Supplementary Material:

An estimate of relative contributions of transcriptional and translational regulation by absolute protein expression profiling

Peng Lu, Christine Vogel, Rong Wang, Xin Yao, Edward M. Marcotte

Abbreviations

APEX – Absolute Protein EXpression index; FDR – false detection rate; GFP – Green Fluorescent Protein; R^2 – squared Pearson correlation coefficient; R_s – Spearman correlation coefficient; SAGE – serial analysis of gene expression; YMD – yeast minimal medium; YPD – yeast rich medium (peptone/dextrose); 2D gel – two-dimensional gel electrophoresis

Contents

Mass-spectrometry based protein quantitation (e.g. spectral counting) is one of the most accurate and easiest methods for high-throughput protein quantitation (http://www.abrf.org/ResearchGroups/Proteomics/Studies/ABRF_Presentation_2006.pdf) . Our paper describes a novel method to derive absolute protein expression (APEX) measurements from MS-based data, and below we describe details of the method.

This document provides further descriptions of technical aspects of our method (**Section 1**) and details of its application to the yeast, *E.coli* and mouse proteome (**Sections 2** and **3**). We demonstrate that APEX is a simple method that can be used in standard MS experiments (**Sections 1.1., 1.3.)**, is highly reproducible (**Sections 1.2., 1.8**.), comparable or better than other approaches (**Sections 1.4., 1.9.**), accurate (**Section 1.5., 1.8.**), robust to different experimental conditions (**Section 1.8.**), and scalable to any organism (**Sections 1.7.**, **2.3., 2.4.**)

1. Establishing and validating APEX

1.1. Confidence in peptide and protein identification

False positive rates (FPR) or false discovery rates (FDR) are an important issue in proteomics. We measure our FDRs using the algorithms in PeptideProphet and ProteinProphet[1]. We have independently checked these values for a number of datasets by performing spectral database analysis against a shuffled version of the proteins (*not shown*).

ProteinProphet[1] provides an error model for the estimated false discovery rate (FDR) of mass spectrometry (MS) protein identification (**Figure S1**). A 5% (10%) FDR requires a minimum ProteinProphet score p_i of 0.78 (0.63) both in minimal (YMD) and rich (YPD) medium, resulting in 454 (555) proteins in YPD and 437 (550) in YMD. In *E. coli*, a minimum ProteinProphet score p_i of 0.60 (5% FDR) results in 504 identified proteins. The FDR calculation is designed to separate true from false positive identifications, regardless of the source of error, i.e. methionine oxidation, other post-translational modifications, or errors in the mass measurements, such as inefficient peptide fragmentation. We explicitly include a quantitative measure of confidence for each protein's identification (p_i) as a parameter in APEX, thus accounting for the variety of errors.

The main text discusses the results obtained from using the high-confidence set of 454 proteins (5% FDR). Below, we also provide the results for the set of 555 proteins (10% FDR). In brief, using 10% FDR ($p_i \geq 0.63$) as cutoff produces very similar correlations between protein abundance and other protein properties. The average of at least two of three protein abundance measurements correlate well with the average of at least two of three different mRNA measurements (R_s=0.84, R^2=0.73 in log-log plot, **Figure S2, A**), which is identical to what is observed with the 5% FDR APEX set (R_S=0.85, R^2=0.73 in log-log plot, Figure 4A main text). Notably, the correlation is very similar when we compare APEX alone against an average of two of three mRNA measurements (R_s=0.76, R^2=0.65 in log-log plot, **Figure S2B**). The average number of proteins per mRNA ranges from ~4,300 (log-log) to ~6,200 (linear-linear).

Further, the codon adaptation index (CAI) correlates with protein abundance at 10% FDR (R_S=0.75, R^2=0.68; **Figure S2C**), similar to what we show for the proteins of 5%FDR (R_S=0.79, R^2=0.70, see main manuscript and **Figure S10**). The APEX-derived protein abundances with 10% FDR have moderate correlation with Western blot and 2D-gel electrophoresis data (R_S=0.60, R^2=0.34 and R_S=0.77, R^2=0.48, respectively, in log-log plot), again similar to what we show for the 5% FDR set (see main text; R_S=0.61, R^2=0.34 and R_S=0.80, R^2=0.52).

Thus, in general, the dataset with higher false discovery rate (10% FDR) produces >100 (22%) more proteins and shows the same trends as the data with 5% FDR.

Note that before a peptide can be associated with a particular protein, as with ProteinProphet[1], each mass spectrum has to be associated with a particular peptide. The accuracy of these peptide assignments is estimated by PeptideProphet[1]**.** In our analysis, PeptideProphet scores were first calculated for each peptide, using the default cutoff of 0.2 which corresponds to <26% FDR for peptide identification. This FDR represents the upper bound of the false positive identification rates – the actual FDR is lower, as ProteinProphet re-calculates the FDR for each peptide conditioned on the positive identification of the protein by other peptides[1]. We also tested a more stringent cutoff for peptide identification, with very similar results: the total number of peptides observed with a peptide score >0.2 correlates with the total number of peptides observed with a peptide score >0.5 (corresponding to 10% FDR) with R^2=0.997. Further, the peptide scores are used by ProteinProphet to calculate protein scores. For this reason, we used only the protein scores, but not the peptide scores for APEX calculations to avoid double-counting.

1.2. Reproducibility

For reliable APEX and Z-score calculations as described in the main manuscript and here, it is important that the probability of observing each peptide in the mass spectrometer is constant between two samples. Figure 2 in our manuscript demonstrates the strong correlation between APEX values from two different replicate sets of experiments. Below we describe further reproducibility tests.

In order to increase protein coverage from a MudPIT experiment, we injected each biological sample several times into the mass spectrometer, and pooled these technical replicates. APEX is highly reproducible between the different injections.

The paper (main text) describes the correlation in abundance for proteins observed in two sets, pooling data from three injections each originating from one experiment (YMD medium). The log-transformed data is correlated with an R^2=0.88, indicatory of high reproducibility.

