UCSF

Title
Identifying substrates of the Pho85-Pcl1 kinase

Permalink
https://escholarship.org/uc/item/7zh073h0

Author
Dephoure, Noah Elias

Publication Date
2006

Peer reviewed|Thesis/dissertation

Identifying Substrates of the Pho85-Pcl1 Kinase

by

Noah Elias Dephoure

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

For me

Acknowledgements

I thank Steve Weinstein, for inviting me to join his lab and giving me my first taste of laboratory science.

I thank Erin O'Shea for her direction, wisdom, excitement and for her exceptional clarity of thought and expression.

I thank the remaining members of my committee, Dave Morgan and Kevan Shokat for their advice and guidance.

I am grateful to Jeff Ubersax whose work and advice were essential to my own success.

I thank the National Institutes of Health and the Howard Hughes Medical Foundation for their financial support of Erin's lab and of this work.

I thank O'Shea Lab members past and present for their constant help, advice, entertainment, and intellectual stimulation.

I thank John Wiley & Sons, Ltd. and the National Academy of Sciences for allowing me to publish manuscripts which appeared as articles in their journals in this dissertation.
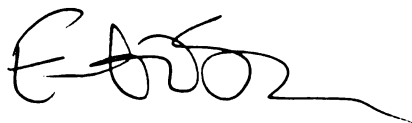
Abstract

Identifying Substrates of the Pho85-Pcl1 Kinase

Noah Elias Dephoure

Phosphorylation is a ubiquitous protein modification important for regulating nearly every aspect of cellular biology. Identifying the substrates of a kinase is essential for understanding its cellular role, but doing so remains a difficult task. Pho85 is a non-essential yeast cyclin dependent kinase which when bound to the cyclin Pho80 it plays a well characterized role in the cellular response to phosphate starvation. However, cells lacking PHO85 have additional phenotypes unrelated to the phosphate starvation pathway, including a role in the G1 phase of the cell-division cycle. Ten different Pho85 cyclins, termed Pcl's, associate with Pho85 and are thought to direct it to phosphorylate distinct substrates involved in different cellular processes. To better understand the role of the Pho85-Pcl1 cyclin dependent kinase, I performed a high-throughput screen of the yeast proteome for kinases substrates. To do so, I helped plan, construct, and verify an epitope tagged protein library composed of 4250 yeast strains each harboring a distinct detectable fusion protein. Using this library and an AS version of Pho85, Pho85(F82G), which functionally substitutes for wild-type Pho85 *in vivo*, and which is able to utilize the ATP analog N^6-benzyl ATP *in vitro*, I developed a high-throughput method to screen for protein kinase substrates and used it to identify 24 direct substrates of the yeast protein kinase Pho85-Pcl1, including the known substrate Rvs167. The power of this method to identify true kinase substrates is strongly supported by functional overlap and

colocalization of candidate substrates and the kinase as well as by the specificity of

Pho85-Pcl1 for some of the substrates over another Pho85-cyclin kinase complex. This

method is readily adaptable to other yeast kinases.

Table of Contents

List of Tables

List of Figures

ix

x

Chapter 1

Methods for Identifying Protein Kinase Substrates

The Importance of Protein Kinases

Since the identification of the first protein kinase by Krebs and Fischer in 1956[1], it has become clear that reversible phosphorylation is a ubiquitous mechanism of cellular regulation. Database homology has identified over 120 kinase in budding yeast [2], and over 500 in the human genome [3]. This constitutes ~2% of the protein coding genes making protein kinases one of the largest gene families [3].

Kinases regulate virtually every cellular process and can affect proteins in a variety of ways. Among the many roles they play are regulating the progression of the cell cycle [4], propagating intracellular signals [5], and coordinating gene expression[6]. Their action can modulate protein stability [7, 8], protein localization [9], enzyme activity[1], and protein-protein interactions[10]. Not surprisingly, defects in protein kinases have been linked to numerous and human diseases [11] including many types of cancer [12]. Deficiencies in human phosphorylase kinase are responsible for glycogen storage disease [13, 14], and misregulation of the Abelson tyrosine kinase is responsible for chronic myelogenous leukemia [15, 16]. Because of their roles in disease, protein kinases are attractive drug targets [17]. Specific protein kinase inhibitors, such as imatinib, commercially known as the cancer drug Gleevec, have already proven to be effective therapeutics [18] and many more compounds have entered clinical trials [17].

Protein kinases are highly evolutionarily conserved [19, 20]. They exist in all known eukaryotic organisms and the kinase domain [19, 20] can be rapidly identified from primary sequence data. Each kinase, or each small set of redundant kinases, has a

2

specific role performed by acting on a specific set of substrates. Identifying a kinase's

substrates is essential for a full understanding of its cellular role. The absence of a

reliable map of kinase-substrate pairs represents an enormous gap in the understanding of

cellular regulation and cell-signaling networks. Though identifying protein kinase genes

is facile, identifying the substrates of a given kinase has proven far more difficult.

Perspectives on Protein Kinase Substrate Identification

How then can we identify the substrates of a given kinase? The first clues about the

substrates of a given kinase usually come from studies of the kinase gene's function. In

many cases this is already known. Data-mining the existing literature including a rapidly

growing number of genome-wide data sets is an important first step. Gene's known to be

required for the same functions in the cell are more likely to be substrates than those that

are not. An elegant example is the case of the yeast kinases Ste7 and Ste11 first

identified in a genetic screen for mutants unresponsive to mating factor [21]. These

kinases were subsequently shown to act sequentially in a MAP kinase signaling pathway

[22]. Where this sort of information is not available classical methods for probing gene

function such as gene deletion, overexpression, genetic screens, mutant analysis, and

chemical inhibition can be used.

The identification of physical interaction partners is a powerful approach to gaining clues

about gene product function and can be used to identify possible kinase substrates.

Rvs167, a substrate of the yeast kinase Pho85 was identified in this way [23].

Techniques like the yeast-two-hybrid screen and affinity purification followed by mass

spectrometric protein identification are now being applied genome-wide [24-26]. The

wealth of data provided by this work can provide clues to substrate identity. But many, if

not most proteins interacting with a given kinase are not substrates. For example, a

search of the BioGRID database (http://www.thebiogrid.org) returns 31 published

proteins that physically interact with Pho85 and another 27 that interact with one or more

of the 10 cyclins that bind Pho85. Of these only eight [23, 27-33] have been

demonstrated by any means to be substrates, even though most of them have been

assayed directly by more than one method [34, 35]. Many kinase-substrate interactions

may be too short-lived to be detected. The inefficiency of this approach makes it of

limited use as a method for substrate identification.

Even when available, the sort of correlative information derived from genetic and

physical interaction studies is rarely sufficient to identify kinase substrates and can bias

the focus of subsequent experiments.

In vitro kinase assays

To identify a kinase substrate it is necessary to demonstrate that the kinase can directly

phosphorylate the substrate and that the substrate is dependent on the kinase activity in

vivo for phosphorylation [36]. In vivo dependency can be extremely difficult to

demonstrate, often involves using large amounts of radioactivity, and ultimately cannot

rule out the action of an intermediate kinase as the true in vivo kinase [36, 37]. Thus

most substrate identification relies on direct assays of the ability of a kinase to

phosphorylate a candidate substrate. This is most commonly done in vitro with purified

4

kinase and substrate proteins in the presence of ATP and Mg^{2+} [37]. The assay is scored

by monitoring the incorporation of phosphate. This method can be problematic because

it generates a large number of false positive results [36, 37]. This is likely due to the

nature and sensitivity of the assay. In vitro kinase reactions often use supraphysiological

concentrations of kinase and substrate proteins [36] in a grossly non-physiological setting

along with radioactively labeled ATP in a fixed time assay. These conditions can result

in detectably phosphorylated proteins even where the reaction is highly inefficient.

(author's unpublished observations and Dave Morgan, Personal Communication)

Despite these concerns, purified component reactions are still widely used to get a

foothold on the identity and character of kinase substrates because of the ease with which

they can be scaled for high-throughput [38]. Recent advances in protein chip technology

have made it possible to perform simultaneous assays on an entire proteome [34, 39].

Chips containing thousands of distinct purified proteins on a single slide are now

commercially available from Invitrogen (Carlsbad, CA) for *Saccharomyces cerevisiae*

and human proteins and likely will become available for additional model organisms.

Such chips can be incubated in a single pool of purified kinase and quickly scanned for

phosphate incorporation [34, 39]. Even under such conditions some kinases maintain

specificity for their cognate substrates. In the most comprehensive study of this type to

date, Pho4 is among the best phosphorylated proteins when assayed with Pho85-Pho80,

but is not detectably phosphorylated by the closely related kinase complexes Pho85-Pcl9,

Pho85-Pcl2, and Pho85-Pcl1 [34]. Thus, despite the caveats, in vitro approaches still

have utility in the identification of protein kinase substrates.

5

Toward Direct in vivo Substrate Identification

The difficulty in identifying protein kinase substrates results in part from the large number of kinases with sometimes overlapping specificities [38]. Given any phosphorylated protein from the cell, it is no simple task to determine which kinase phosphorylated it. The emergence of a novel chemical genetic method for inhibiting and following kinase activity has raised hopes of identifying direct kinase substrates *in vivo* [40-44]. The basis of the technique is the replacement of a conserved bulky hydrophobic residue in the ATP binding pocket that creates a cleft, allowing the kinase to accommodate and use an ATP analog with a bulky adduct. The mutation has little effect on the kinase's ability to use native ATP, and importantly, wild-type kinases are unable to use the analog [40, 45]. Such a mutant kinase is dubbed analog specific (AS). In conjunction with a radio-labeled ATP analog it becomes possible to identify the substrates of the AS kinase in the presence of other kinases. All kinases present may be actively transferring phosphate but only the AS kinase can transfer label from the analog. Any protein labeled with the radioactive phosphate was necessarily phosphorylated by the AS kinase. In theory this technique should allow unambiguous substrate identification *in vivo*. However, introduction of sufficient quantities of labeled analog into intact cells has been difficult to achieve. (Kevan Shokat, personal communication).

In lieu of examining phosphorylation in intact cells Shah and Shokat performed reactions in NIH 3T3 non-denaturing whole cell extracts containing an AS version of v-Src kinase and $[\gamma^{-32}P]$ATP analog [40]. Subsequent separation, autoradiography, and mass

6

spectrometry identified both known and novel substrates of v-Src [40]. Ubersax et. al.

used a similar approach to screen for substrates of yeast Cdk1 in extracts [42]. Rather

than rely on mass spectrometry for substrate identification, they coupled this technique to

the use of an over-expression library of yeast affinity epitope tagged fusion proteins.

From this strain library they chose a directed set of candidate substrates containing Cdk1

consensus sites as well as a control set. They performed reactions in the extracts with

purified AS kinase and labeled ATP analog, purified the tagged proteins, separated them

on polyacrylamide gels, and analyzed the phosphorylation status of the tagged proteins

with a phosphorimager. They identified ~200 candidate substrates and demonstrated that

14 of these are phosphorylated in vivo in a Cdk1 dependent manner [42]. I sought to

improve upon this technique by using endogenous substrate protein levels and by

applying it systematically to the entire proteome.

Pho85 is a non-essential yeast cyclin dependent kinase [46]. When bound to the cyclin

Pho80 it plays a well characterized role in the cellular response to phosphate starvation

[10, 46]. However, cells lacking *PHO85* have additional phenotypes unrelated to the

phosphate starvation pathway, including a role in the G1 phase of the cell-division cycle

[47, 48]. Ten different Pho85 cyclins, termed Pcl's, have been shown to associate with

Pho85 and are thought to direct it to phosphorylate distinct substrates involved in

different cellular processes [30, 49]. To better understand the role of the Pho85-Pcl1

cyclin dependent kinase, I undertook a systematic screen of the yeast proteome for

substrates, making use of an AS version of Pho85, Pho85(F82G), which functionally

7

substitutes for wild-type Pho85 *in vivo*, and which is able to utilize the ATP analog N^6-benzyl ATP *in vitro* [50].

Toward that end, I helped plan, construct, and verify an epitope tagged protein library composed of 4251 yeast strains each harboring a distinct detectable fusion protein. I worked to develop an assay amenable to high-throughput methods for screening protein kinase substrates and used the library to screen for and subsequently identify 24 direct substrates of the yeast protein kinase Pho85-Pcl1, including the known substrate Rvs167. The specificity of Pho85-Pcl1 towards some of the substrates relative to another Pho85-cyclin kinase complex and an overwhelming statistical enrichment among the substrates for certain subcellular localizations and cellular functions suggest that many of these substrates are bona fide. This method can be easily adapted to screen for substrates of nearly all yeast kinases.

Caveats

This technique relies on the production of active recombinant AS kinase which is not always possible or facile. One way around this is to use endogenous kinase as was done by Shah and Shokat [40], or perhaps to overexpress the kinase in the cells to be assayed. Another hurdle is that suitable AS versions of some kinases have been difficult to make. However, additional work in the Shokat laboratory has identified second site mutations that have allowed previously resistant kinases to work with this technique [51, 52].

The future of kinase substrate identification

8

At the heart of any kinase substrate search is the ability to detect phosphorylated proteins. To date the use of radioactive isotopes, the method used in this work, has proven to be the most sensitive and robust. However, the use of radioactive materials is undesirable because of health risks posed to the investigator and can complicate sample handling making certain analyses difficult and/or impractical.

A small number of alternative reagents for phosphoprotein detection exist and more will likely be developed. Antibodies specific to proteins containing phosphorylated tyrosine residues are already in wide use [36]. But, similar antibodies for phosphoserine and phosphothreonine containing proteins have not proven nearly as specific and effective [36]. ProQ Diamond is a gel stain available from Invitrogen that can detect low nanogram quantities of phosphoproteins [53]. PhosTag is another promising reagent that can be used to detect phosphoproteins in gels [54, 55]. These reagents will continue to aid the development of kinase substrate identification.

Recent advances in mass spectrometry have raised hopes that it will become the dominant method for phosphoprotein detection [56-58]. Because of its powerful abilities for high-throughput analysis and the ability to map phosphorylation sites to amino acid residues, it may revolutionize the study of protein phosphorylation [59]. But phosphopeptide analysis by mass spectrometry is still difficult. A 2003 study by the Association of Biomolecular Resources Facilities Proteomics Research Group found that only one of 54 labs surveyed correctly identified both known peptides in a test sample of known identity and only 14 more correctly identified one of the phosphopeptides [60].

Despite the low success rate in this study, the use of mass spectrometric techniques for phosphoprotein analysis is growing rapidly [59, 60] and will likely play a large role in the future of protein kinase substrate identification.

There is currently no straightforward method to determine if a particular substrate is a bona fide in vivo substrate. True validation and confidence in such a claim comes only from a wealth of functional and biochemical evidence to show not only that the substrate is in fact phosphorylated by the kinase, but that phosphorylation has some physiological significance [36]. Only a small fraction of substrates identified by any mechanism meet this standard [36]. It will be many years before this type of work is done on a sufficient number of kinase-substrate pairs to accurately assess the efficacy of current methods of substrate identification.

10

Chapter 2

Construction, Verification, and Experimental Use of Two Epitope-Tagged Collections of

Budding Yeast Strains

Credits

This work is the result of a collaboration between our laboratory and the laboratory of Jonathan Weissman. I worked as a member of a team aiding in the design of the libraries and the development of strategies for their construction and shared in the labor required for the completion, characterization, and verification of the TAP-tagged library.

Construction, Verification, and Experimental Use of Two Epitope-Tagged Collections of

Budding Yeast Strains

Russell Howson[1], Won-Ki Huh[1,3], Sina Ghaemmaghami[2], James V. Falvo[1],
Kiowa Bower[2], Archana Belle[1], Noah Dephoure[1], Dennis D. Wykoff[1],
Jonathan S. Weissman[2] and Erin K. O'Shea[1]*

[1]Department of Biochemistry and Biophysics and [2]Cellular and Molecular Pharmacology
Howard Hughes Medical Institute
University of California at San Francisco
600 16th Street
Genentech Hall, Room GH-S472D
San Francisco, California   94143-2240

[3]Present address:
School of Biological Sciences
Seoul National University
Seoul 151-742, Republic of Korea

*correspondence to:
Erin K. O'Shea
Department of Biochemistry and Biophysics
Howard Hughes Medical Institute
University of California, San Francisco
600 16th Street
Genentech Hall, Room GH-S472D
San Francisco, CA 94143-2240
Phone: 415-476-2212
Fax: 415-514-2073
Email: oshea@biochem.ucsf.edu

13

Running title: Preparation and application of epitope-tagged budding yeast strain collections

Abstract

A major challenge in the post-genomic era is the development of experimental approaches to monitor the properties of proteins on a proteome-wide level. It would be particularly useful to systematically assay protein subcellular localization, post-translational modifications, and protein-protein interactions, both at steady-state and in response to environmental stimuli. Development of new reagents and methods will enhance our ability to do so efficiently and systematically. Here we describe the construction of two collections of budding yeast strains that facilitate proteome-wide measurements of protein properties. These collections consist of strains with an epitope tag integrated at the C-terminus of essentially every open reading frame (ORF), one with the tandem affinity purification (TAP) tag, and one with the green fluorescent protein (GFP) tag. We show that in both of these collections we have accurately tagged a high proportion of all ORFs (approximately 75% of the proteome) by confirming expression of the fusion proteins. Furthermore, we demonstrate the use of the TAP collection in performing high-throughput immunoprecipitation experiments. Building on these collections and the methods described in this paper, we hope that the yeast community will expand both the quantity and type of proteome level data available.

Introduction

The complete sequencing of the *Saccharomyces cerevisiae* genome in 1996[61] enabled a new era of global biological analysis of this organism. Sequence analysis of the genome has provided a wealth of information relevant to many aspects of yeast biology, most recently in comparison to the genomes of other yeast species [62-64] The development of whole genome transcriptional profiling using DNA microarrays has provided tools for assessment of the global transcriptional profile under different experimental conditions [65-67]. This approach has been extremely effective in understanding many biological processes.

There are, however, limitations to microarray analysis. Though the transcriptional profile of an organism is informative, many biological processes do not produce readily interpretable transcriptional readouts. Additionally, the protein effectors of biological processes in the cell cannot be directly monitored through the transcriptional profile. It would be particularly informative to be able to systematically monitor post-translational modifications, localization, and protein-protein interactions in order to understand how the dynamic properties of proteins allow them to carry out biological processes.

Part of what makes microarray analysis possible is the chemical similarity and stability of nucleic acids. Despite the fact that these genes encode proteins of diverse composition, structure and function, the associated nucleic acids have virtually identical chemical properties. As a result, chemical manipulations can be done for all genes in parallel, simply by separating the nucleic acids spatially.

The same type of global analysis has been extremely challenging to achieve for proteins. The diversity of protein structure, chemical composition, and stability makes

16

generalized manipulation difficult. Further, the behavior of proteins in isolation is frequently not identical to their activity in the context of the cellular milieu. Mass spectrometry and two-dimensional electrophoresis have been used with some success in global protein analysis [58], but have been hampered by the complexity of the proteome, and thus far have not been capable of analyses of a truly global nature. In addition, these techniques have a somewhat limited scope in the types of protein characteristics they are capable of measuring. Some groups have created purified protein libraries [68] or protein microarrays [69-71], but these approaches have suffered both from the significant effort involved in purifying the proteins and the inherent caveats of working with proteins *in vitro*. A further problem, which has plagued all of these approaches, has been difficulty in accurately assessing coverage of the proteome.

It would therefore be valuable to have a system that enabled systematic high-throughput analysis of the yeast proteome, both *in vivo* and *in vitro*. One could construct such a system by fusing a constant epitope tag to all proteins, in essence making these proteins more chemically similar to each other. Such similarity would enable systematic manipulation, purification, or analysis of these proteins by a single method.

Here we describe the design and synthesis of a set of oligonucleotide primers useful for the genomic integration of DNA coding for any epitope tag at the C-terminus of every ORF in the yeast genome. Further, we describe the construction of two collections of yeast strains, one with the TAP tag and one with GFP tag, and the high-throughput technology and methodology which enabled the construction of these collections by a small team in a relatively short period of time. Lastly, we discuss the

17

utility of these collections in the systematic execution of high-throughput biochemical and microscopic assays of the yeast proteome.

Materials and Methods

Use of robotics and other high-throughput tools

In designing the oligonucleotides and the collections, as well as in developing the methodology for creation and use of these collections, we made every effort to utilize available high-throughput technologies. The collection was designed in 96-well format to enable efficient oligonucleotide synthesis using a 96-well DNA synthesizer (GeneMachines Polyplex) and subsequent liquid handling by a robotic 96-well pipettor (Beckman Biomek FX). The Biomek FX was used for almost all liquid handling applications, including resuspending oligonucleotides, setting up all PCRs, loading agarose gels, transformations, inoculating cultures and adding lysis buffer to cell pellets. In cases where the Biomek FX could not be used, we instead employed electronic multichannel pipettors, for applications such as loading SDS-PAGE gels and dispensing buffers. In minimizing the number of manual steps performed, we were able to both optimize the efficiency of construction and minimize the opportunity for human error. This enabled the entire process to be completed in-house by a relatively small team.

Oligonucleotide primer design and synthesis

The yeast genome sequence, as well as the coordinates of all ORFs, were obtained by download from the *Saccharomyces* Genome Database [72] on April 17, 2001. We removed all mitochondrial genes, as well as those encoding Ty elements. We then divided the ORFs into two categories: soluble and putative membrane proteins, reasoning that having membrane proteins as a separate group would facilitate any modifications in biochemical assays needed for these proteins. The following criteria were used to put ORFs into the putative membrane category: 1) Any protein experimentally determined to be an integral membrane or membrane-associated protein; 2) Any protein with homology to a membrane or membrane-associated protein; and 3) Any protein with ≥2 putative transmembrane domains which appeared in the microarray data of polysome-associated RNAs [73].

Within the soluble and membrane categories, we ordered genes by size, with largest genes first, and then divided ORFs into groups of 96 to facilitate subsequent manipulations in 96- and 384-well plates. Each ORF is designated by a plate number and coordinates within that plate. All reagents and strains relevant to a given ORF occupy the same unique coordinates. This facilitated subsequent manipulations by the Biomek FX.

We used the "Promoter" program (courtesy of Joe DeRisi, publicly available at **http://derisilab.ucsf.edu**) to extract the last 40 nucleotides (excluding the stop codon) of each ORF, as well as 40 nucleotides of genomic sequence immediately following the stop codon of each ORF. We added the constant forward sequence from the "Pringle" oligonucleotide directed homologous recombination system [74] to the last 40 nucleotides of each ORF to create the F2 oligo sequence, and the reverse complement of

19

the 40 nucleotides following each ORF to the constant reverse sequence to create the R1 oligo sequence. To design the sequence of the unique check primer for each ORF, we utilized Primer 3.0 [75], selecting oligonucleotide primers with melting temperatures of 60°C and which hybridize between 400 and 650 nucleotides upstream of the stop codon for each ORF. We note that this approach was used for all ORFs, including those representing repeated genes. The oligonucleotide sequences are available in the supplementary materials found at http://www.mrw.interscience.wiley.com/suppmat/1531-6912/suppmat/cfg.449.html.

Oligonucleotides were synthesized on a GeneMachines Polyplex 96-well oligonucleotide synthesizer, which was modified to accommodate larger reagent bottles required for 60-mer synthesis. All DNA synthesis reagents used were from Glen Research. We used a protocol (see supplementary materials) optimized for producing full length 60-mers without a need for changing reagent bottles (involving more and longer coupling steps with less volume), enabling us to run the machine overnight, which allowed for the efficient synthesis of the 12,468 60 base and 6234 20 base oligonucleotides required.

Synthesized oligonucleotides were cleaved off the solid synthesis support by incubating 3 times with 100 µl of NH$_4$OH for ten minutes, followed by collection into a deep well 96-well plate with a vacuum manifold (Millipore). The oligos were then deprotected by heating at 65°C for 15-24 hours, and lyophilized in a Sorvall SpeedVac AES2010 to remove the NH$_4$OH. Prior to use, oligonucleotides were resuspended to a concentration of 100 µM with deionized water (typically 200-300 µl, depending on the scale of DNA synthesis).

20

Construction of collections

We performed PCR and transformation in 96-well format as follows: F2 and R1 oligos were combined to working concentrations of 5 μM each. 10 μl of the primer combination were added to 40 μl of a PCR mix [19.5 μl H₂O, 5 μl 10X Pwo buffer, 5 μl 20 mM MgCl₂, 5 μl 2 mM dNTPs, 5 μl plasmid template DNA (~2 ng/ml), 0.5 μl Expand DNA polymerase] aliquotted to 96-well microplates in order to amplify the desired tag. For PCR, we used an MJ Research Tetrad thermal cycler with the following program [94°C 3:00, 10 x (94°C 0:15, 50°C 0:30, 72°C 2:00), 15 x (94°C 0:15, 72°C 2:00 + 5s/cycle), 72°C 10:00]. Following PCR, we checked for a product of correct size using 96-well agarose gels (Amersham Ready to Run system) loaded with the Biomek FX.

These PCR products were then transformed into our base strain [ATCC #201388: S288C, *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*] (Brachmann *et al.*, 1998). Cells were grown to OD₆₀₀ of 0.7-0.8 in YEPD; each transformation in the 96-well plate required 3 ml of starting culture. After being pelleted at ~1000 RCF at room temperature, cells were washed twice in 1/10 of the original culture volume of 100 mM lithium acetate and resuspended in 1/100 of the original culture volume of 100 mM lithium acetate. For each transformation, we added 15 μl of unpurified PCR product to 183 □L aliquots of the following transformation recipe, mixed by vortexing at temperature and scaled appropriately for 96-well format: 100 μl 50% weight/volume PEG (MW 3,350, freshly prepared and filtered at 0.2 □m), 15 μl 1 M lithium acetate, 20 μl salmon sperm carrier DNA (2 mg/ml), 18 μl DMSO, and 30 μl of the yeast cell suspension. We then performed incubations (30°C for 30 minutes followed by 42°C for 15 minutes) for transformation in

21

the thermal cycler. Cells were pelleted in the microplates at ~1000 RCF at room temperature, resuspended in 100 ml water, and plated manually on standard yeast synthetic medium plates lacking histidine (SD–His) to select for genomic integrants.

