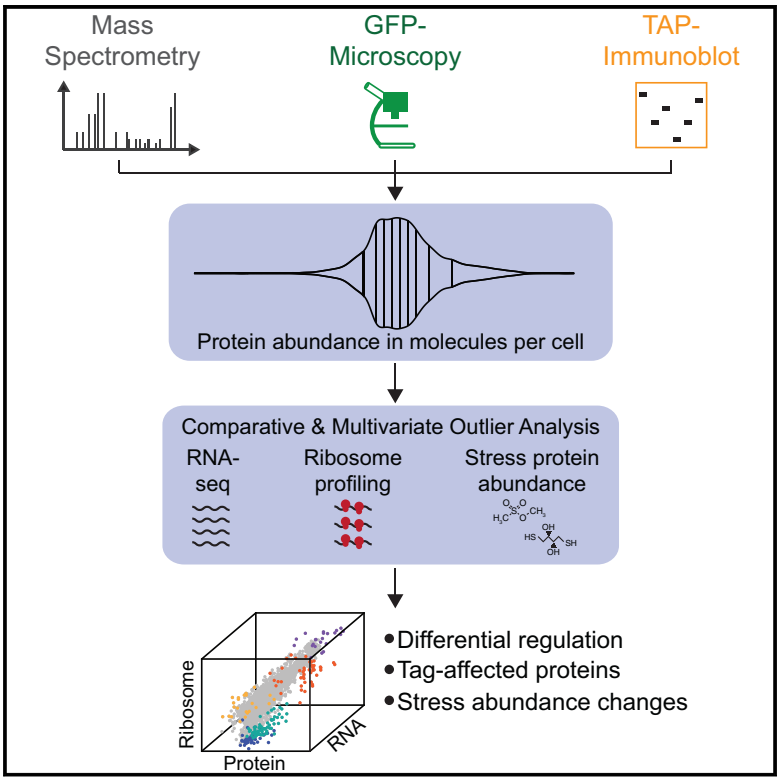
Article

Cell Systems

Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome

Graphical Abstract



Highlights

- Meta-analysis defines the protein abundance distribution of the yeast proteome
- Low- and high-abundance proteins are enriched for biological functions
- Stress-dependent abundance changes reveal functional connections
- Protein fusion tags have a limited effect on native protein abundance

Authors

Brandon Ho, Anastasia Baryshnikova, Grant W. Brown

Correspondence

grant.brown@utoronto.ca

In Brief

By normalizing and converting 21 protein abundance datasets to the intuitive unit of molecules per cell, we provide precise and accurate abundance estimates for 92% of the yeast proteome. Our protein abundance dataset proves useful for exploring the cellular response to environmental stress, the balance between transcription and translation in regulating protein abundance, and the systematic evaluation of the effect of protein tags on protein abundance.

Ho et al., 2018, Cell Systems 6, 192–205
February 28, 2018
https://doi.org/10.1016/j.cels.2017.12.004

CellPress

CellPress

Cell Systems
Article

Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome

Brandon Ho,[1] Anastasia Baryshnikova,[2,3] and Grant W. Brown[1,4,*]
[1]Department of Biochemistry and Donnelly Center, University of Toronto, Toronto, ON M5S 1A8, Canada
[2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[3]Present address: Calico Life Sciences, South San Francisco, CA 94080, USA
[4]Lead Contact
*Correspondence: grant.brown@utoronto.ca
https://doi.org/10.1016/j.cels.2017.12.004

SUMMARY

Protein activity is the ultimate arbiter of function in most cellular pathways, and protein concentration is fundamentally connected to protein action. While the proteome of yeast has been subjected to the most comprehensive analysis of any eukaryote, existing datasets are difficult to compare, and there is no consensus abundance value for each protein. We evaluated 21 quantitative analyses of the *S. cerevisiae* proteome, normalizing and converting all measurements of protein abundance into the intuitive measurement of absolute molecules per cell. We estimate the cellular abundance of 92% of the proteins in the yeast proteome and assess the variation in each abundance measurement. Using our protein abundance dataset, we find that a global response to diverse environmental stresses is not detected at the level of protein abundance, we find that protein tags have only a modest effect on protein abundance, and we identify proteins that are differentially regulated at the mRNA abundance, mRNA translation, and protein abundance levels.

INTRODUCTION

Proteins are one of the primary functional units in biology. Protein levels within a cell directly influence rates of enzymatic reactions and protein-protein interactions. Protein concentration depends on the balance between several processes including transcription and processing of mRNA, translation, post-translational modifications, and protein degradation. Consistent with proteins being the final arbiter of most cellular functions, protein abundance tends to be more evolutionarily conserved than mRNA abundance or protein turnover (Laurent et al., 2010; Christiano et al., 2014). The proteome within a cell is highly dynamic, and changes in response to environmental conditions and stresses. Indeed, protein levels directly influence cellular processes and molecular phenotypes, contributing to the variation between individuals and populations (Wu et al., 2013).

Given the influence that changes in protein levels have on cellular phenotypes, reliable quantification of all proteins present is necessary for a complete understanding of the functions and processes that occur within a cell. The first analyses of protein abundance relied on measurements of gene expression, and due to the relative ease of measuring mRNA levels, protein abundance levels were inferred from global mRNA quantification by microarray technologies (Spellman et al., 1998; Lashkari et al., 1997). Since proteins are influenced by various post-transcriptional, translational, and degradation mechanisms, accurate measurements of protein concentration require direct measurements of the proteins themselves.

The most comprehensive proteome-wide abundance studies have been applied to the model organism *Saccharomyces cerevisiae*, whose proteome is currently estimated at 5,858 proteins (*Saccharomyces* Genome Database, www.yeastgenome.org). In contrast to other organisms, several independent methods for quantifying protein abundance have been applied to budding yeast, including tandem affinity purification (TAP), followed by immunoblot analysis-, mass spectrometry (MS)-, and GFP tag-based methods. Despite the comprehensive nature of existing protein abundance studies, it remains difficult to ascertain whether a given protein abundance from any individual study, independent of other abundance studies, is reliable and accurate. Therefore, aggregating several studies of proteome-wide abundance can provide insight into the precision of protein level estimates. Only six existing datasets quantify protein abundance in molecules per cell (Ghaemmaghami et al., 2003; Kulak et al., 2014; Lu et al., 2007; Peng et al., 2012; Lawless et al., 2016; Lahtvee et al., 2017), and no single study offers full coverage of the proteome. Proteome-scale abundance studies of the yeast proteome in the literature currently number 21 (Ghaemmaghami et al., 2003; Newman et al., 2006; Lee et al., 2007; Lu et al., 2007; de Godoy et al., 2008; Davidson et al., 2011; Lee et al., 2011; Thakur et al., 2011; Nagaraj et al., 2012; Peng et al., 2012; Tkach et al., 2012; Breker et al., 2013; Denervaud et al., 2013; Mazumder et al., 2013; Webb et al., 2013; Kulak et al., 2014; Chong et al., 2015; Lawless et al., 2016; Yofe et al., 2016; Lahtvee et al., 2017; Picotti et al., 2013), providing an opportunity for comprehensive analysis of protein abundance in a eukaryotic cell.

Here we report a unified protein abundance dataset, by normalizing and scaling all 21 yeast proteome datasets to the most intuitive protein abundance unit, molecules per cell. We describe both the accuracy and precision of our dataset, and use it to address interesting biological questions. We find that two-thirds of the proteome is maintained between a narrow

Table 1. Abbreviations Used for Each Dataset

Abbreviation	References	Type of Study	Detection	Abundance Measure	Medium	Growth Phase
LU	Lu et al., 2007	mass spectrometry	label-free spectral counting	absolute	YPD	mid-log
PENG	Peng et al., 2012	mass spectrometry	label-free spectral counting and ion volume-based quantitation	absolute	minimal	early log
KUL	Kulak et al., 2014	mass spectrometry	label-free peak-based spectral counting	absolute	YPD	mid-log
LAW	Lawless et al., 2016	mass spectrometry	stable-isotope labeled internal standards and selected reaction monitoring	absolute	minimal	chemostat
LAHT	Lahtvee et al., 2017	mass spectrometry	SILAC and peak intensity-based absolute quantification	absolute	minimal	chemostat
DGD	de Godoy et al., 2008	mass spectrometry	SILAC and ion chromatogram-based quantification	relative	minimal	mid-log
PIC	Picotti et al., 2009	mass spectrometry	stable-isotope labeled internal standards and selected reaction monitoring	relative	YPD	mid-log
LEE2	Lee et al., 2011	mass spectrometry	isobaric tagging and ion intensities	relative	YPD	mid-log
THAK	Thakur et al., 2011	mass spectrometry	summed peptide intensity	relative	minimal	mid-log
NAG	Nagaraj et al., 2012	mass spectrometry	spike-in SILAC	relative	YPD	mid-log
WEB	Webb et al., 2013	mass spectrometry	label-free spectral counting	relative	YPD	mid-log
TKA	Tkach et al., 2012	GFP microscopy	live cells; confocal	relative	minimal	mid-log
BRE	Breker et al., 2013	GFP microscopy	live cells; confocal	relative	minimal	mid-log
DEN	Denervaud et al., 2013	GFP microscopy	live cells; wide field	relative	minimal	steady state
MAZ	Mazumder et al., 2013	GFP microscopy	fixed cells; wide field	relative	minimal	mid-log
CHO	Chong et al., 2015	GFP microscopy	live cells; confocal	relative	minimal	mid-log
YOF	Yofe et al., 2016	GFP microscopy	N-terminal GFP; live cells; confocal	relative	minimal	mid-log
NEW	Newman et al., 2006	GFP flow cytometry	live cells	relative	YPD	mid-log
LEE	Lee et al., 2007	GFP flow cytometry	live cells	relative	YPD	mid-log
DAV	Davidson et al., 2011	GFP flow cytometry	live cells	relative	YPD	mid-log
GHA	Ghaemmaghami et al., 2003	TAP-immunoblot	SDS extract; immunoblot with internal standard	absolute	YPD	mid-log

range of 1,000–10,000 molecules per cell for cells growing with maximal specific growth rate, and that the global environmental stress response that is evident at the mRNA level is absent at the protein abundance level. Finally, simultaneous analysis of transcription, translation, and protein abundance reveals proteins subject to post-transcriptional regulation, and we describe the effect of C-terminal tags on protein abundance.

RESULTS AND DISCUSSION

Comparisons of Global Quantifications of the Yeast Proteome

With 21 global quantitative studies of the yeast proteome (Table 1), 15 of which are reported in arbitrary units (a.u.), we sought to derive absolute protein molecules per cell for the proteome for each dataset and analyze the resulting data. We extracted the raw protein abundance values from the 21 datasets (Table S1) for the 5,858 proteins in the yeast proteome, and compared the values (absolute abundance or a.u.) from each study with one another, resulting in 210 pairwise correlation plots (Figure 1). The studies agree well with one another, with Pearson correlation coefficients (r) ranging from 0.35 to 0.96. Notably, all studies with abundance measurements derived from GFP fluorescence intensity correlate better with one another than they correlate with the TAP-immunoblot- or MS-based studies. Despite the greater correlations among the GFP-derived datasets, clustering (after normalization and scaling, see below) did not reveal confounding correlations that might mask biological signal (Figure S1). Studies from the same lab, studies using the same medium, studies using the same detection method, and studies using MS did not cluster together exclusively.

Protein Copy Number in *S. cerevisiae*

Normalizing Datasets Reported in a.u.