Figure S3 shows a similar plot using replicate data sets pooled from two and three injections, with cells grown in YPD medium. The two data sets are correlated with an R^2=0.95 for the non-transformed data (one outlier) and R^2=0.75 for the log-transformed data. The proteins were identified with a 5% FDR, resulting in a $p_i \geq 0.78$ and $p_i \geq 0.74$

cutoff for set 1 and 2, respectively. The lower part of **Figure S3** shows the error models for protein identification.

The reproducibility also holds true when comparing data from the five individual injections (from the YPD experiment), as shown in **Table S1** below. Note that the table compares the number of peptides observed per protein (left side of columns), and the calculated APEX values (right side), providing R^2-values as measure of correlation. The upper right half of the matrix compares the non-transformed numbers, the lower left compares the log-transformed numbers. The overall reproducibility for the peptide counts (APEX) is 0.85±0.03 (0.89±0.03) for the log-transformed data, and 0.95±0.03 (0.92±0.03) for the linear data.

In **Section 1.8.,** we discuss the high reproducibility between different experiments using a synthetic mixture of ten proteins of known concentration without and with added cellular lysate (R^2=0.997 on linear-linear, **Figure S24D**; R^2=0.98 on log-log, **Figure 2B**).

Table S1. Reproducibility: Correlation coefficients (R^2-values) for replicate experiments.

R^2	YPD1		YPD2		YPD3		YPD4		YPD5		
	Pept	APEX	Pept	APEX	Pept	APEX	Pept	APEX	Pept	APEX	
YPD1	-		0.98	*0.97*	0.95	*0.93*	0.97	*0.94*	0.95	*0.90*	
YPD2	0.86	*0.88*	-		0.98	0.96	0.94	*0.92*	0.96	*0.92*	
YPD3	0.82	*0.82*	0.85	*0.87*	-		0.88	*0.86*	0.98	*0.95*	**linear**
YPD4	0.86	*0.87*	0.81	*0.86*	0.88	*0.90*	-		0.91	*0.88*	
YPD5	0.81	*0.82*	0.88	*0.89*	0.85	*0.88*	0.87	*0.90*	-		
	log-transformed										

1.3. Training the classifier for learning O_i

The classifier was trained on a set of 4,023 yeast tryptic peptides using machine learning techniques. For each peptide, a vector of 22 features was constructed from the peptide's length, molecular weight, and 20 amino acid frequencies (see example below). In addition, each peptide was flagged as observed (*Obs*) or not (*Not*), considering peptides with up to 2 missed tryptic cleavages (**Supplemental Dataset S4**).

EXAMPLE: 3 sample peptide feature vectors from protein RPS21B (Genbank Accession 6322325):

peptide MENDK: 5, 636.266, 0.000, 0.000, 0.200, 0.200, 0.000, 0.000, 0.000, 0.000, 0.200, 0.000, 0.200, 0.200, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, *Not*

peptide MENDKGQLVELYVPR: 15, 1790.91, 0.000, 0.000, 0.067, 0.133, 0.000, 0.067, 0.000, 0.000, 0.067, 0.133, 0.067, 0.067, 0.067, 0.067, 0.067, 0.000, 0.000, 0.133, 0.000, 0.067, *Not*

peptide GQLVELYVPR: 10, 1173.66, 0.000, 0.000, 0.000, 0.100, 0.000, 0.100, 0.000, 0.000, 0.000, 0.200, 0.000, 0.000, 0.100, 0.100, 0.100, 0.000, 0.000, 0.200, 0.000, 0.100, *Obs*

During training, more than 30 classifiers or variants (i.e, classifiers with bagging) were tested for their performance in separating the observed from non-observed peptides, testing each with 10-fold cross validation and evaluating performance with recall-precision analysis. All classifiers were implemented in the Waikato Environment for Knowledge Analysis (Weka) version 3.4.4, available from http://www.cs.waikato.ac.nz/ml/weka/.

Decision tree-based methods generally out-performed other approaches. Tested classifiers included naive Bayes, logistic regression, J48 trees, Kstar, AD trees, decision stumps, LMT logistic model trees, NB trees, conjunctive rules, PART decision lists, Ridor (Ripple down rule learner), nearest neighbor generator rules, and OneR minimum error attribute classifiers. The best-performing single rule classifiers were based upon length ((length <= 7.5) ==> *Not*) or molecular weight (878.421 > MW or MW > 3675.815 ==> *Not*).

The best-performing classifier, judged by requiring balanced performance in recall-precision analysis on both *Obs*/*Not* peptides, was one based upon bagging with a forest of random decisions trees. The performance of the classifier is shown in **Table S2**, followed by the three next best classifiers.

Table S2. Classifier performance

	Bagging with Random Forest		**Ridor**		**Random Forest**		**Regression with M5P trees**	
Total correct	85.8%		84.4%		84.3%		80.0%	
Total incorrect	14.2%		15.6%		15.7%		20.0%	
TP *'Not'*	0.90		0.91		0.89		0.81	
FP *'Not'*	0.31		0.44		0.36		0.26	
TP *'Obs'*	0.69		0.56		0.64		0.75	
FP *'Obs'*	0.10		0.10		0.11		0.19	
Classified as:	***Not***	***Obs***	***Not***	***Obs***	***Not***	***Obs***	***Not***	***Obs***
Confusion matrix:								
Not	2961	348	2995	314	2932	377	2685	624
Obs	222	492	314	400	256	458	182	532

(TP – true positive; FP – false positive)

Each was trained using re-weighted training instances (cost-sensitive training) to balance performance on the two peptide classes. Note that the classifier based upon regression with M5P trees had a higher TP rate on the observed peptides (0.75 vs. 0.69); however, this classifier showed a large number of not-observed peptides misclassified as observed, as evident in the confusion matrix, leading to a lower overall performance (80% correctly classified instances, versus 85.8%).

Additional performance statistics for the top classifier (Bagging with Random Forest):

Table S3. Performance statistics

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Not	0.895	0.311	0.930	0.895	0.912
Obs	0.689	0.105	0.586	0.689	0.633

(TP – true positive; FP – false positive)

The best-performing classifier used cost-sensitive bagging (each bag 100% of the training set size[2]) to improve performance of the base classifier, which was a random forest of 10 decision trees[2], each constructed while considering 5 random features.