After growth for three days, transformations typically yielded 5 to 100 colonies. We selected up to six individual transformants for each ORF and streaked onto fresh selective medium. After subsequent growth, we performed whole cell PCR on each transformant to determine if the tag had integrated at the correct locus. A small aliquot of freshly grown cells was resuspended in 5 µl of water and boiled in 96-well format (99°C for 5 minutes in the thermal cycler). 5 µl of boiled cells, as well as 2.5 µl of 5 µM unique "check" oligos were added to PCR mix [13 µl H2O, 2.5 µl 10X Taq buffer, 1.5 µl 2 mM dNTPs, 0.25 µl 50 µM "F2CHK" primer, 0.25 µl 5U/µl Taq Polymerase, 0.05 µl 10 mg/ml RNAse], and PCR was performed [94°C 2:30, 35 x (94°C 0:45, 55°C 0:45, 72°C 1:00), 72°C 10:00]. We analyzed the results of these PCRs by 96-well agarose gel electrophoresis, identifying correct integrants appropriate size.

For construction of the GFP collection, we used much the same method, except individual transformants were picked and directly inoculated into 600 µl of SD-His medium for overnight growth. We centrifuged 200 µl of these cultures in 96-well PCR plates to pellet cells, removed the supernatant, and lysed the cells in 20 µl of 0.2% SDS at 99°C for 10 minutes in the PCR machine. We then used 0.6 µl of this lysate as template for a 20 µl PCR [16.2 µl H2O, 2 µl 10X Taq buffer, 0.6 µl 2 mM dNTPs, 0.2 µl each 50 µM oligonucleotide primer, 0.2 µl 5 U/µl Taq polymerase, 0.04 µl 10 mg/ml RNAse] to confirm correct integration of by the presence of a PCR product of the tag. The presence of a PCR product was analyzed by 96-well agarose gel electrophoresis.

22

Assembly and growth

To assemble these strain collections into 96-well plates, we selected two correct integrants (when possible) for each ORF and resuspended cells in 0.5X SD-His + 15% glycerol. The plate number and coordinates were maintained for each ORF, resulting in an "A" and "B" collection for each set of epitope-tagged strains, which we froze at – 80°C. For the GFP strains, we assembled the A and B collections directly from the liquid cultures used for confirmation PCR, mixing saturated cultures with 30% glycerol and freezing.

Subsequent growth was achieved by thawing the glycerol stocks and either inoculating liquid cultures with a Biomek FX robot or spotting onto YEPD plates with a 96-well pinning tool. To grow these cultures in high-throughput format, we used a GeneMachines HiGro growth chamber. Typically, cultures were 1.5-2.0 ml, and cells were grown at 30°C and 500 rpm. Under these conditions, cells grew with the same growth rate as liquid cultures in standard flasks (data not shown). We sometimes used teflon-coated magnetic beads in each well to enhance mixing of the culture. Addition of these beads did not have an effect on growth rate (data not shown), but did prevent settling of cells that began to occur in late log phase.

Immunoblot analysis of the TAP collection

To analyze the TAP collection by immunoblotting, 200 µl aliquots of YEPD medium in a 96 well plate were inoculated with cells from a plate with a 96-well pinning

tool and allowed to grow to saturation overnight. We diluted these cultures to OD_{600} of ~0.1 into 1.8 ml of YEPD in 2 ml deep-well 96 well plates and grew them to logarithmic phase ($0.8 < OD_{600} < 1.0$) at 30°C and 500 rpm in a GeneMachines HiGro Shaker. Log phase cultures were centrifuged to pellet the yeast cells. We removed the medium supernatant and added 50 µl hot SDS lysis buffer [50 mM Tris pH 7.5, 5% SDS, 5% glycerol, 50 mM DTT, 5 mM EDTA, Bromophenol blue, 2 µg/ml leupeptin, 2 µg/ml pepstatin A, 1 µg/ml chymostatin, 0.15 mg/ml benzamidine, 0.1 mg/ml pefabloc, 8.8 µg/ml aprotinin, 3 µg/ml anipatin], and boiled (99°C for 10 minutes in the thermal cycler). These lysates were centrifuged, and the supernatant was kept and frozen at –80°C.

We loaded 13 µl of these lysates on 26-well 4-15% gradient precast Criterion gels (Bio-Rad) with a multi-channel pipettor (Matrix technologies Impact2). Gels were run and transferred to PVDF membranes with a Transblot SD semi-dry blotter (Bio-Rad). Immunoblot analysis was performed using a primary antibody mixture of a rabbit polyclonal affinity-purified antibody to the calmodulin binding peptide (CBP) portion of the TAP tag (1:5000 dilution) and an anti-hexokinase antibody (US Biological, 1:50000) as a loading control and quantitation standard. A horseradish peroxidase (HRP) conjugated goat anti-rabbit antibody (Bio-Rad) was used as a secondary antibody, and the SuperSignal West Femto Maximum Sensitivity ECL substrate (Pierce) was used for detection. Images were collected with a CCD-based imaging system (Alpha Innotech Fluorochem 8800), and analyzed with the FluoroChem FC software (Alpha Innotech).

24

Fluorescence microscopic analysis of the GFP collection

We grew cells from the GFP collection to log phase in the same way, except in SD-His medium. Aliquots of these cultures (100 μL of a 1:10 dilution in SD-His with a final concentration of 1 μg/ml 4',6-diamidino-2-phenylindole (DAPI)) were analyzed in 96-well glass bottom microscope slides (BD Falcon #357311) pre-treated with concanavalin A (50 μg/ml in water) to ensure cell adhesion. Prior to using the slides, we added 100 μL of concanavalin A solution to each well, incubated at room temperature for at least 30 minutes, washed five times with distilled water, removed excess liquid by vigorous shaking, and dried the plates rightside-up and covered by a lid overnight. We imaged cells using a Nikon TE200/300 inverted microscope with an oil-immersed 100X objective (slide bottoms were painted with immersion oil), and made use of scripting functions in the MetaMorph version 4.6r8 imaging software in order to automate most of this process. GFP (Chroma filter set 41020, exciter HQ480/20x, dichroic Q495LP, emitter HQ510/20m) and DAPI (Chroma filter set 86010, exciter S375/20x, dichroic 86010bs, emitter S415/24m) fluorescence images (2 s exposure), as well as differential interference contrast (DIC) images (10 ms exposure), were collected for each strain and analyzed for expression and subcellular localization (Huh *et al.*, 2003). For colocalization analysis (Huh *et al.,* 2003), RFP was visualized using Chroma filter set 86010, exciter S580/20x, dichroic 86010bs, and emitter S630/60m and GFP was visualized using Chroma filter set 86010, exciter S492/18x, dichroic 86010bs, and emitter S530/40m; use of the same dichroic mirror minimized time lapse between fluorescent images.

25

Reorganization of TAP collection

With the TAP immunoblot data in hand [76], we reorganized the TAP collection strains according to abundance. Based on our measured protein expression levels, we divided the strains into 5 different expression level categories designated GS1, GS2, GS3, GS4 and GS5. GS1 is the highest expression category and GS2, GS3 and GS4 have successively lower expression levels, with upper cutoffs of 3.5×10^4, 4.75×10^3 and 1.38×10^3 molecules/cell respectively. The GS5 category includes all ORFs that could be visualized on the blots but whose abundance could not be quantified; either because the expression level was close to background or there were other technical complications with the quantitation (e.g. the protein did not run primarily as a single band). Within each abundance category, the proteins are arranged based on predicted size from largest to smallest. The membrane proteins have been place at the end of each category, also arranged from large to small. The organization of the TAP collection is available in the supplementary materials found at http://www.mrw.interscience.wiley.com/suppmat/1531-6912/suppmat/cfg.449.html.

96-well growth and native extract preparation

To grow cells for extract preparation, we first inoculated 600μl YEPD liquid cultures from a YEPD plate using a 96-well pinning tool. These cultures were allowed to grow to saturation overnight on the benchtop with no agitation. The next morning we diluted these cultures into 6 deep well 96-well plates, with 1.8 ml YEPD medium in each well, to an OD_{600} of between 0.1 and 0.2, and grew to log phase ($0.8 < OD_{600} < 1.0$) at 30°C and 500 rpm in a GeneMachines HiGro Shaker. When cultures had reached log phase, we

26

centrifuged the plates at 3000 rpm for 10 minutes and aspirated the medium. The cell pellets were resuspended in 150 μl cold sorbitol buffer (1.2 M sorbitol, 0.1M KPO₄ pH 7) + 2 μl/ml 2-mercaptoethanol. We then combined the cell suspensions from 6 96-well plates, maintaining the coordinates, and centrifuged and aspirated again. These pellets were then resuspended in 150 μl cold sorbitol buffer with 2 μl/ml 2-mercaptoethanol and 60 μl/ml lyticase (Haswell and O'Shea, 1999), and transferred to a 96-well PCR plate. We incubated these plates at 30°C for 15 minutes in the thermal cycler, then centrifuged gently (1000x g) for 10 minutes. After aspirating the supernatant, the pellets were washed gently in sorbitol buffer, frozen in liquid nitrogen and stored at –80°C.

To lyse the cells, we resuspended the thawed cell pellets in 100 μl hypotonic lysis buffer (50 mM Tris pH 7.5, 5 mM MgCl₂, 5 mM EGTA, 1 mM EDTA, 0.1% Triton X-100, 1 mM β-mercaptoethanol, 2 mM PMSF, 2.5 mM benzamidine, 1 μg/ml leupeptin, 1 μg/ml pepstatin). After incubation on ice for 10 minutes, we added 20 μl of buffer plus 0.9 M NaCl (to make the final buffer 150 mM NaCl), and incubated on ice for another 10 minutes. We then centrifuged at 4000 rpm in a Beckman RC-3B swinging bucket rotor fitted with 96-well plate carriers for 20 minutes to pellet cellular debris. Following centrifugation, we removed 100 μl of lysate to a new 96-well plate containing 25 μl 50% glycerol in each well, and kept a small aliquot to measure protein concentration by Bradford assay. These extracts were frozen in liquid nitrogen and stored at –80°C. Each well contained 125 μl of total extract and typically were ~10 mg/ml.

Multiplexed immunoprecipitations

To perform high-throughput immunoprecipitations, we first prepared native extracts as above, except we combined cell pellets from 6 *different* cultures into one plate. The result is that each well contains a total of 125 µl of ~10 mg/ml extract, but derived from a mixture of 6 different TAP tagged strains. We combined strains with approximately equal expression levels of the fusion protein in order to minimize dominance of well-expressed proteins or loss of minimally expressed proteins.

To perform the immunoprecipitation reactions, we first diluted extracts in 1 ml deep well 96-well plates to a total volume of 560 µl with P buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 5 mM EGTA, 1 mM EDTA, 0.1% Triton X-100, 1 mM β-mercaptoethanol, 2 mM PMSF, 2.5 mM benzamidine, 1 µg/ml leupeptin, 1 µg/ml pepstatin). We then added 3 µg of biotin-conjugated Human IgG (Jackson Immunoresearch), and incubated extracts for 30 minutes at 4°C. 40 µl of a 25% suspension of streptavidin beads (Amersham Pharmacia) was then added to each well. We incubated the plate at 4°C for another 30 minutes, vortexing *very gently* 3-4 times during this period to resuspend the beads.

We then transferred the reactions with a multi-channel pipettor to a filter plate (Orochem catalog #OF1100). This type of plate contains 96 wells, each with a small frit above an opening at the bottom of each well. We placed this plate on a vacuum manifold; applying a vacuum allows for the removal of the supernatant but retention of the beads. Beads were washed four times with 400 µl of PDMS buffer [P buffer + 1% Triton X-100 + 300 mM NaCl]. We then centrifuged the plate briefly (1000x g for 1 minute) to remove any residual liquid remaining in or around each well. 10 µl of sample buffer was added to

each well, and the plate was vortexed to resuspend the beads in the sample buffer. The plate was allowed to stand at room temperature for 5 minutes, and then centrifuged (1000x g for 2 minutes) on top of a shallow 96-well plate to collect the eluate. This process was repeated again, to generate a total elution volume of 20 µl.

To analyze the results of the multiplex immunoprecipitation, the samples were run on 26-well Criterion SDS-PAGE gels (Bio-Rad), and transferred to nitrocellulose in 20 mM NaPO$_4$ pH 6.8 buffer with a BioRad Transblot apparatus. We performed immunoblot analysis by probing with Rabbit Fc (Jackson Immunoresearch, 1:10 000 dilution of 3.8 mg/ml stock in TBST + 5% nonfat dry milk), an HRP-conjugated goat anti-rabbit Fc secondary antibody (Jackson Immunoresearch, 1:50 000 dilution of a 1 mg/ml stock in TBST + 5% milk), and the SuperSignal West Femto Maximum Sensitivity ECL substrate (Pierce). Images were collected with a CCD camera (Alpha Innotech Fluorochem 8800).

Availability of TAP and GFP collections

The TAP collection is available from Open Biosystems (**http://www.openbiosystems.com**) and the GFP collection is available from Invitrogen (**http://www.invitrogen.com**).

29

Figure 1. Schematic of construction and verification process for strain collections. (A) PCR-mediated homologous recombination. Three oligonucleotide primers are designed and synthesized for each ORF: the F2 and R1 oligos and a CHK oligo. The F2 and R1 oligos, containing regions of homology to the ORF of interest, are used to amplify the desired tag and a selectable marker; this PCR product then integrates into the genome at the C-terminus of the ORF of interest. The CHK oligo, as well as an oligo within the tag, is used to verify integration at the desired location. (B) Generalized construction process. We transformed the PCR amplified tag into yeast, plated on selective medium, and then picked individual colonies and verified correct integration of the tag. Correct integrants were then screened for expression of the protein fusion by immunoblot or microscopy and a correctly expressing isolate was selected.

A

ORF X CHK   F2

ORF X

R1

**Amplify tag with F2 and R1 primers
Transform product into yeast**

TAG   HIS3

ORF X

ORF X   TAG   HIS3

ORF X CHK   F2 CHK

**Confirmation PCR with
ORF X CHK and F2 CHK**

B

**transform ORF-specific
PCR product into yeast**

**plate on
selective medium**

**pick individual
colonies**

**check for correct
integration by PCR**

keep 2 positives

**check for expression by
immunoblot/microscopy**

31
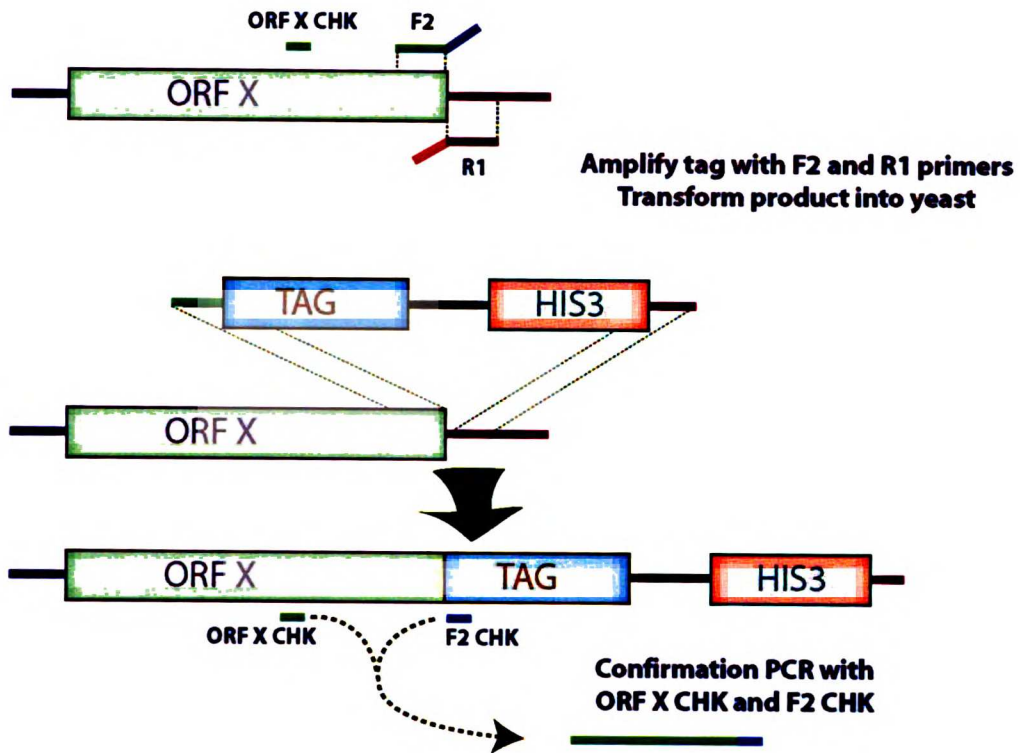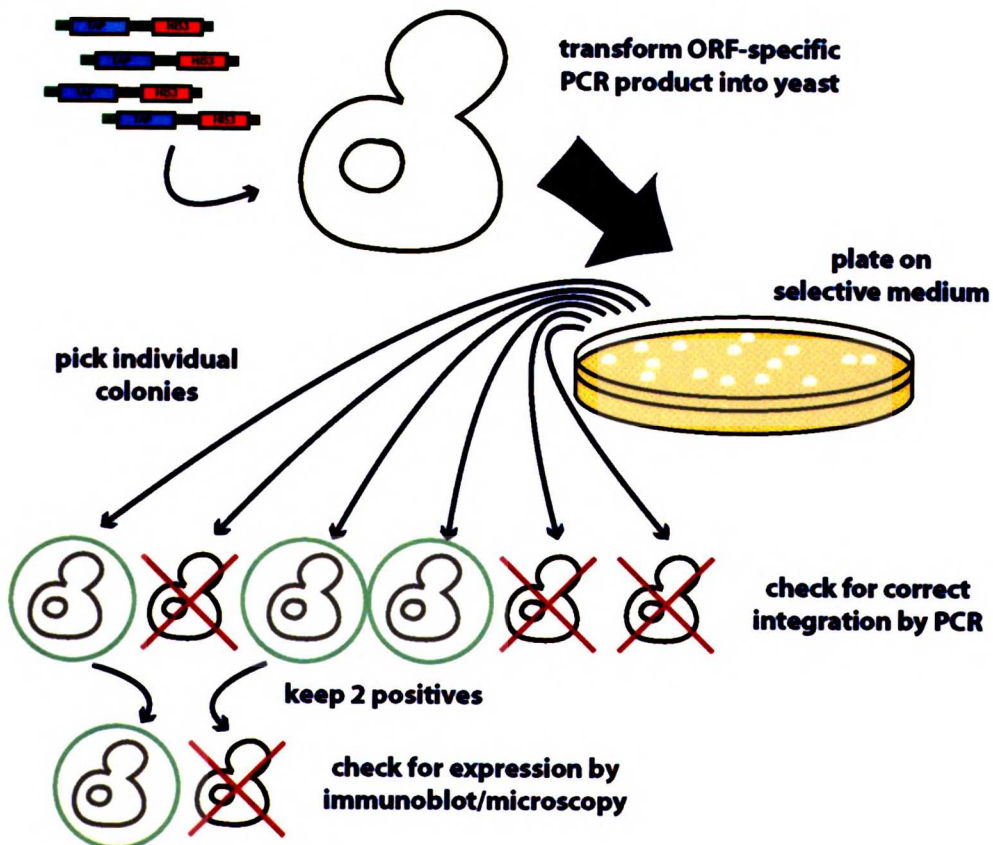
Results and Discussion

Collection design and oligonucleotide synthesis

To create a collection of tagged yeast strains, we utilized oligonucleotide directed homologous recombination (Figure 1A) [74] to integrate an epitope tag at the C-terminus of each ORF. Briefly, 60-mer oligonucleotide primers (the "F2" and "R1" primers) are designed which contain both a variable sequence, homologous to the gene of interest, and a constant sequence, which enables PCR amplification of sequence coding for the desired tag and a nutritional marker used to select for integrants. Homologous sequences in the oligonucleotides direct integration of the tag at the desired location in the genome. Nutritional selection is used to isolate integrants, and integration at the correct locus is then confirmed using PCR with one primer in the tag and one specific to the targeted ORF (Figure 1B, 2A). We chose to tag the C-terminus of each ORF so that the endogenous promoter would remain intact, and to minimize the impact on the signal sequence of secreted and membrane proteins. We therefore designed and synthesized the required oligonucleotides for each ORF in the yeast genome (see methods for details). Though we chose to construct two collections, one with the TAP tag and the other with the GFP tag, this oligonucleotide set could be used to construct a collection of strains with any desired tag integrated at the C-terminus of each ORF.

An important consideration in the design of the collections was the ability to take advantage of high-throughput and automation technologies. We designed the oligos in 96-well format to be efficiently synthesized by a high-throughput synthesizer, and developed methods for construction and use of the collections based around liquid

handling robots and multi-channel pipettors. The use of these tools was imperative in making a project of this nature feasible. After refinement of these methods during construction of the TAP collection, the only manual steps performed in the construction of the GFP collection were plating transformations and picking colonies.

Construction of the TAP and GFP collections

The first round of construction of the TAP collection was accomplished in about four months by a team of six people. We subsequently further refined the protocol for the confirmation PCR by growing individual transformants in liquid culture rather than on solid medium, enabling all manipulations to be performed by the Biomek FX liquid handling robot. With this refinement, as well as others, the efficiency of the construction process was greatly enhanced: the first round of construction of the GFP collection was carried out by two people in under three months. This oligo set and these methods could therefore be useful in the efficient construction of new collections tailored to particular experiments.

After the first round of construction of the TAP collection, we successfully tagged 97% of 6234 ORFs, as assayed by genomic PCR. For the GFP collection, we obtained PCR-positive clones for 99% of ORFs. We also determined that the frequency of obtaining a properly integrated strain (assayed by genomic PCR) was roughly equivalent between essential (93%) and non-essential (98%) ORFs.

Figure 2. Examples of verification of integration and expression.(A) 96-well agarose gel used to analyze check PCRs to confirm proper integration of the tag. Presence of band indicates proper integration, absence indicates improper integration. (B) Immunoblot analysis of TAP collection isolates. Arrow indicates the hexokinase band used as a loading control and quantitation standard. (C) Fluorescence microscopy of GFP collection isolates. Shown is an example in which expression from the A collection did not match expression from the B collection.

A
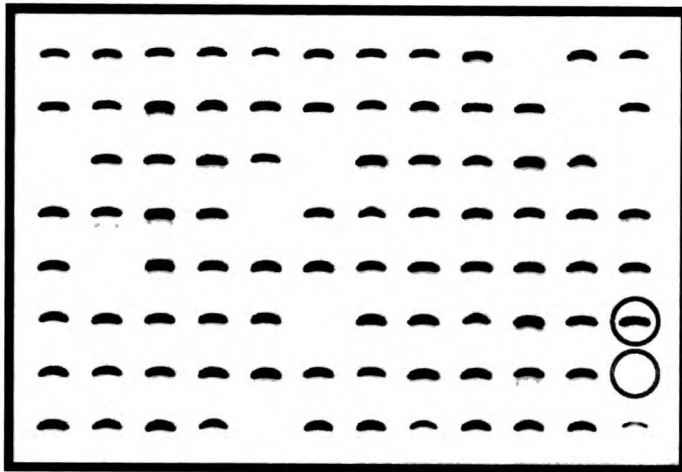
correct integrant
incorrect integrant

B

←HXK

C

	DIC	GFP	DAPI
YJR117W "A"			
YJR117W "B"			
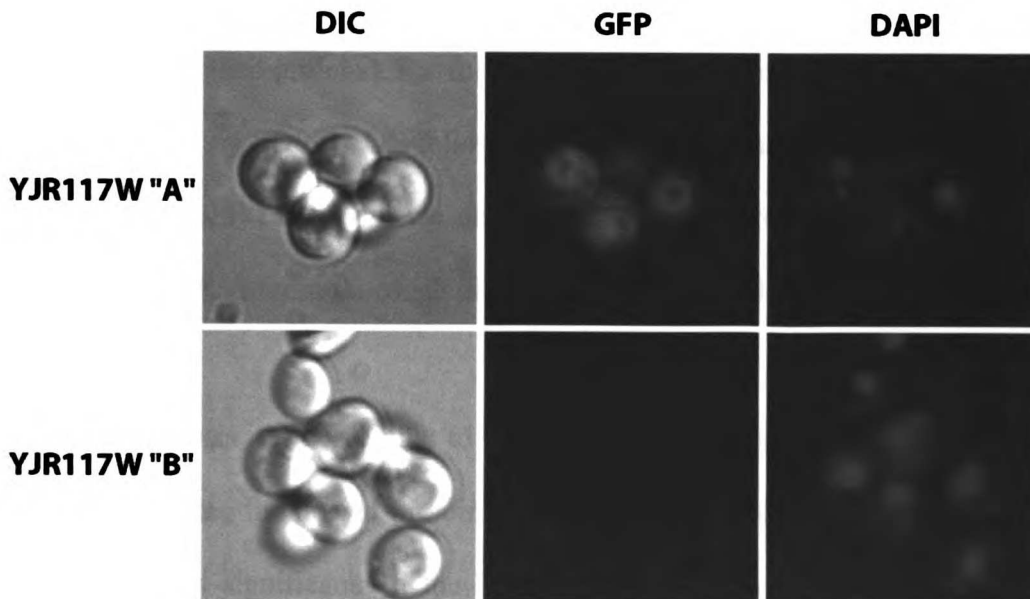
Confirmation of correctly expressed fusion proteins and reconstruction

We collected two PCR-positive clones (when possible) for each ORF, and assembled the strains into an "A" and "B" set for each collection. In beginning to work with these strains, we discovered some inconsistencies in expression of tagged proteins in different isolates from the same tagged ORF. Specifically, we identified some cases in which only one of the two isolates expressed the protein of interest, despite the fact that correct integration was confirmed for both by genomic PCR. After sequence analysis of representative isolates, we identified sequence errors in the junction between the C-terminus of the protein and the tag, or in the tag itself. The source of these errors is presumably errors in the oligonucleotides themselves or in the tag amplification and transformation process. Regardless, we concluded that proper integration of the tag could only be reliably confirmed by analyzing expression of the fusion protein (by immunoblot for the TAP collection, and fluorescence microscopy for the GFP collection).