The most intuitive expression of protein abundance is molecules per cell. To convert all 21 datasets to a common scale of molecules per cell we had to first normalize the datasets before applying a conversion factor to those data not expressed in molecules per cell. The experimental design, data acquisition, and processing for the different global proteome analyses differ between studies. As a result, protein abundance is reported on drastically different scales (Figure S2A). We tested three different methods to normalize the data reported in a.u.: mode shifting, quantile normalization, and center log ratio transformation. The

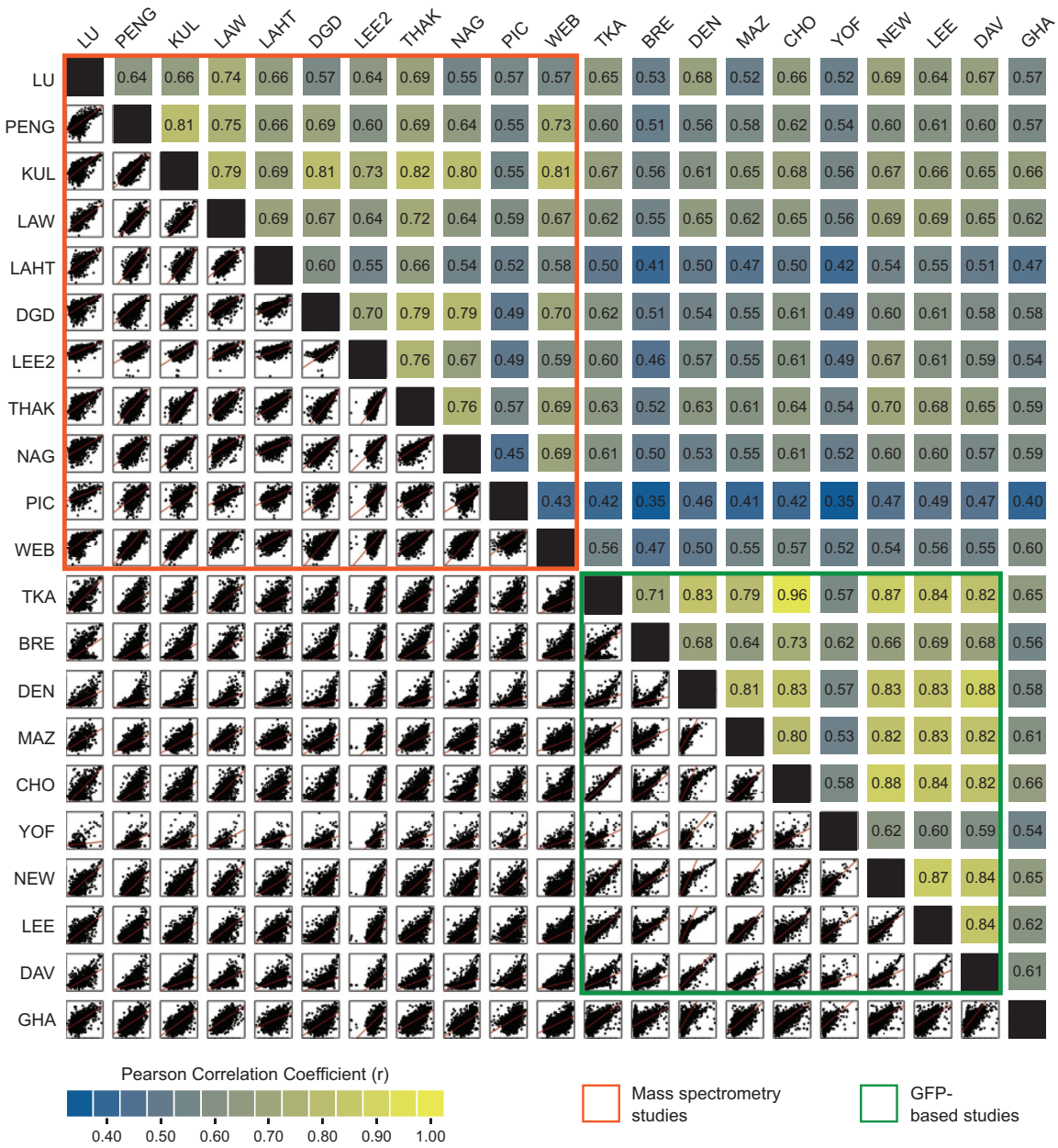
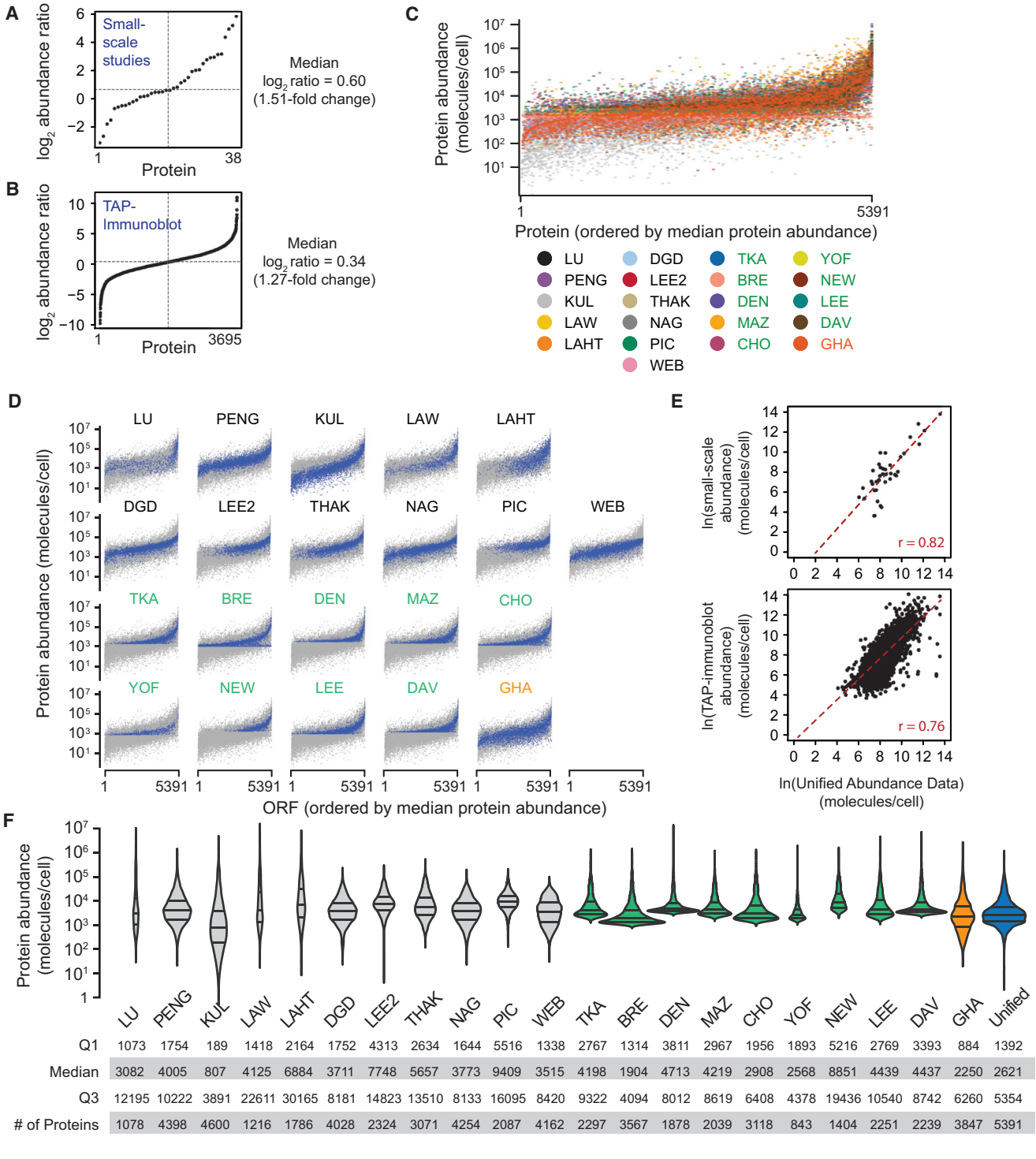


Figure 1. Scatterplot Matrix of Pairwise Comparisons between Protein Abundance Studies

Protein abundance measurements from 21 studies were natural log transformed, and each pairwise combination was plotted as a scatterplot (bottom left). The least-squares best fit for each pairwise comparison is shown (red line). The corresponding Pearson correlation coefficient (r) for each pairwise comparison is shown (top right) and shaded according to the strength of correlation. Mass spectrometry studies are indicated in orange, and GFP-based studies are indicated in green. Each study is indicated by a letter code as described in Table 1.

results of all three methods of normalization correlate very highly with one another (r = 0.93–0.97) indicating that the protein abundance values we calculate are largely independent of the specific normalization technique applied (Figure S2B).

We also considered a normalization scheme where each protein is quantified relative to all other proteins in the dataset, as was done in PaxDb (Wang et al., 2012, 2015). While this relative expression of abundance (parts per million) has the advantage of being independent of cell size and sample volume, it makes comparison between different datasets difficult if the datasets measure different numbers of proteins. Thus, the parts per million normalization alters the pairwise correlations between datasets (Figure S2C). By contrast, normalization by mode shifting or center log ratio transformation allows comparison between datasets by expressing them on a common scale (Figure S2A), and preserves the correlations that are evident in the raw data (Figure S2C). Normalization by mode shifting or center log ratio transformation also allows us to retain proteins whose



	LU	PENG	KUL	LAW	LAHT	DGD	LEE2	THAK	NAG	PIC	WEB	TKA	BRE	DEN	MAZ	CHO	YOF	NEW	LEE	DAV	GHA	Unified
Q1	1073	1754	189	1418	2164	1752	4313	2634	1644	5516	1338	2767	1314	3811	2967	1956	1893	5216	2769	3393	884	1392
Median	3082	4005	807	4125	6884	3711	7748	5657	3773	9409	3515	4198	1904	4713	4219	2908	2568	8851	4439	4437	2250	2621
Q3	12195	10222	3891	22611	30165	8181	14823	13510	8133	16095	8420	9322	4094	8012	8619	6408	4378	19436	10540	8742	6260	5354
# of Proteins	1078	4398	4600	1216	1786	4028	2324	3071	4254	2087	4162	2297	3567	1878	2039	3118	843	1404	2251	2239	3847	5391

Figure 2. Protein Abundance in 21 Datasets, in Molecules per Cell

(A) The $\log_2$ (fold change) between the calibration set and small-scale studies. ORFs are ordered by increasing $\log_2$ ratio. The dotted line represents the median.

(B) The $\log_2$ (fold change) between the calibration set and the TAP-immunoblot study. ORFs are ordered by increasing $\log_2$ ratio. The dotted line represents the median.

(C) The 21 protein abundance datasets were normalized, converted to molecules per cell, and plotted. The proteins are ordered by increasing median abundance on the x axis. Letter codes are as in Table 1.

(D) Proteins from each study are highlighted (blue) and plotted with the abundance measurements from all 21 datasets (gray). Mass spectrometry studies are indicated in black text, GFP-based studies in green, and the TAP-immunoblot study in orange.

(legend continued on next page)

abundance is not reported in all datasets, thereby affording the greatest possible proteome coverage.

Finally, we considered normalization schemes that weight datasets differently. An elegant application of a strategy to weight datasets to minimize variance has been described (Csardi et al., 2015), yet minimizing variance does not necessarily maximize accuracy. There is evidence that some mass spectrometric approaches to quantify absolute protein abundance are more accurate than others (Ahrné et al., 2013), yet we could find no clear metric by which to weight datasets across the entire range of protein abundances and datasets. We tested a matrix of every possible weighting (between 10% and 90%), for the five datasets that measured absolute protein abundance (Lu et al., 2007; Peng et al., 2012; Kulak et al., 2014; Lawless et al., 2016; Lahtvee et al., 2017), and found no measurable improvement in correlations with the small-scale studies or with the TAP-immunoblot study. In the absence of clear evidence that complicated weightings would improve the final dataset, we chose the simpler mode-shifting normalization with equal weighting of the datasets.

Converting a.u. to Molecules per Cell

Currently six protein abundance datasets are reported in molecules per cell, five of which are MS-based studies and one of which used an immunoblotting approach (Lu et al., 2007; Peng et al., 2012; Kulak et al., 2014; Lawless et al., 2016; Ghaemmaghami et al., 2003; Lahtvee et al., 2017). The five MS studies display a range of positive pairwise correlations (r = 0.43–0.81; Figure 1), and all measure native unaltered proteins, and so we reasoned that they could be used to generate a conversion from relative protein abundance in a.u., to molecules per cell. We used the mean of the five datasets as a calibration dataset to convert every other dataset to molecules per cell. Although it is difficult to discern the accuracy of the protein abundance values in the calibration dataset, we find that the median ratio of the calibration dataset values to the protein abundance values reported for 38 proteins in two small-scale, internally calibrated studies (Picotti et al., 2009; Thomson et al., 2011), was 1.51 (Figure 2A; Table S2), suggesting that protein abundance measurements from large-scale studies are similar to those from smaller scale studies. Similarly, the protein abundances in the calibration dataset compare well with the proteome-scale immunoblotting study (Ghaemmaghami et al., 2003): the median ratio of molecules per $\text{cell}_{[\text{calibration set}]}$ to molecules per $\text{cell}_{[\text{immunoblotting set}]}$ is 1.27 (Figure 2B). We conclude that the molecules per cell estimates in the calibration dataset are suitable for use in converting a.u. to molecules per cell.

To identify a model for converting a.u. to molecules per cell, we natural log transformed and compared the normalized arbitrary abundance units to the calibration dataset, for all datasets, for the MS datasets alone, and for the GFP datasets alone (Figure S3A). While the MS datasets have a linear relationship with the calibration set, it is evident that the GFP data contain a number of proteins for which abundance is not linearly related to the calibration set. There is also a sharp cutoff in the GFP data, below which no abundances are reported. The most likely explanation for these phenomena is that background cellular autofluorescence is greater than the fluorescence measured for low-abundance GFP fusion proteins. Indeed, one GFP-based study removed proteins whose fluorescence was close to the background value in their analysis (Chong et al., 2015). We calculated the autofluorescence value of the proteins removed in (Chong et al., 2015), in a.u. after mode-shift normalization (106.56 a.u.), to remove GFP abundance values that are likely due to autofluorescence (Figure S3B). This filter reduced the coverage of our unified dataset from 97% to 92% (5,391 proteins), but yields a slightly higher correlation with the calibration dataset (r = 0.77). The coefficients of variation increase after filtering because values where autofluorescence agrees with autofluorescence are removed, leaving higher variance values that are typical of low-abundance proteins.

