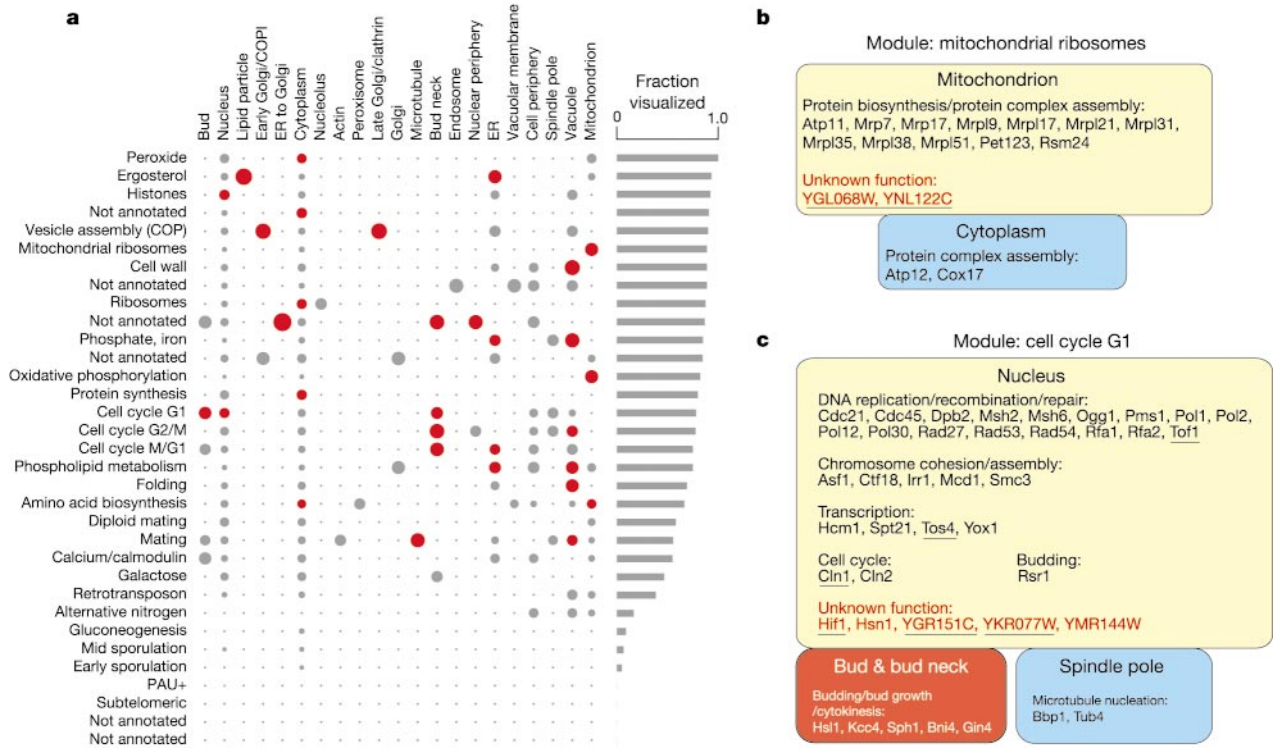

Figure 3 Correlation between transcriptional co-regulation and subcellular localization. **a**, The enrichment of subcellular localizations (top) exhibited by proteins belonging to each transcription module[38] (left) designated by biological function where applicable. Log(enrichment) is proportional to the radius of the circles shown; red circles indicate >95% confidence (one-sided binomial test) that enrichment >1. The fraction of proteins visualized by GFP within each module is indicated by the bar graph (right). **b**, **c**, Diagrams of the proteins of the mitochondrial ribosome (**b**) and G1 (**c**) transcriptional modules divided into principal localizations observed in the GFP library, and further grouped by previously defined biological functions[18]. Proteins newly localized in this study are underlined, and proteins with unknown biological function are noted in red.