We analyzed log phase cultures from the "A" and "B" isolates of both collections for expression of the tagged proteins. For the TAP collection, we made SDS extracts and analyzed these extracts by SDS-PAGE followed by immunoblotting against the TAP tag (Figure 2B). For the GFP collection, we observed the tagged proteins by fluorescence microscopy (Figure 2C). A more thorough discussion of these methods is in the Materials and Methods section.

After this analysis, we were able to identify which isolates of a particular ORF were correctly expressing the tagged protein of interest. We found that the frequency of 'mistagged' ORFs was significant. In the TAP collection, 21% of PCR positive "B" isolates with a corresponding expression positive "A" isolate yielded no detectable

expression by immunoblot. This highlights the importance of expression analysis in confirming the accuracy of these collections.

In cases where expression could not be confirmed for either isolate, we were unable to distinguish between two possibilities. First, it is possible that neither isolate contained the correctly tagged ORF. Alternatively, the protein could be correctly tagged, but not expressed to detectable levels under our growth conditions. We therefore compared our results from both collections to determine if there were cases in which expression of a given ORF was detected in one collection but not the other. Given that these collections were constructed with the same oligo set, one would expect to obtain similar results with the two collections. Though these means of analysis are different and therefore may have different limits of detection, we reasoned that if a tagged protein was detected in one collection but not the other, it was probably not tagged correctly in the collection in which it was not detected.

With this information in hand, we undertook the task of reconstructing those strains (708 for the TAP collection, 759 for GFP) for which the ORF fusion was detected in one collection but not the other. In order to avoid again isolating strains that did not correctly express the tagged protein of interest, we omitted the confirmation PCR and instead analyzed individual transformants by immunoblotting or microscopy. After this process, 457 new positives were obtained for the TAP collection, and 398 new positives were obtained for the GFP collection.

Coverage of the proteome

In the construction of these collections, we monitored our success rate at many steps in order to assess the quality and utility of these collections in performing proteome-wide studies (Table 1). While we were able to obtain a 97% success rate by genetic analysis, our subsequent expression analysis indicated that this is not the most accurate metric for coverage of the proteome. Because of the potential for errors in the oligonucleotides and in the integration process, and the inability of genetic analysis to uncover these errors, we feel that detection of expression of the fusion proteins is the most accurate metric of coverage, especially considering that utility of these collections in monitoring protein characteristics ultimately rests on the ability to detect them.

By this metric, we detect 4251 proteins, or 68% of 6234 annotated ORFs, in the TAP collection, and 4156, or 67% of annotated ORFs, in the GFP collection [76, 77]. Previously, we estimated that 525 of the 6234 annotated ORFs are spurious [76]a result similar to that derived from a comparative genomics study [64]. These observations suggest that there are approximately 5700 protein-coding genes. Therefore, we observe ~75% of the proteome by Western blot analysis of the TAP library and ~73% of the proteome by fluorescence microscopic analysis of the GFP library. A total of 4517 ORFs, or 79% of the proteome, were detected in at least one collection [76]. The overlap between the proteins detected in these collections (over 90% of ORFs detected in the GFP collection were also detected in the TAP collection [76, 77] strengthens our conclusion that the collections represent a large majority of true protein-coding genes. These numbers will undoubtedly improve as more sensitive detection methods are

Table 1. Success rates at various stages of construction. Details of success rate at various stages of construction for the TAP and GFP collections. Success rates for all ORFs and essential ORFs of each collection are displayed, as well as the associated percentage. ND=not determined.

	TAP collection				GFP collection			
	all ORFs		essential ORFs		all ORFs		essential ORFs	
	#	%	#	%	#	%	#	%
Total ORFs	6234	100	1100	100	6234	100	1100	100
Tag amplified with ORF specific PCR	6211	100	1096	100	ND		ND	
≥1 Transformants obtained	6047	97	1014	92	6151	99	1018	93
Positive by check PCR	6040	97	1003	91	6029	97	953	87
Positive by expression (first round)	3811	61	723	66	3758	60	712	65
Positive by expression (after reconstruction)	4251	68	821	75	4156	67	827	75

40

developed (such as immunoprecipitation followed by immunoblot) and as expression in other growth conditions is examined.

However, some strains will certainly need to be reconstructed. We sequenced the ORF-tag junction and the tag for 35 strains for which the tagged ORF is known or strongly predicted to code for a protein, but the fusion protein was not detected in either collection. Of 3 clones for fusions of essential ORFs, 2 had mutations and 1 did not. Of 5 clones for fusions of ORFs with a high codon adaptation index (CAI), 4 had insertions or deletions and 1 had a nonsynonymous substitution in the tag. Of 27 clones for fusions of ORFs coding for proteins with a previously reported localization [72], 20 had mutations while 7 did not. It is possible that the oligonucleotide primers used to generate these strains contained more errors, or that some deleterious consequences of insertion of the tag caused selection against correct integrants.

If we take our success rate for essential proteins as a proxy of our overall success rate (since essential proteins are presumably true ORFs and expressed under normal growth conditions), our collections represent 75% of the proteome for the TAP collection and GFP collections. When looking at both collections, at least one fusion protein was detected for about 80% of essential ORFs. The high percentage of essential ORFs detected also indicates that the fusion protein is likely functional in a high proportion of cases. We conclude that these collections represent useful representations of the proteome for use in global analyses. In characterizing these collections, we were also able to make quantitative measurements of protein expression [76] and describe cellular localization [77] for the majority of the yeast proteome.

"Multiplexed" immunoprecipitations

In constructing these collections, we wished to not only be able to perform descriptive analyses, but also to do experiments systematically on the entire proteome. Our hope was that standard laboratory assays typically performed on a small number of strains or proteins could be applied systematically and efficiently to the entire proteome. The fact that every strain in the TAP collection utilizes the same tag enables a generalized method to be applied to all strains to perform large-scale experiments in parallel.

As a first step, we developed high-throughput methods to efficiently make extracts and immunoprecipitate proteins (Figure 3A). Briefly, this involves growing 2 ml cultures to log phase in 96-well format, combining cell pellets from 6 *different* 96-well plates, and spheroplasting cells with lyticase. We made extracts by osmotic lysis, pelleting cellular debris and keeping the supernatant. Each well of these "multiplex" extracts contains extract from 6 different strains, and therefore 6 different TAP-tagged proteins, facilitating the parallel immunoprecipitation of 6 different proteins in each well of a 96 well plate. Immunoprecipitation in 96-well format, therefore, theoretically enables the simultaneous pulldown of 576 proteins.

Utilizing the quantitative immunoblot data obtained from screening the TAP collection (discussed above), we reorganized the TAP collection according to abundance. We divided the strains into six abundance categories, and within each category ordered the strains by size of the tagged ORF. Because of this reorganization, the six TAP-tagged proteins in each well of the multiplex extracts are of approximately equal abundance in extracts. Because the proteins are ordered by size within each abundance category, the 6

42

Figure 3. High-throughput "multiplex" immunoprecipitations. (A) Schematic of process. Cultures from six 2 ml 96-well plates are pelleted, combined, and used to prepare extracts. IgG biotin and streptavidin beads are added and allowed to bind, and then transferred to a filter plate that allows for retention of beads. Beads are washed, proteins are eluted and analyzed by SDS-PAGE and immunoblot. (B) Example immunoblot of results of multiplex immunoprecipitation.

A

2 mL 96-well culture plates

Combine 6 plates into 1 and prepare extracts

1 mL deep well plate

Add IgG biotin and streptavidin beads; transfer to filter plate

Filter plate: beads are retained while liquid flows through

wash and elute; analyze by SDS-PAGE
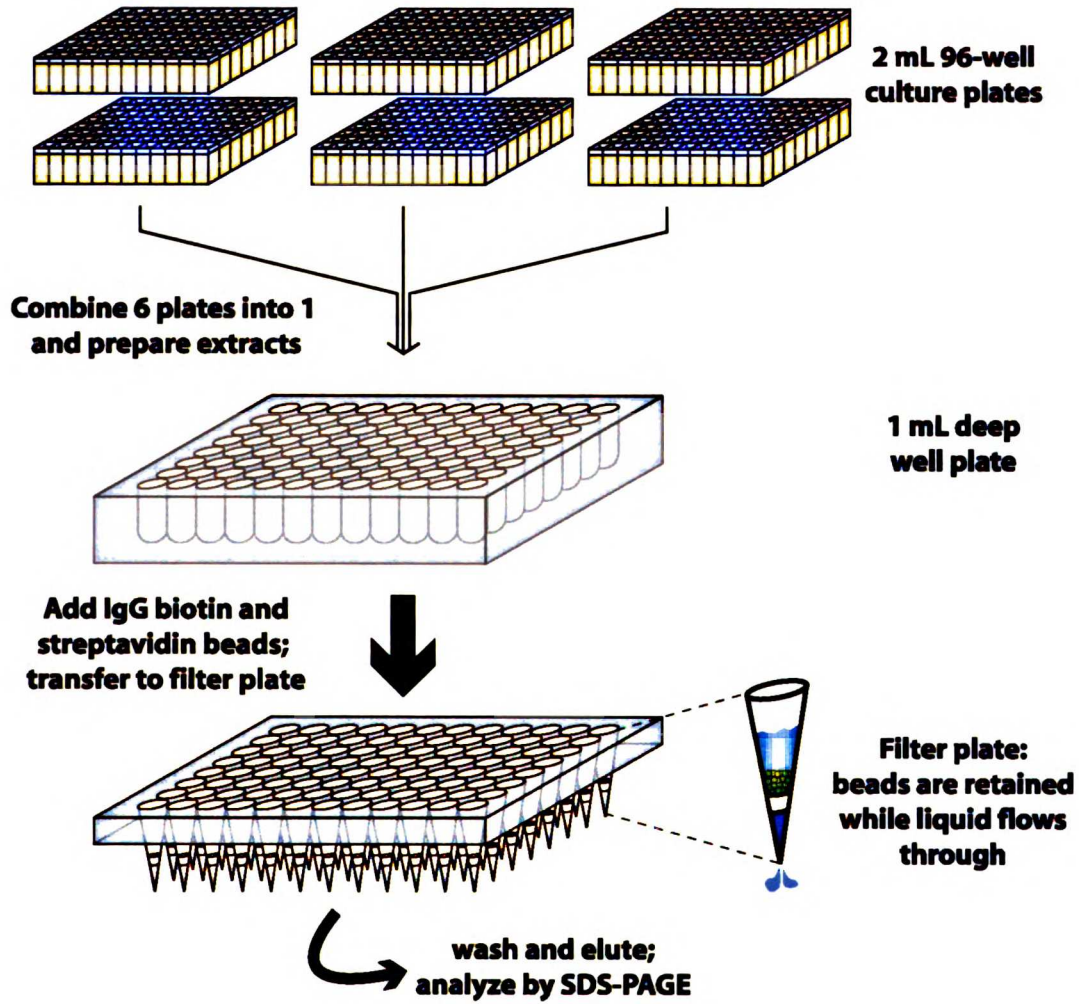
B
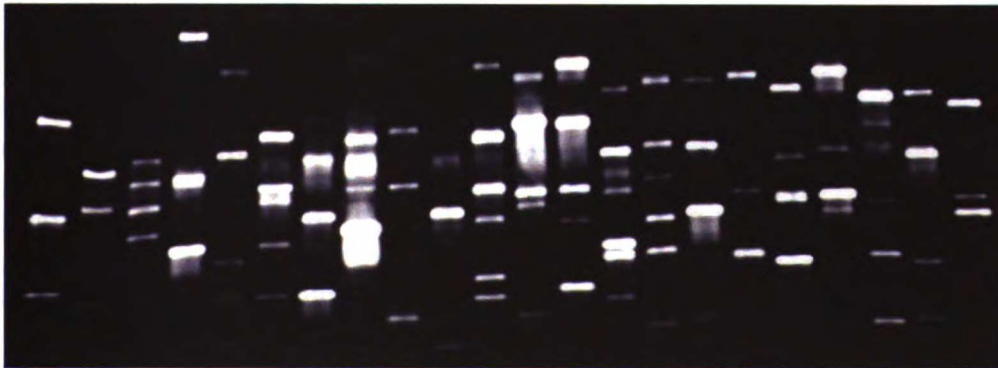
proteins in a given well also represent the maximal possible size distribution within the abundance category. This facilitates the resolution of the individual proteins in subsequent analysis by SDS-PAGE.

To test the feasibility of high-throughput immunoprecipitations, we prepared multiplex extracts and pulled down TAP tagged proteins as described in the methods. In this particular experiment, the last column of the plate was left empty to provide space for controls necessary for a subsequent assay, so the theoretical maximum number of proteins pulled down in this reconstruction is 528 proteins. As shown in Figure 3B, we were successfully able to pull down a large number of proteins with this procedure.

To quantify the efficiency of the multiplex pulldown, we counted the number of distinct bands in each lane, and summed the total for this gel and the entire plate. We cannot discount the possibility that some of these bands may represent breakdown products or modified proteins; nevertheless, due to the distinctness of the bands and the fact that they migrate at the appropriate molecular weight, it is likely that a significant proportion represent full length proteins. It should be noted that because immunoprecipitation efficiency and potential post-translational modification or breakdown product formation is not identical for the six proteins in each lane, the multiplexing technique is not meant for comparisons between proteins within each lane, but rather for comparisons across different lanes (e.g. at different growth conditions or time points) for the same protein. The gel shown contains 22 lanes, or one quarter of the plate, so the theoretical maximum number of immunoprecipitated proteins is 132. We are able to detect 111 distinct bands in this gel, or 84% of possible proteins. In immunoblots for the entire plate, we were able to detect 452 distinct bands, representing 86% of the

528 proteins possible. In separate experiments in which we immunoprecipitated TAP proteins from the collection individually, we were able to successfully pull down proteins with comparable efficiency (1070 proteins pulled down of 1334 attempts, or 79%). We conclude that immunoprecipitations in the multiplex format are an effective method of efficiently purifying TAP fusions from the collection.

Discussion

In this paper, we have described the construction of two collections of yeast strains, both with C-terminal fusions of most ORFs in the yeast genome, one with the TAP tag and one with the GFP tag. The oligonucleotide primer set and the methods discussed could be used to efficiently construct a new collection with any desired C-terminal tag and in any desired genetic background. With the final refinements of our methods, a new collection could be constructed by a small team in a matter of months, depending on the degree of completeness and level of verification desired (expression analysis and reconstruction of our collections took several additional months). Furthermore, many of the methods described could be modified for use with other model organisms that support efficient homologous recombination.

We have confirmed that these collections do indeed represent a significant portion of the proteome, as we have confirmed expression of the ORF-tag fusions in individual transformants, either by immunoblot analysis for the TAP collection or by microscopic analysis for the GFP collection. Importantly, all fusions are under control of the native promoter, minimizing the potential for artifacts due to overexpression. We therefore believe that these collections will be useful tools in performing large-scale proteomic

experiments. Indeed, in the course of confirming expression of the fusion proteins in these collections, we have been able to obtain valuable information about the absolute abundance of proteins (TAP collection) as well as their subcellular localization (GFP collection).

We wished to take proteomic analysis beyond being descriptive, so we developed methods to apply standard biochemical experiments in a parallel manner. We found that multiplexed immunoprecipitation could be performed efficiently from extracts derived from individual yeast cultures as small as 2 ml. This procedure should be easily modifiable in order to perform almost any extract-based assay in a parallel manner on the proteome.

These collections represent exciting possibilities for the future in a number of respects. First, it will be interesting to apply the descriptive methods outlined in this paper to experimental situations. One way in which cells can rapidly respond to environmental stimuli is to alter the localization, abundance, or post-translational modification of proteins, sometimes without any change in transcriptional state. One could use the methods described in this paper to monitor these aspects of many proteins under various growth conditions or in response to environmental insults, either with the entire collections or by examining a specific subset or family of proteins.

Second, it will be exciting to see what types of biochemical and/or microscopic assays are applied in this high-throughput parallel manner. The ability to perform immunoprecipitations in a "multiplex" format allows for efficient screening of the entire proteome with only a handful of experiments. Any number of post-translational modifications can be examined with minor modifications to this assay; it will be

especially exciting to examine the dynamic nature of these modifications in response to environmental stimuli.

Furthermore, the ability to systematically select *MAT*a haploid progeny from large-scale yeast crosses through the use of synthetic genetic array technology [78] enables either of these collections to be efficiently crossed to any desired genetic background. It should be noted that in previous studies, *HIS3* was used for *MAT*a-specific selection in the "magic marker" *MAT*α parent strain [78]; because the TAP and GFP strains were constructed with the *HIS3* marker, an alternate marker (e.g., *LEU2*) must be used in any *MAT*α strains crossed to our collections for SGA assays. This permits the examination of how different mutations impact global localization, abundance, or perhaps post-translational modifications, or could be used to sensitize the strain background to various chemical or environmental stimuli. One could also cross the TAP collection with a strain or strains containing an ORF tagged with a different epitope, thereby enabling high-throughput IP western experiments to identify protein-protein interactions.

Worthy of mention is the ease and accessibility to high-throughput experimentation that these collections provide. Though we employed the use of robotics and other high-throughput equipment to construct these collections, it is possible to efficiently utilize these reagents with simply a handful of multi-channel pipettors and a 96-well pinning tool. We therefore hope that these collections will open the door to systematic proteomic analysis by a wider range of laboratories.

Author's Contributions

RWH, JSW, and EKO conceived of the project. RWH designed and oversaw synthesis of oligonucleotide primers. WKH oversaw construction of the TAP and GFP collections. RWH, SG, KB, ND, AB, JSW and EKO assisted in construction of the TAP collection, and JVF assisted in construction of the GFP collection. SG and KB performed expression analysis of the TAP collection. WKH and JVF performed expression analysis of the GFP collection. RWH performed multiplex immunoprecipitation experiments with the TAP collection; DDW provided data on individual immunoprecipitations. SG, ND, AB, and DDW performed reorganization of the TAP collection.

Acknowledgements

Chapter 3

Combining Chemical Genetics and Proteomics to Identify Protein Kinase Substrates

Credits

This work is the result of a collaboration between our laboratory and the laboratory of Kevan Shokat. Russell Howson performed many of the early experiments and began the library screen. I performed all other experiments. Justin Blethrow synthesized the N^6-benzyl-ADP.

Biological Sciences; Biochemistry

Combining Chemical Genetics and Proteomics to Identify Protein Kinase Substrates

Noah Dephoure[*§], Russell Howson[*§], Justin Blethrow[†], Kevan M. Shokat[†] & Erin K. O'Shea[‡]

[*]Howard Hughes Medical Institute and Departments of Biochemistry and Biophysics, and [†]Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94143. [‡]Howard Hughes Medical Institute and Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138.

[§]These authors contributed equally to this work.

Howard Hughes Medical Institute
Harvard University
Department of Molecular and Cellular Biology
Bauer 307
7 Divinity Avenue
Cambridge MA 02138

Tel: (617) 495-4328
Fax: (617) 496-5425
Email: erin_oshea@harvard.edu

Abbreviations: TEV, tobacco etch virus; TAP, tandem affinity purification

Abstract

Phosphorylation is a ubiquitous protein modification important for regulating nearly every aspect of cellular biology. Protein kinases are highly conserved and constitute one of the largest gene families. Identifying the substrates of a kinase is essential for understanding its cellular role, but doing so remains a difficult task. We have developed a high-throughput method to identify substrates of yeast protein kinases that employs a library of yeast strains each producing a single epitope-tagged protein, and a chemical genetic strategy that permits kinase reactions to be performed in native, whole cell extracts. Using this method, we screened 4250 strains expressing epitope-tagged proteins and identified 24 candidate substrates of the Pho85-Pcl1 cyclin dependent kinase, including the known substrate Rvs167. The power of this method to identify true kinase substrates is strongly supported by functional overlap and colocalization of candidate substrates and the kinase, as well as by the specificity of Pho85-Pcl1 for some of the substrates over another Pho85-cyclin kinase complex. This method is readily adaptable to other yeast kinases.

The general criteria for establishing that a protein is a substrate of a given kinase are the ability of the kinase to phosphorylate the substrate *in vitro* and dependence on the activity of the kinase for phosphorylation of the substrate *in vivo* [36]. *In vivo* validation of kinase substrates continues to be laborious and time-consuming. Thus, it is crucial to be able to efficiently identify candidates before committing to this step. Indirect data, such as genetic and physical interactions, can provide insights into potential substrates, but many interacting proteins and genes are not substrates [36, 38]. More direct approaches using *in vitro* kinase reactions have been attempted in many permutations [36, 38]. Purified component reactions directly measure the ability of a kinase to phosphorylate a particular substrate and can be scaled for high-throughput formats, but such conditions often compromise reaction specificity and produce false positive results [36, 38].

We sought to improve upon these techniques by carrying out a biochemical screen in an environment that more closely resembles the *in vivo* state. To do so, we carried out kinase reactions with near physiological levels of exogenous kinase in non-denatured whole cell extracts. These extracts maintain protein-protein interactions that may affect substrate presentation and preserve a near full complement of cellular proteins, including potential adaptor proteins and cofactors that could affect the reaction. These reaction conditions also preserve the native relative protein abundances and a natural competition among substrates for limited kinase.

A challenge of carrying out protein kinase reactions in whole cell lysates is that one cannot attribute phosphorylation of a protein to a particular kinase in the reaction, since

54

all kinases in the extract are capable of catalyzing phosphotransfer. To permit us to assign phosphorylation of a protein to a particular protein kinase, we made use of a chemical genetic strategy using a mutant kinase (analog specific, or AS) whose nucleotide binding specificity is altered to allow it to utilize a modified form of ATP [79]. An AS version of a kinase is generated by replacing a conserved, bulky hydrophobic group in the ATP-binding pocket with a smaller residue. This allows the AS kinase to utilize an ATP analog that fits into the engineered "hole" in the ATP-binding pocket. In an assay utilizing an AS kinase and a radiolabeled ATP analog, all radiolabeled proteins were necessarily phosphorylated by the mutant kinase, because no other kinase can use the analog. This chemical genetic approach has been applied to identify substrates of v-Src by carrying out a kinase reaction in a cell lysate and identifying radiolabeled protein bands using mass spectrometric approaches [80]. A difficulty with this strategy is that the dynamic range of protein abundance makes it difficult to identify low abundance substrates. More recent studies in budding yeast have circumvented these identification problems by combining this chemical genetic approach with a protein expression library. Cdk1 substrates were identified by screening a set of candidate proteins containing the Cdk1 consensus phosphorylation sequence, which were overexpressed as fusions to glutathione-S-transferase (GST) [42]. The success of this approach motivated us to develop a systematic unbiased strategy that permits screening for substrates of kinases with no known consensus target sequence and which uses proteins present at their natural levels of abundance. Our screen makes use of a collection of yeast strains we generated in which each strain produces a single epitope-tagged protein expressed from its native chromosomal locus, under the control of the endogenous promoter [76, 81]. To identify

55

substrates, we generated extracts from these strains, carried out kinase reactions in the extracts by adding AS kinase and radiolabeled ATP analog, isolated the epitope-tagged proteins by immunopurification, resolved the immunocomplexes on a gel, and assayed for the incorporation of radiolabel.

As a test of this method, we screened for substrates of Pho85, a non-essential yeast cyclin dependent kinase [46]. Pho85, when bound to the cyclin Pho80, plays a well characterized role in the cellular response to phosphate starvation [10, 46]. However, cells lacking *PHO85* have additional phenotypes unrelated to the phosphate starvation pathway, including a role in the G1 phase of the cell-division cycle [47, 48]. Ten different Pho85 cyclins, termed Pcl's, have been shown to associate with Pho85 and are thought to direct it to phosphorylate distinct substrates involved in different cellular processes [30, 49]. To better understand the role of the Pho85-Pcl1 cyclin dependent kinase, we undertook a systematic screen of the yeast proteome for substrates, making use of an AS version of Pho85, Pho85(F82G), which functionally substitutes for wild-type Pho85 *in vivo*, and which is able to utilize the ATP analog N^6-benzyl ATP *in vitro* [50].

Materials and Methods

Yeast strains

All strains are derivatives of BY4741 (available from ATCC) and are from the epitope tagged fusion library described previously [76, 81] (available from Open Biosystems).

Figure 4. Strategy for chemical genetic screening of the yeast proteome for kinase substrates. TAP-tagged strains were grown in 96-well plates, pooled, and lysed. Kinase reactions were performed by adding purified analog-sensitive kinase and radiolabeled ATP analog to the extracts. The tagged proteins were isolated by immunopurification, separated by gel electrophoresis, transferred to a membrane, and analysed on a phosphorimager.