To convert all datasets to molecules per cell, a least-squares linear regression between the natural log transformed calibration dataset (reported in molecules per cell) and each natural log transformed mode-shifted study (reported in a.u.) was generated. The correlation between the calibration dataset and the aggregate mode-shifted dataset was slightly better than for the center log transformed dataset (Figure S3C; r = 0.734 versus 0.732), and had a lower sum of standardized residuals, so we proceeded with normalization by mode shifting. Conversion of all measurements to molecules per cell resulted in a unified dataset covering 97% of the yeast proteome (Table S3), or 92% of the proteome after removing GFP values that likely reflect autofluorescence (Table S4).

In general, there is agreement in the molecules per cell for each protein among the datasets analyzed in our study, with protein abundance ranging from 3 to 5.9×10^5 molecules per cell (Figures 2C, 2D, and 2F; Table S4). The relationship of each dataset to the unified dataset is plotted in Figure 2D, and the distribution and coverage of each dataset is shown in Figure 2F. We again assessed accuracy by comparing our aggregate measurements with the small-scale studies and to the TAP-immunoblot study (Figure 2E), finding correlations of r = 0.82 and 0.76, and median differences of 1.66- and 1.23-fold, respectively.

Of the 5,858 protein proteome, 467 proteins were not detected in any study (Table S5). The 467 proteins are enriched for uncharacterized open reading frames (ORFs) (hypergeometric $p = 7.6 \times 10^{-137}$). The 201 verified ORFs that were not detected are enriched for genes involved in the meiotic cell cycle and in sporulation ($p = 2.4 \times 10^{-25}$ and $p = 1.0 \times 10^{-23}$, respectively). Less than 10% of the yeast proteome is not expressed during mitotic growth in rich medium. Therefore, only a relative handful of proteins are likely to be unneeded in standard laboratory growth conditions.

Variance in Protein Abundance Measurements

A key difference between our comparative analysis and each individual protein abundance study is that we report many

(E) The unified dataset is compared with small-scale measurements (top) and with the TAP-immunoblot study (bottom). The Pearson correlation coefficient is indicated.

(F) The distribution of yeast protein abundance in molecules per cell, with the first quartile (Q1), median, and third quartile (Q3) indicated by horizontal bars. The areas of the violin plots are scaled proportionally to the number of observations. Mass spectrometry-, GFP-, and TAP-immunoblot-based studies are colored in gray, green, and orange, respectively. The unified dataset is colored blue. The number of proteins detected and quantified in each study is indicated.

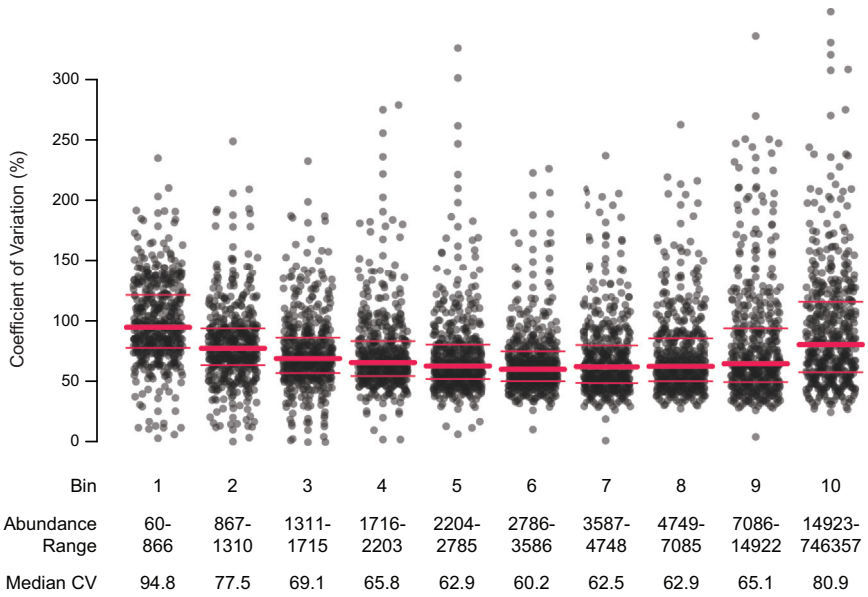


Figure 3. Variability of Each Protein Abundance Measurement

Proteins were ordered by increasing median abundance and binned into deciles. The coefficient of variation was calculated for each protein and plotted. The protein abundance levels associated with each bin are indicated, as is the median CV for each bin. The red lines indicate the third quartile, the median, and the first quartile for each bin.

independent estimates of protein level per ORF in a common unit of molecules per cell. Therefore, we are in a position to explore the variation in reported values for each ORF across 21 datasets. We calculated the coefficient of variation (CV) (SD/mean, expressed as a percentage) across the yeast proteome. In general, the CVs are modest, with 4,048 of 5,065 abundance measurements for which a CV could be calculated having a CV of 100% or less (Table S4). The greatest median CVs (higher than 80%) were exhibited by low-abundance proteins (<866 molecules per cell) and high-abundance proteins (>14,923 molecules per cell) (Figure 3). Interestingly, CV values are, on average, higher for the MS-based measurements than for the GFP-based measurements (65% and 29%, respectively). The lowest CV values (60%–70%) are observed for proteins present at 1,311–14,922 molecules per cell. Therefore, we conclude that the measurement of abundance is most precise for the 62% of the measured proteome that is within this abundance range and that precision is better for the GFP measurements, provided they are above the autofluorescence level of ~1,400 molecules per cell.

The MS-based analyses exhibit the greatest sensitivity, with measurements as low as three molecules per cell, and four studies in particular (Kulak et al., 2014; de Godoy et al., 2008; Thakur et al., 2011; Peng et al., 2012) have both the best proteome coverage (greater than 4,000 proteins) and a large detection range, detecting fewer than 50 to greater than 100,000 molecules per cell. Interestingly, the four studies utilize different quantification methods, and are among the most highly inter-correlated MS studies (r = 0.68–0.82), indicating that distinct approaches can yield similarly sensitive quantifications of the yeast proteome that are in agreement with one another.

Functional Enrichment of Low- and High-Abundance Proteins

We next asked whether particular cellular processes tend to be performed by proteins that are expressed at similar levels. Budding yeast is unique in having a comprehensive map, where genes and pathways have been placed into functional modules (Costanzo et al., 2016) (Figure 4A). We used spatial analysis of functional enrichment (SAFE) (Baryshnikova, 2016) to explore whether regions of the functional cell map (Costanzo et al., 2016) are enriched for high- and low-abundance proteins (Figure 4B). We found that high-abundance proteins were specifically over-represented in regions associated with cell polarity and morphogenesis, and with ribosome biogenesis (Figure 4B, yellow). Low-abundance proteins were over-represented in the region associated with DNA replication and repair, mitosis, and RNA processing (Figure 4B, teal).

Gene ontology term enrichment analysis yielded results consistent with SAFE analysis (Figure 4C). The decile comprising the least-abundant proteins was enriched for response to DNA damage stimulus ($p = 0.0056$), mitotic cell-cycle regulation ($p = 1.1 \times 10^{-5}$), and protein ubiquitination ($p = 6.6 \times 10^{-5}$), perhaps reflecting the importance of restricting the abundance of cell-cycle regulators and DNA repair factors. The most highly expressed proteins tended to be proteins involved in translation in the cytoplasm ($p = 3.0 \times 10^{-140}$) and related processes, consistent with the key role of protein biosynthetic capacity in cell growth and division (Warner, 1999; Volarevic et al., 2000; Jorgensen et al., 2002; Bernstein and Baserga, 2004; Yu et al., 2006; Bjorklund et al., 2006; Teng et al., 2013). Previous analysis of the human proteome, with 73% coverage, indicated functional enrichment for high-abundance proteins, but failed to detect enrichment of function for low-abundance proteins (Beck et al., 2011). One possibility is that the combination of more sparse functional annotation of the human proteome (relative to annotation in yeast) combined with incomplete proteome coverage precluded detection of functional enrichment of low-abundance proteins. However, since the highest abundance categories of human and yeast proteins were similarly enriched for ribosome components there is evidence that relationships between protein function and abundance are evolutionarily conserved.

The Protein Abundance Distribution of the Proteome

The protein abundance distribution of the complete proteome has not been well characterized, therefore what defines a high-abundance protein versus a low-abundance protein is unclear. The abundance of the typical cellular protein is unknown, as is the abundance range that characterizes most cellular

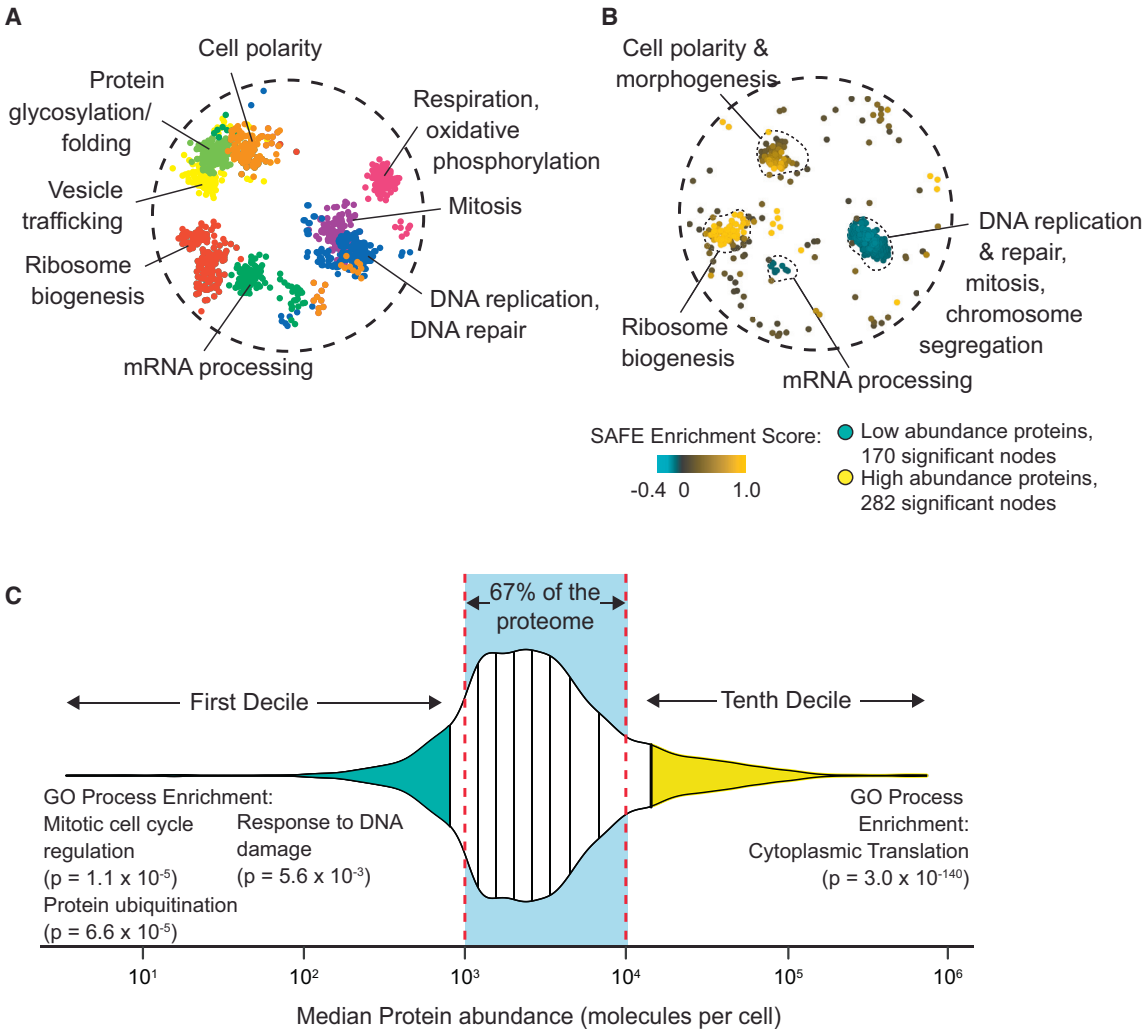


Figure 4. Functional Enrichment of High- and Low-Abundance Proteins

(A) SAFE annotation of the yeast genetic interaction similarity network to identify regions of the network enriched for similar biological processes (Costanzo et al., 2016).