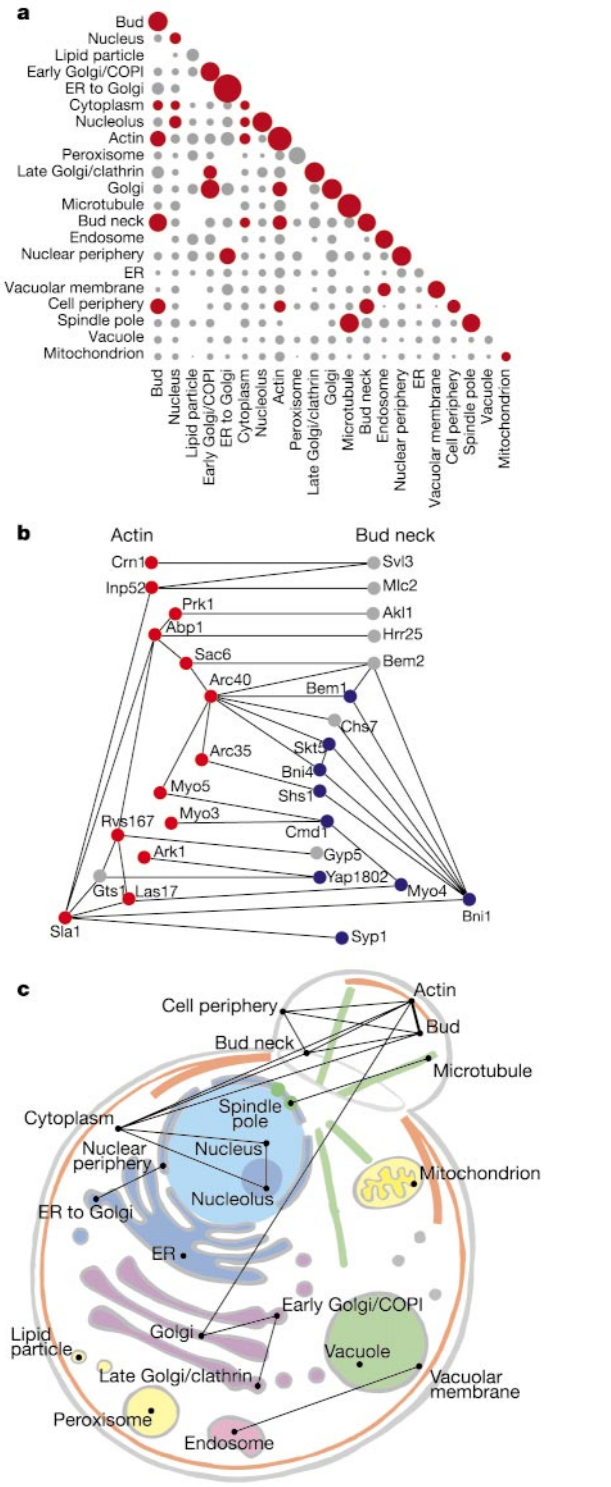

Figure 4 Relationship between genetic and physical interactions and subcellular localization. **a**, Localization pairs were extracted from the localization of interacting partners. Interaction between specific compartments was compared to a randomized interaction set, showing that some compartments interact preferentially with others. Log(enrichment) is proportional to the radii of the circles; red circles have >99.9% confidence (one-sided binomial test) that enrichment >1. ER, endoplasmic reticulum. **b**, Diagram[48] of genetic and protein–protein interactions represented by the off-diagonal enrichment between the actin cytoskeleton (left, red circles) and bud neck (right, blue circles). Grey circles denote proteins not localized previously to either category. **c**, Diagram of the off-diagonal, statistically significant interacting compartments in **a**.

way (Golgi, early Golgi/COPI (coat protein I) and late Golgi/clathrin). Intercompartmental interactions can also reflect close proximity and extensive physical association between localization categories, as is the case for the bud neck, the bud and the actin cytoskeleton. The interaction matrix provides an overview of communication between subcellular compartments as well as a template for evaluating the validity of protein–protein interactions from large-scale experimental or theoretical data sets.