Grow cells in 2 mL 96-well culture plates

Combine 6 plates into 1
and prepare extracts

+ Kinase

ATP

Immunoprecipitate

Transfer to filter plate

Filter plate:
beads retained
liquid flows
through

Wash, elute, analyze by SDS-PAGE

Phosphorimager Analysis
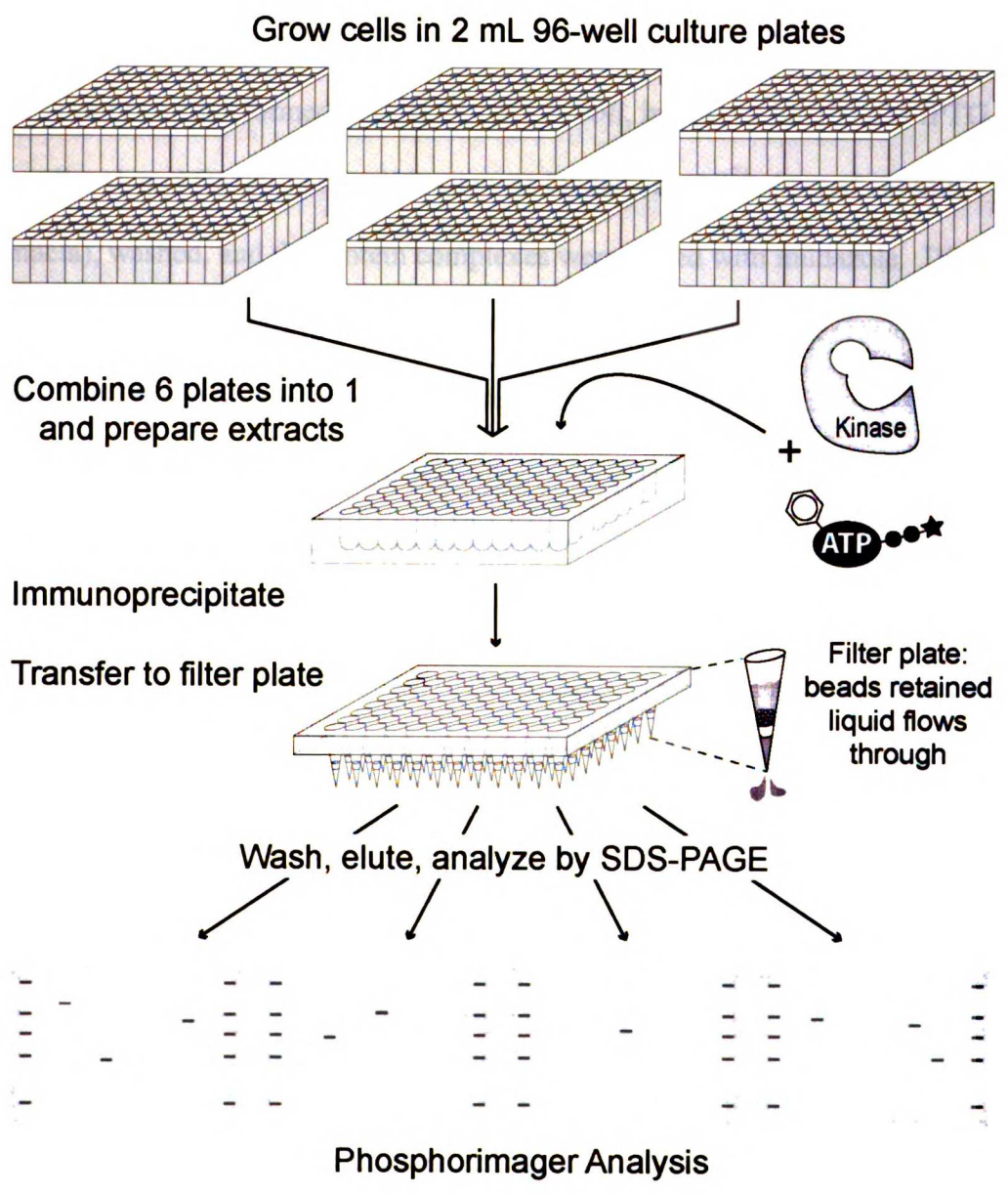
Protein purification

Purification of Pho85-cyclin complexes was performed as described in [82]. Briefly, expression plasmids for Pho85(F82G)-6His (EB1375) and either Pho80 (EB1076) or Pcl1 (EB1613) were co-transformed into BL21 DE3 *E. coli*, induced with 200 μM IPTG for 6 hours, pelleted, and frozen. The cell lysate was loaded on a Ni^{2+} charged column (Pharmacia), washed, and the protein complexes were eluted with imidazole. Dimeric CDK-cyclin was purified away from CDK monomer on a Superose 12 size-exclusion column. Purified kinase concentrations were determined using calculated extinction coefficients and by measuring UV absorption in 6 M guanidine hydrochloride [83]. Kinase activity was confirmed with *in vitro* reactions using a Pho4-peptide substrate and/or full length Pho4. The purification of Pho4 and Pho4-zz has been described[31, 84].

ATP analog synthesis

$[\gamma\text{-}^{32}P]N^6$-benzyl-ATP was prepared enzymatically from N^6-benzyl-ADP using a modified version of the protocol described in [85]. 200 μl of a 1:1 slurry of Co^{2+} charged iminodiacetic acid-sepharose beads in HEPES buffered saline (HBS) (50 mM Hepes, pH 7.4, 150 mM NaCl) were added to an empty 2 ml Bio-Spin column (Bio-Rad), and washed twice with 1 ml HBS. 500 μg 10-His tagged nucleoside diphosphate kinase (NDPK), purified from *E. coli*, was added in 1 ml HBS. The column was washed once with 1 ml PBS, pH 7.5. The NDPK was allowed to autophosphorylate by adding 5 mCi of end-labeling grade $[\gamma\text{-}^{32}P]ATP$ (MP Biomedicals) in 1 ml PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.76 mM KH_2PO_4, pH 7.4) and washed twice with PBS. 25

nmol of N^6-benzyl-ADP [85] was added in 80 μl PBS containing 10 mM MgCl$_2$. The

NDPK catalyzed transfer of the [^{32}P]phosphate to the nucleotide diphosphate, producing

[γ-^{32}P]-labeled triphosphate. The final product was eluted with 250 μl PBS with MgCl$_2$.

Para-amino-benzoic acid was added to 50 mM final concentration as a stabilizer and the

product was stored at 4°C. Product activity and efficiency were monitored by

scintillation counting of flow-through fractions. We routinely recovered >40% of total

activity in the final product.

Cell growth and yeast extract preparation

Individual strains were inoculated from fresh overnight, saturated cultures into 1.8 ml of

YEPD in deep-well 96-well plates (Greiner) at OD600 ~ 0.2-0.3, and grown to late log

phase (0.8 < OD600 < 1.0) at 30°C in a HiGro growth chamber (GeneMachines). The

cells were pelleted and resuspended in 150 μl sorbitol buffer (SB) (1.2M sorbitol, 0.1M

KH$_2$PO$_4$, pH 7.4, 2 μl/ml β-mercaptoethanol (βME). Pellets from 6 plates were

combined, pelleted, resuspended in 150 μl SB along with 15 μl lyticase [86], and

transferred to a 96-well PCR plate. Lyticase treatment proceeded at 30°C for 15 min on a

thermal cycler. Cells were then chilled on ice, pelleted, and washed once with 150 μl SB

without βME, frozen in liquid nitrogen, and stored at -80°C. To lyse the spheroplasted

cells, pellets were thawed on ice and resuspended in 105 μl ice-cold hypotonic lysis

buffer (HLB), containing 50 mM Tris-HCl, pH 7.5, 5 mM MgCl$_2$, 5 mM EGTA, 1 mM

EDTA, 0.1% Triton X-100, 2 mM PMSF, 2.5 mM benzamidine, 1 μg/ml pepstatin A, 1

μg/ml leupeptin, 1 mM βME. After a 10 min incubation on ice, we added 21 μl HLB

with 0.9 M NaCl (final [NaCl] = 150 mM) followed by an additional 10 min incubation

on ice. Lysates were cleared by centrifugation for 45 min at 5000xg, 4°C. 100 μl of

supernatant was transferred to a 1 ml deep-well 96-well plate containing 25 μl 50%

glycerol. Lysates were frozen in liquid nitrogen and stored at -80°C. We routinely

sampled extracts and measured the total protein concentration by Bradford assay. Protein

concentrations were typically ~10 mg/ml. The manipulation of cells and lysates in 96-

well format was performed using a Biomek FX (Beckman).

High-throughput kinase reactions

Reaction solutions were prepared in three parts: the substrate extract, a 3x kinase

solution, and a 3x reaction solution with the radiolabeled ATP analog. Extracts were

thawed on ice and the components were allowed to come to room temperature before

starting the reactions by adding 120 μl 3x kinase solution (30 nM Pho85(F82G)-Pcl1,

0.65 mg/ml pyruvate kinase (Roche), 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 10%

glycerol, 0.1 mg/ml BSA (Roche), 1 mM βME, 0.1% NP-40, 2 mM PMSF, 2.5 mM

benzamidine) followed by 120 μl 3x reaction mix (200 mM Tris-HCl, pH 7.5, 150 mM

NaCl, 10% glycerol, 120 mM phosphoenolpyruvic acid (Sigma), 160 mM β-

glycerophosphate, 30 mM NaF, 3 mM ATP, 10 mM EGTA, 0.3 mg/ml BSA, 25 mM

MgCl$_2$, 300 nM calyculin A, and 30 μCi [γ-^{32}P]N^6-benzyl-ATP). Reactions proceeded

for 15 min at 25°C with occasional gentle vortexing, and were stopped by the addition of

200 μl ice-cold IP buffer (50 mM Tris-HCl, pH 7.5, 300 mM NaCl, 5 mM MgCl$_2$, 5 mM

EGTA, 1 mM EDTA, 1% Triton X-100, 160 mM β-glycerophosphate, 2 mM βME, 10

mM NaF, 1 mM activated Na$_3$VO$_4$, 2 mM PMSF, 2.5 mM benzamidine, 1 μg/ml

pepstatin A, 1 μg/ml leupeptin) with 3 μg Biotin-SP-conjugated human IgG (Jackson

Immunoresearch Labs) and incubated for 30 min at 4°C. Approximately 5 μl packed

volume streptavidin sepharose (Amersham) was added in 40 μl total volume of IP buffer.

The samples were incubated for an additional 90 min at 4°C with occasional light vortexing and transferred to a 96-well fritted filter plate (Orochem, #OF 1100) on a vacuum manifold. The beads were washed 4X with 400 µl of IP buffer. After the final wash, residual buffer was removed with a low speed spin (1000 xg for 1 min). The immunocomplexes were recovered through two serial elutions. For each, 10 µl boiling hot SDS-sample buffer was added, followed by light vortexing, a 5 min incubation at room temperature, and centrifugation (1000 xg for 2 min) into a fresh 96-well plate. 12 µl of each sample was loaded onto a 26-well 4-15% gradient Criterion SDS-PAGE gel (Bio-Rad) along with prestained, Magic Mark (Invitrogen), and ^{14}C-methylated (Amersham) protein molecular weight markers. Gels were run for 1 hour at a constant 200V and transferred to nitrocellulose for 90 min at 1.2 Amps in 20 mM NaH$_2$PO$_4$, pH 6.8. The membranes were exposed to a phosphorimager screen (Molecular Dynamics) for 24-48 hours, scanned on a Storm Phosphorimager, and analysed using Imagequant software.

Kinase reactions in the primary screen exhibited variable levels of background signal (e.g., see fig. 5), and thus had to be scored manually. Subsequent to the primary screen and deconvolution of the pools, we were able to drastically reduce the background signal in our assay (see fig. 7) by employing epoxy-coated magnetic beads (Dynal Biotech) conjugated to human IgG (Sigma).

Pool deconvolution

To deconvolute the pools, kinase assays were performed on each strain in each positive pool. Strains were grown and lysed as described above, but the final lysates came from 10 ml of each strain culture.

Immunoblotting

Immunoblots were probed with an antibody raised against the calmodulin binding peptide region of the TAP tag [76] (Open Biosystems, #CAB1001). We used a horseradish peroxidase linked goat anti-rabbit IgG secondary antibody (Bio-Rad). The membranes were treated with Supersignal West Femto ECL (Pierce) and visualized using a CCD camera (Alpha Innotech).

TEV cleavage kinase assay

Kinase assays were performed on identical aliquots of extract for each candidate as described above, except that the final kinase concentration was 30 nM and immunoprecipitation was performed with epoxy-coated magnetic beads conjugated to human IgG. Following the IP, the beads were resuspended in 15 µl IP buffer with or without 2 µl of TEV protease (Invitrogen, 10 u/µl). Proteolysis proceeded for 30 min at 25°C with constant agitation. Samples were then boiled in sample buffer and analyzed as above.

Kinase specificity assays

30 nM Pho85(F82G)-Pcl1 and Pho85(F82G)-Pho80 were used in parallel kinase reactions on identical aliquots of TAP-tagged substrate strain extracts,

immunoprecipitated with human IgG coupled magnetic beads, and analyzed as described above. Bands were quantitated using Molecular Dynamics Imagequant Software. Similar results were seen when the reactions were performed with 3 nM kinase.

Determining kinetic constants

Because our reactions were carried out in the presence of ATP, the sensitivity of our assay is largely dependent upon the nucleotide selectivity of the kinase. We determined the kinetic constants for wild-type and AS Pho85-Pho80 using both ATP and N^6-benzyl-ATP and found that the mutant kinase has ~6.5-fold preference for N^6-benzyl-ATP (Table 3). A previous study using an AS version of Cdc28 reported a 130-fold preference for the analog over ATP[42], suggesting that other kinases may produce greater sensitivity in this assay than Pho85. Kinase reaction time courses were performed with 1 nM Pho85-Pho80 or Pho85(F82G)-Pho80, 5.2 μM Pho4, and 5 μCi of radiolabeled nucleotide over a range of total nucleotide concentrations at 30°C in 10% Glycerol, 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 0.01% NP-40, 1 mM PMSF, 2.5 mM Benzamidine, 0.1 mg/ml BSA. V_{max} and K_m were determined from Lineweaver-Burk plots of the initial rates of phosphorylation and the nucleotide concentrations. We calculated k_{cat} from V_{max} assuming that 100 % of the enzyme was active.

Determining the limits of detection

Reactions were performed using 10 nM Pho85(F82G)-Pho80 on a dilution series of Pho4-TAP extracts, using extracts from a pho4 deletion strain as a diluent. Parallel

Table 2. Nucleotide Specificity of Pho85(F82G)-Pho80

Kinase	Nucleotide	K_m (µM)	k_{cat} (min^{-1})	k_{cat}/K_m (µM^{-1}min^{-1})
Pho85-Pho80	ATP	99.8 (+/- 17.7)	351 (+/- 57)	3.62 (+/- 0.87)
Pho85-Pho80	N^6-benzyl-ATP	> 2000	ND	ND
Pho85(F82G)-Pho80	ATP	196.0 (+/- 7.9)	341 (+/- 39)	1.74 (+/- 0.17)
Pho85(F82G)-Pho80	N^6-benzyl-ATP	23.8 (+/- 7.1)	258 (+/- 43)	11.3 (+/- 2.5)

ND - At the nucleotide concentrations used, these values could not be determined.

reactions were performed using dilutions of recombinant Pho4-zz and a mutant Pho4-zz that can only be singly phosphorylated.

Results/Discussion

To screen 4250 strains (Fig. 4), each expressing a single protein epitope-tagged with the tandem affinity purification (TAP) tag [87], we grew cells from each strain in 96-well plates to late-log phase, and combined pellets from cultures of six different strains before cell lysis [81]. Kinase reactions were started by adding recombinantly produced Pho85(F82G)-Pcl1 kinase to extracts along with ^{32}P-labeled N^6-benzyl ATP. Tagged proteins were isolated by immunopurification [81], resolved on SDS-polyacrylamide gels, transferred to nitrocellulose, and exposed to a phosphorimager screen for analysis. The same membranes were then immunoblotted for the TAP tag. All bands with radioactive signals above background were scored as positive. The complete screen of all 4250 strains yielded 55 pools with a positive signal (representative positive results are shown in Fig. 5). Analysis of the immunoblots from one plate of reactions revealed that we successfully immunoprecipitated 86% of the library proteins (data not shown), indicating that most proteins including many highly abundant ones, are not phosphorylated in the assay (for examples of negative results from individual strain assays see the bottom panel of Fig. 5).

To eliminate spurious signals and to identify the proteins corresponding to the bands in the pooled extracts, we screened each of the individual strains in the positive pools (Fig. 5). Deconvolution of these 55 pools revealed 34 positive strains.

Figure 5. Kinase screen and pool deconvolution. Kinase reactions were performed in pooled extracts and analyzed as described. Phosphorimager scans of six representative positive signals are shown above. All six strains present in the positive pools were grown, lysed, and assayed separately. Shown below are the phosphorimager scans for the assays of the individual strains present in each pool. The identities of the TAP-tagged proteins in the positive gel lanes are indicated. The multiple bands present in some lanes are degradation products. * These bands were not observed in the individual strain assays.

Screen in Pools

Rom2   Reg1   Shs1   Yfr017c   Sfb3 Ksp1
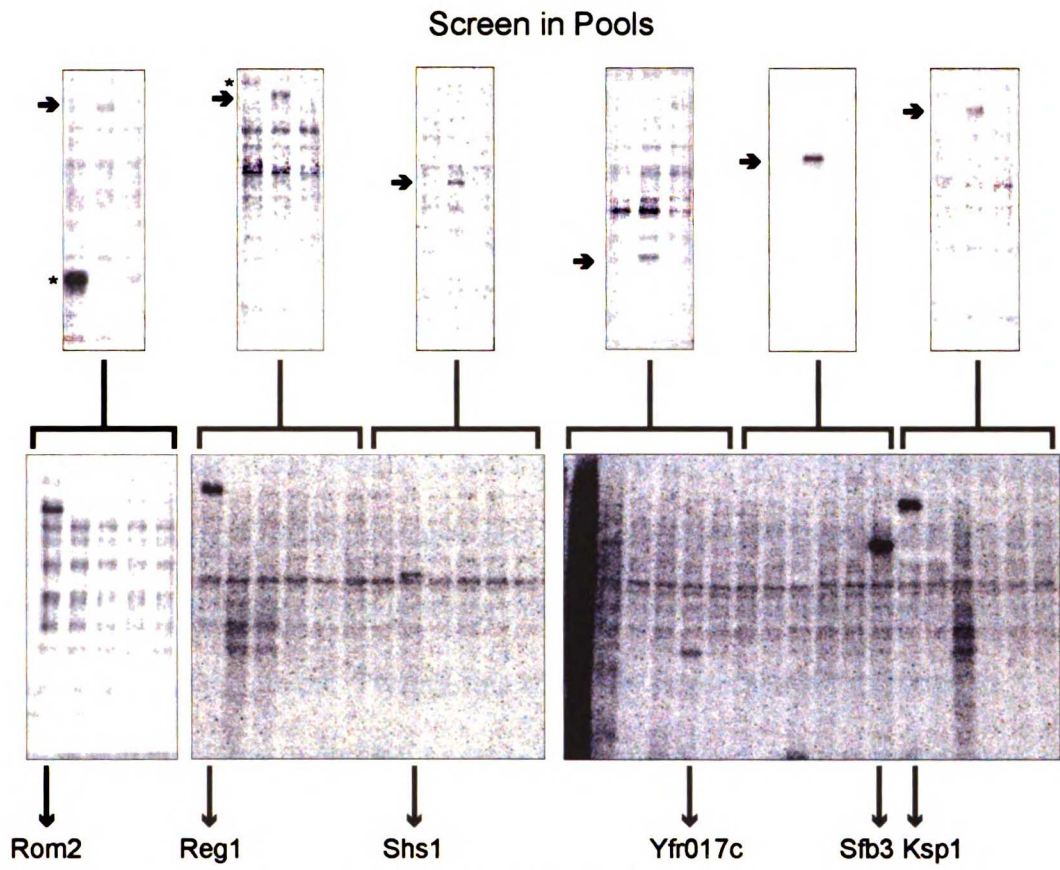
Screen Individual Strains
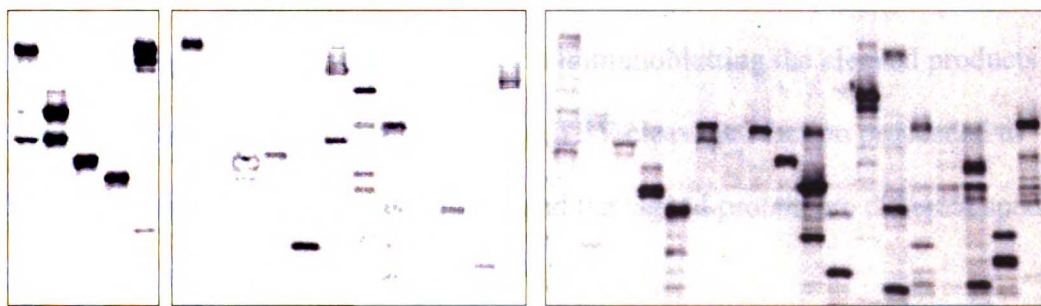
Immunoblots

Because the immunopurifications were performed in non-denaturing conditions, it was not possible to distinguish between signals arising from the tagged protein and those from co-precipitating, untagged proteins. To determine which of these proteins were direct substrates of Pho85-Pcl1, we took advantage of the protease cleavage sequence in the TAP tag [87]. This sequence is specifically recognized and cleaved by the tobacco etch virus (TEV) protease and lies amino terminal to the dual Protein A domains in the tag. If the radiolabeled protein visible from the phosphorimager scan is the TAP tagged protein, cleavage and release of the Protein A domains will result in an increase in electrophoretic mobility visible on the phosphorimager scan. We repeated reactions for the 34 positive strains in duplicate, adding TEV protease to one set of reactions after immunopurification and mock treating the others. Twenty-four of our 34 targets showed a change in electrophoretic mobility upon TEV protease treatment (Fig. 6 and data not shown), indicating that they were direct targets of Pho85-Pcl1 (Table 2). Additionally, for these 24 targets, the signals from the radiolabeled band and those detected by immunoblotting co-migrated. The ten strains showing no shift in the radiolabeled band upon TEV cleavage displayed clearly shifted signals upon immunoblotting the cleaved products (Fig. 6 and data not shown), indicating that the TEV cleavage reaction proceeded to completion and that the radiolabeled protein and the tagged protein are different species.

The Pho85 cyclins are believed to direct Pho85 kinase to different substrates [88]. In support of this, none of the pcl's can substitute for the function of Pho80 in phosphorylating the transcription factor Pho4 *in vivo* [31]. This specificity is recapitulated *in vitro* in extract kinase assays as well as in reactions with purified Figure

Figure 6. TEV protease cleavage assay to identify direct substrates. Kinase reactions and immunopurifications were performed in duplicate and either treated with TEV protease (right lane of each pair) or mock-treated (left lane) after immunopurification. The reaction products were separated on polyacrylamide gels, transferred to a membrane, scanned on a phosphorimager (left), and immunoblotted for the TAP tag (right). A shift in the electrophoretic mobility of the radiolabeled band upon protease treatment indicates that the phosphorylated protein contains the TEV recognition sequence.

Phosphorimager

Immunoblot
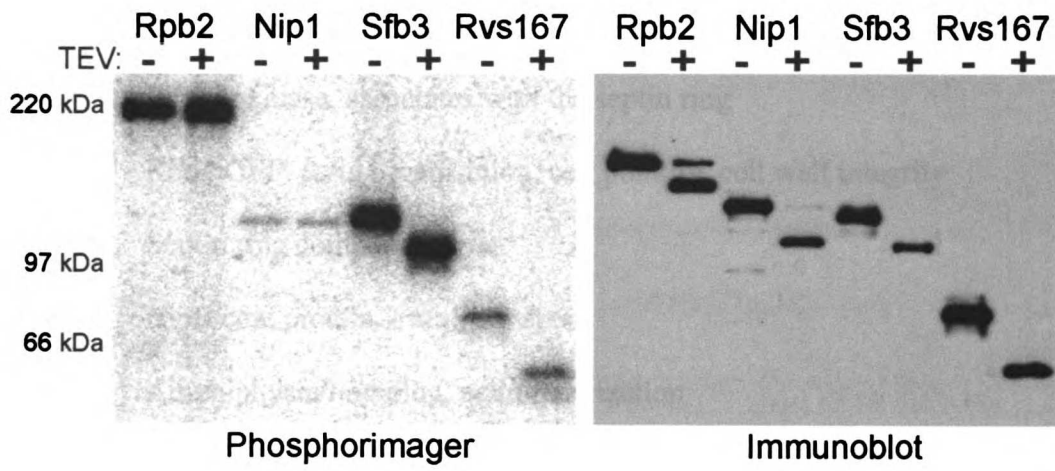
Table 3. TEV cleavage confirmed substrates

Protein	Description (Yeast Proteome Database [89])

Pho85-Pcl1 specific substrates

Protein	Description
Dma1	Regulation of spindle position and septin ring assembly
Isr1	Protein kinase, inhibits staurosporine resistance
Pan3	Pab1 dependent ribonuclease
Kcc4	Protein kinase, associates with the septin ring
Rom2	Rho1 GEF, actin organization, cell polarity, cell wall integrity
Shs1	Septin ring component
Sfb3	copII coat protein, Pma1 transport
Rvs167	Amphiphysin homolog, actin polarization

Pho85-Pho80 specific substrates

Protein	Description
Pho81*	Pho85/Pho80 cyclin dependent kinase inhibitor
Glc8*	Glc7 phosphatase regulatory subunit
Pho4*	Transcription factor, phosphate starvation response
Vip1	Actin organization and biogenesis
Smp2	Respiration, plasmid maintenance

Substrates phosphorylated by both kinases

Protein	Description
Ame1	COMA complex member, functions in kinetochore assembly
Bcp1	Export of Mss4 and 60S ribosomal subunit components
Gts1	Regulator of metabolic oscillations
Cdc19	Pyruvate kinase

Ksp1	Protein kinase, suppressor of *prp20-10*
Npl3	mRNA binding protein, mRNA nuclear export
Pam1	Multi-copy suppressor of *pph21-Δ, pph22-Δ, sit4-Δ* lethality
Reg1	Glc7 phosphatase regulatory subunit
Rpg1	eIF3 translation initiation factor subunit
Sap185	Sit4 phosphatase regulatory protein
Ssd1	Cell wall organization and biogenesis, suppressor of *sit4-Δ*
Ugp1	UDP-glucose phosphorylase, glycogen and glucan synthesis
YFR017C	Unknown
YNR047W	Protein kinase, suppresses alpha-factor induced arrest

*Pho81, Glc8, and Pho4 were identified in a pilot screen using Pho80/85

components. In both cases we observe that Pho80-Pho85 phosphorylates Pho4 with

dramatically higher efficiency than do Pcl1-Pho85 and Pcl7-Pho85 (Fig. 7, Fig 8, and

N.D. unpublished results). Pho4 interacts directly with Pho80, which recruits this

substrate to the kinase [31, 90]. Rvs167, a putative substrate for Pcl2-Pho85 , also

physically interacts with Pcl2 [23], suggesting that direct physical interaction between the

cyclin and substrate may mediate substrate specificity. This mechanism for enhancing

kinase specificity is also used by other CDKs involved in cell cycle progression [91]. It

is also likely that subcellular localization of Pcl-Pho85 complexes contributes to the

selection of correct substrates. Pho80-Pho85 is localized exclusively to the nucleus,

where it phosphorylates its substrate, Pho4 [92]. Pcl1 is localized partially to the nucleus,

but is also found at the incipient bud site [93]. The closely related cyclin Pcl2 is found at

both of these sites as well as at the growing bud tip and bud neck [93].