(B) The protein abundance enrichment landscape is plotted on the genetic interaction profile similarity network. Colored nodes represent the centers of local neighborhoods enriched for high- or low-abundance proteins, shaded according to the log of the enrichment score. The outlines of the gene ontology (GO)-based functional domains of the network where protein abundance enrichment is concentrated are shown.

(C) Violin plot of the distribution of abundances for the yeast proteome. The first decile and tenth decile are shaded in teal and yellow, respectively. The blue shaded area represents 67% of all protein abundance measurements.

proteins. Large-scale quantifications of yeast protein levels suggest that few proteins are very highly expressed, but these analyses relied on limited data covering only ~70% of the proteome (Ghaemmaghami et al., 2003; Kulak et al., 2014). With our unified dataset, we find that yeast protein abundance, when logarithmically transformed, is skewed toward high-abundance proteins (Figure 4C). Protein abundance ranges from zero to 7.5×10^5 molecules per cell, the median abundance is 2,622 molecules per cell, and 67% of proteins quantified exist between 1,000 and 10,000 molecules per cell (Figure 4C). Low-abundance proteins, the first decile, have abundances ranging from 3 to 822 molecules per cell, while high-abundance proteins, the tenth decile, have abundances ranging from 1.4×10^5 to 7.5×10^5. Our data suggest that protein copy number is maintained within a narrow range from which only a small portion of the proteome deviates.

Total Protein Content of a Yeast Cell

An estimate of yeast cell protein content can be derived from the cellular protein mass per unit volume and the mass of the average protein (Milo, 2013). Using a density of 1.1029 g/mL (Bryan et al., 2010), a water content of 60.4% (Illmer et al., 1999), and a protein fraction of dry mass of 39.6% (Yamada and Sgarbieri, 2005), typical of yeast in standard growth conditions, we calculate 0.17 g of protein per mL. With an average protein mass of 54,580 Da, and mean logarithmic phase cell volume of 42 μm^3 (Jorgensen et al., 2002), we calculate 7.9×10^7 protein molecules per cell. Adding the median abundances of

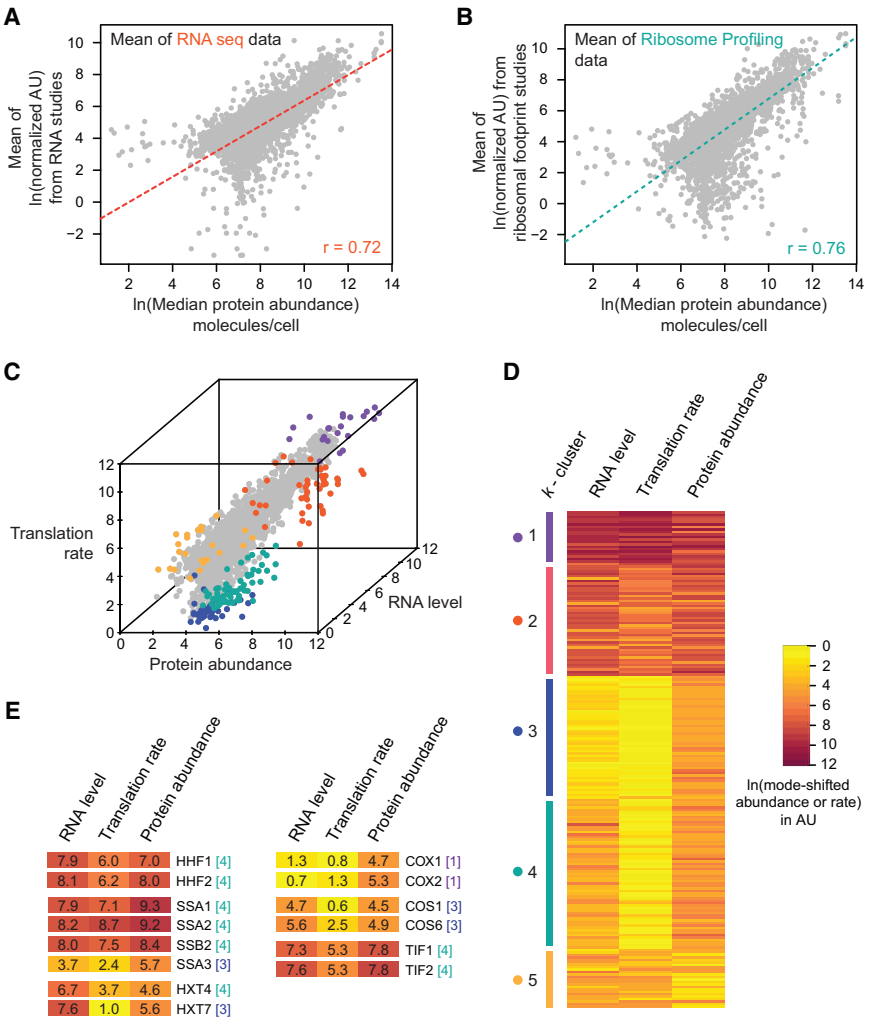


Figure 5. Identification of Proteins with Post-Transcriptional and Post-Translational Regulation

(A) Protein abundance compared with mRNA levels measured by RNA sequencing.

(B) Protein abundance compared with ribosome footprint abundance from an aggregate of five ribosome-profiling analyses.

(C) A three-dimensional scatterplot of RNA transcript level, ribosome density, and protein abundance, with outliers colored by *k*-cluster as defined in (D).

(D) *k*-Means clustering of outliers (Mahalanobis distance >12.84) from comparison of mRNA abundance, translation rate (ribosome density), and median protein abundance. Each row corresponds to a gene, and is colored according to ln(mode-shifted abundance or rate) in a.u..

(E) Example outliers are shown with their associated cluster and colored according to ln(mode-shifted abundance or rate) in a.u.. The a.u. values are also indicated.

all detected proteins in our unified abundance dataset, we arrive at a total of 4.2×10^7 protein molecules per yeast cell, or 0.53 of the calculated estimate. Total protein content estimates derived from individual studies agree well with our estimate (4.5×10^7 [Ghaemmaghami et al., 2003], 5.3×10^7 [von der Haar, 2008], and 5×10^7 [Futcher et al., 1999]), and also tend to be lower than the calculated estimate of 7.9×10^7 molecules per cell. We infer that our aggregate abundance estimates are likely accurate within 2-fold, on average.

RNA Expression and Translation Rate Both Capture Variance in Protein Abundance

The degree to which mRNA levels can explain protein abundance remains unclear (Vogel and Marcotte, 2012), as mRNA and protein concentrations have been reported to correlate well in some studies (Csardi et al., 2015; Franks et al., 2015), and poorly in others (Ingolia et al., 2009; Lahtvee et al., 2017). We reasoned that a more complete view of the relationship between transcript and protein abundance could be obtained with our comprehensive protein abundance dataset. We compared protein molecules per cell with mRNA levels from three microarray and three RNA sequencing (RNA-seq) datasets (Roth et al., 1998; Causton et al., 2001; Lipson et al., 2009; Nagalakshmi et al., 2008; Yassour et al., 2009). Between 37% and 56% of the variance in protein abundance that we observe is explained by mRNA abundance, as measured by microarray (r = 0.61–0.68) and RNA-seq (r = 0.67–0.75), with both mRNA methodologies performing similarly in estimating protein levels (p = 0.21, two-tailed t test; Figure S4A). Higher correlations between mRNA and protein abundance have been reported (r = 0.66–0.82) (Futcher et al., 1999; Greenbaum et al., 2003; Franks et al., 2015) in studies using less comprehensive protein abundance datasets (2,044 proteins at most), and more sophisticated analysis indicates that the true correlations could be higher, due to experimental noise (Csardi et al., 2015). Our protein abundance dataset correlates similarly with translation rates measured in ribosome-profiling studies (r = 0.67–0.75; Figure S4B) (Ingolia et al., 2009; Brar et al., 2012; Albert et al., 2014; Pop et al., 2014; Weinberg et al., 2016). When the mRNA abundance and ribosome-profiling datasets are aggregated, we find that ribosome profiling captures only slightly more of the protein abundance variance than does mRNA abundance (Figures 5A and 5B). Our data indicate that in unperturbed conditions mRNA abundance and ribosome footprint analysis explain similar fractions of protein abundance variance, in agreement with previous analysis (Csardi et al., 2015). Indeed, when we compare the aggregate of three RNA-seq studies of mRNA abundance with five independent studies of protein synthesis by ribosomal profiling, they correlate well (r = 0.89; Figure S4C).

The Balance between Transcriptional and Translational Regulation

Although mRNA abundance and ribosome profiling capture similar fractions of the variance in protein abundance, it is likely

that there is a complex interplay between transcriptional and translational regulation for individual proteins. We sought to capture this complexity by comparing protein abundance, mRNA abundance, and ribosome density simultaneously (Figure 5C). We calculated Mahalanobis distances, a metric used for identifying multivariate outliers, reasoning that outliers are proteins that are differentially regulated at either the transcriptional, translational, or protein level. A total of 200 proteins were identified as outliers and were *k*-means clustered to reveal patterns of co-regulation (Figure 5D; Table S6). The outliers were enriched for cytoplasmic translation (33 proteins, $p = 8.43 \times 10^{-29}$) and glucose metabolic processes (11 proteins; $p = 2.2 \times 10^{-6}$). We find several instances where proteins with similar function cluster with one another (Figure 5E). For example, histone H4 subunits (*HHF1* and *HHF2*) cluster together (cluster 4), having high mRNA expression, lower translation rates, and high protein levels, suggestive of co-regulation. Additional examples include *COX1*/*COX2*, *COS1*/*COS6*, and *TIF1*/*TIF2*.

We also find cases of proteins within the same family whose expression pattern are not covariant, perhaps revealing functional differences. Three members of the *HSP70* gene family, *SSA1*, *SSA2*, and *SSB1*, are found in a different cluster than *SSA3* (cluster 4 versus 3; Figure 5E), indicating differential regulation. It has been noted that *SSA3* has a greater role in Hsp104-independent acquired thermotolerance during heat-shock stress in comparison with other protein family members, and thus its expression may be regulated differently (Hasin et al., 2014). The glucose transport genes *HXT4* and *HXT7* also lie in different groups (cluster 4 versus 3). Both have high transcript and protein levels, and lower than expected translation rates. However, *HXT4* appears to be engaged by ribosomes more frequently than *HXT7*. These proteins are functionally distinct based on their affinity for glucose, which may explain differences in their regulation. While transcriptional regulation of yeast hexose transporters has been extensively studied, our data suggest that detailed analysis of the translational component of hexose transporter regulation could be fruitful, especially in conditions of varying glucose levels.

Among the five clusters, we find functional enrichment only in cluster 4 (cytoplasmic translation; $p = 8.43 \times 10^{-29}$), which is characterized by high protein and transcript levels, but lower translation rates, indicating a role for negative regulation of translation in controlling ribosomal protein abundance. Indeed, further downregulation of translation of ribosome biogenesis gene transcripts is apparent upon starvation (Ingolia et al., 2009), and protein turnover data indicate that ribosome protein synthesis is tightly coordinated in budding yeast (Christiano et al., 2014). A non-linear relationship between ribosomal protein mRNA abundance and ribosomal protein abundance has been noted in fission yeast (Marguerat et al., 2012), and is consistent with the lower than expected translation rates that we find in budding yeast.

Proteins in cluster 5 have lower protein abundance than expected, given the RNA level and translation rates. Half-life measurements have been reported for several proteins in cluster 5, and five of six proteins measured (Belle et al., 2006), and three of three proteins measured (Christiano et al., 2014), have half-lives lower than the median, consistent with the lower than expected protein abundances in cluster 5.

The improvement in proteome coverage of our abundance dataset facilitates more detailed analyses of the relationships between transcription, translation, and protein abundance and could be useful in making functional predictions from patterns of similar transcriptional/translational regulation, and to explore other genes where transcription, translation, and protein abundance are not co-directional.

