nature biotechnology

Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation

Peng Lu[1–3], Christine Vogel[1,2], Rong Wang[1,2], Xin Yao[1] & Edward M Marcotte[1]

We report a method for large-scale absolute protein expression measurements (APEX) and apply it to estimate the relative contributions of transcriptional- and translational-level gene regulation in the yeast and *Escherichia coli* proteomes. APEX relies upon correcting each protein's mass spectrometry sampling depth (observed peptide count) by learned probabilities for identifying the peptides. APEX abundances agree with measurements from controls, western blotting, flow cytometry and two-dimensional gels, as well as known correlations with mRNA abundances and codon bias, providing absolute protein concentrations across approximately three to four orders of magnitude. Using APEX, we demonstrate that 73% of the variance in yeast protein abundance (47% in *E. coli*) is explained by mRNA abundance, with the number of proteins per mRNA log-normally distributed about ~5,600 (~540 in *E. coli*) protein molecules/mRNA. Therefore, levels of both eukaryotic and prokaryotic proteins are set per mRNA molecule and independently of overall protein concentration, with >70% of yeast gene expression regulation occurring through mRNA-directed mechanisms.

Although routine, large-scale measurement of expressed cellular proteins has yet to be realized[1], shotgun proteomics is perhaps closest to reaching this goal[2,3]. This technology involves proteolysis of protein mixtures, followed by analysis of the peptides generated using chromatography and mass spectrometry (MS). In shotgun proteomics—for example, MudPIT, based on multidimensional chromatography (two-dimensional (2D) high performance liquid chromatography (HPLC)) with in-line tandem mass spectrometry (MS/MS)[2]—evidence for individual proteins accumulates through observations of component peptides. This technique is mature enough to observe ~500–1,000 different proteins from a cell lysate (e.g., see Peng *et al.*[4]). However, this approach is not generally thought to be quantitative because the efficiency with which peptides ionize and enter the mass spectrometer depends upon both their composition and the local chemical environment[1], producing variation in the MS signal intensity, that is, peak height. Several approaches quantify peptides by introducing internal reference standards (e.g., see Silva *et al.*[5]), typically by mixing in isotopically labeled samples[6]. These reference peptides derive either from cells grown in labeled medium (Stable Isotope Labeling with Amino acids in Cell culture, SILAC[7]), by derivatizing natural samples (Isotope Coded Affinity Tags, ICAT[8]) or by doping in synthetic peptides, as in isotope dilution (e.g., Absolute Quantification of proteins, AQUA[9]). The first two approaches result in protein quantification relative to the isotopically labeled reference sample. In contrast, AQUA provides absolute quantification because the amounts of added reference peptides are known. However, owing to the expense and difficulty of synthesizing thousands of isotopically labeled peptides, this approach has yet to be applied on a proteomic scale.

Research in quantifying MS proteomics data has mainly focused on measuring peak heights rather than using other information, such as peptide counts. Several peptides may be observed for each protein, some of these many times (**Fig. 1**). Both the coverage of unique peptides (that is, percentage of possible peptides per protein actually observed) and the total number of repeat observations of peptides provide rough approximations of protein abundance (e.g., see refs. 10–12). However, both of these measures have distinct shortcomings. Coverage of unique peptides is a poor measure of abundance, saturating at 100% and limiting the dynamic range. It is normalized for protein size but ignores different sampling depths between experiments. By contrast, approximating abundance from the number of repeat peptide observations per protein ignores protein size—as large proteins contribute more peptides than small ones, their abundance will be overestimated unless the data are normalized[13,14]. Like peptide coverage data, data from this method are not directly comparable between experiments with different sampling depths. Most importantly, neither approach includes any prior expectations of which peptides are observed in the mass spectrometer, although such trends can in part be predicted from a peptide's composition[15–17].

We derive a simple but robust measure of protein expression that can be calculated from peptide sampling depth. This provides a strategy for rapid, highly reproducible and accurate absolute

[1]Center for Systems and Synthetic Biology, Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, 2500 Speedway, University of Texas, Austin, Texas 78712, USA. [2]These authors contributed equally to this work. [3]Present address: Department of Genetics and Genomics, Roche, Palo Alto S3-1, 3431 Hillview Ave., Palo Alto, California 94304, USA. Correspondence and requests for materials should be addressed to E.M.M. (marcotte@icmb.utexas.edu).

Received 19 September; accepted 26 October; published online 24 December 2006; doi:10.1038/nbt1270

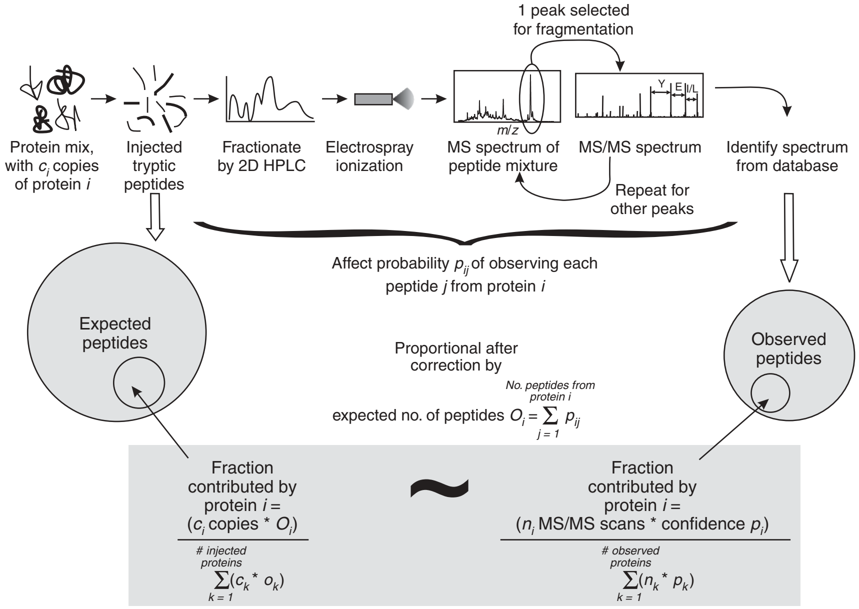

Figure 1 Absolute protein expression (APEX) profiling exploits the proportionality between the fractions of peptides expected and observed from a given protein. Proteins are analyzed by standard shotgun proteomics, beginning with tryptic digest of a protein mixture, liquid chromatographic separation of the mixture (2D HPLC), analysis of peptide masses by mass spectrometry (MS) and fragmentation of peptides and subsequent analysis of the fragmentation spectra (MS/MS). Each step introduces bias into the peptides ultimately interpreted from the analysis, thereby affecting the probability p_{ij} of observing each peptide j from protein i. APEX involves training a classifier to estimate O_i, the prior estimate of the number of unique peptides expected from a given protein during such an experiment. By correcting for O_i, the number of peptides observed per protein thereby provides an estimate of the protein's abundance. HPLC, high-performance liquid chromatography.

quantification in shotgun proteomics. APEX provides a measure of absolute protein abundance that correlates well with measurements from control mixtures, from 2D gel electrophoresis[18,19] and from the only other large-scale protein quantification approaches for which absolute expression measurements are available: high-throughput analysis of fusion proteins by western blotting[20] or flow cytometry[21]. We apply APEX to characterize the yeast and *E. coli* proteomes and their transcriptional and translational regulation.