Discussion

By creating a GFP-tagged yeast strain collection and database that covers three-quarters of the proteome and over two-thirds of previously unlocalized proteins, we have provided an experimental and informational resource to the scientific community. Although we have presented an analysis of the yeast proteome in a nominal resting state, the GFP library serves as a starting point for understanding the complex state of flux in the eukaryotic proteome that underlies the survival and development of an organism. The library provides a tool for analysing the global dynamics of the proteome in response to specific external stimuli or growth conditions over a selected period of time. Similarly, the library can be used in combination with high-throughput strain construction techniques[13] to assay the effects of deletion or mutation of a protein of interest on global protein localization. Complex regulatory networks responsible for targeting proteins to specific cellular compartments can thus be systematically dissected.

We have shown that the combination of high-resolution, high-accuracy, proteome-wide localization information with data from other proteomics-scale studies provides an independent dimension of information that reveals patterns not visible within a single data set. The localization data from the GFP library can confirm and extend predictions based on trends within a single data set; if proteins grouped together in a given data set have a common localization, the prediction of common function is strengthened. This is particularly useful in the case of proteins for which little functional data exists. The localization data also make it possible to subdivide groups of proteins related by genome-wide trends in other data sets, indicating that one group may be composed of subsets of proteins with even more specific, separate biological roles. A comparative proteomics approach promises to reveal important features of basic cellular processes, improving our understanding of *S. cerevisiae* and of the proteins and pathways conserved among eukaryotes. □

Methods

Construction of GFP-tagged yeast strains

To construct a chromosomally GFP-tagged library, 6,234 pairs of gene-specific oligonucleotide primers were synthesized, each of which had been designed to share complementary sequences to the GFP tag-marker cassette at the 3′ end and contain 40 base pairs (bp) of homology with a specific gene of interest to allow in-frame fusion of the GFP tag at the C-terminal coding region of the gene. Gene-specific cassettes containing a C-terminally positioned GFP tag were then generated by PCR using as a template pFA6a–GFP(S65T)–His3MX, which contains the *Schizosaccharomyces pombe his5*$^{+}$ gene and permits selection of transformed strains in histidine-free media[19]. The haploid parent yeast strain (ATCC 201388: *MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) was transformed with the PCR products, and strains were selected in SD medium (synthetic medium plus dextrose, Difco) lacking histidine. Insertion of the cassette by homologous recombination was verified by genomic PCR of samples from individual colonies with a primer internal to the GFP tag and a separate set of ORF-specific primers designed to produce a product of approximately 500 bp. Strains representing 6,029 ORFs were successfully tagged with GFP (Table 1), and independent strains from two to six selected colonies from each ORF were analysed by fluorescence microscopy.

Microscopic imaging of GFP-tagged strains

Aliquots of strains grown to mid-logarithmic phase in SD medium lacking histidine were analysed in 96-well glass-bottom microscope slides (BD Falcon) pre-treated with concanavalin A (50 μg ml^{-1}) to ensure cell adhesion. Cells were incubated in SD medium containing 1 μg ml^{-1} 4′,6-diamidino-2-phenylindole (DAPI) as a marker for the nucleus and mitochondria, and analysed by multiple wavelength fluorescence and visible light microscopy with a digital imaging-capable Nikon TE200/300 inverted microscope using an oil-immersed objective at ×100 magnification. Using a script in MetaMorph version

4.6r8 imaging software (Universal Imaging Corporation), fluorescence microscopy images for GFP, DAPI and Nomarski/DIC (differential interference contrast) images were taken in rapid succession, and the stage was automatically advanced between wells on the 96-well slide.

Localization category refinement by co-localization

Subcellular localizations that could not be assigned readily by GFP fluorescence alone—typically classified as punctate or non-uniform nuclear—were resolved by mating the GFP-tagged strains to strains expressing reference proteins (Table 2) fused to mRFP[20]. The coding sequence for mRFP was amplified by PCR from its parent vector (pRSET–mRFP1)[20] and inserted into pFA6a–KanMX6, which carries the *Escherichia coli* kanamycin-resistance gene[19], to create the plasmid pFA6a–mRFP–KanMX6. This vector was then used to generate gene-specific cassettes to yield reference strains expressing C-terminally mRFP-tagged proteins. The haploid parent strain (ATCC 201389: *MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) was transformed, selected in the presence of G418 sulphate (200 $\mu g\,ml^{-1}$), and analysed for positive RFP signal by fluorescence microscopy as described above. The GFP-tagged strains were then mated with at least one of the mRFP-tagged reference strains in SD medium lacking lysine and methionine, and the resulting diploid strains were analysed by microscopy to generate GFP, RFP, DIC and GFP–RFP merged images. Haploid strains exhibiting potential non-uniform mitochondrial GFP patterns were subjected to the same microscopic analysis using the mitochondrion-specific dye MitoTracker red CMXRos (Molecular Probes).