The Effect of Protein Fusion Tags on Native Protein Abundance

Protein fusion tags are utilized extensively, yet the effect of tags on protein abundance has not been assessed systematically. The 4,502 yeast strains used to measure protein abundance by GFP fluorescence all express proteins with C-terminal fusions to GFP (Huh et al., 2003), with the exception of the Yofe et al. (2016) study, which analyzed N-terminal GFP fusions. Fusion to GFP sequences adds an extra 27 kDa to the native protein, alters the identity of the C terminus, and changes the DNA sequence of the 3′ UTR of the gene. We reasoned that proteins whose expression differs greatly between mass spec datasets (which measure native proteins) and GFP datasets are likely affected by the presence of the tag. We compared the median ln abundance between MS- and GFP-based abundance studies, applying a t test to define outliers ($p < 0.05$; Figure 6; Table S7). A total of 716 proteins were identified, with 281 proteins showing at least 2-fold (and as much as 50-fold) lower abundance when GFP-tagged (Figures 6A and 6B). Of the 281 proteins, 259 have been assessed as C-terminal fusions to the 21 kDa TAP tag (Ghaemmaghami et al., 2003). Of the 259, 141 proteins also had reduced abundance (by at least 2-fold) when TAP tagged, suggesting that these proteins are either destabilized by the presence of any protein tag at the C terminus, or require their native 3′ UTR for mRNA stability. The 118 proteins that decreased in abundance when GFP-tagged but not when TAP tagged could represent GFP-specific protein destabilization, or protein-specific issues with fluorescence detection (Waldo et al., 1999). Interestingly, 57 of the 281 proteins with reduced abundance when C-terminal GFP-tagged were also assessed as N-terminal GFP fusions (Yofe et al., 2016), with 31 having reduced abundance (by at least 2-fold), irrespective of the location of the GFP tag. We also observed 259 proteins that had at least 2-fold greater abundance when tagged with GFP (to as much as 67-fold), indicating that in some cases GFP could stabilize its protein fusion partner.

Together, our data indicate that changes in protein abundance can occur upon adding additional sequences to the C terminus, and we find that 12% of the 4,502 proteins measured with C-terminal GFP tags have statistically supported abundance changes of greater than 2-fold when tagged with GFP. Since 1,356 proteins are absent from the C-terminal GFP datasets, it is possible that additional proteins are affected by the presence of a tag. Of these, 467 proteins were not detected by any method and so are likely not expressed during normal mitotic growth. Two hundred and thirteen proteins showed no abundance change > 2-fold when detected with a TAP tag. We infer that at most an additional 676 proteins, for a total of 22% of the detectably expressed proteome, could be affected by tagging. Thus, the proteins absent from existing GFP datasets are unlikely to affect the general

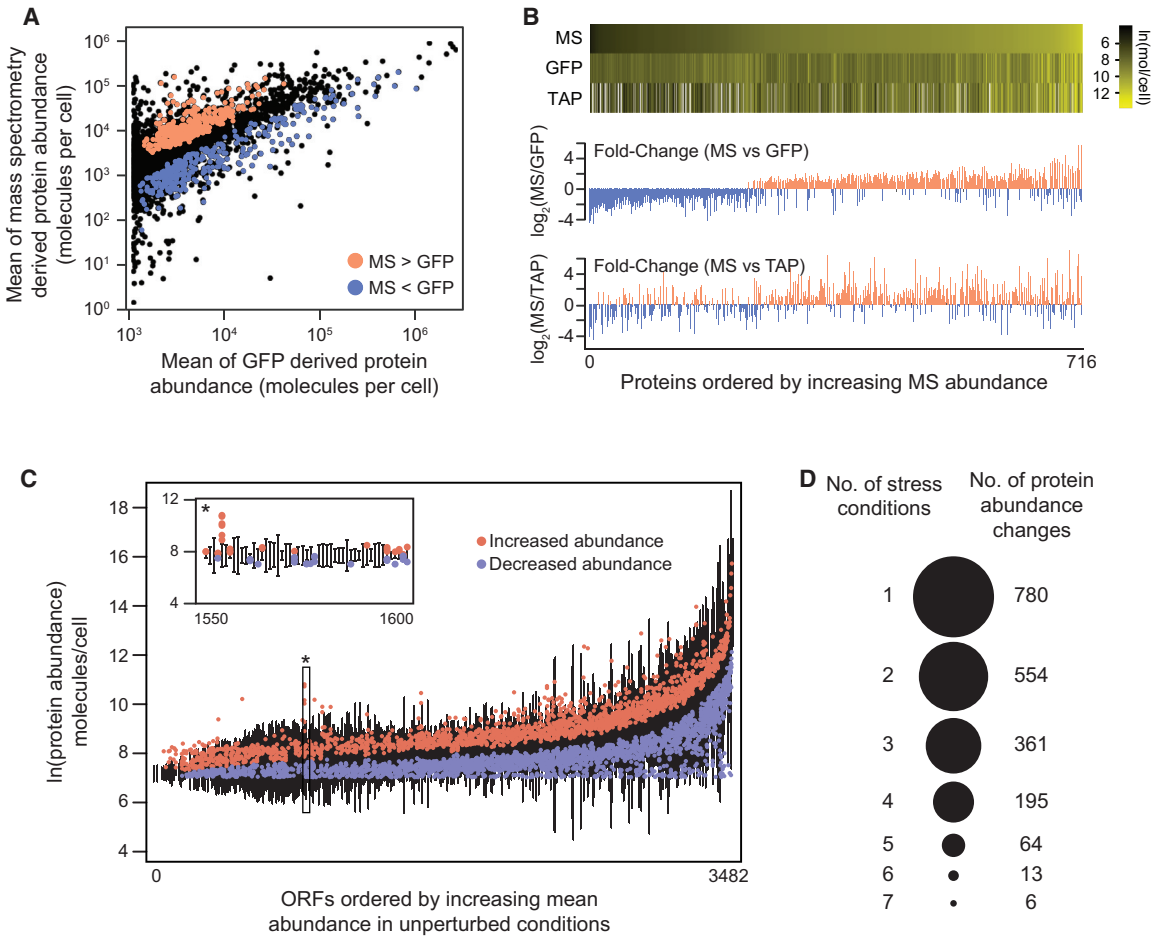


Figure 6. Identification of Proteins Whose Expression Is Influenced by Protein Fusion Tags or by Stress Conditions

(A) Means of mass spectrometry (MS) abundance values are plotted against means of GFP abundance values. GFP-tagged proteins with lower abundance or greater abundance compared with MS measurements are indicated in orange and blue, respectively (t test, $p < 0.05$).

(B) Mean MS, TAP-immunoblot (TAP), and GFP protein abundance values for each identified outlier are compared. Proteins are ordered by increasing MS abundance, with each bar representing a single protein (top). The $\log_2$ ratio of the MS abundance to the GFP abundance (middle) or TAP abundance is displayed for each outlier (bottom).

(C) Proteins are ordered by increasing mean abundance in unperturbed conditions. Proteins that increase or decrease in abundance in a stress condition are colored red or blue, respectively, and gray bars span 2 SDs of abundance in unperturbed conditions.

(D) The number of proteins that change in abundance in the given number of stress conditions is indicated, with the area of the circles proportional to the number of proteins that change in abundance. The stress conditions considered are MMS, HU, rapamycin, H_2O_2, DTT, nitrogen starvation, and quiescence.

conclusion that the yeast proteome can tolerate C-terminal tags well.

Changes in Protein Abundance under Environmental Stresses

Given that protein concentration directly influences cellular processes and function, we were interested to use our molecules per cell dataset to determine absolute protein abundance following stress. External stressors can perturb cellular processes and activate the environmental stress response, a mechanism for cells to protect themselves from fluctuating conditions in the environment (Gasch et al., 2000, 2001; Gasch and Werner-Washburne, 2002; Causton et al., 2001). The environmental stress-response genes were identified through microarray analyses, but have not been studied at the protein level. High-throughput fluorescence microscopy and MS have enabled large-scale analyses of the proteome after exposure to diverse stresses, including quiescence, DNA replication stress conditions, oxidative stress, nitrogen starvation, reductive stress, and rapamycin treatment, providing an opportunity to compare changes in protein levels across studies investigating condition-dependent protein abundance changes and to elucidate a protein core stress response. To simplify the comparisons, we focused on GFP-based studies, which are available for hydroxyurea, methyl methanesulfonate, oxidative stress, reductive stress, nitrogen starvation, rapamycin treatment, and quiescence (Davidson et al., 2011; Tkach et al., 2012; Breker et al., 2013; Denervaud et al., 2013; Mazumder et al., 2013; Chong et al., 2015). MS datasets are available for diploid cells, heat shock, high salt, quiescence, different temperatures, ethanol, and 13 different carbon sources (de Godoy et al., 2008; Nagaraj et al., 2012; Lee et al., 2011; Webb et al., 2013; Usaite et al.,

2008; Paulo et al., 2015, 2016; Lahtvee et al., 2017), but are not considered here.

Since the majority of proteins do not change in abundance in any given stress condition, we normalized GFP intensities from each study by the mode-shifting method and applied the same linear regressions used previously to convert a.u. to protein molecules per cell (Table S8). We consider a protein to change in abundance in stress if the molecules per cell value is more than 2 SDs from the mean of the abundance measurements without stress (Figure 6C; Table S9). At this cutoff, 1,973 of the 4,100 proteins assessed change in abundance in at least one stress, and 616 of the 1,973 proteins have a fold change greater than 2. The abundance changes range up to 109-fold for increases and to 49-fold for decreases (Table S9). Proteins that increased or decreased in abundance during stress tended to be of higher abundance in unperturbed cells than the proteome median (Figure S5), and low-abundance proteins do not show a tendency to be upregulated in response to stress (the first decile, hypergeometric $p = 1$).

Unexpectedly, 68% of the abundance changes were observed in only one or two conditions, suggesting that most condition-dependent regulation of protein abundance levels could be stress specific (Figure 6D). Six proteins (Afg3, Rpt6, Isa1, Rax2, Dat1, and Rna1) were the most universal stress responders, increasing in abundance in all seven stress conditions. By contrast, the mRNA environmental stress response includes some 900 transcripts that are differentially expressed in multiple distinct stress conditions (Gasch, 2007; Gasch et al., 2000, 2001; Gasch and Werner-Washburne, 2002; Causton et al., 2001), including heat shock, oxidative stress, reductive stress, nutrient starvation, DNA damage, and pH. Focusing on oxidative stress, reductive stress, nitrogen starvation, and DNA damage, which are represented in the protein abundance data, we find only 26 proteins that change in abundance in all 4 conditions. Our data suggest that changes in mRNA expression of general stress-response genes are not reflected rapidly at the protein abundance level.

Ribosome biogenesis proteins are strongly over-represented in the proteins that decrease in abundance during stress ($p = 1.37 \times 10^{-18}$, 210 proteins), and are also over-represented in the mRNA environmental stress response (Gasch et al., 2000, 2001; Causton et al., 2001). By contrast to the mRNA response, however, the decrease of ribosome biogenesis proteins is specific to stresses that cause G1 delay (rapamycin, nitrogen starvation, and quiescence), and does not extend to oxidative, reductive, and DNA damage stress. Thus, we see no clear evidence for a global protein abundance response to stress. The apparent absence of a global protein response likely reflects the short half-life of a typical mRNA (~32 min) (Geisberg et al., 2014) compared with the considerably longer median protein half-life (8.8 hr [Christiano et al., 2014]; 2.0 hr [Martin-Perez and Villen, 2015]). In addition, diverse post-transcriptional regulation modes can be brought to bear on protein function, including regulation of translation, protein degradation, protein modification, and intracellular localization changes, such that protein function need not be altered at the level of abundance alone.

The availability of four protein abundance datasets for MMS treatment (Lee et al., 2007; Tkach et al., 2012; Denervaud et al., 2013; Mazumder et al., 2013) allows us to assess the variation among different abundance change analyses. When the four datasets (Table S8) are compared, there are 128 proteins with a statistically supported change (Student's t test, $p < 0.05$), ranging from a 2.3-fold decrease to an 8.5-fold increase. Within this higher-confidence protein abundance increase cohort that we identified, functional enrichment for ubiquitin-mediated proteolysis ($p = 8.1 \times 10^{-5}$) and response to oxidative stress ($p = 1.8 \times 10^{-3}$) was evident. We previously detected the connection between MMS treatment and oxidative stress-response proteins (Tkach et al., 2012), but failed to identify the upregulation of proteasome components and ubiquitination enzymes when analyzing a single stress-response dataset. Thus, meta-analysis of protein abundance studies provides a path to identification of new functional connections among cellular stress-response pathways.

In conclusion, we provide a comprehensive view of protein abundance in yeast by normalizing and combining 21 abundance datasets, collected by MS, GFP fluorescence flow cytometry, GFP fluorescence microscopy, and western blotting. Since cellular autofluorescence interferes with detection of GFP fluorescence, we find that the lower limit for reliable detection of GFP proteins corresponds to ~1,400 molecules per cell. Above this threshold we found less variation among GFP-based studies than among MS studies, suggesting that although MS analyses provide the greatest sensitivity and dynamic range for protein measurements, the GFP-based measurements have greater precision. Collectively, our analyses indicate that protein abundance in the yeast proteome ranges from 3 to 7.5×10^5 molecules per cell, with a median abundance of 2,622 molecules per cell. We define the lowest abundance proteins as those present at 866 or fewer copies, and the highest abundance proteins as those with 14,938 or more copies.