RESULTS

APEX estimates absolute protein concentration per cell from the proportionality between the abundance of a protein and the number of its peptides observed in a MudPIT experiment. APEX is corrected by the background expectation of observing each peptide in the experiment, the total sampling depth and the confidence in protein identification. In an experiment (**Fig. 1**), a mixed pool of peptides derived from many proteins of varying concentrations is injected into the mass spectrometer ('injected peptides'). Owing to differential ionization, molecular weight, solubility and other properties, only some fraction of these peptides is ionized, analyzed by the mass spectrometer and identified ('observed peptides'). An individual protein will account for some fraction of the total number of peptides in the injected peptide pool, and it will account for some fraction of the interpreted peptide mass spectra. The key to APEX is the introduction of appropriate correction factors that make these fractions proportional to one another. We estimate the protein's abundance from the fraction of peptide mass spectra associated with one protein, corrected by the prior expectation of observing each peptide.

Validating APEX with other measures of protein abundance

To be useful for expression profiling, the accuracy of APEX measurements must be validated. First, APEX-derived abundances match known abundances in simple protein mixtures over ~2.5 orders of magnitude, with mean abundance differences of 2.3 ± 1.1- and 3.0 ± 3.5-fold for mixtures comprising ten and five different proteins, respectively. A more than tenfold difference for individual proteins was never observed (**Fig. 2a** and **Supplementary Notes** online). Second, the protein abundances determined by APEX are highly reproducible between replicate experiments, both in small-scale (**Fig. 2b**) and proteome-scale (Spearman rank correlation R_s = 0.95, squared Pearson correlation R^2 = 0.88, **Fig. 2c** and **Supplementary Notes**) experiments. Third, for high-complexity samples, APEX compares favorably to other approaches. As other MS-derived protein expression measurements (e.g., ICAT and SILAC) provide only relative expression changes and are not comparable with APEX, we used technologies (2D gels and high-throughput fusion protein analysis) that produce comparable large-scale absolute measurements of protein abundance.

We applied APEX to yeast growing in rich medium to measure absolute abundances of 454 proteins with <5% false discovery rate. We compared APEX-derived measurements with 3,869 measurements from an analysis by western blotting[20] (**Fig. 2d**), 2,214 measurements by flow cytometry of GFP-tagged fusion proteins[21] (**Fig. 2e**) and 71 measurements by 2D gels[18] (**Fig. 2f**). APEX data provided measurements of an additional 76 proteins omitted by the other approaches. APEX-derived values correlated better with western and 2D gel measures (R_s = 0.61 and 0.80; R^2 = 0.34 and 0.52; respectively) than they correlated with each other (R_s = 0.30, R^2 = 0.02). Good correlation with GFP fusion protein quantities (R_s = 0.69, R^2 = 0.49) confirms that APEX is comparable to other approaches.

Given that all datasets are noisy, we assume that data points with similar values in two independent measurements are reasonably reliable. In all comparisons, the measurements correctly fell along the diagonal, estimating absolute protein abundance across three to four orders of magnitude. The results were similar when analyzing APEX abundances for 555 yeast proteins with a 10% false discovery rate (**Supplementary Data 1** online). Most (95%) of the APEX-observed protein abundances fell between ~1,300 and ~710,000 molecules/cell. The least abundant proteins observed were YHL009W-B and BRE4, at 482 and 851 molecules/cell, respectively, whereas the most abundant were enolase 2 (ENO2; ~2,500,000 molecules/cell) and translation elongation factor 1 alpha (TEF2; ~1,200,000 molecules/cell). Although the expression levels of ENO2 and TEF2 have been confirmed by independent estimates[22,23], both represent outliers in the western blot data (**Fig. 2d**), indicating that APEX produces more typical values. Lower detection limits for APEX can be achieved through higher sampling depth (**Supplementary Notes**).