To investigate the specificity of Pcl1-Pho85 for the candidate substrates, we compared

the relative activity of Pho85-Pho80 and Pho85-Pcl1 towards each substrate in parallel

kinase assays on identical aliquots of extracts derived from TAP-tagged strains (Fig. 7

and Fig. 8). We observed three classes of substrates: Pho85-Pcl1 specific; those

phosphorylated equally well by both kinases; and Pho85-Pho80 specific. Eight proteins

were better phosphorylated by Pho85-Pcl1. Two proteins, Vip1 and Smp2, were better

phosphorylated by Pho85-Pho80. Three additional proteins, identified in a pilot screen

using Pho85-Pho80 kinase, were also tested in this assay and showed a marked

Figure 7. Kinase – substrate specificity. Parallel reactions using Pho85(F82G)-Pho80 (left lane of each pair) or Pho85(F82G)-Pcl1 kinase (right lane) were performed on identical aliquots of the indicated TAP-library strains and analyzed as before. Shown are phosphorimager scans of a representative experiment.
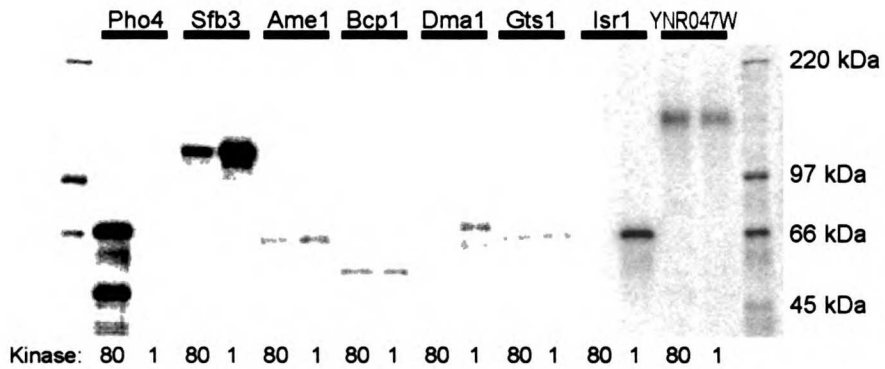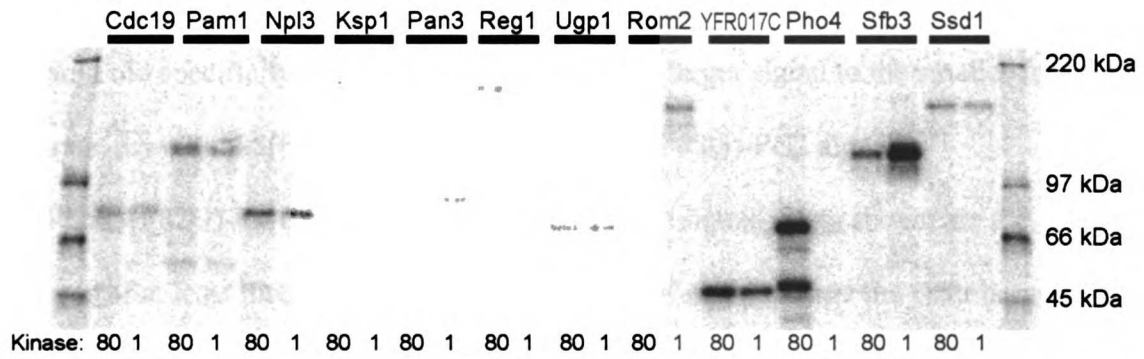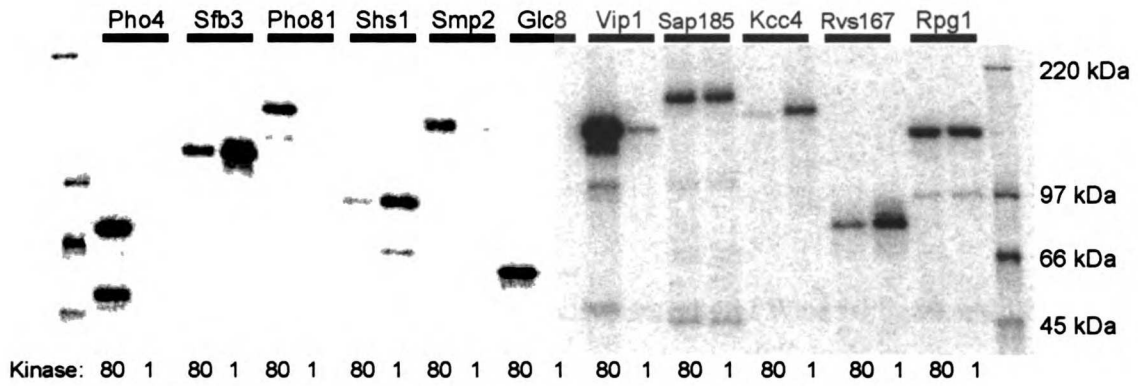
Figure 8. Kinase – substrate specificity; relative activity of Pho85-Pho80 and Pho85-Pcl1 towards each substrate. Kinase reactions were performed on identical aliquots of each extract with Pho85(F82G)-Pho80 or Pho85(F82G)-Pcl1 kinase and analysed as before. Fold specificity is defined as the ratio of the larger signal to the smaller for each substrate. ☐ = Pho85(F82G)-Pho80 signal / Pho85(F82G)-Pcl1 signal; ■ = Pho85(F82G)-Pcl1 signal / Pho85(F82G)-Pho80 signal. Data shown are the averages of at least three experiments for each pair of reactions and the error bars represent the mean +/- standard deviation.

Fold Specificity for Pcl1/Pho85

Fold Specificity for Pho80/Pho85
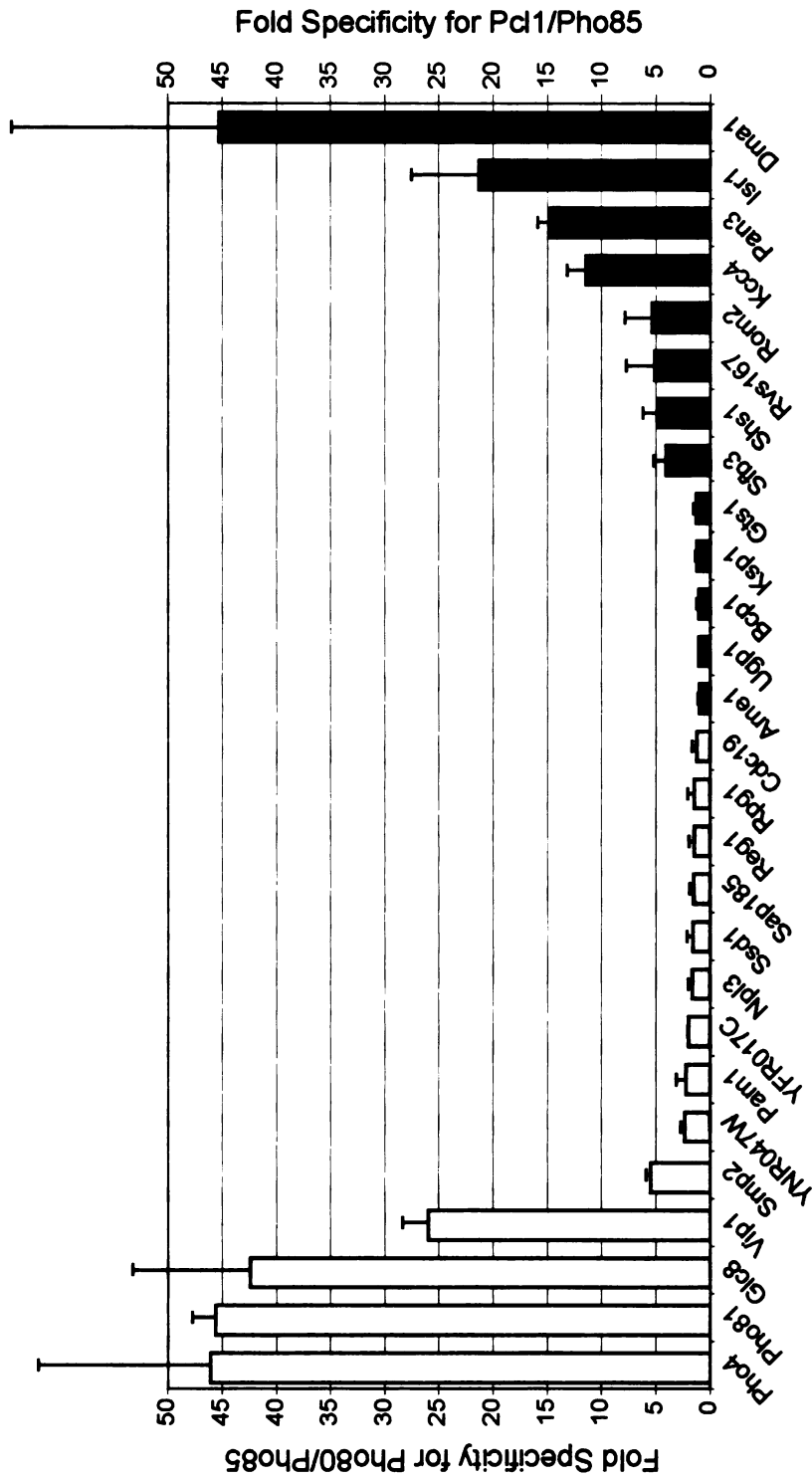
specificity for Pho85-Pho80. Fourteen of our candidates showed no significant specificity for either kinase. *In vivo*, conditions not preserved in our assay, such as subcellular localization may contribute to specificity or it may be that these proteins are more specifically phosphorylated by another Pho85 kinase complex. The high degree of specificity observed for many of the newly-identified substrates suggests that they may be bona fide *in vivo* targets of Pho85-Pcl1 and Pho85-Pho80.

The cyclins Cln1, Cln2, Pcl1, and Pcl2 are essential for proper septin ring dynamics and cellular polarization [93]. These functions are consistent with the localization of Pcl1 and Pcl2 proteins to the incipient bud site, the growing bud tip, and the bud neck. In a recent genome wide localization analysis, 124 of the 4156 (3%) proteins assigned a localization were found at the bud neck and/or in the bud [77]. Twenty-two of the candidate substrates we identified have known cellular localizations. Five of them (Shs1, Pam1, Kcc4, Rpg1, and Rom2) (~23%), are found at one or both of these regions [77, 94-97], greater than seven-fold more than would be expected by random sampling. Among the five bud neck and bud localized substrates are three of the six Pcl1-specific substrates, including a structural component of the septin ring and known regulators of septin dynamics and actin polarization. An additional target, Dma1, plays a role in septin ring assembly [98]. Co-localization of these substrates and Pcl1-Pho85 further suggests that the substrates we have identified are specific for this kinase.

Synthetic lethal screens with *pho85* have revealed a number of functional groups of Pho85-interacting genes, implying roles for Pho85 in the processes these gene products

control [99, 100]. Genes whose products are involved in cell wall formation, maintenance, and regulation, including members of the Pkc/MAP kinase cell integrity sensing pathway, constitute the largest of the Pho85 synthetic lethal gene classifications. Of the 24 direct substrates, (Table 2) seven have known roles in cell wall formation, maintenance, and regulation [101-105] and/or interact with elements of the Pkc/MAP kinase pathway [26, 106-110] (Fig.s 9A and 9B). The second largest group of genes synthetically lethal with *pho85* act in polarized cell growth. Eight of the 24 substrates we identified are among 314 genes classified in the MIPS Functional Catalogue under the closely related heading 'budding, cell polarity, and filament formation' [111] (Fig. 9A). It is also notable that 4 of the 8 'budding, cell polarity and filament formation' candidates (Kcc4, Rom2, Rvs167, and Shs1) are Pcl1-specific substrates and that Pcl1 localizes to sites of polarized cell growth [93]. In all, 12 of the 24 substrates fall into just two Pho85-related functional categories, further suggesting that many of them will prove to be true substrates. Further characterization of these targets will likely help to better define the role of Pho85-Pcl1 in the cell cycle and in cell polarization. For example, Kcc4, Rom2, and Shs1 may represent new downstream effectors of Pho85-Pcl1 that help it carry out its essential function in the absence of Cln1 and Cln2.

One of the challenges in adapting methods to a proteomic scale is maintaining sufficient sensitivity. Despite using only ~2OD units of cells, we were able to detect many proteins of extremely low abundance, including seven present in less than 1000 copies per cell and an 8th whose expression was too low to be detected without immunoprecipitation [76]. Despite our ability to identify a number of low abundance proteins, we likely

missed some substrates that fell below the limits of detection or that were obscured by the variable background signal. Four proteins have been reported to be phosphorylated in a Pho85-dependent manner and can be phosphorylated by Pho85-Pcl1 kinase in vitro (Rvs167, Gcn4, Cdc24, Sic1). Only one of these, Rvs167 was identified in our screen. We assayed the others and observed no signal for Gcn4 and very weak phosphorylation for Cdc24 and Sic1 (data not shown). These proteins may have been missed because of their low abundance. Alternatively, they may not be bona fide Pho85-Pcl1 substrates. During the course of our work, we were able to improve our assay by using a different bead system that nearly eliminated the background signal (compare the gels in fig. 5 and fig. 7). Using this system and dilutions of Pho4-TAP extract, recombinant Pho4-zz, and a recombinant Pho4^{SA1234}-ZZ mutant that can only be singly phosphorylated[10], we measured the detection limit of our assay. The assay can reliably detect singly phosphorylated proteins that are present in 800 copies per cell (data not shown). Proteins that can be multiply phosphorylated, such as Pho4, can be detected at even lower copy numbers. Using this new bead system, finding ways to forgo the use of pools, and increasing the amount and/or quality of extract will further increase sensitivity in future screens.

We have demonstrated a rapid and sensitive method to identify protein kinase substrates using full-length proteins at natural abundance levels. The ability of this assay to identify known substrates of Pho85 kinases, the functional overlap between the identified targets and Pho85-Pcl1, and the co-localization of the targets and kinase, suggest that many of

Figure 9. Functional and spatial overlap between Pho85 and substrates.

(a) Venn diagram showing some of the substrates that function in Pho85-related

processes and/or that colocalize with Pcl1, Pcl2. Additional genetic and physical

interactions are indicated. (b) Network diagram depicting interactions between Pho85, its

substrates, and the Pkc/MAP kinase pathway. o = Pho85 substrate; ● = non-substrate

protein; ▬▬ = physical interaction; .... = genetic interaction; ——▶ = kinase-substrate

interaction.

A.

Budneck/budtip localization

Cell wall formation, maintenance, and regulation
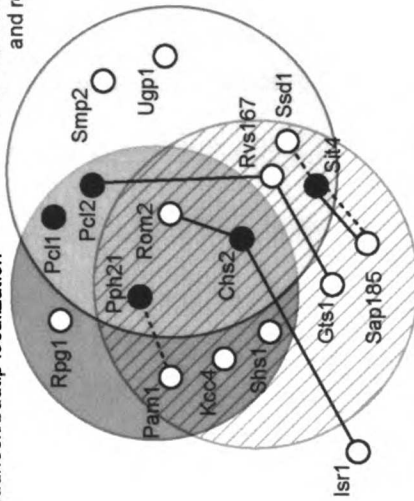
Budding, cell polarity and filament formation

B.

Pkc1/MAPK
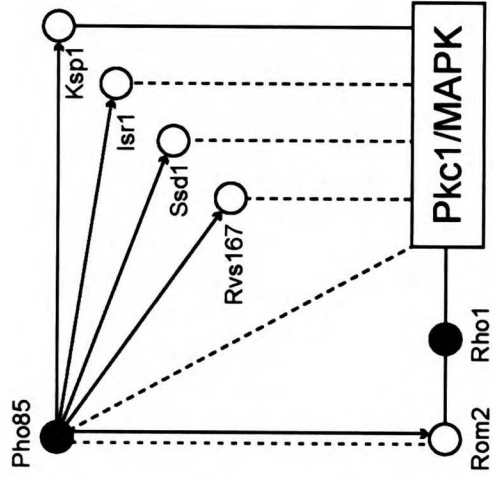
them will prove to be bona fide. This method is readily adaptable to other yeast kinases and with modifications can be used to screen for substrates in other systems.

Acknowledgments

References

1. Krebs, E.G., and Fischer, E.H. (1956). The phosphorylase b to a converting enzyme of rabbit skeletal muscle. Biochim Biophys Acta 20, 150-157.

2. Hunter, T., and Plowman, G.D. (1997). The protein kinases of budding yeast: six score and more. Trends Biochem Sci 22, 18-22.

3. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science 298, 1912-1934.

4. Morgan, D.O. (1997). Cyclin-dependent kinases: engines, clocks, and microprocessors. Annu Rev Cell Dev Biol 13, 261-291.

5. Hunter, T. (2000). Signaling--2000 and beyond. Cell 100, 113-127.

6. Karin, M., and Hunter, T. (1995). Transcriptional control by protein phosphorylation: signal transmission from the cell surface to the nucleus. Curr Biol 5, 747-757.

7. Schneider, B.L., Yang, Q.H., and Futcher, A.B. (1996). Linkage of replication to start by the Cdk inhibitor Sic1. Science 272, 560-562.

8. Nash, P., Tang, X., Orlicky, S., Chen, Q., Gertler, F.B., Mendenhall, M.D., Sicheri, F., Pawson, T., and Tyers, M. (2001). Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. Nature 414, 514-521.

9. 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.

10. Komeili, A., and O'Shea, E.K. (1999). Roles of phosphorylation sites in regulating activity of the transcription factor Pho4. Science *284*, 977-980.

11. Cohen, P. (2001). The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. Eur J Biochem *268*, 5001-5010.

12. Blume-Jensen, P., and Hunter, T. (2001). Oncogenic kinase signalling. Nature *411*, 355-365.

13. Hug, G., Schubert, W.K., and Chuck, G. (1966). Phosphorylase kinase of the liver: deficiency in a girl with increased hepatic glycogen. Science *153*, 1534-1535.

14. Hendrickx, J., and Willems, P.J. (1996). Genetic deficiencies of the glycogen phosphorylase system. Hum Genet *97*, 551-556.

15. Daley, G.Q., Van Etten, R.A., and Baltimore, D. (1990). Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. Science *247*, 824-830.

16. Heisterkamp, N., Jenster, G., ten Hoeve, J., Zovich, D., Pattengale, P.K., and Groffen, J. (1990). Acute leukaemia in bcr/abl transgenic mice. Nature *344*, 251-253.

17. Cohen, P. (2002). Protein kinases--the major drug targets of the twenty-first century? Nat Rev Drug Discov *1*, 309-315.

18. Capdeville, R., Buchdunger, E., Zimmermann, J., and Matter, A. (2002). Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. Nat Rev Drug Discov *1*, 493-502.

19. Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci *27*, 514-520.

20. Hanks, S.K., and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. Faseb J *9*, 576-596.

21. Hartwell, L.H. (1980). Mutants of Saccharomyces cerevisiae unresponsive to cell division control by polypeptide mating hormone. J Cell Biol *85*, 811-822.

22. Neiman, A.M., and Herskowitz, I. (1994). Reconstitution of a yeast protein kinase cascade in vitro: activation of the yeast MEK homologue STE7 by STE11. Proc Natl Acad Sci U S A *91*, 3398-3402.

23. Lee, J., Colwill, K., Aneliunas, V., Tennyson, C., Moore, L., Ho, Y., and Andrews, B. (1998). Interaction of yeast Rvs167 and Pho85 cyclin-dependent kinase complexes may link the cell cycle to the actin cytoskeleton. Curr Biol *8*, 1310-1321.

24. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

25. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C.,

Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

26.   Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-183.

27.   Wanke, V., Pedruzzi, I., Cameroni, E., Dubouloz, F., and De Virgilio, C. (2005). Regulation of G0 entry by the Pho80-Pho85 cyclin-CDK complex. Embo J *24*, 4271-4278.

28.   Nishizawa, M., Kawasumi, M., Fujino, M., and Toh-e, A. (1998). Phosphorylation of sic1, a cyclin-dependent kinase (Cdk) inhibitor, by Cdk including Pho85 kinase is required for its prompt degradation. Mol Biol Cell *9*, 2393-2405.

29.   Tan, Y.S., Morcos, P.A., and Cannon, J.F. (2003). Pho85 phosphorylates the Glc7 protein phosphatase regulator Glc8 in vivo. J Biol Chem *278*, 147-153.

30. Huang, D., Moffat, J., Wilson, W.A., Moore, L., Cheng, C., Roach, P.J., and Andrews, B. (1998). Cyclin partners determine Pho85 protein kinase substrate specificity in vitro and in vivo: control of glycogen biosynthesis by Pcl8 and Pcl10. Mol Cell Biol *18*, 3289-3299.

31. Kaffman, A., Herskowitz, I., Tjian, R., and O'Shea, E.K. (1994). Phosphorylation of the transcription factor PHO4 by a cyclin-CDK complex, PHO80-PHO85. Science *263*, 1153-1156.

32. Shi, X.Z., and Ao, S.Z. (2002). Analysis of phosphorylation of YJL084c, a yeast protein. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai) *34*, 433-438.

33. Measday, V., McBride, H., Moffat, J., Stillman, D., and Andrews, B. (2000). Interactions between Pho85 cyclin-dependent kinase complexes and the Swi5 transcription factor in budding yeast. Mol Microbiol *35*, 825-834.

34. Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., McCartney, R.R., Schmidt, M.C., Rachidi, N., Lee, S.J., Mah, A.S., Meng, L., Stark, M.J., Stern, D.F., De Virgilio, C., Tyers, M., Andrews, B., Gerstein, M., Schweitzer, B., Predki, P.F., and Snyder, M. (2005). Global analysis of protein phosphorylation in yeast. Nature *438*, 679-684.

35. Dephoure, N., Howson, R.W., Blethrow, J.D., Shokat, K.M., and O'Shea, E.K. (2005). Combining chemical genetics and proteomics to identify protein kinase substrates. Proc Natl Acad Sci U S A *102*, 17940-17945.

36. Berwick, D.C., and Tavare, J.M. (2004). Identifying protein kinase substrates: hunting for the organ-grinder's monkeys. Trends Biochem Sci *29*, 227-232.

37. Johnson, S.A., and Hunter, T. (2005). Kinomics: methods for deciphering the kinome. Nat Methods *2*, 17-25.

38. Manning, B.D., and Cantley, L.C. (2002). Hitting the target: emerging technologies in the search for kinase substrates. Sci STKE *2002*, PE49.

39. 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.

40. Shah, K., Liu, Y., Deirmengian, C., and Shokat, K.M. (1997). Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. Proc Natl Acad Sci U S A *94*, 3565-3570.

41. Bishop, A.C., Shah, K., Liu, Y., Witucki, L., Kung, C., and Shokat, K.M. (1998). Design of allele-specific inhibitors to probe protein kinase signaling. Curr Biol *8*, 257-266.

42. Ubersax, J.A., Woodbury, E.L., Quang, P.N., Paraz, M., Blethrow, J.D., Shah, K., Shokat, K.M., and Morgan, D.O. (2003). Targets of the cyclin-dependent kinase Cdk1. Nature *425*, 859-864.

43. Habelhah, H., Shah, K., Huang, L., Burlingame, A.L., Shokat, K.M., and Ronai, Z. (2001). Identification of new JNK substrate using ATP pocket mutant JNK and a corresponding ATP analogue. J Biol Chem *276*, 18090-18095.

44. Bishop, A.C., Ubersax, J.A., Petsch, D.T., Matheos, D.P., Gray, N.S., Blethrow, J., Shimizu, E., Tsien, J.Z., Schultz, P.G., Rose, M.D., Wood, J.L., Morgan, D.O., and Shokat, K.M. (2000). A chemical switch for inhibitor-sensitive alleles of any protein kinase. Nature *407*, 395-401.

45.   Bishop, A.C., Buzko, O., and Shokat, K.M. (2001). Magic bullets for protein kinases. Trends Cell Biol *11*, 167-172.

46.   Lenburg, M.E., and O'Shea, E.K. (1996). Signaling phosphate starvation. Trends Biochem Sci *21*, 383-387.

47.   Measday, V., Moore, L., Ogas, J., Tyers, M., and Andrews, B. (1994). The PCL2 (ORFD)-PHO85 cyclin-dependent kinase complex: a cell cycle regulator in yeast. Science *266*, 1391-1395.

48.   Espinoza, F.H., Ogas, J., Herskowitz, I., and Morgan, D.O. (1994). Cell cycle control by a complex of the cyclin HCS26 (PCL1) and the kinase PHO85. Science *266*, 1388-1391.

49.   Moffat, J., Huang, D., and Andrews, B. (2000). Functions of Pho85 cyclin-dependent kinases in budding yeast. Prog Cell Cycle Res *4*, 97-106.

50.   Carroll, A.S., Bishop, A.C., DeRisi, J.L., Shokat, K.M., and O'Shea, E.K. (2001). Chemical inhibition of the Pho85 cyclin-dependent kinase reveals a role in the environmental stress response. Proc Natl Acad Sci U S A *98*, 12578-12583.

51.   Zhang, C., Kenski, D.M., Paulson, J.L., Bonshtien, A., Sessa, G., Cross, J.V., Templeton, D.J., and Shokat, K.M. (2005). A second-site suppressor strategy for chemical genetic analysis of diverse protein kinases. Nat Methods *2*, 435-441.