Protein abundance directly influences cellular processes and phenotypes. The plasticity of the proteome in stress conditions has been extensively investigated in yeast. Our normalization scheme allowed us to unify the available data and report protein abundance in a single common unit of molecules per cell, in both unperturbed cells and in response to stress. This method can be applied to other abundance datasets in other stress conditions or to other organisms for comparative studies. Our unified protein abundance dataset provides a useful resource for further analysis of the dynamic regulation of the proteome.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Data Collection and Processing
  - Data Transformation and Assessing Correlation
  - Normalization of Arbitrary Unit Abundance Values
  - Converting Protein Abundance from Arbitrary Units to Molecules per Cell
  - Clustering Analysis
  - Calculating Coefficients of Variation

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, nine tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.cels.2017.12.004.

ACKNOWLEDGMENTS

This work was supported by the Canadian Cancer Society Research Institute (Impact grant 702310 to G.W.B.), an Ontario Government Scholarship and a Natural Sciences and Engineering Research Council of Canada CGS-M award (to B.H.), and a Lewis-Sigler Fellowship (to A.B.). We thank Simon Hubbard, Maya Schuldiner, Allan Drummond, Helena Friesen, Tina Sing, Xanita Saayman, and Raphael Loll-Krippleber for helpful discussions and careful reading of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, B.H., A.B., and G.W.B.; Methodology, B.H., A.B., and G.W.B.; Formal Analysis, B.H. and G.W.B.; Writing – Original Draft, B.H. and G.W.B.; Writing – Review & Editing, B.H., A.B., and G.W.B.

Received: May 24, 2017
Revised: October 10, 2017
Accepted: December 8, 2017
Published: January 17, 2018

REFERENCES

Ahrné, E., Molzahn, L., Glatter, T., and Schmidt, A. (2013). Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics *13*, 2567–2578.

Albert, F.W., Muzzey, D., Weissman, J.S., and Kruglyak, L. (2014). Genetic influences on translation in yeast. PLoS Genet. *10*, e1004692.

Baryshnikova, A. (2016). Systematic functional annotation and visualization of biological networks. Cell Syst. *2*, 412–421.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol. Syst. Biol. *7*, 549.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., and O'Shea, E.K. (2006). Quantification of protein half-lives in the budding yeast proteome. Proc. Natl. Acad. Sci. USA *103*, 13004–13009.

Bernstein, K.A., and Baserga, S.J. (2004). The small subunit processome is required for cell cycle progression at G1. Mol. Biol. Cell *15*, 5038–5046.

Bjorklund, M., Taipale, M., Varjosalo, M., Saharinen, J., Lahdenpera, J., and Taipale, J. (2006). Identification of pathways regulating cell size and cell-cycle progression by RNAi. Nature *439*, 1009–1013.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder – open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics *20*, 3710–3715.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science *335*, 552–557.

Breker, M., Gymrek, M., and Schuldiner, M. (2013). A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. J. Cell Biol. *200*, 839–850.

Bryan, A.K., Goranov, A., Amon, A., and Manalis, S.R. (2010). Measurement of mass, density, and volume during the cell cycle of yeast. Proc. Natl. Acad. Sci. USA *107*, 999–1004.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. (2001). Remodeling of yeast genome expression in response to environmental changes. Mol. Biol. Cell *12*, 323–337.

Chong, Y.T., Koh, J.L., Friesen, H., Duffy, S.K., Cox, M.J., Moses, A., Moffat, J., Boone, C., and Andrews, B.J. (2015). Yeast proteome dynamics from single cell imaging and automated analysis. Cell *161*, 1413–1424.

Christiano, R., Nagaraj, N., Frohlich, F., and Walther, T.C. (2014). Global proteome turnover analyses of the yeasts *S. cerevisiae* and *S. pombe*. Cell Rep. *9*, 1959–1965.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. *2*, 2366–2382.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420.

Csardi, G., Franks, A., Choi, D.S., Airoldi, E.M., and Drummond, D.A. (2015). Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. PLoS Genet. *11*, e1005206.

Davidson, G.S., Joe, R.M., Roy, S., Meirelles, O., Allen, C.P., Wilson, M.R., Tapia, P.H., Manzanilla, E.E., Dodson, A.E., Chakraborty, S., et al. (2011). The proteomics of quiescent and nonquiescent cell differentiation in yeast stationary-phase cultures. Mol. Biol. Cell *22*, 988–998.

de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254.

Denervaud, N., Becker, J., Delgado-Gonzalo, R., Damay, P., Rajkumar, A.S., Unser, M., Shore, D., Naef, F., and Maerkl, S.J. (2013). A chemostat array enables the spatio-temporal analysis of the yeast proteome. Proc. Natl. Acad. Sci. USA *110*, 15842–15847.

Franks, A.M., Csárdi, G., Drummond, D.A., and Airoldi, E.M. (2015). Estimating a structured covariance matrix from multi-lab measurements in high-throughput biology. J. Am. Stat. Assoc. *110*, 27–44.

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.

Gasch, A.P. (2007). Comparative genomics of the environmental stress response in ascomycete fungi. Yeast *24*, 961–976.

Gasch, A.P., Huang, M., Metzner, S., Botstein, D., Elledge, S.J., and Brown, P.O. (2001). Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol. Biol. Cell *12*, 2987–3003.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241–4257.

Gasch, A.P., and Werner-Washburne, M. (2002). The genomics of yeast responses to environmental stress and starvation. Funct. Integr. Genomics *2*, 181–192.

Geisberg, J.V., Moqtaderi, Z., Fan, X., Ozsolak, F., and Struhl, K. (2014). Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. Cell *156*, 812–824.

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.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. *4*, 117.

Hasin, N., Cusack, S.A., Ali, S.S., Fitzpatrick, D.A., and Jones, G.W. (2014). Global transcript and phenotypic analysis of yeast cells expressing Ssa1, Ssa2, Ssa3 or Ssa4 as sole source of cytosolic Hsp70-Ssa chaperone activity. BMC Genomics *15*, 194.

Hodge, V.J., and Austin, J. (2004). A survey of outlier detection methodologies. Artif. Intell. Rev. *22*, 85–126.

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

Illmer, P., Erlebach, C., and Schinner, F. (1999). A practicable and accurate method to differentiate between intra- and extracellular water of microbial cells. FEMS Microbiol. Lett. *178*, 135–139.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Jorgensen, P., Nishikawa, J.L., Breitkreutz, B.J., and Tyers, M. (2002). Systematic identification of pathways that couple cell growth and division in yeast. Science *297*, 395–400.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Lahtvee, P.J., Sanchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., and Nielsen, J. (2017). Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. Cell Syst. *4*, 495–504.e5.

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. USA *94*, 13057–13062.

Laurent, J.M., Vogel, C., Kwon, T., Craig, S.A., Boutz, D.R., Huse, H.K., Nozue, K., Walia, H., Whiteley, M., Ronald, P.C., et al. (2010). Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics *10*, 4209–4212.

Lawless, C., Holman, S.W., Brownridge, P., Lanthaler, K., Harman, V.M., Watkins, R., Hammond, D.E., Miller, R.L., Sims, P.F.G., Grant, C.M., et al. (2016). Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. Mol. Cell Proteomics *15*, 1309.

Lee, M.V., Topper, S.E., Hubler, S.L., Hose, J., Wenger, C.D., Coon, J.J., and Gasch, A.P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Mol. Syst. Biol. *7*, 514.

Lee, M.W., Kim, B.J., Choi, H.K., Ryu, M.J., Kim, S.B., Kang, K.M., Cho, E.J., Youn, H.D., Huh, W.K., and Kim, S.T. (2007). Global protein expression profiling of budding yeast in response to DNA damage. Yeast *24*, 145–154.

Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P., and Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. Nat. Biotechnol. *27*, 652–658.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. *25*, 117–124.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bahler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell *151*, 671–683.

Martin-Perez, M., and Villen, J. (2015). Feasibility of protein turnover studies in prototroph *Saccharomyces cerevisiae* strains. Anal. Chem. *87*, 4008–4014.

Mazumder, A., Pesudo, L.Q., McRee, S., Bathe, M., and Samson, L.D. (2013). Genome-wide single-cell-level screen for protein abundance and localization changes in response to DNA damage in *S. cerevisiae*. Nucleic Acids Res. *41*, 9310–9324.

Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays *35*, 1050–1055.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. Mol. Cell Proteomics *11*, M111.013722.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature *441*, 840–846.

Paulo, J.A., O'Connell, J.D., Everley, R.A., O'Brien, J., Gygi, M.A., and Gygi, S.P. (2016). Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 *S. cerevisiae* proteins across 10 carbon sources. J. Proteomics *148*, 85–93.

Paulo, J.A., O'Connell, J.D., Gaun, A., and Gygi, S.P. (2015). Proteome-wide quantitative multiplexed profiling of protein expression: carbon-source dependency in *Saccharomyces cerevisiae*. Mol. Biol. Cell *26*, 4063–4074.

Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A.J. (2012). Protease bias in absolute protein quantitation. Nat. Methods *9*, 524–525.

Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B., and Aebersold, R. (2009). Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell *138*, 795–806.

Picotti, P., Clement-Ziza, M., Lam, H., Campbell, D.S., Schmidt, A., Deutsch, E.W., Rost, H., Sun, Z., Rinner, O., Reiter, L., et al. (2013). A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature *494*, 266–270.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol. Syst. Biol. *10*, 770.

Qiu, X., Wu, H., and Hu, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC Bioinformatics *14*, 124.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. *16*, 939–945.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

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.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.

Teng, T., Mercer, C.A., Hexley, P., Thomas, G., and Fumagalli, S. (2013). Loss of tumor suppressor RPL5/RPL11 does not induce cell cycle arrest but impedes proliferation due to reduced ribosome content and translation capacity. Mol. Cell. Biol. *33*, 4660–4671.

Thakur, S.S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., and Mann, M. (2011). Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol. Cell Proteomics *10*, M110.003699.

Thomson, T.M., Benjamin, K.R., Bush, A., Love, T., Pincus, D., Resnekov, O., Yu, R.C., Gordon, A., Colman-Lerner, A., Endy, D., et al. (2011). Scaffold number in yeast signaling system sets tradeoff between system output and dynamic range. Proc. Natl. Acad. Sci. USA *108*, 20265–20270.

Tkach, J.M., Yimit, A., Lee, A.Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J.A., Ou, J., Moffat, J., Boone, C., et al. (2012). Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat. Cell Biol. *14*, 966–976.

Usaite, R., Wohlschlegel, J., Venable, J.D., Park, S.K., Nielsen, J., Olsson, L., and Yates Iii, J.R. (2008). Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression

Saccharomyces cerevisiae strains: the comparison of two quantitative methods. J. Proteome Res. *7*, 266–275.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Volarevic, S., Stewart, M.J., Ledermann, B., Zilberman, F., Terracciano, L., Montini, E., Grompe, M., Kozma, S.C., and Thomas, G. (2000). Proliferation, but not growth, blocked by conditional deletion of 40S ribosomal protein S6. Science *288*, 2045–2047.

von der Haar, T. (2008). A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst. Biol. *2*, 87.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. *17*, 691–695.

Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics *15*, 3163–3168.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S.P., Hengartner, M.O., and von Mering, C. (2012). PaxDb, a database of protein abundance averages across all three domains of life. Mol. Cell Proteomics *11*, 492–500.

Warner, J.R. (1999). The economics of ribosome biosynthesis in yeast. Trends Biochem. Sci. *24*, 437–440.

Webb, K.J., Xu, T., Park, S.K., and Yates, J.R., 3rd (2013). Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. J. Proteome Res. *12*, 2177–2184.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Rep. *14*, 1787–1799.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature *499*, 79–82.

Yamada, E.A., and Sgarbieri, V.C. (2005). Yeast (*Saccharomyces cerevisiae*) protein concentrate: preparation, chemical composition, and nutritional and functional properties. J. Agric. Food Chem. *53*, 3931–3936.

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., et al. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. USA *106*, 3264–3269.

Yofe, I., Weill, U., Meurer, M., Chuartzman, S., Zalckvar, E., Goldman, O., Ben-Dor, S., Schutze, C., Wiedemann, N., Knop, M., et al. (2016). One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. Nat. Methods *13*, 371–378.