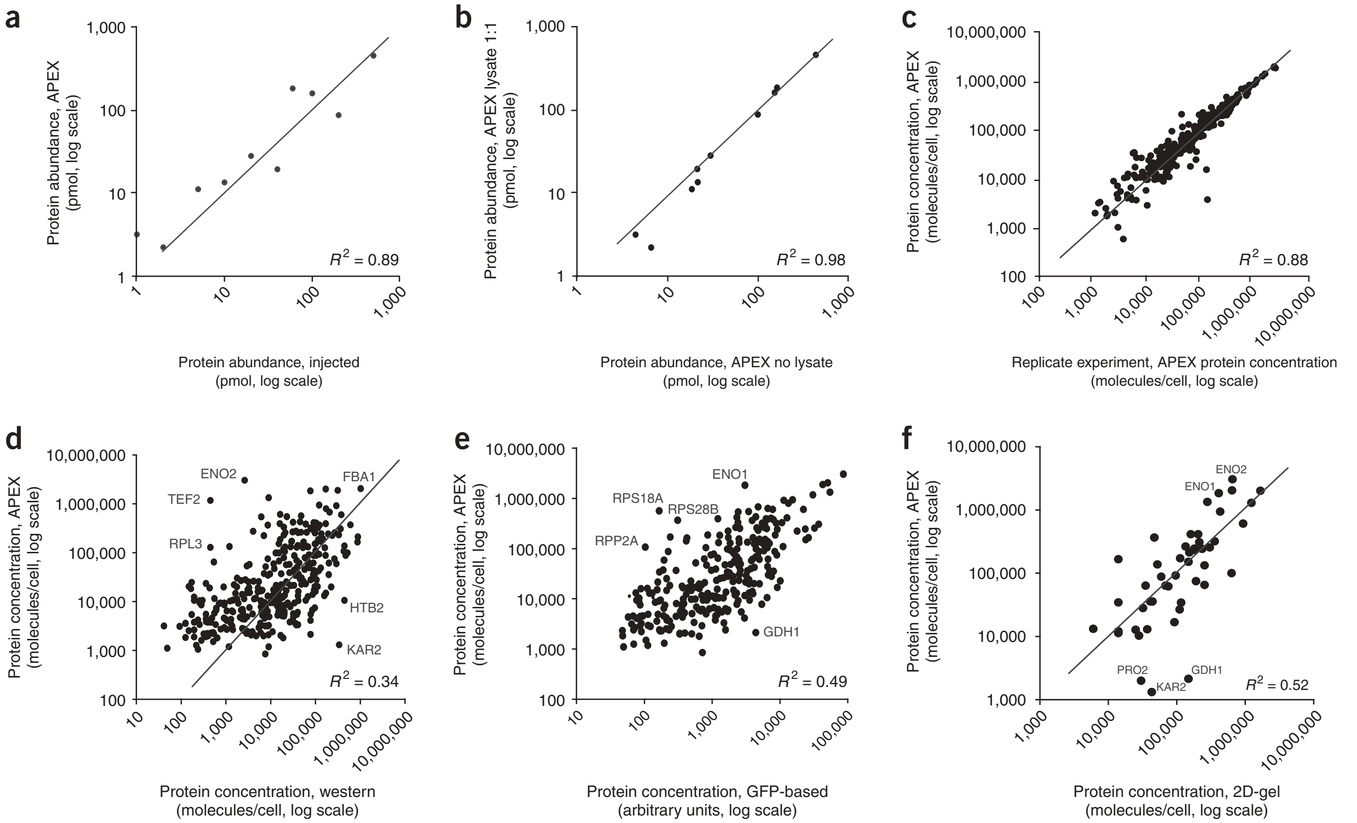

Figure 2 APEX measurements are both reproducible and consistent with other abundance measurements. (**a**) APEX measurements are accurate to approximately twofold, on average, for a control set of ten proteins spiked into yeast cell lysate in a mass ratio of 1:1 (see **Supplementary Notes** for more control experiments) ($R_s = 0.93$, $R^2 = 0.84$ for linear data, $R^2 = 0.89$ for log-transformed data). (**b**) APEX measurements are reproducible, even under changing experimental backgrounds, as shown by comparing abundances derived for the control mixture of ten proteins analyzed as a distinct set without lysate and as spiked into yeast cell lysate in a mass ratio of 1:1 ($R_s = 0.98$, $R^2 = 0.997$ for linear data, $R^2 = 0.98$ for log-transformed data). (**c**) Proteome-scale APEX measurements are highly reproducible ($R_s = 0.95$, $R^2 = 0.97$ for linear data, $R^2 = 0.88$ for log-transformed data), as shown for 278 yeast proteins measured in independent shotgun proteomics experiments on cells grown in minimal (YMD) medium (two sets of data pooled from three injections each). (**d**–**f**) APEX measurements of absolute protein abundance per cell are also correlated with abundances measured by western blot[20] ($R_s = 0.61$, $R^2 = 0.34$; 340 proteins) (**d**), flow cytometry[21] ($R_s = 0.69$, $R^2 = 0.49$; 308 proteins) (**e**) and 2D gel[18] ($R_s = 0.80$, $R^2 = 0.52$; 48 proteins) (**f**). In **a**–**d** and **f**, the line indicates the diagonal of the plot (omitted for **e** as the data are reported in arbitrary units), demonstrating that APEX is generally correct as to magnitude of absolute abundance.

Further, we found good correlation ($R_s = 0.77$, $R^2 = 0.68$; **Supplementary Notes**) when the APEX-derived yeast protein abundances were compared with published measurements of absolute expression of the corresponding mRNAs. Comparisons of APEX-derived protein abundances to other properties revealed that the APEX-based protein levels were well correlated with codon bias ($R_s = 0.80$, $R^2 = 0.69$), codon adaptation indices ($R_s = 0.79$, $R^2 = 0.70$), the frequency of optimal codons ($R_s = 0.80$, $R^2 = 0.69$) and protein molecular weight ($R_s = –0.67$, $R^2 = 0.28$). Also, consistent with previous observations[20,24], APEX-derived protein abundances were not correlated with protein isoelectric point (pI; $R_s = 0.12$, $R^2 = 0.04$), aromaticity (frequency of aromatic amino acids; $R_s = –0.21$, $R^2 = 0.04$) and hydropathicity[25] ($R_s = 0.10$, $v^2 = 0.01$) (**Supplementary Notes**).

Validating APEX with differentially expressed proteins

To demonstrate the sensitivity to expression changes under different conditions, we compared APEX measurements of 626 proteins observed from yeast grown in either rich or minimal media. As expected, the changes in expression predominantly reflect differential expression of metabolite biosynthetic enzymes (**Fig. 3**). We require a statistical framework for deciding which proteins are significantly differentially expressed, adapting statistics developed for SAGE (Serial Analysis of Gene Expression) mRNA expression profiling[26,27].

Using this framework, 80 proteins were significantly induced ($Z > 2.58$, corresponding to 99% confidence, as defined in Methods) in minimal medium relative to rich medium. These minimal medium–induced proteins are statistically significantly enriched for proteins of metabolism ($P < 2 \times 10^{-14}$; 70 proteins) and biosynthesis ($P < 9 \times 10^{-14}$; 36 proteins)[28], as expected for cells forced to manufacture all amino acids and nucleotides from glucose. Twenty out of 23 significantly enriched Gene Ontology (GO) 'biological process' categories involve small molecule metabolism, for example, amino acid biosynthesis ($P < 10^{-14}$), or metabolism of the aspartate family ($P < 4 \times 10^{-14}$), glutamine family ($P < 2 \times 10^{-10}$), methionine ($P < 3 \times 10^{-9}$), sulfur amino acids ($P < 5 \times 10^{-8}$), branched chain family amino acids ($P < 2 \times 10^{-7}$), aromatic compounds ($P < 2 \times 10^{-6}$), glutamate ($P < 4 \times 10^{-7}$) and lysine, aminoadipic pathway ($P < 2 \times 10^{-6}$). More specifically, targets of transcription factor GCN4 (ref. 29) are significantly upregulated in

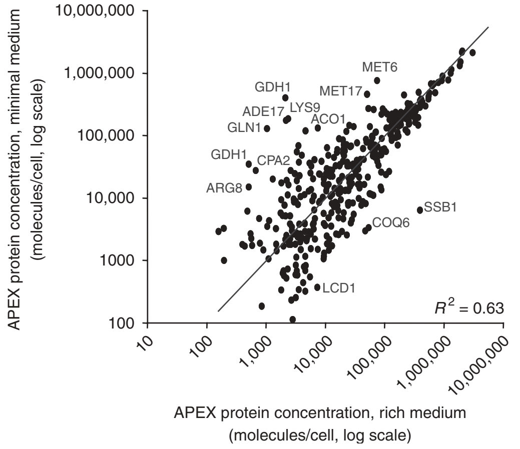

Figure 3 APEX is a sensitive measure of differential expression. APEX abundances are plotted for 265 proteins from yeast grown in rich and minimal media; an additional 361 proteins are seen in only one condition. Although the measurements generally agree ($R_s = 0.91$, $R^2 = 0.63$), 146 of the 626 proteins observed in either of the two media are significantly differentially expressed ($|Z| > 2.58$, 99% confidence). Proteins induced in minimal medium are predominantly involved in biosynthesis of amino acids and nucleotides, consistent with expectation. We obtain similar results in two additional analyses of differential protein expression (**Supplementary Notes** and **Supplementary Data 2** online).