The top classifier's performance can be seen most clearly by plotting the frequency histograms observed and non-observed peptides from the 4,023 peptide set, as shown in **Figure S4A,B**. The classifier and all training data are available as supplemental data to this paper, and is distributed via the Open Proteomics Database[3].

Further, we assume that the probability of observing a peptide is constant from condition to condition. As the chemical identities of the peptides are identical, this is possible, with the exception of a mechanism such as differential ion suppression, where 'neighboring' peptides in the separation differentially affect the ionization of the peptide of interest. However, we note that this is precisely the factor that the MudPIT strategy is designed to overcome, and analyses of ion suppression in complex mixtures[4, 5] demonstrate that this issue diminishes with increasing stages of up-front sample fractionation. Thus, this is experimentally addressable and not a problem with APEX itself. As we expect this tendency to be distributed more or less randomly across various peptides associated with a given protein, the total count of peptides observed from a protein will be reasonably robust, see also **Section 1.6** for further discussion. **Figure S24** confirms that this is the case, with reproducible APEX measurements on a 10-protein control set in the presence and absence of yeast cell lysate.

1.4. Performance of the classifier for learning O_i

Probabilities of observing peptides in a shotgun proteomics experiment are learned from >4,000 peptides derived from 40 well-sampled proteins (**Figure S4**). Classifier-assigned probabilities are plotted for 714 true positive peptides (peptides actually observed from 40 well-sampled proteins) and 3,309 true negative peptides (peptides not observed from the same proteins). Thus, only 18% of peptides are expected to be observed in general, reflected by the smaller histogram (black bars, **Figure S4B**). Thus, it is not surprising

that, for example for yeast peptides, the expected number of peptides O_i is generally smaller than the number of unique peptides (**Figure S4C**).

The classifier, trained using 10-fold cross validation, effectively separates these from the non-observed peptides. A probability threshold of 0.5 gives a 69% true positive rate on observed peptides and 90% true positive rate on non-observed peptides, for a total accuracy of 86%.

Applying a prior expectation of observing peptides from a protein *i, i.e.* O_i, substantially improves our estimates of protein abundance, as is shown in **Table S4** below. **Table S4** lists R^2 values of the correlation between APEX calculated with O_i values and without (O_i=1) and protein and mRNA abundance data. Protein abundance estimates were obtained from Western blot[6] and 2D gel[7] analysis; estimates of mRNA abundance are averages of at least two of three measurements (Affymetrix microarrays[8], SAGE[9], dual channel microarrays with genomic DNA reference[10]). The prior expectation of observing peptides from protein *i*, O_i, is part of APEX' core.

When O_i is included in APEX calculations of protein abundance, the correlation with other estimates of protein or mRNA abundance improves by 5-33% (**Table S4**).

Tables S4. O_i significantly improves estimates of protein abundance. R^2-values of correlations of APEX and peptide counting.

Yeast	**Log(APEX) R^2**	**Log(APEX, O_i=1) R^2**
Log[protein; Western]	0.34	0.22
Log[protein; 2D-gel]	0.52	0.33
Log[protein; flow cytometry]	0.49	0.42
Log[mRNA; average]	0.68	0.47
E. coli	**Log(APEX) R^2**	**Log(APEX, O_i=1) R^2**
Log[protein; 2D-gel]	0.21	0.16
Log[mRNA; average]	0.47	0.34
Synthetic protein mixtures	**Log(APEX) R^2**	**Log(APEX, O_i=1) R^2**
5-protein mixture without lysate	0.85	0.52
10-protein mixture without lysate	0.89	0.79
10-protein mixture with lysate 1:10	0.82	0.58
10-protein mixture with lysate 1:1	0.84	0.77

1.5. Validation of APEX-based yeast protein abundances with other large-scale measurements

In the main text, we focus on the analysis of APEX-based protein abundances of yeast grown in rich (YPD) and minimal (YMD) medium. *Saccharomyces cerevisiae* DBY8724 cells[11] (*MATa GAL2 ura3 bar1::URA3*) were grown with aeration at 30 °C until O.D. 0.6-0.8 in either rich YPD medium (2 % yeast extract, 1 % peptone, 2 % glucose) or synthetic YMD minimal medium (0.7 % yeast nitrogen base without amino acids and ammonium sulfate (DIFCO Bacto), 2 % glucose, 5 g/L ammonium sulfate, and 20 mg/L uracil), pelleted, washed, resuspended in buffer (20 mM Tris-HCl, 100 mM NaCl) containing 1% protease inhibitor cocktail (Calbiochem, CA) and lysed with glass beads. Soluble protein extracts were diluted to 4 mg/ml into digestion buffer (50 mM Tris HCL pH 8.0, 1.0 M Urea, 2.0 mM $CaCl_2$), denatured at 95 °C for 10 min, and digested with sequencing grade trypsin (Sigma, MO) at 37 °C for ~20 hours. The trypsin digested protein extract was then further analyzed in tandem LC/LC/MS/MS as described in the main manuscript, and the APEX-based protein abundances calculated from the observed spectra.

Yeast APEX-based protein abundances are consistent with estimates of protein abundance from other experiments for yeast growing in rich medium, i.e. high-throughput Western blot[6], 2D gel[7] and flow cytometry analysis of GFP-tagged protein[12]. In fact, APEX correlates better with each of these datasets (**Figure 2**, main text) than the three other data sets correlate with each other (**Figure S5**, Western-2D: R^2=0.05; Western-GFP: R^2=0.43; 2D-GFP: R^2=0.28).

The GFP-tagged protein analysis is also available for yeast growing in minimal medium[12]. Again, APEX compares well to this data set, with R_S=0.64 (R^2=0.41, **Figure S6A**) as well as for the subset of 102 proteins that are YMD-specific (R^2=0.40, **Figure S6B**). However, in contrast to the method by Newman et al., MS-based methods (like APEX) do not require labor-intensive establishment of GFP-tagged gene libraries, but can be conducted with standard mass spectrometry.