Database features

We have designed a publicly available web-based user interface to the localization database at http://yeastgfp.ucsf.edu. At this site, users can perform searches using a number of criteria, including ORF name, gene name, subcellular localization, cell cycle, cell morphology, cell–cell brightness variability and subcellular signal heterogeneity. Searches retrieve full-sized, lossless compressed images that were used to assign localizations in this study; specific cells used to justify localizations are indicated in the images.

Comparison with other data sets

The distribution of subcellular localizations exhibited by the test ORF set (groups of transcriptionally co-regulated genes[38,40]) was assessed in comparison to a reference set (the localization distribution seen in all ORFs characterized in this study). The identity of genes in the modules can be found at http://barkai-serv.weizmann.ac.il/modules/page/details.html using a threshold cutoff of 4.0. The frequency with which each subcellular localization is observed in the test and reference set was calculated; the ratio of these frequencies is reported as the enrichment. Individual binomial tests were performed for each subcellular localization to accept or reject the null hypothesis that the measured enrichment occurred due to chance. A one-tailed P-value <0.05 is taken to be statistically significant and is indicated by red circles in Fig. 3a. Distribution of subcellular localization of interacting partners was assessed by comparison to that which would occur by random association of ORFs, giving an enrichment of interactions between localizations. Individual binomial tests confirm that enrichment for certain localization pairs, indicated in Fig. 4a by red circles, is not the product of sampling error ($P < 0.001$).