minimal versus rich medium (15/50; $P < 4 \times 10^{-6}$), as expected for the amino acid starvation response. By contrast, no amino acid or nucleotide biosynthetic GO categories were enriched among the 66 rich medium–induced proteins ($Z < -2.58$). Instead, six GO categories for rapid growth were seen: protein biosynthesis ($P < 7 \times 10^{-7}$), macromolecule biosynthesis ($P < 9 \times 10^{-7}$), biosynthesis ($P < 5 \times 10^{-7}$), metabolism ($P < 4 \times 10^{-6}$), alcohol metabolism ($P < 3 \times 10^{-6}$) and cell growth and/or maintenance ($P < 2 \times 10^{-6}$). The significant expression changes ($|Z| > 2.58$) range from as little as ~1.3-fold for highly abundant proteins (CDC19, EFT2, ADH1, PGK1) to ~60- or even ~190-fold changes (SSB1, GDH1) (**Supplementary Notes** and **Supplementary Data 1**).

We obtained similar, biologically meaningful results when analyzing data from an independent yeast study and from mouse T-cell lymphoma nuclear and cytoplasmic fractions (**Supplementary Notes** and **Supplementary Data 2** online), confirming that MS-derived peptide counts, and thus APEX, can be used for sensitive protein quantification in any species.

The number of proteins per mRNA is log-normally distributed

To improve our understanding of proteome dynamics, we compared APEX-based protein abundances with absolute expression levels of the

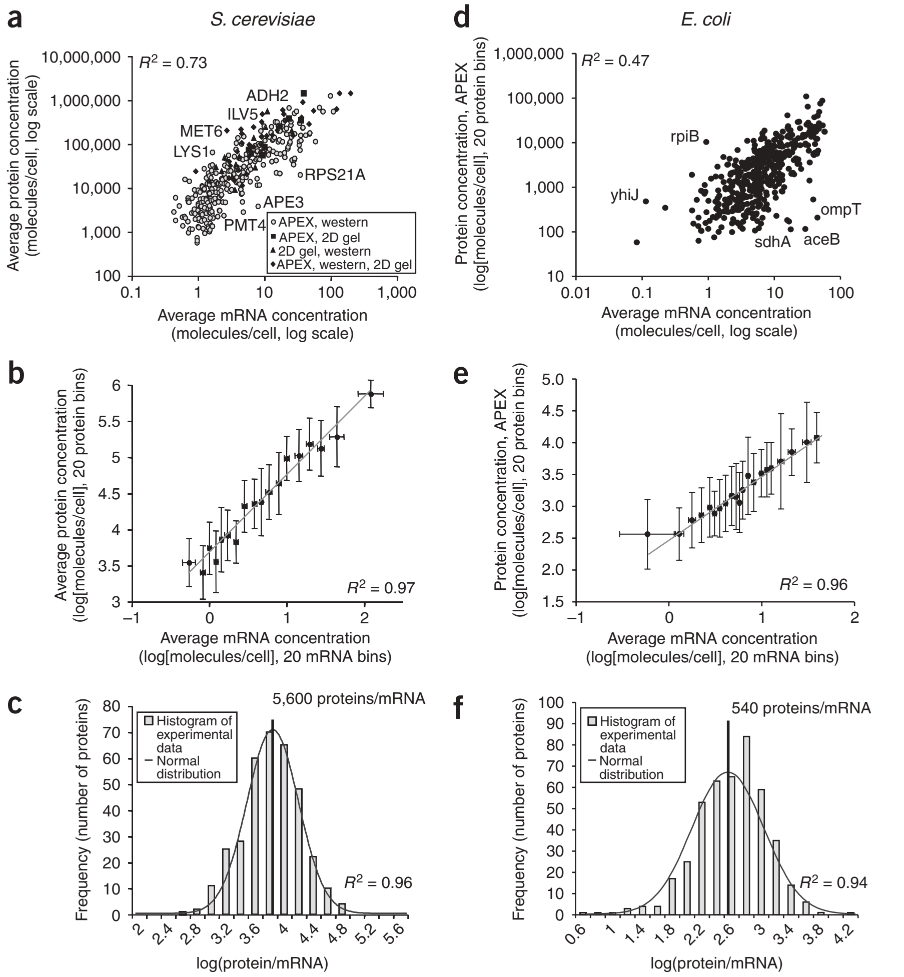

Figure 4 mRNA abundance explains over 70% of variance in yeast protein abundance and about half of variance in *E. coli* protein abundance. Abundances of 346 yeast proteins, calculated as the average of at least two of three independent proteomics measurements (APEX, 2D gel[18], western[20]), correlate very well ($R_s = 0.85$, $R^2 = 0.73$) with absolute mRNA abundances, calculated as the average of at least two of three independent mRNA expression measurements (SAGE[27], single-channel microarrays[33], dual-channel microarrays[34]). APEX-derived abundances of 437 *E. coli* proteins show moderate correlation ($R_s = 0.69$, $R^2 = 0.47$) with absolute mRNA abundance (average of at least two of three independent measurements[30–32]). (**a**,**d**) Plots of individual protein and mRNA abundances. (**b**,**e**) Binned measurements, calculated by rank-ordering mRNAs by expression level and calculating average protein and mRNA expression levels per bin of 20 genes, with error bars indicating ± 1 s.d. Lines correspond to the power law relationships between protein and mRNA abundances (yeast: $\log[protein] = 1.08 \times \log[mRNA] + 3.67$; *E. coli*: $\log[protein] = 0.96 \times 0.96 \times \log[mRNA] + 2.53$), well approximated by the equations $[protein] = 5{,}600 \times [mRNA]$ and $[protein] = 540 \times [mRNA]$, with the exact proportion varying between ~4,000–7,000 (yeast) and ~300–600 (*E. coli*), depending on method of calculation and on estimates of the total number of molecules/cell. The histograms of protein abundances per mRNA, plotted in (**c**,**f**), are well-fit as log-normal distributions.