52.   Ulrich, S.M., Kenski, D.M., and Shokat, K.M. (2003). Engineering a "methionine clamp" into Src family kinases enhances specificity toward unnatural ATP analogues. Biochemistry *42*, 7915-7921.

53.   Steinberg, T.H., Agnew, B.J., Gee, K.R., Leung, W.Y., Goodman, T., Schulenberg, B., Hendrickson, J., Beechem, J.M., Haugland, R.P., and Patton,

W.F. (2003). Global quantitative phosphoprotein analysis using Multiplexed Proteomics technology. Proteomics *3*, 1128-1144.

54.   Kinoshita, E., Yamada, A., Takeda, H., Kinoshita-Kikuta, E., and Koike, T. (2005). Novel immobilized zinc(II) affinity chromatography for phosphopeptides and phosphorylated proteins. J Sep Sci *28*, 155-162.

55.   Kinoshita, E., Kinoshita-Kikuta, E., Takiyama, K., and Koike, T. (2006). Phosphate-binding tag, a new tool to visualize phosphorylated proteins. Mol Cell Proteomics *5*, 749-757.

56.   Ficarro, S.B., McCleland, Mark L., Stukenberg, P. Todd , Burke, Daniel J., Ross, Mark M., Shabanowitz, Jeffrey, Hunt, Donald F., White, Forest M. (2002). Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol *20*, 301-305.

57.   Domon, B., and Aebersold, R. (2006). Mass spectrometry and protein analysis. Science *312*, 212-217.

58.   Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198-207.

59.   Loyet, K.M., Stults, J.T., and Arnott, D. (2005). Mass spectrometric contributions to the practice of phosphorylation site mapping through 2003: a literature review. Mol Cell Proteomics *4*, 235-245.

60.   Arnott, D., Gawinowicz, M.A., Grant, R.A., Neubert, T.A., Packman, L.C., Speicher, K.D., Stone, K., and Turck, C.W. (2003). ABRF-PRG03: phosphorylation site determination. J Biomol Tech *14*, 205-215.

61. Goffeau, A. (2000). Four years of post-genomic life with 6,000 yeast genes. FEBS Lett *480*, 37-41.

62. Brachat, S., Dietrich, F.S., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T., and Philippsen, P. (2003). Reinvestigation of the Saccharomyces cerevisiae genome annotation by comparison to the genome of a related fungus: Ashbya gossypii. Genome Biol *4*, R45.

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

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

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

66. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol *14*, 1675-1680.

67. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

68. Martzen, M.R., McCraith, S.M., Spinelli, S.L., Torres, F.M., Fields, S., Grayhack, E.J., and Phizicky, E.M. (1999). A biochemical genomics approach for identifying genes by the activity of their products. Science *286*, 1153-1155.

69. MacBeath, G., and Schreiber, S.L. (2000). Printing proteins as microarrays for high-throughput function determination. Science *289*, 1760-1763.

70. Phizicky, E., Bastiaens, P.I., Zhu, H., Snyder, M., and Fields, S. (2003). Protein analysis on a proteomic scale. Nature *422*, 208-215.

71. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M., and Snyder, M. (2001). Global analysis of protein activities using proteome chips. Science *293*, 2101-2105.

72. (1997). The yeast genome directory. Nature *387*, 5.

73. Diehn, M., Eisen, M.B., Botstein, D., and Brown, P.O. (2000). Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat Genet *25*, 58-62.

74. Longtine, M.S., McKenzie, A., 3rd, Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P., and Pringle, J.R. (1998). Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast *14*, 953-961.

75. Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol *132*, 365-386.

76. Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-741.

77. Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686-691.

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

79. Liu, Y., Shah, K., Yang, F., Witucki, L., and Shokat, K.M. (1998). Engineering Src family protein kinases with unnatural nucleotide specificity. Chem Biol *5*, 91-101.

80. Shah, K., and Shokat, K.M. (2002). A chemical genetic screen for direct v-Src substrates reveals ordered assembly of a retrograde signaling pathway. Chem Biol *9*, 35-47.

81. Howson, R., Huh, W.K., Ghaemmaghami, S., Falvo, J.V., Bower, K., Belle, A., Dephoure, N., Wykoff, D.D., Weissman, J.S., and O'Shea, E.K. (2005). Construction, verification and experimental use of two epitope-tagged collections budding yeast strains. Comparative and Functional Genomics *6*, 2-16.

82. Jeffery, D.A., Springer, M., King, D.S., and O'Shea, E.K. (2001). Multi-site phosphorylation of Pho4 by the cyclin-CDK Pho80-Pho85 is semi-processive with site preference. J Mol Biol *306*, 997-1010.

83.   Edelhoch, H. (1967). Spectroscopic determination of tryptophan and tyrosine in proteins. Biochemistry *6*, 1948-1954.

84.   Kaffman, A., Rank, N.M., and O'Shea, E.K. (1998). Phosphorylation regulates association of the transcription factor Pho4 with its import receptor Pse1/Kap121. Genes Dev *12*, 2673-2683.

85.   Kraybill, B.C., Elkin, L.L., Blethrow, J.D., Morgan, D.O., and Shokat, K.M. (2002). Inhibitor scaffolds as new allele specific kinase substrates. J Am Chem Soc *124*, 12118-12128.

86.   Haswell, E.S., and O'Shea, E.K. (1999). An in vitro system recapitulates chromatin remodeling at the PHO5 promoter. Mol Cell Biol *19*, 2817-2827.

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

88.   Carroll, A.S., and O'Shea, E.K. (2002). Pho85 and signaling environmental conditions. Trends Biochem Sci *27*, 87-93.

89.   Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C., Roberg-Perez, K.J., Tillberg, M., Brooks, J.E., and Garrels, J.I. (2000). The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res *28*, 73-76.

90.   Byrne, M., Miller, N., Springer, M., and O'Shea, E.K. (2004). A distal, high-affinity binding site on the cyclin-CDK substrate Pho4 is important for its phosphorylation and regulation. J Mol Biol *335*, 57-70.

91.   Loog, M., and Morgan, D.O. (2005). Cyclin specificity in the phosphorylation of cyclin-dependent kinase substrates. Nature *434*, 104-108.

92.   Kaffman, A., Rank, N.M., O'Neill, E.M., Huang, L.S., and O'Shea, E.K. (1998). The receptor Msn5 exports the phosphorylated transcription factor Pho4 out of the nucleus. Nature *396*, 482-486.

93.   Moffat, J., and Andrews, B. (2004). Late-G1 cyclin-CDK activity is essential for control of cell morphogenesis in budding yeast. Nat Cell Biol *6*, 59-66.

94.   Mino, A., Tanaka, K., Kamei, T., Umikawa, M., Fujiwara, T., and Takai, Y. (1998). Shs1p: a novel member of septin that interacts with spa2p, involved in polarized growth in saccharomyces cerevisiae. Biochem Biophys Res Commun *251*, 732-736.

95.   Barral, Y., Parra, M., Bidlingmaier, S., and Snyder, M. (1999). Nim1-related kinases coordinate cell cycle progression with the organization of the peripheral cytoskeleton in yeast. Genes Dev *13*, 176-187.

96.   Hasek, J., Kovarik, P., Valasek, L., Malinska, K., Schneider, J., Kohlwein, S.D., and Ruis, H. (2000). Rpg1p, the subunit of the Saccharomyces cerevisiae eIF3 core complex, is a microtubule-interacting protein. Cell Motil Cytoskeleton *45*, 235-246.

97.   Manning, B.D., Padmanabha, R., and Snyder, M. (1997). The Rho-GEF Rom2p

        localizes to sites of polarized cell growth and participates in cytoskeletal functions

        in Saccharomyces cerevisiae. Mol Biol Cell *8*, 1829-1844.

98.   Fraschini, R., Bilotta, D., Lucchini, G., and Piatti, S. (2004). Functional

        characterization of Dma1 and Dma2, the budding yeast homologues of

        Schizosaccharomyces pombe Dma1 and human Chfr. Mol Biol Cell *15*, 3796-

        3810.

99.   Lenburg, M.E., and O'Shea, E.K. (2001). Genetic evidence for a morphogenetic

        function of the Saccharomyces cerevisiae Pho85 cyclin-dependent kinase.

        Genetics *157*, 39-51.

100.   Huang, D., Moffat, J., and Andrews, B. (2002). Dissection of a complex

        phenotype by functional genomics reveals roles for the yeast cyclin-dependent

        protein kinase Pho85 in stress adaptation and cell integrity. Mol Cell Biol *22*,

        5076-5088.

101.   Daran, J.M., Dallies, N., Thines-Sempoux, D., Paquet, V., and Francois, J. (1995).

        Genetic and biochemical characterization of the UGP1 gene encoding the UDP-

        glucose pyrophosphorylase from Saccharomyces cerevisiae. Eur J Biochem *233*,

        520-530.

102.   Lussier, M., White, A.M., Sheraton, J., di Paolo, T., Treadwell, J., Southard, S.B.,

        Horenstein, C.I., Chen-Weiner, J., Ram, A.F., Kapteyn, J.C., Roemer, T.W., Vo,

        D.H., Bondoc, D.C., Hall, J., Zhong, W.W., Sdicu, A.M., Davies, J., Klis, F.M.,

        Robbins, P.W., and Bussey, H. (1997). Large scale identification of genes

involved in cell surface biosynthesis and architecture in Saccharomyces cerevisiae. Genetics *147*, 435-450.

103.   Kaeberlein, M., and Guarente, L. (2002). Saccharomyces cerevisiae MPT5 and SSD1 function in parallel pathways to promote cell wall integrity. Genetics *160*, 83-95.

104.   Bickle, M., Delley, P.A., Schmidt, A., and Hall, M.N. (1998). Cell wall integrity modulates RHO1 activity via the exchange factor ROM2. Embo J *17*, 2235-2245.

105.   Breton, A.M., Schaeffer, J., and Aigle, M. (2001). The yeast Rvs161 and Rvs167 proteins are involved in secretory vesicles targeting the plasma membrane and in cell integrity. Yeast *18*, 1053-1068.

106.   Uesono, Y., Toh-e, A., and Kikuchi, Y. (1997). Ssd1p of Saccharomyces cerevisiae associates with RNA. J Biol Chem *272*, 16103-16109.

107.   Ozaki, K., Tanaka, K., Imamura, H., Hihara, T., Kameyama, T., Nonaka, H., Hirano, H., Matsuura, Y., and Takai, Y. (1996). Rom1p and Rom2p are GDP/GTP exchange proteins (GEPs) for the Rho1p small GTP binding protein in Saccharomyces cerevisiae. Embo J *15*, 2196-2207.

108.   Miyahara, K., Hirata, D., and Miyakawa, T. (1998). Functional interaction of Isr1, a predicted protein kinase, with the Pkc1 pathway in Saccharomyces cerevisiae. Biosci Biotechnol Biochem *62*, 1376-1380.

109.   Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Menard, P., Munyana, C., Parsons, A.B., Ryan,

O., Tonikian, R., Roberts, T., Sdicu, A.M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H., and Boone, C. (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808-813.

110.   Bon, E., Recordon-Navarro, P., Durrens, P., Iwase, M., Toh, E.A., and Aigle, M. (2000). A network of proteins around Rvs167p and Rvs161p, two proteins related to the yeast actin cytoskeleton. Yeast *16*, 1229-1241.

111.   Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., and Mewes, H.W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res *32*, 5539-5545.

112.   Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., and Davis, R.W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A *94*, 13057-13062.

113.   Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer, M.S. (2003). A vision for the future of genomics research. Nature *422*, 835-847.

114.   SGD.

115.   Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Davis,

R.W., and et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901-906.

116.   Schwob, E., Bohm, T., Mendenhall, M.D., and Nasmyth, K. (1994). The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae. Cell *79*, 233-244.

117.   Grandin, N., and Reed, S.I. (1993). Differential function and expression of Saccharomyces cerevisiae B-type cyclins in mitosis and meiosis. Mol Cell Biol *13*, 2113-2125.

118.   Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell *9*, 3273-3297.

119.   Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization of budding yeast.

120.   Harrison, P.M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res *30*, 1083-1090.

121.   Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Bienkowska, J., Adams, R.M., Smith, T.F., and Lindelien, J. (1997). Biology's new Rosetta stone. Nature *385*, 29-30.

122. Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M.R., and Cebrat, S. (1999). Total number of coding open reading frames in the yeast genome. Yeast *15*, 1031-1034.

123. Zhang, C.T., and Wang, J. (2000). Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Res *28*, 2804-2814.

124. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. Nat Genet *31*, 370-377.

125. Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys *67*, 031902.

126. Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol Cell Biol *19*, 1720-1730.

127. Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S., and Garrels, J.I. (1999). A sampling of the yeast proteome. Mol Cell Biol *19*, 7357-7368.

128. Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E., and Yates, J.R., 3rd (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *100*, 3107-3112.

129. Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol *19*, 242-247.

130. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

131. Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D., and Brown, P.O. (2002). Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A *99*, 5860-5865.

132. Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res *15*, 1281-1295.

133. Grantham, R., Gautier, C., and Gouy, M. (1980). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res *8*, 1893-1912.

Appendix 1

Global Analysis of Protein Expression

Credits

This work is the result of a collaboration between our laboratory and the laboratory of

Jonathan Weissman. This work complements and depends entirely upon the resources

developed in the work of Chapter 1. However I did not perform any of the experiments

described, thus it is presented here as an appendix.

Reprinted by permission of the authors from Nature, Volume 425, Sina

Ghaemmaghami, Won-Ki Huh, Kiowa Bower, Russell W. Howson, Archana Belle,

Noah Dephoure, Erin K. O'Shea, and Jonathan S. Weissman, Global Analysis of

Protein Expression in Yeast, 2005. The copyright rests with the authors.

Global Analysis of Protein Expression in Yeast

Sina Ghaemmaghami*, Won-Ki Huh, Kiowa Bower*, Russell W. Howson, Archana

Belle, Noah Dephoure, Erin K. O'Shea, and Jonathan S. Weissman*

Howard Hughes Medical Institute

Department of Cellular & Molecular Pharmacology* and Biochemistry & Biophysics

University of California, San Francisco
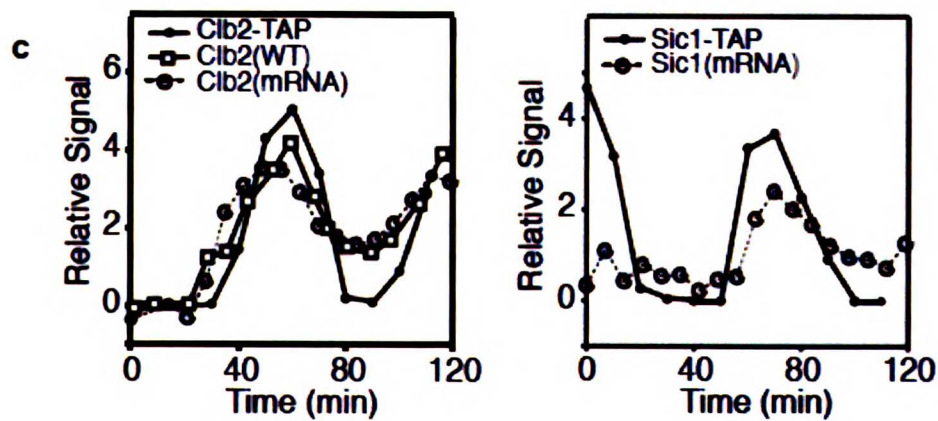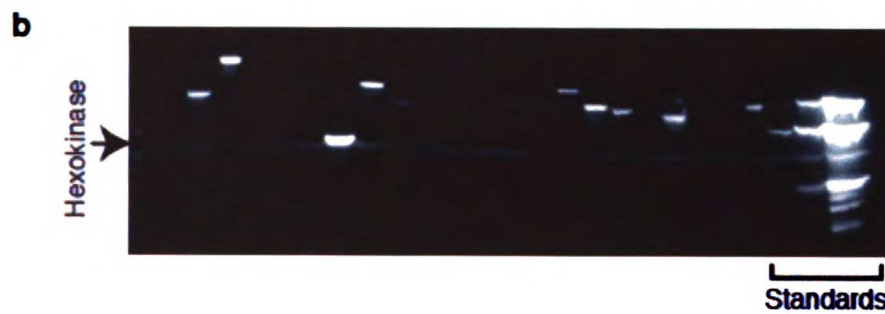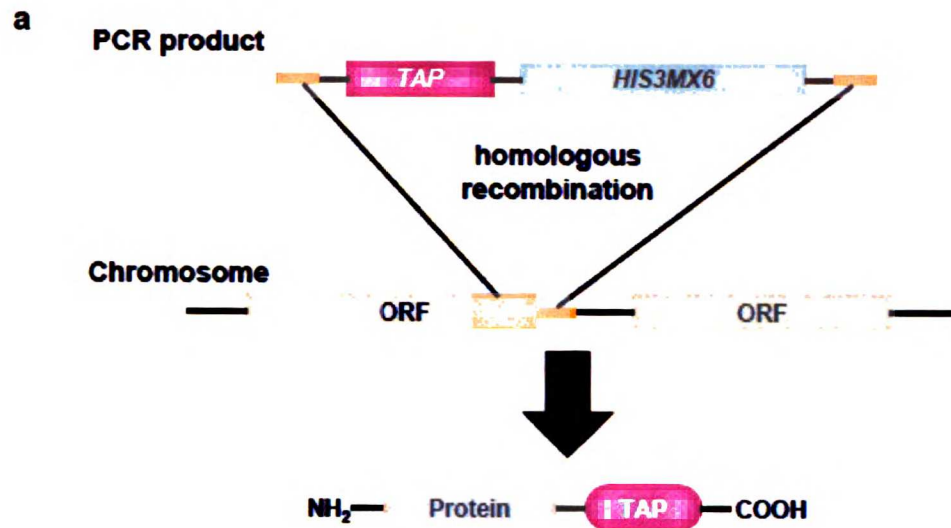
San Francisco, CA 94143-2240

E-mail: jsw1@itsa.ucsf.edu

The availability of complete genomic sequences and technologies that allow comprehensive analysis of global mRNA expression profiles[66, 112, 113] have greatly expanded our ability to monitor the internal state of a cell. Yet ultimately biological systems must be explained in terms of the activity, regulation, and modification of proteins. The ubiquitous use of post-transcriptional regulation makes mRNA an imperfect proxy for such information. To facilitate global protein analyses, we have created a *Saccharomyces cerevisiae* fusion library where each ORF is tagged with a high-affinity epitope and expressed from its natural chromosomal location. Through immunodetection of the common tag, we provide a census of proteins expressed during log-phase growth and quantify their absolute levels. We find that ~80% of the proteome is expressed during normal growth conditions and, using additional sequence information, systematically identify misannotated genes. The abundance of proteins ranges from fewer than 50 to more than 10^6 molecules/cell, with many, including essential proteins and most transcription factors, having levels not readily detectable by other proteomic techniques nor predictable by mRNA levels or codon bias measurements.

The diverse chemical nature of proteins makes the development of globally applicable proteomic assays highly challenging. We have overcome this obstacle in the yeast *S. cerevisiae* by individually tagging each of its annotated open reading frames (ORFs) with a high affinity epitope tag so that the resulting fusion proteins are expressed under the control of their natural promoters. The fusion library allows the immunodetection and immunopurification of the entire yeast proteome using a single

antibody, enabling the development of a range of high-throughput functional assays. To allow for the facile construction of epitope-tagged yeast fusion libraries, we synthesized 6,234 pairs of ORF specific oligonucleotide primers. Each of the oligonucleotide pairs have shared 3' ends that allow for PCR amplification of a common insertion cassette, as well as gene-specific 5' ends that allow for the precise introduction, through homologous recombination, of the amplified insertion cassettes as a perfect in-frame fusion at the C-terminal end of the coding region of each gene[74] (Fig. 10a). The insertion cassettes contained the coding region for a modified version of the TAP (Tandem Affinity Purification) tag[25, 87], which consists of a calmodulin binding peptide, a TEV cleavage site and two IgG binding domains of *Staphylococcus aureus* protein A, as well as a selectable marker (see Supplementary Information). In total, we obtained successful integrants for 98% of all ORFs annotated in the Saccharomyces genome database (as of April 2001)[114], including 93% of all essential ORFs[115] in haploid yeast.

Western blot analysis, using an antibody that specifically recognizes the TAP tag, demonstrated that the large majority (>95%) of detected fusion proteins migrate predominantly as a single band of the approximate expected molecular weight (Fig. 10b). Furthermore, analysis of two known cell-cycle regulated proteins, Clb2 and Sic1[116, 117], indicated that the tagging does not hinder their regulated proteolysis by the ubiquitin/proteasome degradation system and that the TAP tag itself is rapidly destroyed during the targeted degradation of the fusion protein (Fig. 10c). These and other data[25] suggest that the function, regulation, and stability of most, but not all (see Supplementary Information), of the proteome is uncompromised by the fused tag.
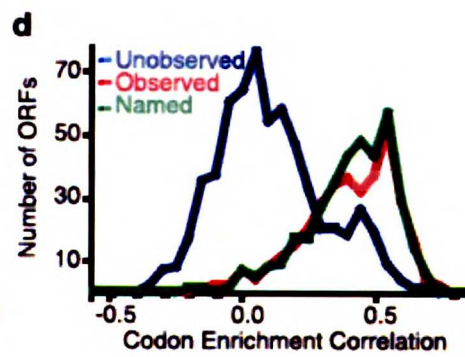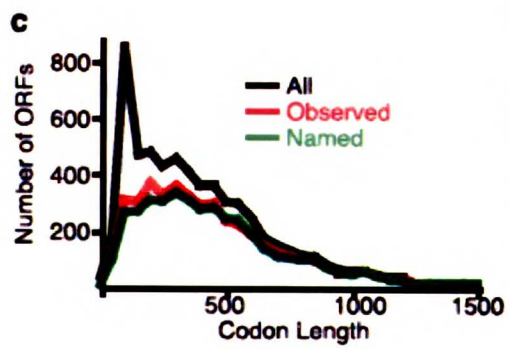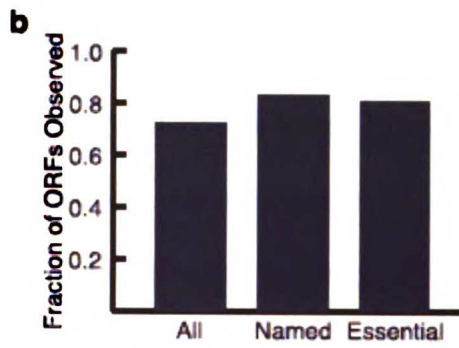
Figure 10. Tagging and detection of the yeast proteome. **a,** Schematic diagram of tagging strategy. **b,** Detection of tagged proteins. Extracts containing TAP-fusion proteins were prepared and analyzed by Western blots using an anti-CBP antibody (See Supplementary Information). Immunodetectection of an endogenous protein (hexokinase) provided a loading control. Serial dilutions of TAP-tagged proteins provided an internal abundance standard (right). **c,** Monitoring dynamic protein levels for two cell cycle regulated proteins. Strains expressing Clb2- and Sic1-TAP fusions were grown to log-phase and arrested in G1 by α-factor treatment. The cell-cycle was resumed by α-factor removal, aliquots were taken at 7 minute intervals and levels of the tagged proteins were quantified using Western blot analysis (closed circle). For comparison, we include mRNA levels of the two proteins obtained by an earlier microarray analysis[118] (open circle) as well as changes in untagged Clb2 protein (open square) levels obtained using an antibody against the endogenous protein in an untagged strain.

a

PCR product

TAP   HIS3MX6

homologous
recombination

Chromosome

ORF   ORF

NH₂— Protein —TAP—COOH

b

Hexokinase →

Standards

c

Clb2-TAP
Clb2(WT)
Clb2(mRNA)

Sic1-TAP
Sic1(mRNA)

Relative Signal

Time (min)

We observed a protein product for 4,251 of the TAP-tagged ORFs by comprehensive Western blot analysis. This set of proteins shows excellent overlap (>90%) with the set of GFP fusion proteins detected by fluorescence microscopy[119] (Fig. 11a) and together indicate that at least 4,517 proteins are expressed during log-phase growth in rich media. We detect 79% of all essential proteins and 83% of gene products corresponding to ORFs with assigned gene names. By contrast, only 73% of all annotated ORFs expressed a detectable protein product (Fig. 11b). This discrepancy largely results from the presence of spurious ORFs in the annotated yeast genome database stemming from well-known difficulties in distinguishing actual coding regions from fortuitous short open reading frames[61, 120]. For the original annotation of the yeast genome, an arbitrary cutoff of 100 codons was used to qualify ORFs as potential genes, leading to an anomalous peak centered between 100-150 amino acids in the sequence length distribution (Fig. 11c, black)[121] of the genome that is not present in the length distribution of the subset of named genes (Fig. 11c, green). Importantly, although we tagged and analyzed all potential ORFs, the length distribution of the subset of observed proteins did not contain the above artifactual peak (Fig. 11c, red), indicating that our analysis of expressed genes has a very low false positive rate (see also Supplementary Information).