Received 28 July; accepted 1 September 2003; doi:10.1038/nature02026.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).
2. Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88,** 243–251 (1997).
3. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl Acad. Sci. USA* **99,** 5860–5865 (2002).
4. Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286,** 1153–1155 (1999).
5. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293,** 2101–2105 (2001).
6. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
7. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
8. Gavin, A.-C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
9. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183 (2002).
10. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804 (2002).
11. Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402,** 413–418 (1999).
12. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).
13. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294,** 2364–2368 (2001).
14. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418,** 387–391 (2002).
15. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
16. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16,** 707–719 (2002).
17. Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem.* **67,** 509–544 (1998).
18. Dolinski, K. *et al.* Saccharomyces *Genome Database* ⟨http://www-genome.stanford.edu/Saccharomyces⟩ (2003).
19. Longtine, M. S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14,** 953–961 (1998).
20. Campbell, R. E. *et al.* A monomeric red fluorescent protein. *Proc. Natl Acad. Sci. USA* **99,** 7877–7882 (2002).
21. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
22. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
23. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651 (2000).
24. Wigge, P. A. *et al.* Analysis of the *Saccharomyces* spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. *J. Cell Biol.* **141,** 967–977 (1998).
25. Bhattacharya, S., Chen, L., Broach, J. R. & Powers, S. Ras membrane targeting is essential for glucose signaling but not for viability in yeast. *Proc. Natl Acad. Sci. USA* **92,** 2984–2988 (1995).
26. van Berkel, M. A., Caro, L. H., Montijn, R. C. & Klis, F. M. Glucosylation of chimeric proteins in the cell wall of *Saccharomyces cerevisiae*. *FEBS Lett.* **349,** 135–138 (1994).
27. Gould, S. J. *et al.* Peroxisomal protein import is conserved between yeast, plants, insects and mammals. *EMBO J.* **9,** 85–90 (1990).
28. Pelham, H. R., Hardwick, K. G. & Lewis, M. J. Sorting of soluble ER proteins in yeast. *EMBO J.* **7,** 1757–1762 (1988).
29. Taylor, S. W., Fahy, E. & Ghosh, S. S. Global organellar proteomics. *Trends Biotechnol.* **21,** 82–88 (2003).
30. Shou, W. *et al.* Exit from mitosis is triggered by Tem1-dependent release of the protein phosphatase Cdc14 from nucleolar RENT complex. *Cell* **97,** 233–244 (1999).
31. Visintin, R., Hwang, E. S. & Amon, A. Cfi1 prevents premature exit from mitosis by anchoring Cdc14 phosphatase in the nucleolus. *Nature* **398,** 818–823 (1999).
32. San-Segundo, P. A. & Roeder, G. S. Pch2 links chromatin silencing to meiotic checkpoint control. *Cell* **97,** 313–324 (1999).
33. Rabitsch, K. P. *et al.* Kinetochore recruitment of two nucleolar proteins is required for homolog segregation in meiosis I. *Dev. Cell* **4,** 535–548 (2003).
34. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12,** 1–11 (2002).
35. Hodges, P. E. *et al.* Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics. *Nucleic Acids Res.* **30,** 137–141 (2002).
36. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686 (1997).
37. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728 (1998).
38. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nature Genet.* **31,** 370–377 (2002).
39. Drawid, A., Jansen, R. & Gerstein, M. Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* **16,** 426–430 (2000).
40. Bergmann, S., Ihmels, J. & Barkai, N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E* **67,** 031902 (2003).
41. Koc, E. C. *et al.* The large subunit of the mammalian mitochondrial ribosome. *J. Biol. Chem.* **276,** 43958–43969 (2001).
42. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403 (2002).
43. Breitkreutz, B.-J., Stark, C. & Tyers, M. The GRID: the General Repository for Interaction Datasets. *Genome Biol.* **4,** R23 (2003).
44. Madden, K. & Snyder, M. Cell polarity and morphogenesis in budding yeast. *Annu. Rev. Microbiol.* **52,** 687–744 (1998).
45. Trilla, J. A., Duran, A. & Roncero, C. Chs7p, a new protein involved in the control of protein export from the endoplasmic reticulum that is specifically engaged in the regulation of chitin synthesis in *Saccharomyces cerevisiae*. *J. Cell Biol.* **145,** 1153–1163 (1999).
46. DeMarini, D. J. *et al.* A septin-based hierarchy of proteins required for localized deposition of chitin in the *Saccharomyces cerevisiae* cell wall. *J. Cell Biol.* **139,** 75–93 (1997).
47. Cope, M. J., Yang, S., Shang, C. & Drubin, D. G. Novel protein kinases Ark1p and Prk1p associate with and regulate the cortical actin cytoskeleton in budding yeast. *J. Cell Biol.* **144,** 1203–1218 (1999).
48. Breitkreutz, B.-J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biol.* **4,** R22 (2003).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank F. Lam for designing the original relational data model and for assistance with database organization; M. Springer for sharing microscopy expertise; J. Newman for assistance with FACS analysis; N. Barkai, S. Ghaemmaghami and K. Bower for discussions and sharing results; R. Tsien for providing the pRSET–mRFP1 plasmid; M. Levin for assistance with preparation of RFP-tagged strains; R. Marion for contributing GFP-tagged strains of CAK1, PHO4, RTS2, SET1 and STP4; A. DePace for preparing the illustration in Fig. 4c; D. Ahern, F. Sanchez, A. Belle and M. Liku for primer synthesis and other technical assistance; and S. Emr and members of the O'Shea and Weissman laboratories for critical discussion of the work. This work was supported by the Howard Hughes Medical Institute and the David and Lucile Packard Foundation (J.S.W. and E.K.O.). J.V.F. is the recipient of a Ruth L. Kirschstein National Research Service Award. The database and strain collection information is accessible at http://yeastgfp.ucsf.edu.

Authors' contributions Strain construction and analysis was performed by W.-K.H. and J.V.F., bioinformatics analysis and database management by L.C.G., with additional database management by A.S.C., and oligonucleotide primer design by R.W.H.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.K.O. (oshea@biochem.ucsf.edu).