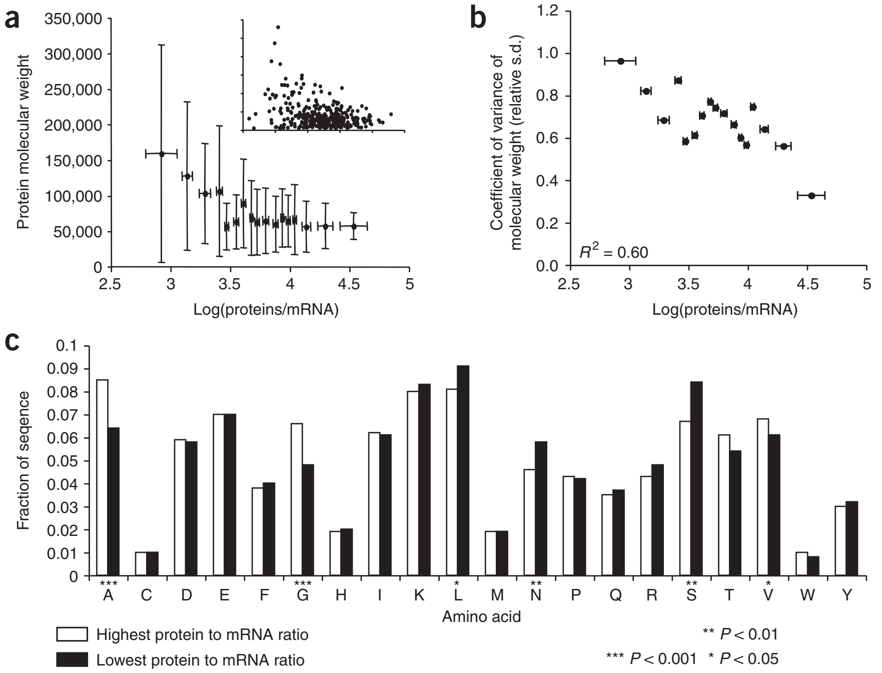

Figure 5 Yeast protein per mRNA ratios are correlated with amino acid frequencies and variance in molecular weight. (**a**) Protein molecular weight is poorly correlated with protein per mRNA levels, shown for bins of 20 proteins each as well as for individual proteins ($R^2 = 0.10$ and 0.76 for logarithmic relationships with raw (inset) and binned data, respectively). Much of this trend is accounted for by the variance in the data. (**b**) The coefficient of variance of molecular weight (100 x σ_{MW} / μ_{MW}, the s.d. of molecular weight divided by the average molecular weight within a bin of 20 proteins) is strongly anti-correlated to protein per mRNA levels ($R_s = -0.67$, $R^2 = 0.60$, for the logarithmic relationship plotted). We find $\sigma_{\mathrm{MW}}/\mu_{\mathrm{MW}}$ decreases as the number of proteins per mRNA increases, indicating that proteins with a wide range of molecular weights are found with low steady-state protein per mRNA levels, but that proteins with high ratios of protein to mRNA show considerably less variance and tend to be smaller in molecular weight, perhaps reflecting constraints of the translational apparatus to express large proteins at high protein to mRNA levels. (**c**) Biases in the amino acid frequencies in proteins with the 50 highest and 50 lowest levels of proteins per mRNA. Significantly different frequencies, calculated by *Z*-score test statistic, are indicated by asterisks. ***, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$.

corresponding mRNAs, measured by single channel DNA microarrays[30–33], SAGE[27] and microarrays using genomic DNA as reference[34].

We assume that each proteomics technology exhibits intrinsic measurement biases and stochastic error. For instance, fusion protein analysis can underestimate protein abundance because of destabilization by epitope tags, and western blotting signal may saturate at high concentrations. Shotgun proteomics may introduce bias through differential peptide isolation, solubilization, and ionization, although APEX is explicitly designed to correct the latter two biases. 2D gel quantification may be affected by signal saturation and multiple proteins per spot. Thus, averaging measurements from different technologies can improve estimates of steady-state protein levels.

Average yeast protein concentrations from at least two of the three technologies correlated extremely well with average mRNA expression levels from at least two of the three methods ($R_s = 0.85$, $R^2 = 0.73$; 346 proteins; **Fig. 4a**). This correlation was considerably higher than previously observed ($R^2 = 0.58$ (ref. 18)) for high-abundance proteins only; $R_s = 0.21$–0.58 (refs. 20, 35–37), suggesting that averaging across technologies does indeed remove technology-specific errors. We also observed a strong correlation if only APEX-based measures were compared with the average of at least two out of three yeast mRNA expression measures ($R_s = 0.77$, $R^2 = 0.68$; **Supplementary Notes**), and if we selected a high-confidence set of 58 proteins for which the relative s.d. for both average protein and average mRNA abundance is ≤0.5 ($R_s = 0.87$, $R^2 = 0.77$; **Supplementary Notes**). The correlation coefficient R^2 indicates the amount of variance in protein levels explained by mRNA levels. Given remaining measurement errors, the true correlation between mRNA and protein abundances is likely to be even higher (>73%) and the contribution of factors other than mRNA levels even lower (<27%).

Further, binning the measurements (**Fig. 4b**) indicates that the relationship between the numbers of protein and mRNA molecules has the form of a power law with an exponent close to one. Therefore, the distribution of individual proteins is also well fit by a linear relationship of [*protein*] = 5,600 × [*mRNA*], implying ~5,600 proteins present per mRNA, which is somewhat higher than previous estimates[20,35]. The logarithm of the ratio of proteins per mRNA is well modeled ($R^2 = 0.98$) by a normal distribution (**Fig. 4c**).

The *E. coli* proteome revealed a similar trend—APEX-based protein abundances correlated with average absolute mRNA abundances from at least two of three experiments[30–32] ($R_s = 0.69$, $R^2 = 0.47$; **Fig. 4d**). Again, the relationship is well described by a linear relationship [*protein*] = 540 × [*mRNA*], with ~300–600 proteins/mRNA (**Fig. 4e**), depending upon details of normalization (**Supplementary Notes** and **Supplementary Data 3**). As with yeast, the logarithm of the ratio of proteins per mRNA is well fit ($R^2 = 0.94$) by a normal distribution (**Fig. 4f**).

Further characteristics of yeast protein abundance per mRNA

The log-normal distribution of protein per mRNA suggests that a systematic search for significantly different protein/mRNA ratios using a *Z*-score should reveal post-transcriptional gene regulation. Using this approach in yeast, we identified 16 proteins with $|Z| > 1.96$ (95% confidence threshold; **Fig. 4**). These include protein per mRNA ratios either higher (ADH2, ALD6, ILV5, MET6, LYS1, BMH2) or lower (RPS21A, APE3, TOM1, TOS4, YLR422W, SPO14, PMT4, YCF1, YKR089C, TIF34) than expected. Examination of these proteins rationalizes their post-transcriptional regulation. For example, ADH2 mRNA associates substantially more with polysomes than RPS21A (**Supplementary Notes**). ADH2 protein levels are also catabolite repressed by multiple unexplained mechanisms[38]. ILV5 exhibits strong codon bias[39] and is regulated by leucine levels[40], suggesting possible post-transcriptional regulation similar to CPA1 (ref. 41). Some variation in protein per mRNA may arise from technical factors, for example, differences between strains, cell populations and laboratories. We expect that future systematic comparisons of mRNA and protein levels will identify additional examples of post-transcriptional regulation.