In general, each of the approaches, Western blotting, 2D gels, GFP-tagging, and MS-based technologies, have their own advantages and disadvantages. For example, high-throughput Western blot or the GFP-tag based approach are currently applicable to lower abundance proteins, however, they only apply to yeast, which is the only organism with a TAP- and GFP-tagged collection of strains available. Also, to measure the ~3,800 proteins required growing up ~3,800 different cultures of yeast cells, a non-trivial experiment.

By contrast, APEX can be performed routinely for fewer proteins in a single day's experiment. 2D gels offer a similar promise, but in practice are difficult to run and require extensive automated sample preparation and mass spectrometry in order to identify fewer proteins than we present with APEX. Further, when using APEX, the number of identified proteins and sensitivity to lower concentrations can be increased by future experiments using more sensitive MS methods, e.g. new high-resolution mass spectrometry proteomics technology (OrbiTrap). Thus, to measure many proteins without

expensive labeling, and to do it in any organism other than yeast, APEX is a reasonable choice.

1.6. Normal distribution of the number of peptides per protein

We use Z-scores to estimate the significance of differentially expressed proteins as measured by APEX. To do so, the probability f_i to observe a protein in all MS/MS spectra must be normally distributed. **Figure S7** shows that this is the case.

The number of MS/MS spectra n_i associated with a given protein (out of all MS/MS spectra N) can be seen as Bernoulli trial in which an MS/MS spectrum is either associated with a particular protein i or not. As the total number of spectra N is different for each experiment, it is more useful to compare the fraction $f_i = n_i / N$. With N being very large, the binomial distribution can be approximated by a normal distribution. As each experiment comprised at most six injections (replicates), only 6 or fewer data points are available for each protein to test normality. We tested normality for the ten most and least abundant proteins (**Figure S7**). While there are two exceptions, FBA1 and ILV2, 18 out of the 20 tests confirm normality ($p>0.05$ under the Shapiro-Wilk test).

1.7. Analysis of *E. coli* protein abundance data.

As further validation, we conducted similar MS analysis in a different organism, *Escherichia coli*, grown in minimal medium. Prior to these analyses, we calculated *Oi* values for all *E. coli* proteins using the same classifier and same procedures as described above (**Supplementary Dataset S5**). We estimated *E. coli* protein abundances using APEX and compared these with protein abundance data from 2D-gel electrophoresis experiments[13], three mRNA expression dataset[14-16] and information on codon bias.

Wild type *E. coli* strain K12 N3433 was grown aerobically in MOPS, a minimal medium supplemented with Glucose (0.4%)[17]. An overnight culture growing in exponential phase (i.e., 0.5 OD600) was used to inoculate MOPS media and grown to early logarithmic phase in 37°C (i.e., 0.3 OD600). Cells were harvested by centrifugation at 4,000g for 30 minutes at 4°C and washed three times with cold PBS buffer (0.137 mM NaCl, 2.7 mM KCl, 8.0 mM Na_2HPO_4, 1.5 mM KH_2PO_4, pH 7.5). These conditions are identical to those used to produce the 2D-gel data[13]. Tryptic peptides were prepared and analyzed with LC/LC/MS/MS as described in the main manuscript.

We identified 504 proteins (FDR<5%), and calculated protein abundances using APEX with *E. coli* protein-specific O_i values (**Supplementary Dataset S2**) normalizing by an estimated total number of 2.6×10^6 molecules in the cytoplasm (http://redpoll.pharmacy.ualberta.ca/CCDB/cgi-bin/STAT_NEW.cgi, and references therein). The least abundant proteins are *lhr* and *gph* with fewer than 35 copies/cell, the most abundant proteins are *tufA* and *tufB* with >70,000 copies/cell.

Similar to what we observed in yeast, the correlation with independently published protein abundance data using 2D-gel electrophoresis[13] is only moderate ($R^2=0.21$, $R_S=0.47$, N=210, **Figure S8A**). However, this result is unlikely due to strong errors

exclusively in APEX measurements alone as APEX outperforms the 2D-gel data in the analyses described below.

When comparing protein abundance measurements with mRNA abundance data from three individual studies using similar experimental conditions[14-16], we observe the same trends as for yeast, although with lower correlation coefficients. Protein and mRNA abundances correlate in a power-law relationship with ~300 to ~600 protein/mRNA. The correlation is better when comparing the average of three mRNA data sets vs. APEX alone (R^2=0.47, R_S=0.69, N=437, **Figure 4** main text, **Figure S8B**), than when using the 2D-gel data (R^2=0.21, **Figure S8C**) or the average of 2D-gel data and APEX (R^2=0.35).

When correlating the mRNA abundance from individual datasets[14-16] with APEX, the correlation coefficients are similar to that of the average mRNA abundance (see **Supplementary Dataset S2**): in contrast to the yeast data, averaging over several data sets does not improve the observed trends. However, as discussed in the main text, the protein/mRNA ratios are log-normally distributed around their mean of 540 proteins/mRNA suggesting a multiplicative error model (**Figure 4**).

We also investigated the relationship between protein abundance and codon adaptation index (CAI)[18] which often correlate with each other. Surprisingly, for APEX alone, the 2D-gel data alone and the average of the mRNA data sets, the correlation coefficients are very low (R^2=0.33, R^2=0.30, R^2=0.32, respectively; **Figure S9A**). Similarly, while molecular weight and expression levels often correlate inversely, there is only a very weak such trend when comparing APEX-based protein abundance with gene length (R^2=0.18) or the average mRNA abundance with gene length (R^2=0.01)(**Figure S9B**).

Overall, while the general trends are the same, the relationship between protein and mRNA measures appears weaker for *E. coli* than for yeast. A definitive answer should await *E. coli* protein abundance data from other laboratories and techniques.

1.8. Analysis of synthetic protein mixtures

As validation of APEX' ability to estimate accurate protein concentrations, we analyzed several samples of purified proteins mixed in known concentrations. The proteins were trypsinized and analyzed by LC/MS/MS as described in the main text but without SCX fractionation. O_i values were calculated with the same classifier described previously, and APEX abundances were estimated, normalizing for the measured total protein concentration.