A number of bioinformatics approaches, including recent analyses of the genomic sequences of a number of related yeast species, have been used to distinguish between the real and missannoted ORFs[63, 64, 122, 123] although the true number and identity of the spurious ORFs remains unclear. Our results offer experimental verification for a

Figure 11. Analysis of proteins expressed during log-phase growth. **a,** Venn diagram comparing sets of proteins detected by Western blot of TAP-tagged strains (red), fluorescence microscopy of GFP-tagged strains[119] (green) and both (yellow). **b,** Fraction of the indicated set of ORFs observed in either the TAP-tagged or GFP-tagged libraries. **c,** Size distribution of ORFs, binned by length using 50 codon intervals. The number of ORFs per bin is plotted for the indicated sets of ORFs. **d,** Codon Enrichment Correlation (CEC) distribution of small ORFs. CECs were calculated for ORFs with lengths from 100 to 150 codons. ORFs were binned according to CEC values using intervals of 0.05 units. Number of ORFs in each bin is plotted for the indicated sets of ORFs. Note, observed proteins have a positive CEC value characteristic of named genes whereas unobserved ORFs show a major peak centered near a zero value expected for random sequences.

a TAP GFP

363 3888 266

Not Observed: 1717

b

Fraction of ORFs Observed

1.0
0.8
0.6
0.4
0.2

All Named Essential

c

Number of ORFs

800
600
400
200

— All
— Observed
— Named

500 1000 1500
Codon Length

d

Number of ORFs

70
50
30
10

— Unobserved
— Observed
— Named

-0.5 0.0 0.5
Codon Enrichment Correlation
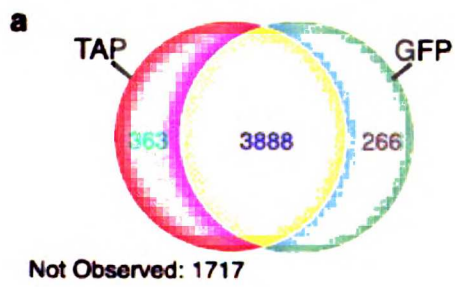
large number of hypothetical genes (we observed 1018 protein products belonging to functionally uncharacterized ORFs) and yields a large, experimentally validated set to evaluate the success of computational methods for identifying falsely annotated genes. By combining a novel metric, termed the codon enrichment correlation (CEC), that evaluates the patterns of codon usage in potential ORFs, with our protein expression data, we identified a set of 525 potentially spurious ORFs (listed in Supplementary Information) that have codon compositions not characteristic of genuine genes and did not yield detectable protein products (Fig. 11d, See Methods). Based on the CEC distribution of genuine ORFs, we estimate that this list is contaminated by ~20 genuine coding sequences. Our proteomics-based approach complements the comparative genomics strategy for identifying spurious ORFs[64]. The large majority (all but seven) of the 496 spurious ORFs suggested by Kellis et al.[64] were not observed in our TAP and GFP studies. The set of spurious ORFs we identified overlaps well with those detected by this cross-species genome study (381 genes were identified as spurious by both studies), and expands the set by 144 ORFs. Among these 144 ORFs are a large number of sequences that overlap with real genes on the opposite strand and therefore are difficult to distinguish through homology analysis.

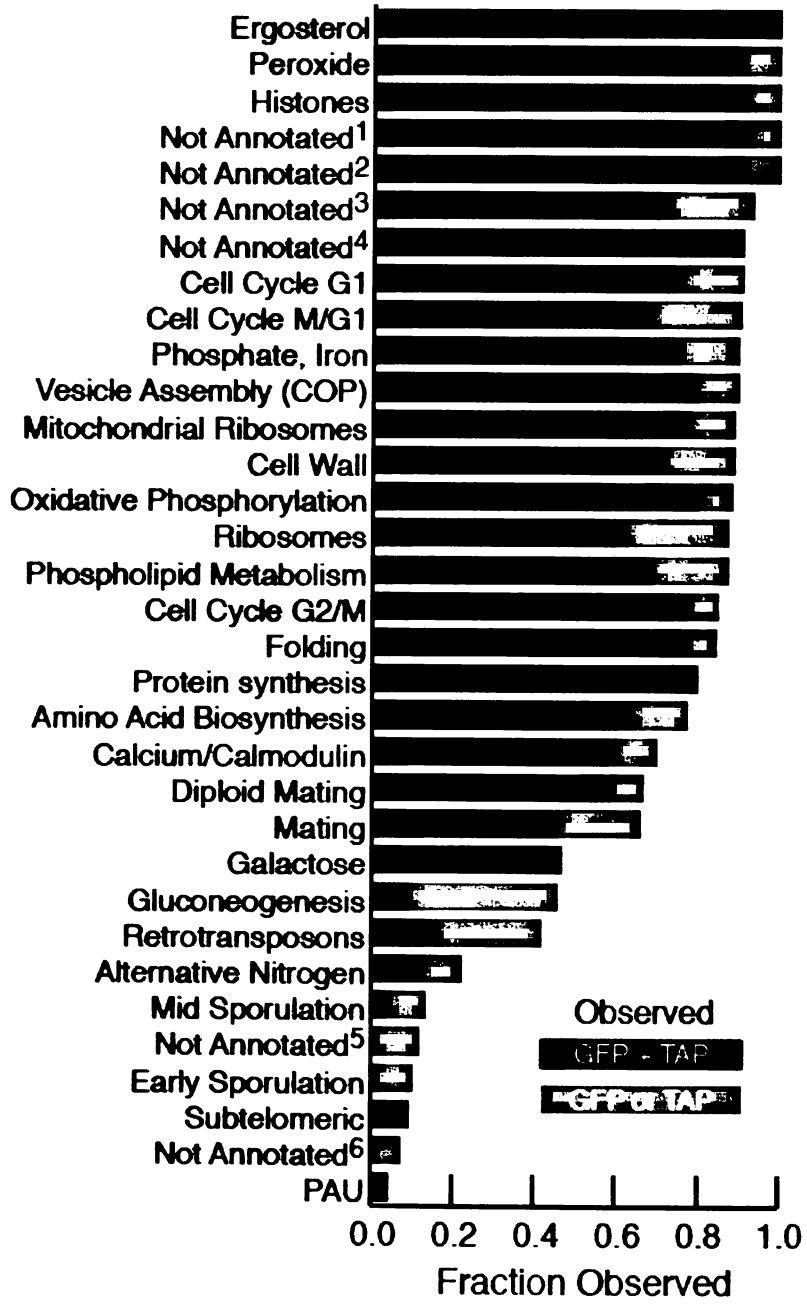After discounting the spurious ORFs, there remain ~1,000 genuine coding regions that did not produce a detectable protein product. To determine if the unobserved proteins belong to classes of genes that are not transcribed during normal log-phase growth conditions, we compared our results with global transcriptional array data. A recent analysis of mRNA expression profiles from a ~1,000 published microarray

experiments allowed for the identification of 33 "modules" of transcriptionally co-regulated genes[124, 125] . For modules that are expressed in log phase (e.g., those coding for house keeping functions, such as ergosterol and amino acid biosynthesis and cell cycle), we were able to detect the large majority of the protein products (Fig. 12). By contrast, modules composed of genes involved in functions required only under specialized conditions (e.g., meisosis/sporulation and alternative nitrogen utilization) generally produced few detectable proteins.

We took advantage of the fact that all gene products were detected using the same epitope/antibody interaction to measure the absolute abundance of each the tagged proteins using quantitative Western blot analyses. This effort was facilitated by the inclusion of internal standards in each gel (Fig. 10b). We find that the levels of different proteins show an enormous dynamic range varying from fewer than 50 to more than 1×10^6 molecules per cell (Fig. 13a and 13b). The results show that previous efforts to quantitate protein levels using two-dimensional gel electrophoresis or mass spectrometry were strongly biased towards the detection of abundant proteins (Fig. 13a)[126-129]. For example, a recent study using mass spectrometry and isotope labeling succeeded in quantitatively monitoring changes in the abundance of 688 yeast proteins[128]. For the most abundant proteins (>50,000 molecule/cell) the coverage was excellent (~60%) whereas for the 75% of the proteome that is present at fewer than 5,000 molecules/cell, only 8% of the proteins were observed. Another mass-spectrometry effort that focused on detecting, without directly quantitating, the complement of proteins in log-phase yeast[129] observed a larger number (1484) of proteins although it was also biased
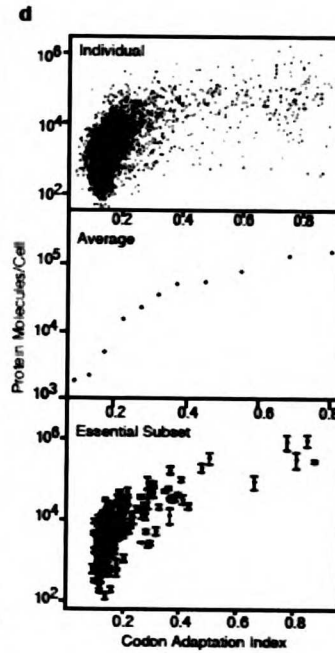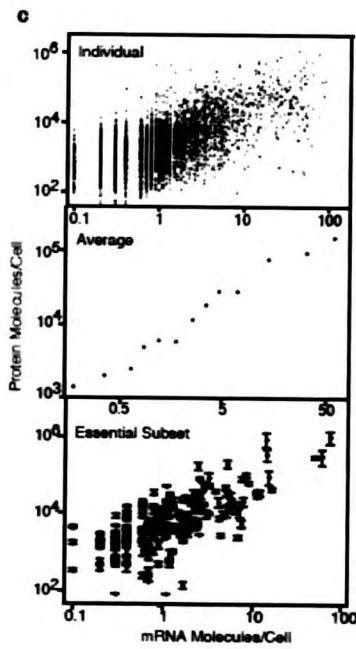
Figure 12. Functional categorization of proteins expressed during log-phase growth in rich medium. 33 modules of co-expressed, functionally related genes were identified by global analysis of ~1000 microarray data sets[124][125]. Plotted is the fraction of the ORFs in each module that produced a detectable protein product by TAP Western analysis or GFP microscopy[119] alone (gray), or both methods (black). Where possible, modules are annotated by function. The gene composition of the modules can be obtained at http://barkai-serv.weizmann.ac.il/modules/ using a cutoff threshold of 4.0. 'Not annotated 1-6' correspond to modules containing YHR025W, YER103W, YPL016W, YPL180W, YER039C-A and YCL076W, respectively.

Ergosterol
Peroxide
Histones
Not Annotated[1]
Not Annotated[2]
Not Annotated[3]
Not Annotated[4]
Cell Cycle G1
Cell Cycle M/G1
Phosphate, Iron
Vesicle Assembly (COP)
Mitochondrial Ribosomes
Cell Wall
Oxidative Phosphorylation
Ribosomes
Phospholipid Metabolism
Cell Cycle G2/M
Folding
Protein synthesis
Amino Acid Biosynthesis
Calcium/Calmodulin
Diploid Mating
Mating
Galactose
Gluconeogenesis
Retrotransposons
Alternative Nitrogen
Mid Sporulation
Not Annotated[5]
Early Sporulation
Subtelomeric
Not Annotated[6]
PAU

Observed
GFP - TAP
GFP or TAP

0.0 0.2 0.4 0.6 0.8 1.0

Fraction Observed

117

towards abundant proteins (90% of the proteome present at >50,000 molecules/cell was detected whereas only 19% of the proteome present at fewer than 5,000 molecules/cell was observed). Our validated list of expressed proteins will help evaluate future advances in mass spectrometry approaches[58].

Overall, we observe a significant relationship between mRNA levels, as measured by an earlier microarray analysis of log-phase yeast[130], and protein levels (Spearman rank correlation coefficient r_s=0.57). Very abundant mRNAs generally encode for abundant proteins and the average protein per mRNA ratio remains remarkably constant throughout the full range of mRNA abundances (Fig. 13c, middle and Fig. 17). The average protein per mRNA ratio is 4800 using this measure of mRNA levels and 4200 using an alternate mRNA abundance measurements based on a microarray analysis comparing mRNA to genomic DNA levels[131] (Fig. 17). However, individual genes with equivalent mRNA levels can result in large differences in protein abundances (Fig. 13c, top). To assess if this variability was primarily caused by protein measurement error and/or disruption of protein function by the TAP tag, we preformed further triplicate measurements of protein abundances on a subset of 206 essential, soluble proteins (See Supplementary Information); the selected strains grew robustly ensuring that the tagged proteins were functional. This subset also shows a high degree of protein to mRNA variability relative to our measurement error indicating that the large differences in individual protein to mRNA ratios are not due primarily to noise in the protein abundance measurements or disruption of the protein by the tag (Fig. 13c, bottom). However, the

Figure 13. Abundance distribution of the yeast proteome. **a,** Distribution of yeast proteins observed by TAP/Western-blot (red), LC-MS(MUDPIT) analysis focusing on comprehensive detection[129] (purple) and quantitative analysis[128] (green) and combined results from two 2D-Gel analyses[126, 127] (blue). The bins are Log2 increments with upper boundaries indicated. **b,** Normalized abundance distribution of observed proteins (red), essential proteins (purple) and transcription factors (dashed line). **c,** The relationship between steady-state mRNA and protein levels. Top plot. Abundance of each protein is plotted against its mRNA level determined by microarray analysis[130]. Middle plot. ORFs are sorted according to mRNA levels and binned into successive groups with cutoffs of 0.25,0.5,0.75,1.0,1.5, 2.0,3.0,4.0,5.0,10,20,50 and 100 molecules/cell. For each bin the mean protein abundance is plotted against the mean mRNA level. Bottom plot. Protein versus mRNA relationship for a subset of essential soluble proteins (See Supplementary Information). Errors represents the standard deviation of three measurements. **d,** Relationship between codon adaptation index (CAI) and protein levels. Individual and averaged protein values are plotted against CAI[132]. In the middle plot, the values are binned using CAI cutoffs of 0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.50, 0.60, 0.70, 0.80 and 1.0.

a

TAP Western
LC/MS (MudPIT) Detection
LC/MS (MudPIT) Quantitation
2D Gels

Number of Proteins

Log$_2$ (Molecules/Cell)

b

Transcription Factors

Normalized Number of Proteins

All
Essential

Log$_2$ (Molecules/Cell)

c

Individual

Average

Essential Subset

Protein Molecules/Cell

mRNA Molecules/Cell

d

Individual

Average

Essential Subset

Protein Molecules/Cell

Codon Adaptation Index

120

correlation between mRNA and protein levels is somewhat greater (Rs =0.66) suggesting the disruption of protein by the TAP tag or difficulty in analyzing membrane proteins may have contributed to some of the variation. We also observed a significant relationship (r_s=0.55) between protein abundance and codon usage as measured by the codon adaptation index (CAI)[132]. Protein abundances drop rapidly for genes with CAI values <0.2, explaining the difficulty that previous proteomic approaches have typically had in detecting these proteins[128]. But, on an individual gene basis there is great variability that is also present in the subset of more carefully measured essential, soluble proteins (Fig. 13d).

A number of observations argue that the full range of abundances detected in this study, including the very low expression levels, represent functionally significant amounts of the proteins. First, the analysis of transcription modules (Fig. 12) indicates that within groups of genes that are turned off during log-phase growth the corresponding proteins are not observed, even at residual levels. Second, the abundance distribution profile of the entire yeast proteome (Fig. 13b, red) is similar to the profile of the portion of the proteome whose function is required for survival under standard growth conditions (Fig. 13b, purple). This suggests that in general functional proteins are not under-represented amongst low-abundance proteins. Third, there are entire classes of functionally important proteins, such as transcription factors (Fig. 13b, line) and cell-cycle proteins (Fig. 18) that are present at very low expression levels. Thus the low abundance proteins detected and quantitated in the present study represent a large and

functionally important portion of the yeast proteome that is almost entirely invisible to systematic quantitative analysis by other proteomic methods.

The TAP-tagged library now makes it feasible to monitor dynamically the abundance of the yeast proteome through basic cellular events such as the cell cycle and meiosis and will allow the determination of protein lifetimes. In addition, important subsets of proteins, such as transcription factors, can be readily studied under a more comprehensive set of conditions. This protein-based data will provide critical information for efforts to understand the logic of cellular regulatory circuits, and by comparison to mRNA levels, the data will give insight into the nature and extent of post-transcriptional regulation.

Methods.

Quantitation of protein levels. 1.7ml cultures of tagged strains were grown in 96-well format to log-phase, and total cell extracts were examined by SDS-PAGE/Western blot analysis as described in the Supplementary Information. The bands corresponding to the tagged proteins were detected using chemiluminescence and a CCD camera (Alpha Innotech FluorChem 8800). To control for variation in extraction and loading, each blot was probed with an antibody against endogenous hexokinase in addition to the TAP specific anti-CBP antibody. Extracts whose hexokinase signals varied by greater than a factor of ~2 from the expected value were re-grown and re-analyzed. A standard containing a mixture of three TAP-tagged proteins (Pgk1, Cdc19, Rpl1A) were included in each gel at one-, ten- and 100-fold dilutions. Proteins whose chemiluminescence signals was approaching saturation were reexamined by performing the Western blot

analysis using a 10-fold dilution of the extract and/or lower exposure times during detection. Prior to the quantitative SDS-PAGE/Western blot analysis, strains were ordered based on estimates of TAP abundance from a preliminary dot-blot analysis. In order to provide a standard for the conversion of Western signals to absolute protein levels, a TAP tagged protein (*E. coli* initiation factor A [INFA]) was overexpressed in *E. coli* and purified to homogeneity. Yeast extracts containing serial dilutions of INFA ranging from 500 atamoles (which was the limit of detection, see Supplementary Fig. S1) to 25 picomoles were run on a gel along with extracts from 25 different yeast TAP-tagged strains representing the full range of observed protein signals (a second TAP-tagged protein [initiation factor B] was also analyzed to ensure that the observed TAP signal was not influenced by the fusion protein). Comparison of the signals generated by these 25 proteins to the known standards allowed the creation of a conversion factor between the observed Western blot signals and absolute protein levels. Based on the number of cells ($\sim 1 \times 10^7$) used for the SDS-PAGE/Western blot analysis, the protein levels were then converted to measurements of protein molecules/cell. In order to assess the error in our quantitation, a set of 33 proteins with a range of protein abundances were grown in duplicate cultures, separately extracted and analyzed on different gels. The replicate signals showed a linear correlation coefficient of R=0.94 with the pairs of proteins having a median variation of a factor of 2.0. This error analysis does not account for potential alterations in the endogenous levels of the proteins caused by the fused tag which may be particularly disruptive for small proteins (Supplementary Information) or difficulty in analyzing some polytopic membranes proteins by SDS-PAGE. For dynamic measurements protein levels (e.g., the cell cycle dependence of Clb2 and Sic1 levels

shown in Fig. 10c or triplicate measurements in Fig. 13c, d) much smaller errors can be obtained by running the samples being compared side-by-side on a single gel. For quantitiation in the triplicate measurements shown in bottom figures of 13c and 13d, serial dilutions of extracts containing purified TAP-tagged INFA were run on each gel.

Codon enrichment correlation (CEC) and identification of spurious ORFs. Codon usage in genuine protein-coding regions deviates systematically from randomly generated ORFs due to both preferences in amino acid composition and biases in the usage of synonymous codons[133] and the CEC provides a measure of this deviation. To calculate CEC values, we first determined the relative prevalence of the 61 amino acid specifying codons in the 3753 named ORFs (Table 4). The codon usage expected in random sequences was then calculated based on the approximate prevalence of 30% T, 30% A, 20%C and 20%G nucleotides in the yeast genomes. The enrichment of each codon for the positive set is given by dividing its prevalence among the named ORFs by its expected prevalence in random sequences (Table 4). Codon enrichments were similarly calculated for each test ORF. The CEC is the linear correlation coefficient (r) between the codon enrichments of the test ORF and the positive set (for examples, see Fig. 15). ORFs were designated as spurious if they failed to be detected by both the TAP and GFP analyses and they had CEC values below a cutoff of 0.25, 0.16, 0.07, or 0.06 for ORFs of size 0-150, 151-200, 201-250, and 251-300 codons, respectively. For ORFs >150 amino acids these values were chosen so that <4.5% of the ORFs falling below these cutoffs that are not detected by the GFP or TAP analyses are genuine coding sequences. The number of genuine coding sequences contaminating our list of spurious ORFs was estimated for each size range and CEC cutoff by the following equation:

124

$N_{real}=N_{obs} \bullet R$. Where N_{obs} is the number of detected ORFs that have a CEC value below the cutoff, and R is the ratio of unobserved to observed ORFs, as determined by the probability of detecting named ORFs for the given size range.

Competing Interest Statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S.W. (jsw1@itsa.ucsf.edu)

Supplementary Information

Construction of the TAP-tagged yeast strain collection. PCR reactions, yeast culture and transformations, gel electrophoresis, and other large-scale manipulations were automated in 96-well format using a Biomek® FX Laboratory Workstation (Beckman Coulter), and yeast strains were grown in 96-well format in a HiGro temperature-controlled shaker (GeneMachines). To construct a chromosomally TAP-tagged library, 6,234 pairs of gene-specific oligonucleotide primers were synthesized on a PolyPlex oligo synthesizer (Gene Machines), each of which had been designed to share complementary sequences to the TAP tag-marker cassette at the 3' end and contain 40 base pairs of homology with a specific gene of interest to allow in frame fusion of the TAP tag at the C-terminal coding region of the gene. Gene-specific cassettes containing a C-terminally positioned TAP tag were then generated by PCR using as a template pFA6a-TAP*-His3MX (the complete sequence of the insert is given below), which in addition to TAP also encodes the *Schizosaccharomyces pombe his5*$^+$ gene and permits selection of transformed strains in histidine-free media, and the primer pairs described above that correspond to 6,234 ORFs annotated in the SGD. PCR reactions were performed in 96-well format with PTC-225 DNA Engine Tetrad gradient cyclers (MJ Research) and products were resolved on Ready-to-Run 96-well agarose gels and electrophoretic apparatus (Amersham Pharmacia). The haploid parent yeast strain (ATCC 201388: *MATa 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

126

verified by genomic PCR of samples from individual colonies with a primer internal to the TAP tag and a separate set of ORF-specific primers designed to produce a product of approximately 500 base pairs. Two to six colonies from strains representing the 6,106 ORFs verified by PCR were analyzed by Western blots. A strain collection for public distribution (which will be available through Open Biosystems) was prepared comprising all ORFs detectable by Western blots in this study.

Culture growth, extract preparation and Western blot analysis. 1.7 mL cultures of successfully tagged strains were grown in YEPD media within 96-well deep well plates. A Teflon coated ball bearing was added to each well in order to keep cultures in suspension and provide proper aeration during shaking. Control experiments indicated that the growth rate of the cultures in this 96-well format was equivalent to growth rates obtained by shaking in conventional flasks, both with glucose or a non-fermentable carbon source. Cultures were grown to OD of ~0.7 at 30 °C and pelleted cells were lysed by the addition of 50 μL of a boiling SDS solution (50 mM Tris , pH 7.5, 5% SDS, 5% glycerol, 50 mM DTT, 5mM EDTA, Bromophenol Blue, 2μg/mL Leupeptin, 2 μg/mL Pepstatin A, 1μg/mL Chymostatin, 0.15 mg/mL Benzamidine, 0.1 mg/mL Pefabloc, 8.8 μg/mL Aprotanin, 3μg/mL Anitpain). Lysed cells were centrifuged and the supernatant extract was stored at –80 °C. 13 μL aliquots of the SDS-lysed extracts were loaded on 26 well, 4-15% gradient acrylamide Tris-HCl Criterion precast gels (Bio Rad). The gels were run at 200 Volts for 70 minutes and transferred using Trans-blot SD semi-dry transfer cell onto PVDF membranes at a constant current of 160 mA per gel for 120 minutes. Before transfer, the activated PVDF membranes and the gels were soaked in 39 mM Glycine, 48 mM Tris 0.375% SDS with and without 20% methanol respectively in

order to facilitate transfer while preventing bleed through. Analysis of a number of randomly chosen blots indicated that the large majority of the proteins samples were transferred onto the PVDF membrane. The blots were probed using an affinity purified rabbit polyclonal antibody raised against the calmodulin binding peptide. This antibody can detect the TAP tag with great sensitivity as it can bind CBP as well as the Protein A segment of the tag through interaction with its Fc region. The blots were subsequently probed with a horse radish peroxidase (HRP) conjugated Goat secondary antibody (Jackson ImmunoResearch) against rabbit IgG and reacted with SuperSignal West Femto Maximum Sensitivity Substrate ECL (BioRad) and the chemiluminescence of the bands corresponding to the tagged proteins were detected and quantified using a CCD camera (Alpha Innotech). Transfer efficiency was monitored by Ponceau S staining all membranes and including "Magic Mark" (invitrogen) molecular weight standards, which contain IgG binding domains allowing visualization by Western blots, on all gels.

Estimate of false positive rates: The list of spurious genes identified by Kellis et al. provides an estimate of our false positive rate. Of the 496 spurious ORFs identified by Kellis et al., only 7 (three of which are good candidates for being genuine ORFs based on strong expression levels and high CEC values) were observed by the GFP and TAP analyses. Even if we assume that all of the genes identified by Kellis et al are spurious, this yields a false positive rate of less than 1.5%.

The effect of the TAP tag on the function, regulation and degradation of proteins.

1) Activity

In haploid yeast, we were able to tag 93% and observe a protein product for 78% of all proteins that are essential for growth. Comparison to the percentage of all named genes that were observed (83%) indicates that the majority of the tagged proteins retain their function. However, the efficiency of tagging does decrease for small essential proteins, due in part to difficulty in tagging ribosomal proteins, suggesting that smaller proteins have a higher likelihood of being rendered non-functional by the TAP tag. In addition, microscope analysis of a similarly constructed C-terminal GFP fusion library (Huh, W. - K. *et al.* Global analysis of protein localization of budding yeast. *Submitted.*) found excellent (~80%) overlap with previously published localization arguing that the C-terminal tag does not generally disturb proper subcellular localization although, as discussed in Huh *et al.*, there are subsets of proteins that do require intact C-termini for proper localization.

2) Regulated protein degradation

As part of systematic effort to measure the lifetime of proteins in the yeast proteome, we have analyzed the degradation rate of a number of proteins that are known to be regulated by ubiquitin-dependant proteolysis by monitoring the change in their levels after inhibiting translation by the addition of cycloheximide (A. B., unpublished data). For the large majority of the proteins that are known to be short-lived, a rapid decrease in protein level is observed indicating that the tag is not inhibiting their proteolysis. Furthermore, many known cell-cycle regulated proteins (e.g. Clb2 and Sic1 in Figure 10C) have been found to have the expected fluctuations during the course of the cell cycle (S.G., unpublished data).

3) Degradation of the TAP tag

Our data suggests that the *in vivo* cleavage of the TAP tag from the fusion protein is not a general problem. In our analysis of Western blots during log phase growth or after cycloheximide shutoff (A.B., unpublished data), a band corresponding to the molecular weight of the TAP tag is almost never observed. Furthermore, Western blots performed on extracts from strains containing 5 tagged protein (Pho4,Pgk1,Sup35,Hxk2,Dpm1) using an antibody against enodgenous protein (instead of the TAP-tag) failed do detect a band corresponding to the untagged protein. This indicates that the TAP tag is not clipped from the protein and rapidly degraded.