Having shown that the level of mRNA explains >70% of the yeast protein levels, we examined factors that might affect translation efficiency in order to explain the remaining variance. Surprisingly, codon bias (R_s = 0.80, R^2 = 0.69) and codon adaptation indices (R_s = 0.79, R^2 = 0.70) correlated well with overall protein levels, but not with protein per mRNA levels (**Supplementary Notes**). This suggests that codon choice is important on an evolutionary time scale, but not on the kinetic time scale of protein synthesis. Neither transcription nor translation rates explain additional protein per mRNA variance (**Supplementary Notes**). However, protein per mRNA levels correlated negatively with the variance of the proteins' molecular weights (R_s = –0.67, R^2 = 0.60; **Fig. 5a,b**). We also observed amino acid composition biases as a function of protein per mRNA levels (**Fig. 5c**).

DISCUSSION

APEX is a robust and rapid method to quantify absolute protein abundance, without requiring construction of fusion protein libraries, labeling or internal standards. Given the simplicity with which it can be used for large datasets, APEX may have important applications in biomarker discovery or serum profiling. The ability to associate abundance measurements with proteins from historical shotgun proteomics experiments emphasizes the importance of public deposition of proteomics data[42].

We illustrated the biological relevance of APEX-identified protein abundances by comparison with other measures, such as mRNA levels and analysis of differentially expressed proteins. Protein levels correlate well with mRNA abundance data obtained using SAGE and DNA microarrays. This suggests that >70% of yeast and about half of *E. coli* protein levels are determined by transcriptional regulation, with the protein per mRNA levels log-normally distributed. The weaker correlation in *E. coli* may stem from bacterial operon structure, in which genes are cotranscribed but often differentially translated.

Log-normal distributions typically arise from multiplicative random effects (the product of many small independent factors) when the growth over a time step is normally distributed and independent of the total size. Log-normal distributions occur frequently in natural systems. Here, this distribution implies that the logarithm of the amount of protein maintained per mRNA per time step can be modeled as a normally distributed random variable centered on ~4,000–7,000 proteins/mRNA for yeast and ~300–600 proteins/mRNA for *E. coli*. Note that proteins at the distribution tails are far from these values. Nonetheless, both eukaryotic and prokaryotic steady state protein levels appear to be primarily set on a per mRNA molecule basis, independent of total protein concentration.

We observe that proteins present at high copies per mRNA are of low molecular weight; proteins present at low copies per mRNA show no such constraint. These results may indicate a ceiling on the capacity of the cell to maintain high ratios of protein to mRNA levels for large proteins, and possibly a limit on the capacities of the translational or degradative apparatus. However, it is consistent with the findings that ribosome density on mRNAs decreases with increasing gene lengths[43], and that longer mRNAs have disproportionately lower ribosome initiation rates[44].

Lastly, availability of absolute protein abundance data enables a variety of future analyses. For example, we can use it to estimate protein degradation rates, which are hard to measure systematically[45]. Indeed, the least abundant yeast proteins have an increased occurrence of the PEST (proline, glutamic acid, serine, threonine) ubiquitinylation signal[46] compared to the most abundant proteins. However, this does not hold true for the protein per mRNA ratio (**Supplementary Notes**). Investigators have analyzed amino acids at the N-terminal end of proteins with respect to their influence on molecular stability and determined a set of stabilizing and a set of destabilizing amino acids, known as the N-end rule[47]. Although we find only minor bias with respect to amino acid occurrences at protein N termini (**Supplementary Notes**), we find stronger signal when we analyze whole sequences. The least abundant proteins per mRNA have significant protein-wide surpluses of serine, leucine and asparagine, whereas the most abundant proteins per mRNA have significant surpluses of valine, alanine and glycine (**Fig. 5c**). This is largely consistent with the N-end rule, suggesting that protein degradation contributes to protein per mRNA levels. By such analyses, we can describe and compare the influence of various factors, such as mRNA levels, growth conditions, molecular weight or sequence characteristics, on absolute protein expression levels and thus complete the picture of *in vivo* transcriptional and translational regulation.

METHODS

Derivation of the absolute protein expression index (APEX). If each protein *i* is present in c_i copies in the injected sample, then the expected fraction of the observed peptide pool accounted for by protein *i* of all injected proteins is:

$$\frac{c_i \times O_i}{\sum_{k=1}^{\#\text{injected proteins}} (c_k \times O_k)}, \text{ where } O_i = \sum_{j=1}^{\#\text{peptides from protein } i} p_{ij},$$

where O_i is the expected number of unique peptides observed for protein *i*, and p_{ij} is the probability of observing peptide *j* from protein *i* through the course of the MS experiment, which is a function of the peptide's ionization efficiency, solvent conditions, appropriate mass-to-charge ratio for analysis, and other factors.

Likewise, the fraction of the observed peptide mass spectra accounted for by protein *i* is:

$$\frac{n_i \times p_i}{\sum_{k=1}^{\#\text{observed proteins}} (n_k \times p_k)},$$

where n_i is the total number of redundant MS/MS scans observed from peptides of protein *i* through the course of the experiment, and p_i represents the probability of correctly identifying the protein. Assuming the maximum likelihood estimate of proportionality between these two fractions and solving for the concentration of protein *i* gives:

$$c_i \propto \left(\frac{n_i \times p_i}{O_i}\right) \times \left(\frac{\sum_{k=1}^{\#\text{injected proteins}} (c_k \times O_k)}{\sum_{k=1}^{\#\text{observed proteins}} (n_k \times p_k)}\right).$$

The second term captures the ratio of the total number of expected peptides to the total number of observed MS/MS spectra. As this term is constant for all proteins in a given experiment, it can be divided out to create a normalized protein score. Based on this derivation, we define the absolute protein expression index $APEX_i$ of protein *i* as:

$$APEX_i = \frac{n_i \times p_i}{O_i \times \sum_{k=1}^{\#\text{observed proteins}} \frac{n_k \times p_k}{O_k}} \times C,$$

where *C* is an estimate of the total concentration of protein molecules in the sample, approximately 5 × 10^7 molecules/cell for a typical yeast cell[18] and 2–3 × 10^6 molecules/cell for *E. coli*[48], serving to convert a normalized expression measure to an absolute protein number per cell. Under different experimental conditions, this number can, of course, be replaced by the measured total protein concentration.