The first sample contained 5 purified proteins from organisms other than *E.coli* (chicken egg white lysozyme, 277 pmol; bovine erythrocyte carbonic anhydrase, 77 pmol; bovine serum albumin, 16 pmol; horse heart myoglobin, 4 pmol; bovine liver glutamate dehydrogenase, 1 pmol). While the MS/MS sampling was relatively shallow in this single-injection experiment (54 total observations of peptides), APEX abundances are within just 2.5-fold of the correct concentration in all cases but one (median 2.3-fold difference)(**Figure S23**) and show a significant improvement over the case with flat priors (O_i = 1)(R^2=0.90 vs. R^2=0.48). The one protein with substantially higher difference between expected and observed concentration is glutamate dehydrogenase which was added in extremely low concentration (1 pmol).

The second sample contained 10 purified proteins mixed in known concentrations from organisms other than yeast (chicken egg white ovalbumin 100 pmol; chicken egg white conalbumin 5 pmol; bovine liver catalase 20 pmol; bovine milk lactoperoxidase 2 pmol; bovine milk β-Casein 10 pmol; bovine milk β-lactoglobulin 500 pmol; hourse heart myoglobin 60 pmol; bovine erythrocyte carbonic anhydrase 200 pmol; human apo-transferrin 1 pmol; bovine serum albumin 40 pmol). In an extension of the second experiment, we also added the 10-protein mixture to yeast protein cell lysate in concentration ratios 1:1 and 1:10. The proteins were trypsinized and analyzed by LC/MS/MS as described in the main text. The 1735 MS/MS spectra obtained from a three-injection experiment were identified by comparison to a database of *Saccharomyces cerevisiae* proteins plus the 10 protein sequences. O_i values were calculated with the same classifier described previously, and APEX abundances were estimated, normalizing by the calculated total protein concentration.

APEX performs very well and is robust to different experimental conditions (**Figure S24, Table S6, 7**). Independent of the concentration of yeast cell lysate, APEX-based protein concentrations correlate with the known protein concentrations with R^2>0.82 at both linear and logarithmic scale (**Figure 2A, Figure S24A-C**). The median fold change ranged from 1.8 to 2.2 (mean 1.9-2.4).

APEX is also highly reproducible: the estimates of protein concentrations in different experiments (no lysate versus lysate 1:1) are virtually identical (**Figure 2B** and **Figure S24D**). The known concentrations have been added according to what we expected to be natural concentrations, with as little as 1pmol. In addition, there is no indication in **Figure S24** that APEX measurements saturate at higher concentrations (500pmol), thus we expect our method to be easily scalable to more than 3 orders of magnitude.

1.9. Comparison with existing methods of protein quantitation

APEX-based protein concentrations are more accurate than those derived from existing methods, *i.e*. Protein Abundance Index (PAI)[19] and the exponentially modified PAI (emPAI)[20]. **Table S5** summarizes the relationship between parameters used in the different methods. PAI is calculated as *n_obs_uniq/n_exp_uniq* where *n_obs_uniq* and *n_exp_uniq* are the number of experimentally observed (within the respective mass range) or the theoretically observable (expected) peptides per protein, respectively. emPAI is a modified version of PAI, calculated as $emPAI = 10^{PAI} - 1$ (**Table S5**). While PAI and emPAI employ unique peptide counts, APEX employs redundant peptide counts, allowing for an intrinsically larger dynamic range that does not saturate at 100% coverage of unique peptides. Also, APEX uses the classifier O_i,, which is an estimate of the number of expected unique peptides as calculated from their probability of being observed (see **Section 1.3, 1.4**).

Table S5. Relationship between different parameters in MS experiments.

	Observed peptide spectra from protein *i*	**Expected peptide spectra from protein *i***
Unique	$n\_obs\_uniq_i$ (used for PAI)	$n\_exp\_uniq_i$ (used for PAI) O_i (corrected $n\_exp\_uniq_i$; used for APEX)
Redundant	n_i (used for APEX)	-

We compared APEX' performance to that of PAI and emPAI, using the synthetic mix of 10 proteins described in **Section 1.8**. The correlation coefficients (R^2; linear-linear) between the true concentrations and the concentrations calculated by the three methods are summarized in **Table S6**.

Table S6. Comparison of APEX with PAI and emPAI.

	APEX R^2 **(linear-linear)**	**PAI** R^2 **(linear-linear)**	**emPAI** R^2 **(linear-linear)**
Synthetic 10 protein mix – without cell lysate	0.88	0.41	0.42
Synthetic 10 protein mix – with cell lysate 1:10	0.88	0.34	0.34
Synthetic 10 protein mix – with cell lysate 1:1	0.84	0.38	0.38

Table S7 and **Figure S25** provide a more detailed comparison of concentrations estimated by APEX, PAI and emPAI in the protein mixture without cell lysate added. The other two methods, PAI[19] and emPAI[20], perform well in their estimation of protein abundance, but show saturation at higher protein concentrations. The mean, median and maximum fold difference to the true (injected) protein concentrations is lower in APEX than in the other two methods.

Table S7. Example comparison (10 protein mix without lysate) of APEX, PAI and emPAI

	Injected concentration (pmol)	APEX (pmol)	PAI (pmol)	emPAI (pmol)
Apotransferrin	1.0	4.2	27.2	24.5
Lactoperoxidase	2.0	6.3	80.8	77.5
Conalbumin	5.0	17.8	58.2	54.4
β-casein	10.0	20.7	50.1	46.3
Catalase	20.0	28.8	100.2	98.3
Albumin	40.0	20.6	61.0	57.1
Myoglobin	60.0	159.6	132.3	135.0
Ovalbumin	100.0	149.6	116.0	116.0
Carbonic anhydrase	200.0	97.5	172.2	184.4
β-Lactoglobulin	500.0	432.9	140.2	144.5
Mean fold change		**2.4**	**9.9**	**9.3**
Median fold change		**2.1**	**4.3**	**4.0**
Minimum fold change		**1.2**	**1.2**	**1.1**
Maximum fold change		**4.2**	**40.4**	**38.7**

2. Characteristics of yeast protein expression levels

2.1. Comparison with other protein properties

Yeast protein expression levels are known to correlate with several other properties of proteins (**Figure S10**). In this section, we verify that APEX-derived measurements show the appropriate trends. Each plot shows comparisons involving 454 proteins identified from yeast growing in rich medium, with a false positive protein identification rate of ~5% (ProteinProphet[1] $p_i \geq 0.78$). First, protein concentrations measured by APEX are inversely correlated ($R_s = -0.67$, $R^2 = 0.28$) with protein molecular weight, as noted for measurements of protein expression by other techniques, *e.g.* see Ghaemmaghami *et al.* [6] or Coghlan *et al.* [21]. Second, yeast protein concentrations, measured by APEX, are positively correlated with protein codon adaptation indices ($R_s = 0.79$, $R^2 = 0.70$)[18] and with codon bias ($R_s = 0.80$, $R^2 = 0.69$), again as noted for measurements by other techniques[6, 7, 21].