Yu, L., Castillo, L.P., Mnaimneh, S., Hughes, T.R., and Brown, G.W. (2006). A survey of essential gene function in the yeast cell division cycle. Mol. Biol. Cell *17*, 4736–4747.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Denervaud et al protein abundance	https://github.com/opencb/cellbase	N/A
Causton et al mRNA abundance	http://younglab.wi.mit.edu/environment/	N/A
Brar et al ribosome profiling	Gene Expression Omnibus	GSE34082
Albert et al ribosome profiling	Gene Expression Omnibus	GSM1335348
Pop et al ribosome profiling	Gene Expression Omnibus	GSE63789
Other datasets were extracted from the supplementary material of the indicated publications	N/A	N/A
Software and Algorithms		
R v3.3.1	N/A	https://www.r-project.org
Cytoscape v3.4.0	(Cline et al., 2007; Shannon et al., 2003)	http://www.cytoscape.org/
REViGO	(Supek et al., 2011)	http://revigo.irb.hr/
GO TermFinder	(Boyle et al., 2004)	http://go.princeton.edu/
ProteinAbundance_Figure1.R	This paper	Data S1
ProteinAbundance_Figure2.R	This paper	Data S1
ProteinAbundance_Figure3.R	This paper	Data S1
ProteinAbundance_Figure4.R	This paper	Data S1
ProteinAbundance_Figure5.R	This paper	Data S1
ProteinAbundance_Figure6.R	This paper	Data S1
ProteinAbundance_Figure7.R	This paper	Data S1
ProteinAbundance_PartOne_Functions.R	This paper	Data S1
ProteinAbundance_PartTwo_Functions.R	This paper	Data S1

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Grant W. Brown (grant.brown@utoronto.ca).

METHOD DETAILS

Data Collection and Processing

We gathered 21 data sets from published studies measuring protein abundance across the yeast proteome, either reported in arbitrary units or in molecules per cell (Ghaemmaghami et al., 2003; Newman et al., 2006; Lee et al., 2007; Lu et al., 2007; de Godoy et al., 2008; Davidson et al., 2011; Lee et al., 2011; Thakur et al., 2011; Nagaraj et al., 2012; Peng et al., 2012; Tkach et al., 2012; Breker et al., 2013; Denervaud et al., 2013; Mazumder et al., 2013; Webb et al., 2013; Kulak et al., 2014; Chong et al., 2015; Lawless et al., 2016; Yofe et al., 2016; Lahtvee et al., 2017; Picotti et al., 2009). Throughout our analysis, we used designated codes to refer to each study (Table 1). Unperturbed measurements derived from (Chong et al., 2015) are the mean of the three technical replicates in their study, and for stress conditions the 160 minute data for hydroxyurea and the 700 minute data for rapamycin were used. Measurements in unperturbed cells derived from (Denervaud et al., 2013) were accessed at https://github.com/opencb/cellbase and are the mean of all time points prior to their treatment condition. For (Peng et al., 2012), the average of all data was used. For (Webb et al., 2013), the average of the emPAI values from the three micro MudDPIT replicates was used. For (Yofe et al., 2016), abundance measurements obtained from proteins expressed from their native promoters were used.

For the purposes of our analysis, we consider the yeast proteome to consist of 5858 proteins (*Saccharomyces* Genome Database, www.yeastgenome.org, accessed October 28, 2016), encoded by 5157 verified ORFs and 701 uncharacterized ORFs (only 2 verified ORFs, *GPC1* and *ANY1*, have been added since). We excluded 746 dubious ORFs, as defined in the *Saccharomyces* Genome Database, from our analysis. Although some had peptides detected by mass spectrometry as annotated in the PeptideAtlas (http://www.peptideatlas.org) and Global Protein Machine (GPM, http://www.thegpm.org) databases, only 6 had good evidence for expression as

defined by GPM. Proteins encoded by transposable elements, although readily detected, are not included in our analysis because most do not map to a unique ORF. Abundance data were called out of each of the 21 datasets using the 5858 protein ORFeome (Table S1).

Data Transformation and Assessing Correlation

The natural logarithm was taken for each data set, since this is approximately normally distributed and thus suitable for linear regression analyses. All analyses and calculations were performed on natural log transformed data, unless specified otherwise. Pearson correlation coefficient (r) was used for all correlation analyses.

Normalization of Arbitrary Unit Abundance Values

Mode shift normalization was applied to all studies that measured relative protein abundance and reported values in arbitrary units (Table 1). Each study that required normalization was natural log transformed and divided into 50 bins of equal abundance range. The median value of the bin with the greatest number of observations (values reported) was defined as the mode of the distribution. A scalar value was applied to each study to shift the mode to an arbitrarily chosen value of 100 arbitrary units. Mode shift normalized values were used for the remainder of the analysis.

For comparison to the mode shift normalization, studies were also quantile normalized and center log ratio transformed. For quantile normalization, proteins with a reported measurement from every study were retained for analysis, and quantile normalization was performed as described (Qiu et al., 2013). To normalize data sets by the center log ratio transformation method, arbitrary abundance measurements for each protein from each study were divided by the geometric mean and $\log_{10}$ transformed:

$$center\ log\ ratio = log_{10}\ (X_i/geometric\ mean\ (X)).$$

Converting Protein Abundance from Arbitrary Units to Molecules per Cell

Mean protein abundance for each ORF was calculated for the five mass spectrometry-based studies reporting absolute protein abundance (Lu et al., 2007; Peng et al., 2012; Kulak et al., 2014; Lawless et al., 2016; Lahtvee et al., 2017). We used the mean value for each protein as our calibration set (Table S1) so that the contribution of each study was weighted equally. The calibration set was natural log transformed, as was each normalized protein abundance dataset. Prior to calculating protein molecules per cell from arbitrary units, a filter was applied to remove GFP-based protein measurements that were likely within autofluorescence levels. Mode-shift normalized measurements for proteins labeled as autofluorescent in (Chong et al., 2015) were extracted from their dataset. The maximum value (106.549 arbitrary units) within this subset of proteins was defined as autofluorescence in our dataset. Any protein in the other GFP datasets with a normalized value less than 106.549 arbitrary units was labeled as 'autofluorescence' and removed from the remainder of the analysis. A linear least-squares regression was applied to model the relationship between the calibration set and each abundance dataset. The resulting equations were then applied to each protein abundance dataset to convert arbitrary units to molecules per cell.

Clustering Analysis

Clustering analyses were performed on natural log transformed median protein abundance in molecules per cell for the 21 studies. Hierarchical agglomerative clustering was performed using complete linkage clustering on the dissimilarity matrix measuring the similarity between each of the studies analyzed. Each dataset is its own independent cluster and iteratively combined by Euclidean distance, with the distance between each cluster being recomputed using the Lance-Williams algorithm, into larger clusters.

Clustering by *k*-means was performed with six defined centres. To determine an appropriate number of *k* clusters, the total within-cluster sum of squares was measured with increasing numbers of clusters. A total of six clusters were determined suitable for *k*-means clustering analysis since increasing the number of clusters did not provide better modelling of the data, as determined by measuring the decrease in total within-cluster sum of squares.

Calculating Coefficients of Variation

For each ORF, the coefficient of variation (CV) was calculated by:

$$CV = 100 \times \frac{SD_{ORF}}{Mean}$$

The CV was calculated for each ORF when at least two measurements were reported.

Gene Ontology Term Enrichment

GO term analysis was performed using the GO term finder tool (http://go.princeton.edu/) using a p-value cutoff of 0.01 and applying Bonferroni correction, querying biological process or component enrichment for each gene set. After removing high frequency terms (>10% of background gene set), GO term enrichment results were further processed with REViGO (Supek et al., 2011) using the "Medium (0.7)" term similarity filter and simRel score as the semantic similarity measure.

Spatial Analysis of Functional Enrichment (SAFE)

Functional annotation of the aggregated median protein abundance measurements on available genetic similarity networks constructed by Costanzo et al. (2016) was performed as previously described (Baryshnikova, 2016), using Cytoscape v3.4.0 (Cline et al., 2007; Shannon et al., 2003).

Identifying Abundance Differences between GFP and Mass Spectrometry Studies

An unpaired, two-tailed t-test was performed between the 11 mass spectrometry and 8 GFP studies for each protein. Abundance differences were considered to be statistically supported if the p-value was less than 0.05.

RNA Level, Ribosome Profiling, and Protein Abundance Comparison

RNA transcript levels in arbitrary units from RNA-seq datasets (Yassour et al., 2009; Lipson et al., 2009; Nagalakshmi et al., 2008), translation rates in arbitrary units from ribosomal profiling (Albert et al., 2014; Brar et al., 2012; Ingolia et al., 2009; Pop et al., 2014; Weinberg et al., 2016), and median protein abundance in molecules per cell from this study were mode-shift normalized and natural log transformed. The mean was determined for the natural log transformed RNA and ribosomal profiling data. Mahalanobis distances have been previously used in multivariate outlier detection analysis (Hodge and Austin, 2004). Therefore, we identified multivariate outliers by calculating Mahalanobis distances for each gene/protein. Proteins with Mahalanobis distances greater than 12.84 were considered outliers (Chi-squared distribution, alpha level = 0.005, degrees of freedom = 3).

Changes in Protein Abundance in Stress Conditions

For each study considered for analysis, unperturbed and stress measurements were mode-shift normalized, filtered for autofluorescence, and converted to molecules per cell as described above. The standard deviation was calculated for each protein from the seven GFP studies for the unperturbed condition. Any protein observation from any single study with an abundance measurement in a stress condition that was greater than 2 or less than 2 standard deviations from the mean was considered to be a protein with changed abundance.

Proteins with significant abundance changes in the MMS treatment conditions compared to the unperturbed condition were identified using an unpaired, two-tailed t-test ($p < 0.05$).

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analysis, data manipulation, and data visualization was performed in R (https://www.r-project.org). All of the details of data analysis can be found in the Results and Method Details sections.

DATA AND SOFTWARE AVAILABILITY

Datasets are provided in the Tables S1, S2, S3, S4, S5, S6, S7, S8, and S9. The R scripts used for data analysis are provided in the Data S1.

Cell Systems, Volume *6*

Supplemental Information

Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome

Brandon Ho, Anastasia Baryshnikova, and Grant W. Brown

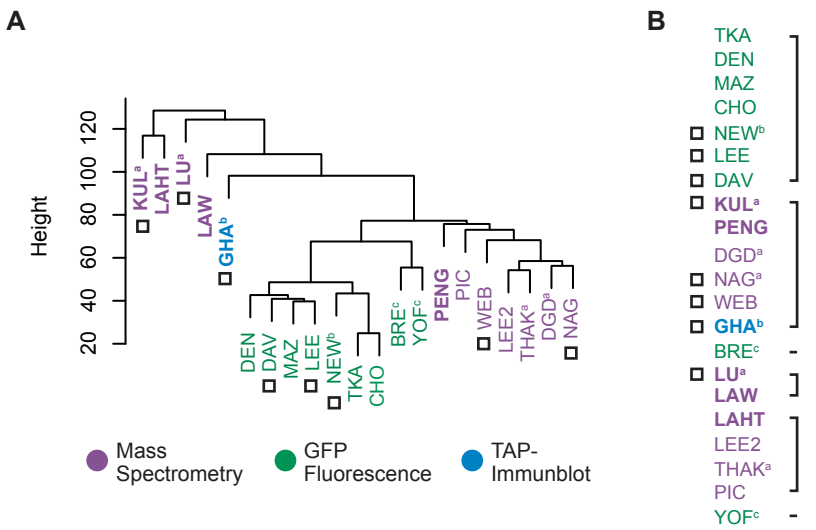


Figure S1. Related to Figure 1. Hierarchical and k-means clustering of normalized and scaled data sets. Twenty-one protein abundance data sets were subjected to hierarchical clustering (A) or k-means clustering (B). Codes for each study are as in Table 1. Data sets are colour-coded by type (purple for mass spectrometry, green for GFP fluorescence, blue for immunoblotting). Studies from the same lab are designated (a, b, c). Studies where cells were grown in rich media are designated with squares (the rest used minimal media). Data sets that measured absolute abundance are indicated in bold.