Although both n_i and p_i are experimentally measured variables (here, calculated by ProteinProphet[49]), O_i is not directly available. This parameter provides the expected contribution of a protein to the pool of observed peptides, and it captures a broad set of trends, such as the total number of possible peptides generated from the proteins under the given experimental conditions (e.g., by tryptic digest) and the probability of each peptide ionizing and ultimately being analyzed by the mass spectrometer. A simple first approximation of O_i for a typical shotgun proteomics run can be calculated as the number of possible tryptic peptides of protein i that fall within the mass/charge window examined by the mass spectrometer. We obtained a more accurate estimate of O_i by training a classification algorithm to predict the observed tryptic peptides from a given protein based upon peptide length and amino acid composition. Although, in theory, all peptides from the same protein occur stoichiometrically, not all are observed; our classifier O_i captures trends leading to differential observation. In other words, O_i accounts for both protein- and peptide-specific sequence characteristics that bias observation in the mass spectrometer. O_i is a key feature of APEX that improves estimation of protein abundance by up to ~30% (see **Supplementary Notes**).

In this manner, we can correct the abundance for proteins with unusual amino acid sequences. For example, we observe the yeast protein FLO1 to have a particularly low O_i value: only 11 of the 64 predicted FLO1 tryptic peptides (or 17%) are deemed likely to be observed in an 2D HPLC-MS/MS experiment. The low value results from the unusual amino acid composition: FLO1 has an extremely high percentage of serine and threonine residues (41%). These residues are often sites of post-translational modifications, for example, glycosylation, which change peptide fragmentation patterns and therefore reduce the interpretability of the MS/MS spectra. Similarly, the protein RPL39 is only 51 residues long, but one-third of these are lysine and arginine residues, the sites of trypsin cleavage specificity. Thus, the majority of RPL39 tryptic peptides are too short for reliable detection. We expect to see only 3 of 38 tryptic peptides and the O_i value is correspondingly low. By contrast, the protein glyceraldehyde 3-phosphate dehydrogenase (TDH3) has one of the highest O_i values—39 of the 103 tryptic peptides (38%) are of a length and composition likely to be observed. Running the classifier on the set of all predicted tryptic peptides from yeast proteins provided an estimate of O_i for each protein. Normalizing the number of observed tryptic peptides by the prior expectation O_i improves estimates of each protein's abundance (**Supplementary Notes**).

A classifier for predicting observed peptides. To estimate O_i, we first derived a benchmark set of tryptic peptide amino acid sequences from the 40 most abundant (and therefore well sampled) proteins observed in a shotgun analysis of the yeast proteome. All training data are provided (**Supplementary Data 4** and **5** online). All possible tryptic peptides with at most two missed trypsin cleavages were predicted from these 40 proteins, keeping the 4,023 peptides in the molecular weight range 250–7,500 with at least three amino acids. Peptides ranged from 3 to 69 amino acids in length, with the average ~19 amino acids. Of these, 714 were observed in the shotgun proteomics experiment; the remaining 3,309 were not observed. For each peptide, a feature vector was constructed from the frequencies of each amino acid, the peptide length and molecular weight. Diverse classification algorithms, implemented in Weka Explorer v.3.4.4 (http://www.cs.waikato.ac.nz/ml/weka/), were tested for their performance in differentiating the 'observed'/'non-observed' peptides based upon these properties, including Bayesian classifiers, support vector machines, logistic regression, instance-based learners and decision trees. As simply guessing that all peptides are 'non-observed' is correct 82% of the time, cost-sensitive classifiers were used to balance the performance across the two peptide categories (**Supplementary Notes**).

The best performance (as judged by optimizing classifier precision and recall in tenfold cross-validated tests and requiring balanced performance on the two sets) was shown by a cost-sensitive classifier based upon bagging with a forest of random decision trees, with a final performance of 86% correct classifications on the cross-validated training set and true positive rates of 69% on observed peptides and 90% on non-observed peptides. Using the learned model, the classifier (**Supplementary Data 4** and **5**) was applied to the set of tryptic peptides from all yeast or *E. coli* proteins, predicting the likelihood of each peptide to be observed in a shotgun proteomics experiment. The value of O_i was calculated for each protein as the sum of probabilities for observing each tryptic peptide derived from that protein, and can be interpreted as the maximum number of unique peptides likely to be observed from the protein in a shotgun proteomics experiment conducted similarly to those described here. Values of O_i for all yeast and *E. coli* proteins are provided in **Supplementary Data 1** and **3**.

Analysis of differential protein expression using APEX. Given a shotgun proteomics experiment, we calculate the fraction f_i of interpreted peptides accounted for by protein i in the experiment as n_i/N, where n_i is the number of peptides from protein i, and N is the total number of interpreted peptides in that experiment. At typical values of N (~5,000–30,000), we find n_i to be well-approximated by a normal distribution of mean f_i and s.d. $\sqrt{f_i(1-f_i)}$ (see **Supplementary Notes** online for the test). We make the assumption that the probability of observing each peptide in the mass spectrometer, p_{ij}, is constant between two samples and can be ignored. We then calculate the test statistic for differential expression of a protein as:

$$Z = \frac{f_{i,1} - f_{i,2}}{\sqrt{f_{i,0}(1-f_{i,0})/N_1 + f_{i,0}(1-f_{i,0})/N_2}},$$

where the numerator represents the difference in sampled proportions of protein i in two shotgun proteomics experiments, and the denominator represents the standard error of the difference under the null hypothesis in which the two sampled proportions are drawn from the same underlying distribution with the overall proportion $f_{i,0} = (n_{i,1} + n_{i,2})/(N_1 + N_2)$.

Shotgun analysis of the yeast proteome. Yeast, *E. coli*, and mouse T-cell lymphoma cells, growth, protein extraction and proteolysis are described in the **Supplementary Notes** online. Tryptic peptide mixtures were separated by automated 2D HPLC. Chromatography was performed at 2 μl/min with all buffers acidified with 0.1% formic acid. Chromatography salt step fractions were eluted from a strong cation exchange column with a continuous 5% acetonitrile (ACN) background and 10-min salt bumps of 0, 20, 60 and 900 mM ammonium chloride. Each salt bump was eluted directly onto a reverse phase C18 column and washed free of salt. Reversed-phase chromatography was run in a 125-min gradient from 5% to 55% ACN, and then purged at 95% ACN. Peptides were analyzed online with electrospray ionization ion trap mass spectrometry using a ThermoFinnigan Surveyor/DecaXP+ instrument. In each MS spectrum, the five tallest individual peaks were fragmented by collision-induced dissociation (CID) with helium gas to produce MS/MS spectra. Gas phase fractionation was used to achieve maximum proteome coverage[50]: each tryptic peptide mixture was analyzed by three sequential 2D HPLC-MS/MS analyses, in each case examining a different mass/charge (*m/z*) range (300–650, 650–900 and 900–1500 *m/z*) for data-dependent precursor ion selection for CID; fragmentation data from the three runs were then combined for analysis by BioWorks (ThermoFinnegan). In total, 246,820 MS/MS scans were collected for yeast rich medium (YPD) data, 241,288 scans for yeast minimal medium (YMD) data, ~144,000 scans for *E. coli* and ~384,000 for mouse. The probability of observing each protein and the total number of observed peptides were calculated using ProteinProphet[49], selecting proteins above a 5% false discovery rate for protein identification threshold. Proteins identified for yeast, mouse and *E. coli* are provided in the **Supplementary Data 1**, **2** and **3**, respectively. Control protein mixtures were analyzed in a similar fashion (**Supplementary Notes**).