Yeast protein expression levels, as measured by APEX, show no significant correlation with protein isoelectric point (pI), hydrophobicity (Gravy[22] scores), or aromaticity (frequency of the aromatic amino acids Phe, Tyr, and Trp)(**Figure S11**), implying both that shotgun proteomics/APEX-based quantitation shows no systematic sampling bias for these properties and that steady state protein expression levels are largely independent of these properties.

2.2. Comparison with estimated protein synthesis rates

We find that absolute protein expression levels are well-correlated with estimates of translation rates derived from association of mRNAs with polysomes[23] ($R^2 = 0.65$), and partially correlated with transcription rates estimated from measurements of mRNA half-life[10] ($R^2 = 0.31$) (**Figure S12**).

We find an even stronger correlation ($R^2 = 0.73$) between the measured protein abundances (average of at least two of three measurements and the protein production rates estimated by multiplying relative translational levels by the number of copies of each mRNA[23], indicating excellent agreement between observed protein levels and those predicted from translation alone. This correlation suggests that 73% of the variance in steady state protein levels can be explained by variation in protein production rates, with the remainder explained by experimental errors, protein degradation rate variation, and other factors.

2.3. Differential protein expression in yeast rich vs. minimal medium

Absolute protein expression measurements can be applied to a variety of biological questions. In the main text, we describe how MS-based measurements are able to extract proteins that are differentially expressed in yeast rich versus minimal medium.

This procedure is highly sensitive. It can detect significant changes (|Z|>2.58, 99% confidence) in protein expression that are less than 1.2-fold in highly abundant proteins (CDC19, EFT2, ADH1, PGK1) up to ~60- or even ~190-fold changes (SSB1, GDH1), or involve fewer than 10,000 molecules/cell expression difference (PEP1, SCP160, GLT1). Many peptides are measured in numbers large enough to enable sensitive detection of only small changes in expression: many of the significant fold-changes are less than 10-fold (**Figure S13**).

Note that while expression differences are most meaningful to discuss in terms of protein abundances (**Figure 3**, main text), we actually calculated the Z-score based on peptides, as illustrated in **Figure S14**. **Figure S14A** shows significant proteins in their peptide abundance (n_i counts) in rich vs. minimal medium, and the 5% confidence intervals. **Figure S14B** shows the same significantly differently expressed proteins, but in their protein abundances (APEX-based counts of molecules/cell). The conversion of peptide counts n_i to APEX-based protein abundances involves the ProteinProphet score p_i, the classifier O_i, and a constant describing the total number of molecules per cell and the total expected number of peptides (see main text, derivation of APEX). Thus, it can happen that two proteins with very similar abundances have different significance of their differential expression, as the peptide counts and other parameters vary from protein to protein (e.g. COQ6 and LYS7).

The sensitivity of MS-based data in measuring changes in protein expression is confirmed when using an independent data set published by Zybailov *et al.* [24]. In order to use these authors' LTQ MS data to calculate protein abundances with the APEX method, we should ideally re-calculate the O_i values for the LTQ MS technology used by Zybailov *et al.* [24]. However, comparison of the significantly up- or down-regulated proteins (|Z|>2.58; 99% confidence) is still valid as the Z-score does not rely on use of the classifier O_i.

Protein abundance measurements from the two datasets both capture differences known for cells growing in different media (**Figure S15**). The overall correlation between Z-scores of differentially expressed proteins in our (Lu *et al*.) and Zybailov *et al*.'s data is moderate (R^2=0.28) and reasonable for differentially expressed proteins (R^2=0.58 for |Z|>2.58). There is a significant overlap in proteins up-regulated in minimal medium (p<5e-12), and in proteins up regulated rich medium (p<3e-2). For example, MET6 is the single most strongly up-regulated protein in both YMD datasets, with a 10- to 11-fold increase in expression from ~48-75,000 copies to ~550-758,000 copies/cell. Proteins in the overlap, i.e. that are up-regulated in minimal medium in both the Lu and Zybailov dataset, are significantly enriched for targets of the transcription factor GCN4[25] (14/30; p<1.1e-8) which is expected for amino acid starvation response.

The differential expression of some proteins is specific to either the Lu or Zybailov dataset (**Figure S15**), and such differences can be explained by the genetic background of the two yeast strains. Lu-specific proteins that are up-regulated in minimal medium are enriched for proteins of purine synthesis (ADE-genes), *i.e.* targets of BAS1 (8 of 37 known targets[26], p<1.3e-8). Up-regulation of genes of purine (and also histidine) biosynthesis may be caused by an imbalance in nucleotide metabolism due to the URA3 marker gene in the strain. In contrast, Zybailov set-specific genes up-regulated in minimal medium are involved in glucose metabolism, possibly caused by deletion of

galactose-transporter gal2, maltose/mellibiose metabolism genes mal and mel in Zybailov's yeast strain. The proteins are also enriched for glycosylation and glucoprotein metabolism, possibly caused by deletion of flo1 and flo8-1. Five BAS1 targets, i.e. proteins of purine synthesis, and four HAP1 targets[26], *i.e*. proteins of anaerobic growth, including several ERG-genes that are part of ergosterol biosynthesis, are down-regulated in minimal medium in Zybailov's data ($p<6.4e-4$ and $p<0.08$). The latter case of differential expression may be due to mutation in the hap1 gene in the yeast strain used for Zybailov's data.