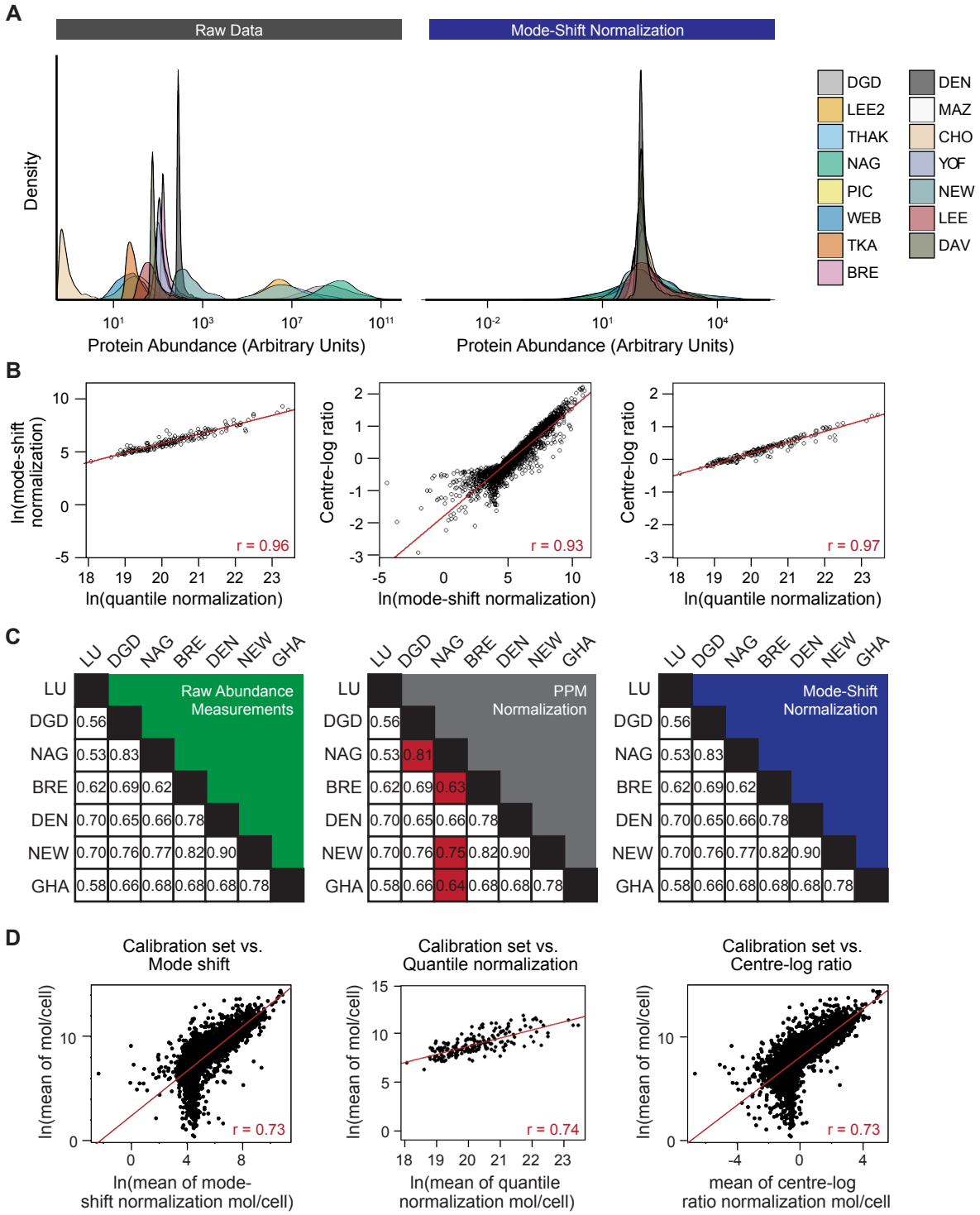


Figure S2. Related to Figures 1 and 2. Normalization methods and comparisons to the calibration abundance data set. (A) Raw protein abundance measurements from studies reporting arbitrary units (left) were mode-shift normalized (right). (B) Protein abundance measurements were normalized using mode-shift, quantile, or centre-log ratio normalization methods. The mean abundance for each protein was calculated following normalization, and each was compared to the others. (C) Pearson correlation coefficients were calculated for each pairwise comparison of the seven studies indicated. Correlation coefficients were calculated for the raw abundance measurements (left), parts per million normalized datasets (middle), or mode-shifted datasets (right). Boxes shaded in red indicate correlations that are not equivalent to correlations among the original raw datasets. (D) Each normalization method was compared with the mean abundance from the calibration data set. Pearson correlation coefficients (r) are indicated in each plot.

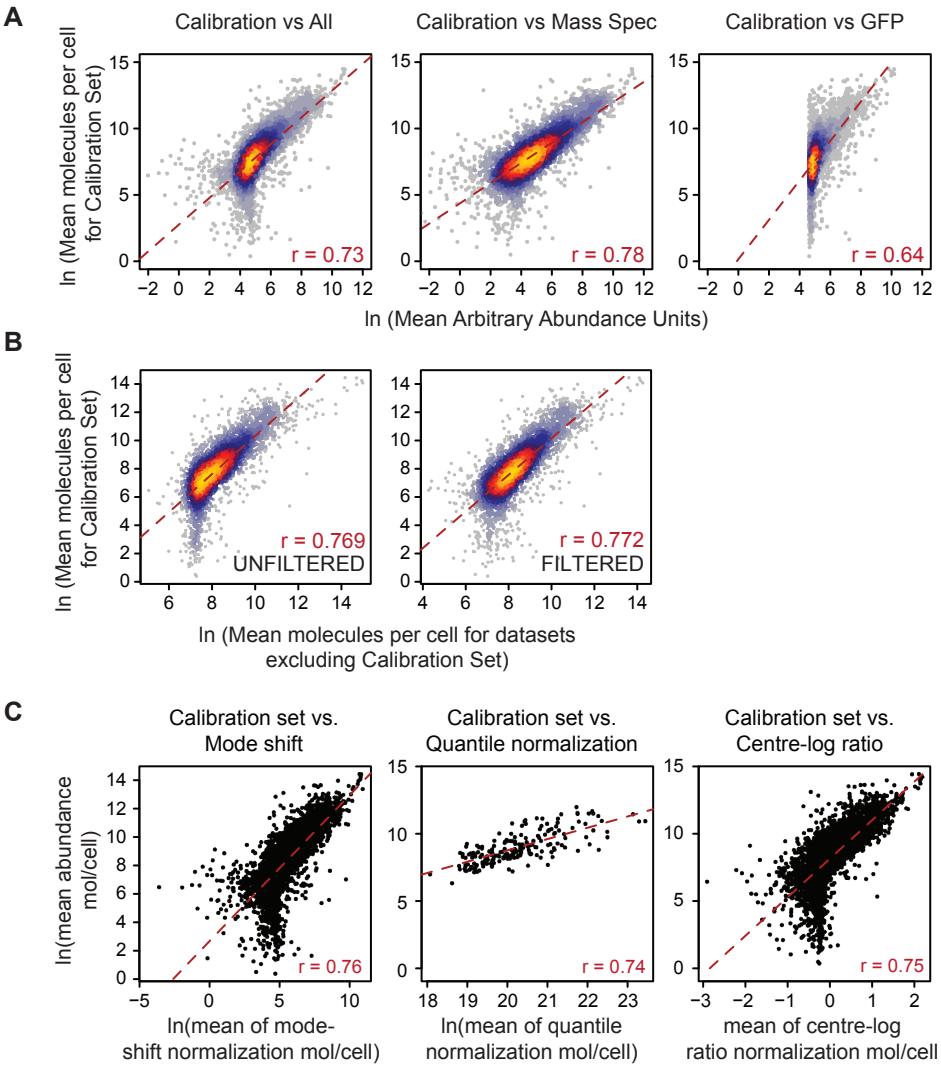


Figure S3. Related to Figure 2. Removing GFP autofluorescence from the GFP data sets. (A) Protein abundances in the calibration data set are compared to those in all arbitrary abundance units datasets, those in mass spectrometry arbitrary abundance units data sets, and those in GFP fluorescence arbitrary abundance units data sets. Pearson correlation coefficients (r) are shown. (B) Protein abundances in the calibration data set are compared to those in the remaining 16 datasets (excluding the calibration data set studies), either before (left) or after (right) removing GFP values below the cellular autofluorescence value. (C) Each normalization method was compared with the mean abundance from the calibration data set. Pearson correlation coefficients (r) are indicated in each plot.

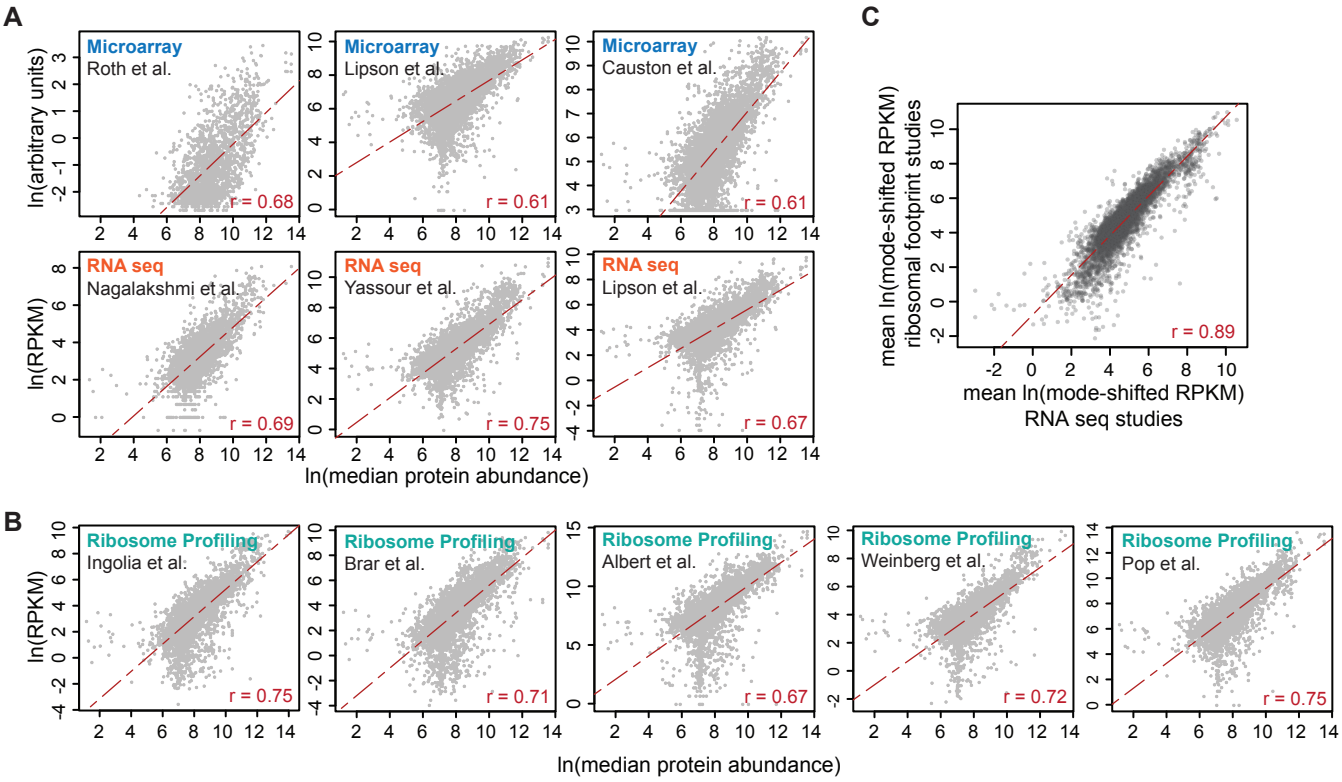


Figure S4. Related to Figure 5. Comparison of protein abundance with mRNA levels. (A) Protein abundance (natural log of the median number of molecules per cell) is compared with mRNA levels measured by microarray analyses from three independent studies (natural log of arbitrary units), and with mRNA levels measured by RNA sequencing analysis (natural log of reads per kilobase of transcript per million mapped reads (RPKM)). (B) Protein abundance (natural log of median molecules per cell) is compared with ribosome footprint abundance (natural log of RPKM) from ribosome profiling analysis. (C) Three RNA seq data sets (Nagalakshmi et al. 2008; Lipson et al. 2009; Yassour et al. 2009) and five ribosomal profiling data sets (Ingolia et al. 2009; Brar et al. 2012; Albert et al. 2014; Pop et al. 2014; Weinberg et al. 2016) were mode-shift normalized as described in the methods and subsequently natural log transformed. The mean of the transformed datasets was calculated for each ORF, and ribosomal profiling quantification is compared with mRNA abundance by RNA seq. The least squares linear regressions (red line) and the Pearson correlation coefficients (r) are indicated.

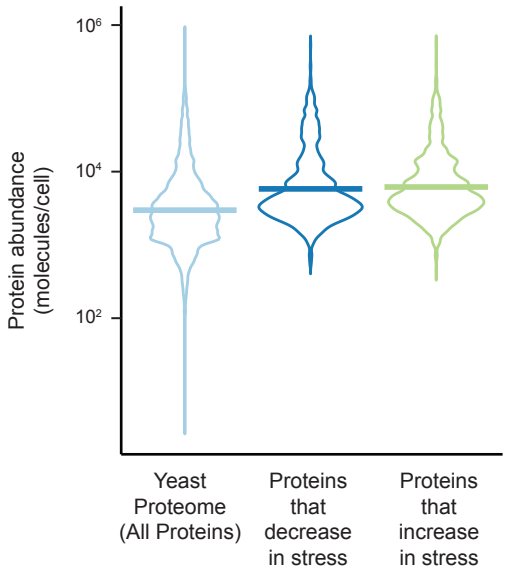


Figure S5. Related to Figure 6. Abundance distributions of the proteome and proteins that change in stress. The protein abundance distribution in molecules per cell is presented as a violin plot for all proteins in the proteome, and for proteins that decrease or increase in abundance by at least two standard deviations from the mean abundance in unperturbed cells, in at least one stress condition. The horizontal bars represent the medians, and violins are scaled such that all have the same area.