Quantitative properties of the yeast proteome. Protein-derived data, including aromaticity, hydrophobicity and codon adaptation index, frequency of optimal codons and codon usage, were downloaded from the Saccharomyces Genome Database (SGD). mRNA expression data were taken from SAGE data[27] as reported in the SGD database and from DNA microarray data[30–34]. Protein expression data were taken from western blot analyses[20], flow cytometry[21] and from 2D gel electrophoresis-based quantification[18,19].

All raw shotgun proteomics data are freely available for download from the Open Proteomics Database[42] at http://bioinformatics.icmb.utexas.edu/OPD under accession numbers opd00038_YEAST – opd00042_YEAST, opd00047_YEAST – opd00051_YEAST, opd00098_YEAST, opd00095_ECOLI – opd00097_ECOLI, and opd00087_MOUSE – opd00094_MOUSE. The training set and classifier used to assign O_i values to proteins are available as **Supplementary Data 4** and **5** and are suitable for use in assigning APEX-based abundances to proteins from organisms other than yeast.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank John Prince and Aleksey Nakorchevskiy for valuable discussion and help with computational analysis of peptide fragmentation. This work was supported by grants from the Welch (F-1515) and Packard Foundations, the National Science Foundation and National Institutes of Health. C.V. acknowledges support by the International Human Frontier of Science Program.

AUTHOR CONTRIBUTIONS

E.M.M. designed the project; P.L., R.W., X.Y. conducted the experiments; P.L., C.V., R.W., E.M.M. analyzed the data and wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Steen, H. & Pandey, A. Proteomics goes quantitative: measuring protein abundance. *Trends Biotechnol.* **20**, 361–364 (2002).
2. Washburn, M.P., Wolters, D. & Yates, J.R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
3. Hunt, D.F., Yates, J.R., III, Shabanowitz, J., Winston, S. & Hauer, C.R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* **83**, 6233–6237 (1986).
4. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. & Gygi, S.P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
5. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P. & Geromanos, S.J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5**, 144–156 (2006).
6. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596 (1999).
7. Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
8. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
9. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* **100**, 6940–6945 (2003).
10. Gao, J., Friedrichs, M.S., Dongre, A.R. & Opiteck, G.J. Guidelines for the routine application of the Peptide hits technique. *J. Am. Soc. Mass Spectrom.* **16**, 1231–1238 (2005).
11. Liu, H., Sadygov, R.G. & Yates, J.R., III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
12. States, D.J. *et al.* Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* **24**, 333–338 (2006).
13. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4**, 1265–1272 (2005).
14. Rappsilber, J., Ryder, U., Lamond, A.I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**, 1231–1245 (2002).
15. Craig, R., Cortens, J.P. & Beavis, R.C. The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **19**, 1844–1850 (2005).
16. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
17. Tang, H. *et al.* A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22**, e481–e488 (2006).
18. Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. & Garrels, J.I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).
19. Lopez-Campistrous, A. *et al.* Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol. Cell. Proteomics* **4**, 1205–1209 (2005).
20. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
21. Newman, J.R. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* (2006).
22. Fievet, J. *et al.* Assessing factors for reliable quantitative proteomics based on two-dimensional gel electrophoresis. *Proteomics* **4**, 1939–1949 (2004).
23. Thiele, D. *et al.* Elongation factor 1 alpha from *Saccharomyces cerevisiae*. Rapid large-scale purification and molecular characterization. *J. Biol. Chem.* **260**, 3084–3089 (1985).
24. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117–117.8 (2003).
25. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
26. Kal, A.J. *et al.* Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell* **10**, 1859–1872 (1999).
27. Velculescu, V.E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
28. Robinson, M.D., Grigull, J., Mohammad, N. & Hughes, T.R. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* **3**, 35–40 (2002).
29. Natarajan, K. *et al.* Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol. Cell. Biol.* **21**, 4347–4368 (2001).
30. Allen, T.E. *et al.* Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.* **185**, 6392–6399 (2003).
31. Corbin, R.W. *et al.* Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl. Acad. Sci. USA* **100**, 9232–9237 (2003).
32. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
33. Holstege, F.C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
34. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA* **99**, 5860–5865 (2002).
35. Beyer, A., Hollunder, J., Nasheuer, H.P. & Wilhelm, T. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell. Proteomics* **3**, 1083–1092 (2004).
36. Washburn, M.P. *et al.* Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA* **100**, 3107–3112 (2003).
37. Griffin, T.J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae. Mol. Cell. Proteomics* **1**, 323–333 (2002).
38. Sloan, J.S., Dombek, K.M. & Young, E.T. Post-translational regulation of Adr1 activity is mediated by its DNA binding domain. *J. Biol. Chem.* **274**, 37575–37582 (1999).
39. Petersen, J.G. & Holmberg, S. The ILV5 gene of *Saccharomyces cerevisiae* is highly expressed. *Nucleic Acids Res.* **14**, 9631–9651 (1986).
40. Holmberg, S. & Petersen, J.G. Regulation of isoleucine-valine biosynthesis in *Saccharomyces cerevisiae. Curr. Genet.* **13**, 207–217 (1988).
41. Werner, M., Feller, A., Messenguy, F. & Pierard, A. The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression. *Cell* **49**, 805–813 (1987).
42. Prince, J.T., Carlson, M.W., Wang, R., Lu, P. & Marcotte, E.M. The need for a public proteomics repository. *Nat. Biotechnol.* **22**, 471–472 (2004).
43. Arava, Y. *et al.* Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA* **100**, 3889–3894 (2003).
44. Allemeersch, J. *et al.* Benchmarking the CATMA microarray. A novel tool for *Arabidopsis* transcriptome analysis. *Plant Physiol.* **137**, 588–601 (2005).
45. Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E.K. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. USA* **103**, 13004–13009 (2006).
46. Rogers, S., Wells, R. & Rechsteiner, M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* **234**, 364–368 (1986).
47. Bachmair, A., Finley, D. & Varshavsky, A. In vivo half-life of a protein is a function of its amino-terminal residue. *Science* **234**, 179–186 (1986).
48. Neidhardt, F.C. & Umbarger, H.E. in *Escherichia coli and Salmonella Typhimurium: Cellular and Molecular Biology*, edn. 2, vol. 1 (eds. Neidhardt, F.C. *et al.*) 13–16 (ASM Press, Washington, DC, 1996).
49. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
50. Yi, E.C. *et al.* Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **23**, 3205–3216 (2002).