2.4. Differential protein expression in mouse T-lymphoma cells

Protein abundance measurements using mass spectrometry are scalable and can easily be applied to higher eukaryotes, as we also demonstrate for mouse T-lymphoma cells.

Approximately $3x10^7$ mouse T-lymphoma BW5147 cells were harvested, washed in PBS, and resuspended in 5ml buffer (10mM Hepes pH7.9, 1.5 mM $MgCl_2$, 10mM KCl, 1mM DTT, protease inhibitors) for 10 min. Cells were pelleted (1,000g for 10 min) and resuspended in 2ml of the same buffer. Cells were then lysed using 10 strokes in a homogenizer and nuclei were pelleted (1,000g for 10 min), the supernatant was retained as the cytoplasmic protein sample. The wash and centrifugation steps were repeated once, centrifuging at 30,000g for 20 min. Nuclei were resuspended in 1ml buffer (20 mM Hepes pH7.9, 25% glycerol 0.42 M NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT, protease inhibitors), and homogenized (~30 strokes), and stirred on a magnetic stirrer for 30-60 min. The lysed nuclei were centrifuged at 30,000g for 20 min) and the supernatant dialyzed against 150 volumes of buffer (20mM Hepes pH7.9, 20% glycerol, 100mMKCl, 0.2mM EDTA, 1mM DTT, protease inhibitors) for 3-4 hours. Nuclear extracts were centrifuged (30,000g for 20 min) and the supernatants collected for analysis.

The protein mixtures were diluted in digestion buffer (50mM Tris HCL pH8.0, 1.0M Urea, 2.0mM $CaCl_2$), trypsin digested, and analyzed by LC/LC/MS/MS as described in the main text. Three sequential LC/LC/MS/MS analyses were performed, and the fragmentation spectra were analyzed using the program TurboSequest/ BioWorks 3.1 and ProteinProphet[1]. For protein identification, we downloaded the database of 25,371 mouse (*Mus musculus*) proteins from Entrez Genome (http://www.ncbi.nih.nlm.gov).

In total, we identified 1391 proteins (**Supplementary Dataset S3**) across all experiments with a false identification rate of ≤5%. All proteins were then assigned a Z-score based on the frequency of total peptides identified per protein in the nuclear and cytosolic protein samples; the distribution of Z-scores is shown in **Figure S16**. Proteins with $Z > 1.96$ or $Z < -1.96$ are significantly enriched in the nucleus or cytoplasm, respectively (95% confidence). Of the mouse proteins identified, 180 and 192 proteins are known to be localized to nucleus and cytoplasm, respectively, as annotated by the DAVID webserver at http://david.niaid.nih.gov/david/version2/index.htm). **Figure S16** shows that the higher the Z value, the higher the fraction of the identified proteins known to be nuclear proteins. Thus, proteins of the two different cellular localizations can be identified using our approach, clearly distinguishing nuclear proteins from contaminating cytoplasmic proteins.

The identified nuclear proteins include histone deacetylase, DNA helicase, DNA methyl-transferase, HMGB2, radixin, methyl-CpG binding protein, acidic nuclear phosphoprotein 32 family member B, tumor rejection antigen gp96, and TBP-interacting protein. Importantly, we also identified cell division cycle 2 homolog A, transcription factor Swi, and transcription elongation factor. In addition, we detected abundant HnRNP proteins (A3, L, M, H2, R, I, U, K, H1), splicing factor 3b, snRNP (A, E, U2, B, D1), ubiquitin-conjugating enzyme, and valyl-tRNA synthetase 2. These results confirm that even low abundance proteins such as transcription factors can successfully be identified by shotgun proteomics of the nucleus.

Translation initiation factor 5A and translation elongation factor 1 were observed to be abundant in both nucleus and cytoplasm. Some cytoplasmic proteins were also identified in nuclear sample, and this may be due to cross-contamination or that certain proteins are present in both nucleus and cytoplasmic pools, reflecting their ability to translocate between these compartments. These proteins are ELAV (embryonic lethal abnormal vision), proliferating cell nuclear antigen, endoplasmic reticulum protein Pdia3, zinc-finger proteins, matrin, septin, actin, protein phosphatase (Ppp1ca, Ppp2cb, Anp32b, Pnkp), cytochrome C, protein disulfide isomerase-related protein, vemintin, golgi coil-coiled protein Gcc1, heat shock proteins, dynein, and spectrin.

3. Correlation of yeast protein and mRNA levels

3.1. Comparisons with mRNA levels

Protein expression levels estimated by APEX are well correlated with mRNA expression levels from Affymetrix microarrays[8], SAGE[9], and dual channel microarrays with genomic DNA reference[10], as well as with the average of at least two mRNA measurements, over approx. 3-4 orders of magnitude of protein concentration and 3 of mRNA concentration **(Figure S17)**.

The correlation between protein and mRNA levels is also confirmed when we examine the high-confidence set of 58 proteins which have less than 50% relative standard deviation across the measurements of protein and mRNA abundance (**Figure S18**). Protein and mRNA levels correlate with R^2=0.77 in a power-law relationship.

3.2. Analysis of protein per mRNA ratios

To ensure that the log-normal relationship of protein to mRNA identified was not an effect of averaging protein expression levels from different platforms, we performed the same analysis for single techniques. Each produces comparable results, as shown here for a comparison of Western[6] and APEX protein expression levels to the average of two measures of mRNA levels (DNA microarrays[8] and SAGE[9]) (**Figure S19**). As described in the paper, we calculated the logarithm of the ratio of each protein's abundance divided by its corresponding mRNA levels, and then calculated the histogram of these values and the fit to a normal distribution. Each distribution is well described by a log-normal curve (R^2 = 0.93 and 0.94, respectively).