TAP Amino Acid Sequence

GRRIPGLINPWKRRWKKNFIAVSAANRFKKISSSGALDYDIPTTASENLYFQGEFGLAQHDEAVDNKFNKE
QQNAFYEILHLPNLNEEQRNAFIQSLKDDPSQSANLLAEAKKLNDAQAPKVDNKFNKEQQNAFYEILHLPN
LNEEQRNAFIQSLKDDPSQSANLLAEAKKLNDAQAPKVDANHQZ

TAP Insertion Cassette DNA Sequence (CBP-TEV-ZZ-His3MX6)

(common 20 bases in F2 forward primer)
GGTCGACGGATCCCCGGGTTAATTAATCCATGGAAGAGAAGATGGAAAAAGAATTTCATAGCCGTCTCAGC
AGCCAACCGCTTTAAGAAAATCTCATCCTCCGGGGCACTTGATTATGATATTCCAACTACTGCTAGCGAGA
ATTTGTATTTTCAGGGAGAATTCGGCCTTGCGCAACACGATGAAGCCGTGGACAACAAATTCAACAAAGAA
CAACAAAACGCGTTCTATGAGATCTTACATTTACCTAACTTAAACGAAGAACAACGAAACGCCTTCATCCA
AAGTTTAAAAGATGACCCAAGCCAAAGCGCTAACCTTTTAGCAGAAGCTAAAAAGCTAAATGATGCTCAGG
CGCCGAAAGTAGACAACAAATTCAACAAAGAACAACAAAACGCGTTCTATGAGATCTTACATTTACCTAAC
TTAAACGAAGAACAACGAAACGCCTTCATCCAAAGTTTAAAAGATGACCCAAGCCAAAGCGCTAACCTTTT
AGCAGAAGCTAAAAAGCTAAATGATGCTCAGGCGCCGAAAGTAGACGCGAATCATCAGTGAGGCGCGCCAC
TTCTAAATAAGCGAATTTCTTATGATTTATGATTTTTATTATTAAATAAGTTATAAAAAAAATAAGTGTAT
ACAAATTTTAAAGTGACTCTTAGGTTTTAAAACGAAAATTCTTATTCTTGAGTAACTCTTTCCTGTAGGTC
AGGTTGCTTTCTCAGGTATAGTATGAGGTCGCTCTTATTGACCACACCTCTACCGGCAGATCCGCTAGGGA
TAACAGGGTAATATAGATCTGTTTAGCTTGCCTCGTCCCCGCCGGGTCACCCGGCCAGCGACATGGAGGCC
CAGAATACCCTCCTTGACAGTCTTGACGTGCGCAGCTCAGGGGCATGATGTGACTGTCGCCCGTACATTTA
GCCCATACATCCCCATGTATAATCATTTGCATCCATACATTTTGATGGCCGCACGGCGCGAAGCAAAAATT
ACGGCTCCTCGCTGCAGACCTGCGAGCAGGGAAACGCTCCCCTCACAGACGCGTTGAATTGTCCCCACGCC
GCGCCCCTGTAGAGAAATATAAAAGGTTAGGATTTGCCACTGAGGTTCTTCTTTCATATACTTCCTTTTAA
AATCTTGCTAGGATACAGTTCTCACATCACATCCGAACATAAACAACCATGGGTAGGAGGGCTTTTGTAGA
AAGAAATACGAACGAAACGAAATCAGCGTTGCCATCGCTTTGGACAAAGCTCCCTTACCTGAAGAGTCGA
ATTTTATTGATGAACTTATAACTTCCAAGCATGCAAACCAAAAGGGAGAACAAGTAATCCAAGTAGACACG
GGAATTGGATTCTTGGATCACATGTATCATGCACTGGCTAAACATGCAGGCTGGAGCTTACGACTTTACTC
AAGAGGTGATTTAATCATCGATGATCATCACACTGCAGAAGATACTGCTATTGCACTTGGTATTGCATTCA
AGCAGGCTATGGGTAACTTTGCCGGCGTTAAAAGATTTGGACATGCTTATTGTCCACTTGACGAAGCTCTT
TCTAGAAGCGTAGTTGACTTGTCGGGACGGCCCTATGCTGTTATCGATTTGGGATTAAAGCGTGAAAAGGT
TGGGGAATTGTCCTGTGAAATGATCCCTCACTTACTATATTCCTTTTCGGTAGCAGCTGGAATTACTTTGC
ATGTTACCTGCTTATATGGTAGTAATGACCATCATCGTGCTGAAAGCGCTTTTAAATCTCTGGCTGTTGCC
ATGCGCGCGGCTACTAGTCTTACTGGAAGTTCTGAAGTCCCAAGCACGAAGGGAGTGTTGTAAAGAGTACT
GACAATAAAAAGATTCTTGTTTTCAAGAACTTGTCATTTGTATAGTTTTTTTATATTGTAGTTGTTCTATT
TTAATCAAATGTTAGCGTGATTTATATTTTTTTTCGCCTCGACATCATCTGCCCAGATGCGAAGTTAAGTG
CGCAGAAAGTAATATCATGCGTCAATCGTATGTGAATGCTGGTCGCTATACTGCTGTCGATTCGATACTAA
CGCCGCCATCCAGTTTAAACGAGCTCGAATTCATCGA
(common 20 bases in R1 reverse primer)

131

Table 4. Enrichment of codons in the positive protein-coding standard set (named genes)

Codon	Enrichment	Codon	Enrichment
CGG	0.21	ACA	0.98
CGA	0.23	TTA	0.99
CGC	0.30	TTC	1.01
TGC	0.37	TCA	1.02
CTC	0.42	CAG	1.03
TGT	0.44	ACC	1.05
CGT	0.53	ATG	1.05
CAC	0.63	CCT	1.12
CCG	0.64	ACT	1.13
GTA	0.64	ATT	1.13
ATA	0.64	TCC	1.19
CTT	0.65	GGC	1.21
ACG	0.66	AGA	1.21
TAT	0.68	GTT	1.25
TCG	0.70	TCT	1.32
GGG	0.74	AAT	1.34
CTA	0.74	GCA	1.36
CAT	0.74	AAC	1.39
GCG	0.78	CCA	1.52
AGG	0.78	TTG	1.54
AGT	0.79	CAA	1.54
AGC	0.80	AAA	1.57
TAC	0.80	GCC	1.61
CCC	0.84	GAG	1.63
TGG	0.85	GAC	1.70
CTG	0.87	AAG	1.75
GGA	0.89	GCT	1.79
GTG	0.91	GGT	2.05
ATC	0.95	GAT	2.16
TTT	0.95	GAA	2.62
GTC	0.97		

Table 5. This supplementary table is available as an excel file at

http://www.nature.com/nature/journal/v425/n6959/extref/nature02046-s2.xls.

Comprehensive list of detected proteins, measured abundances, CEC calculations and annotated spurious ORFs. Column 1. Systematic ORF names. Column 2. Detected expression in TAP Western experiments (TAP), GFP microscopy (GFP), both analyses (Both) or no detected expression (None). Column 3. Measured protein levels in terms of molecules/cells, (-) indicates no detected expression and (#) indicates a detected band that was unquantifiable either because of extremely low signal or experimental problems with the Western blot. Column 4. For a subset of 206 essential proteins, measurements were made in triplicate and the standard deviation (SD) is included in this column. For this subset, column 3 lists the average of the three measurements. (-) indicates that only a single measurement was made and no SD is available. Column 5. Calculated CEC values. Column 6. Spurious ORFs, as identified by the CEC/expression analysis are marked with '1'.

Figure 14. Detection limit of the TAP antibody. Wild type yeast extracts were made at the same concentrations used in our experimental samples. The extracts were spiked with known amounts of a purified TAP-tagged protein (INFA). Western blots were conducted as described in the methods and the chemiluminescence was detected at one and five minute exposure times. The blots indicate that as little as 1 femtomoles of the protein can be detected without any cross-reactivity with endogenous yeast proteins. Using 1.7 mL cultures grown to an OD_{600} of 0.7, this sensitivity allows us to detect proteins present at levels of 50 molecules/cell or greater.

1 minute exposure

5 minute exposure

Lane	Amount
1	200 fmol
2	100 fmol
3	50 fmol
4	20 fmol
5	10 fmol
6	5 fmol
7	2 fmol
8	1 fmol
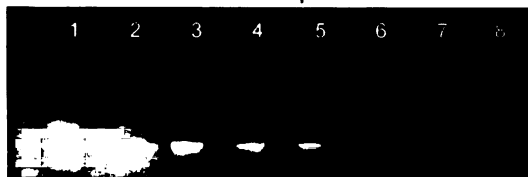
Figure 15. CEC Analysis of two hypothetical ORFs. YDL121C and YKR047W are two small, uncharacterized ORFs that are listed as 'hypothetical' in the yeast genome database. The enrichments of the 61 codons within the ORF are plotted against the positive protein-coding set. The linear correlation coefficients (CEC values) are 0.70 and −0.33 for YDL121C and YKR047W, respectively. A protein product was observed for YDL121C using both the TAP and GFP fusion libraries. Conversely, no protein product was observed for YKR047W. The lack of an expressed protein along with a low CEC value marks YKR047W as a spuriously annotated ORF.

YDL121C

YKR047W
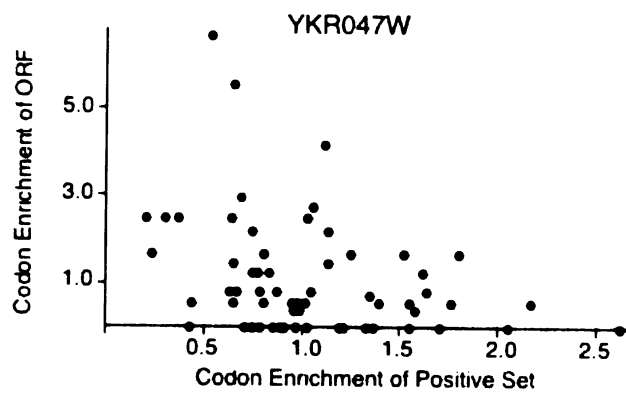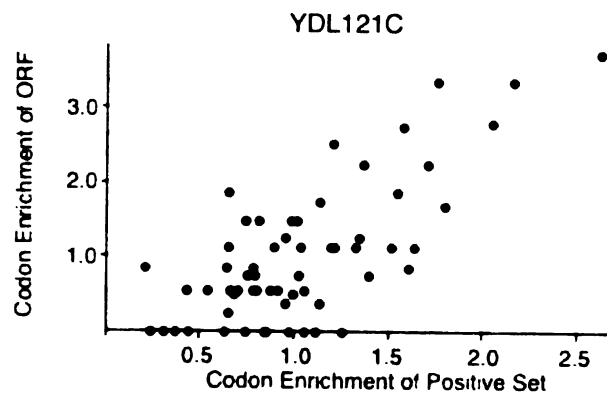
Figure 16. Codon Adaptation Index (CAI) distribution profile of small ORFs. Similar to Figure 2d, CAIs were calculated for the set of ORFs with lengths ranging from 100 to 150 codons according to the method of Sharp, et al (Sharp, P.M. & Li, W.H. Nucleic Acids Res. 15, 1281-1295 (1987)). The ORFs were binned according to CAI using intervals of 0.025 units. The number of ORFs in each bin is plotted for all ORFs for which no expression was observed (blue), all proteins with detectable expression (red) and all named genes (green). Comparison to Figure 1D indicates that CEC is more effective at distinguishing protein-coding ORFs from their non-coding counterparts.
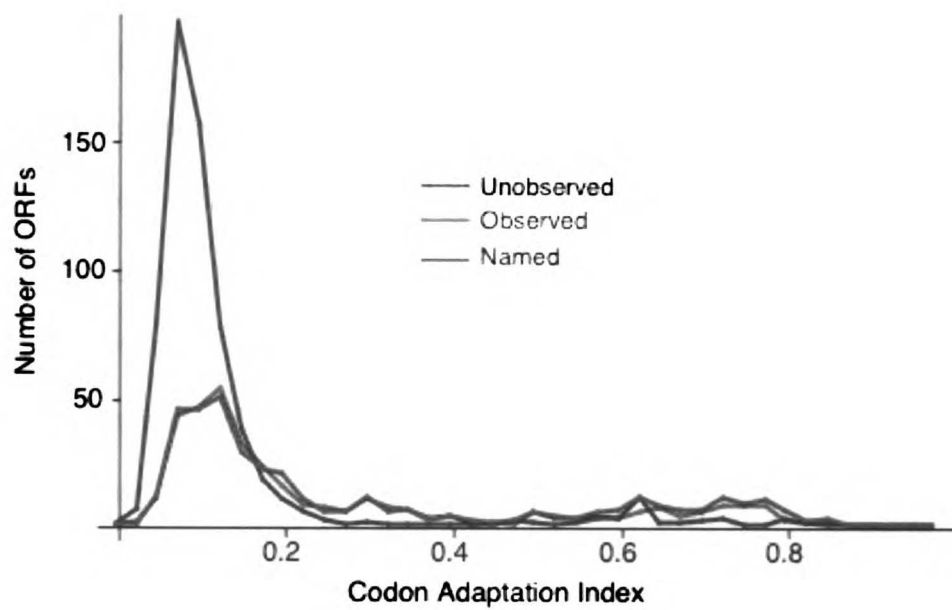
Figure 17. The relationship between steady-state protein levels and mRNA levels as measured by cDNA microarray. In the top plot, the measured log-phase abundance of each protein is plotted against its mRNA level as determined by a recent cDNA microarray analysis in which genomic DNA was used to normalize detected mRNA levels (Wang, Y. et al. *Proc Natl Acad Sci U S A* 99, 5860-5 (2002)). In the middle plot, all ORFs are sorted according to mRNA levels and binned into successive groups with mRNA cutoff levels of 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 10, 20, 50 and 100 molecules per cell. For each bin the mean of the protein abundances is plotted against the mean mRNA level. The bottom plot shows the protein versus mRNA relationship for a subset of essential soluble proteins. The error represents the standard deviation of three measurements.
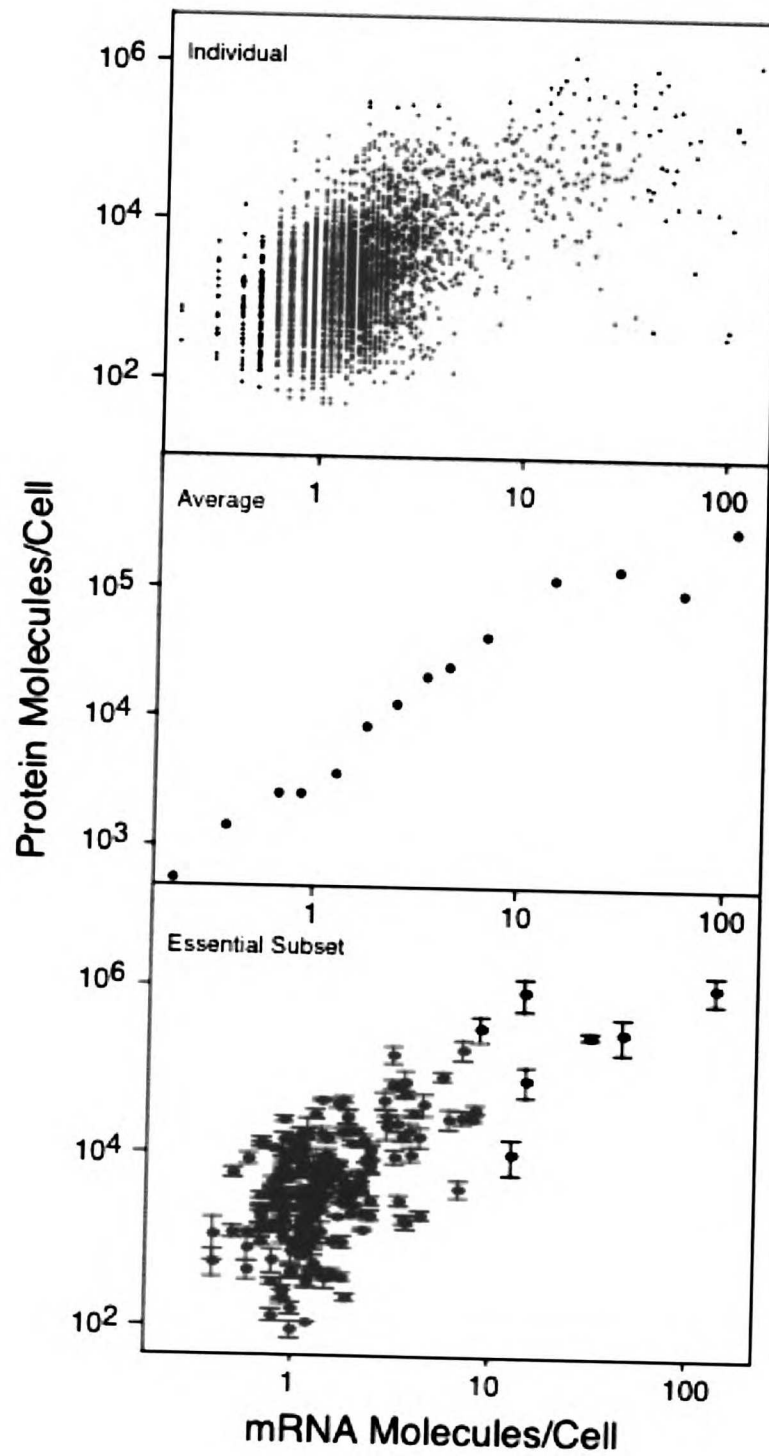
Figure 18. Average protein levels and protein per mRNA ratios of MIPS functional categories. Functional groupings of ORFs were obtained from the MIPS database: (http://mips.gsf.de/genre/proj/yeast). For each functional category, the mean protein level (black) and the mean protein/mRNA ratio (gray), using mRNA levels obtained from microarray analysis (Holstege, F. C. et al. *Cell* 95, 717-28 (1998)), was calculated. The results indicate significant protein enrichment for a number of functional categories, most notably for proteins involved in protein synthesis and energy production. However, the protein/mRNA ratio is relatively constant amongst the different categories.
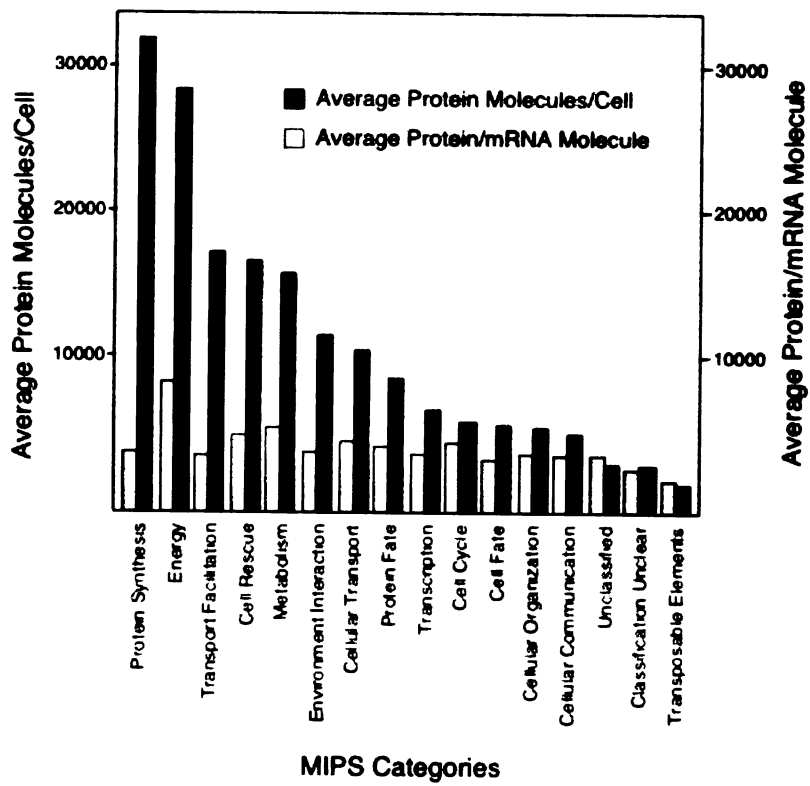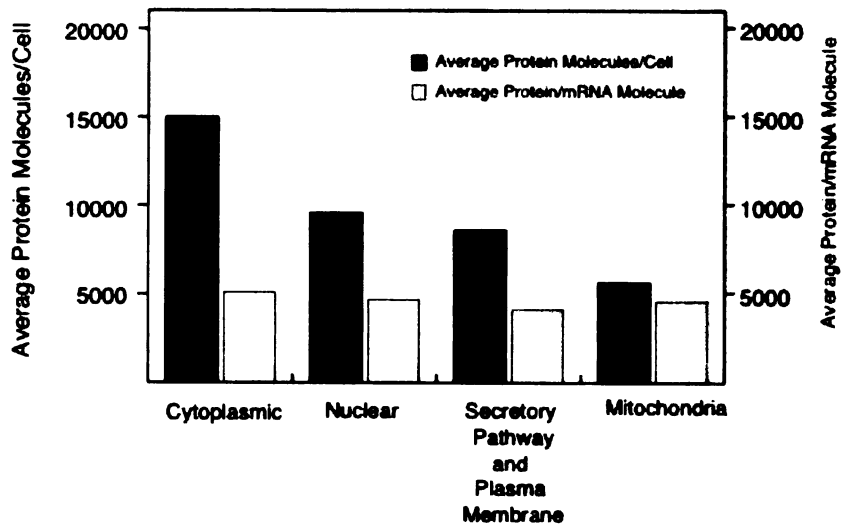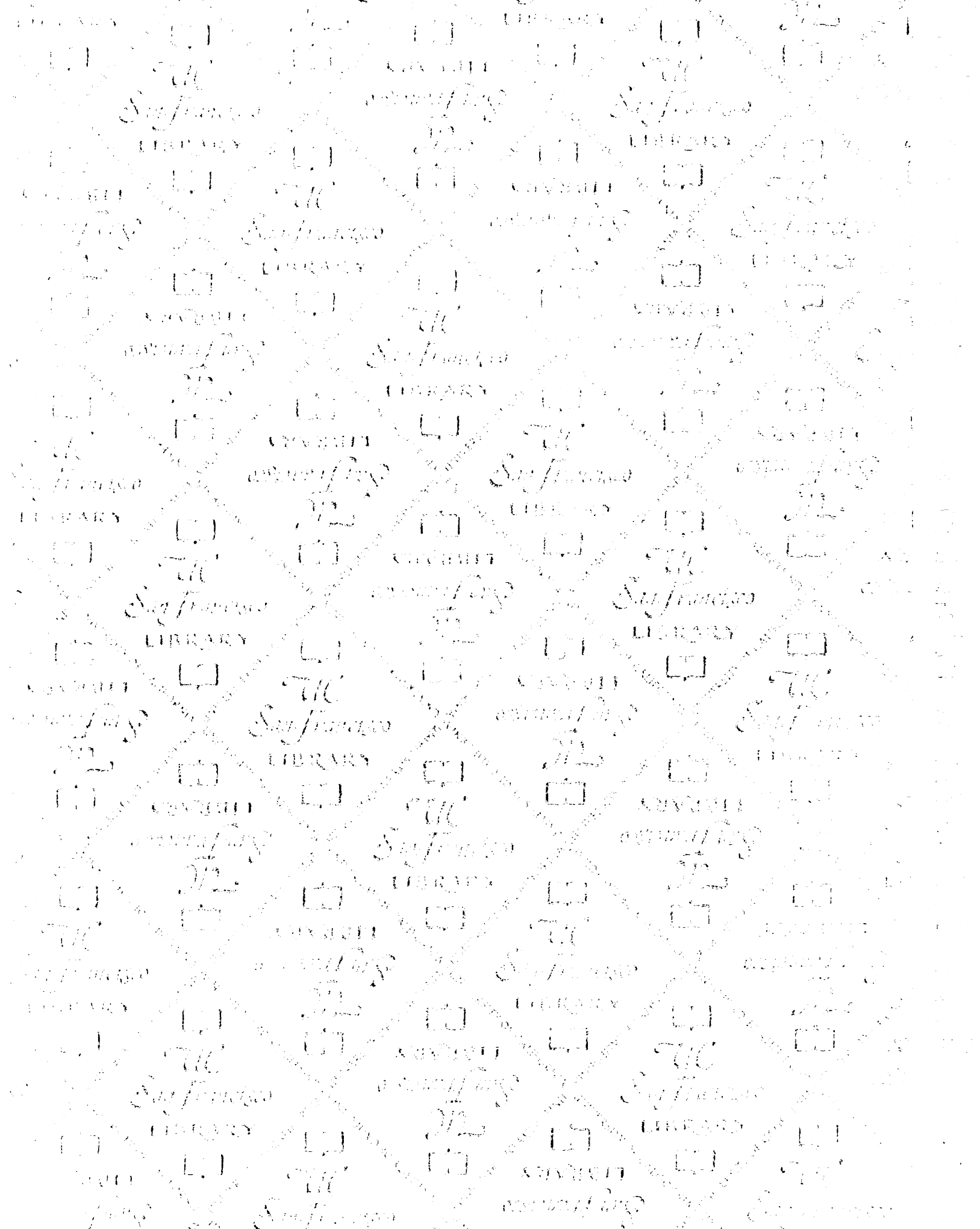
Figure 19. Average protein levels and protein per mRNA ratios of localization categories. ORFs localized to cytoplasmic, nuclear, mitochondria and secretory pathway and plasma membrane compartments by the analysis of Huh et al. (Huh, W. K. et al. Global analysis of protein localization of budding yeast. (2003), *Submitted*) were grouped together. For each group, the mean protein level (black) and the mean protein/mRNA ratio (gray) was calculated as in Figure S5. The results show protein enrichment for the cytoplasmic proteins. However, the protein/mRNA ratio is relatively constant amongst the different categories.