3.3. Unusual protein per mRNA ratios

A comparison of the unusual protein per mRNA ratios with ribosomal loading of the corresponding transcripts, using RNA-polysome association data published by Arava *et al*. [23, 27], indicates that the under-translated protein RPS21A exhibits significantly lower association with polysomes than the over-translated protein ADH2 (**Figure S20**). RPS21A is one of the few *S. cerevisiae* genes with introns[28], a factor that may contribute to its lower availability to ribosomes, or that at least might lower the effective mature mRNA concentration relative to the total quantity of RPS21A mRNA.

3.4. Comparison with protein properties

In attempting to explain the variance of protein per mRNA ratios, we examined trends likely to control levels of translation (**Figure S21**). In particular, measures of codon choice (CAI, codon bias, and frequency of optimal codons) are largely uncorrelated with

the ratio of protein to mRNA. Note that in comparison, all these measures are correlated with the protein concentration, as discussed for APEX measurements above and shown for average protein abundance in **Figure S21**. As explained in the main text, molecular weight correlates both with protein concentration and, to a far lesser extent, with the protein/mRNA ratio.

3.5. Comparison with sequence characteristics

We analyzed the set of 331 genes with at least two measurements for both protein and mRNA levels for sequence characteristics that relate to protein abundance and to the protein per mRNA ratio (**Figure S22**). To do so, we compared the 50 most abundant proteins to the 50 least abundant proteins, and the 50 proteins with highest protein/mRNA ratios to those 50 with the lowest ratios.

The PEST sequence is known to be a ubiquitinylation signal that can trigger protein degradation[29]. Indeed, we find that the 50 most abundant proteins have a slightly lower fraction of PEST residues than the 50 least abundant proteins; this is true to a much lesser extent for the protein/mRNA ratios.

Further, it has been observed that while some amino acids at the N-terminal end of proteins have destabilizing effects (RKFLWHAQY), other amino acids have stabilizing effects on proteins (CASTGVM)[30]. As discussed in the main text, we can generally confirm these observations: there are significant biases with respect to amino acid choice at the N-terminal end of highly abundant proteins or those with a high protein/mRNA ratio. There is one obvious exception: leucine is overrepresented in both highly abundant proteins and those with a high protein/mRNA ratio, but it is known to be destabilizing[30].

We observe similar results when counting amino acids in a sliding window (not shown).

References

1. 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).
2. Breiman, L. Bagging predictors. *Machine Learning* **24**, 123-140 (1996).
3. 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).
4. Choi, B.K., Hercules, D.M. & Gusev, A.I. LC-MS/MS signal suppression effects in the analysis of pesticides in complex environmental matrices. *Fresenius J Anal Chem* **369**, 370-377 (2001).
5. Choi, B.K., Hercules, D.M. & Gusev, A.I. Effect of liquid chromatography separation of complex matrices on liquid chromatography-tandem mass spectrometry signal suppression. *J Chromatogr A* **907**, 337-342 (2001).
6. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).
7. 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).
8. Holstege, F.C. et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-728 (1998).
9. Velculescu, V.E. et al. Characterization of the yeast transcriptome. *Cell* **88**, 243-251 (1997).
10. Wang, Y. et al. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* **99**, 5860-5865 (2002).
11. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-3297 (1998).
12. Newman, J.R. et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* (2006).
13. 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).
14. 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).
15. 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).

16. Corbin, R.W. et al. Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc Natl Acad Sci U S A* **100**, 9232-9237 (2003).

17. Neidhardt, F.C., Bloch, P.L. & Smith, D.F. Culture medium for enterobacteria. *J Bacteriol* **119**, 736-747 (1974).

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

19. Rappsilber, J., Ryder, U., Lamond, A.I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res* **12**, 1231-1245 (2002).

20. 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).

21. Coghlan, A. & Wolfe, K.H. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* **16**, 1131-1145 (2000).

22. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132 (1982).

23. Arava, Y. et al. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **100**, 3889-3894 (2003).

24. Zybailov, B., Coleman, M.K., Florens, L. & Washburn, M.P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* **77**, 6218-6224 (2005).

25. 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).

26. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).

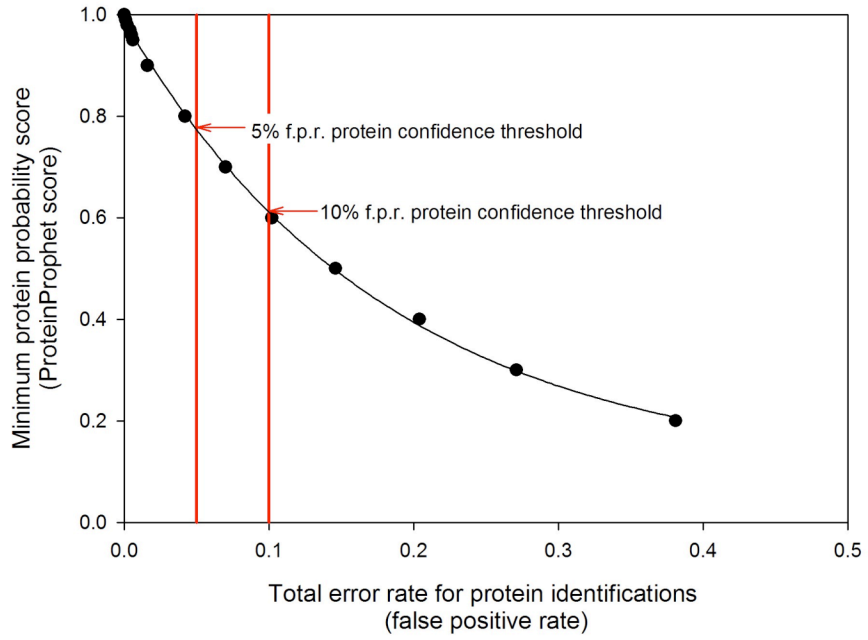
27. Arava, Y., Boas, F.E., Brown, P.O. & Herschlag, D. Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Res* **33**, 2421-2432 (2005).

28. Suzuki, K. & Otaka, E. Cloning and nucleotide sequence of the gene encoding yeast ribosomal protein YS25. *Nucleic Acids Res* **16**, 6223 (1988).

29. Rogers, S., Wells, R. & Rechsteiner, M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* **234**, 364-368 (1986).

30. 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).

A. YMD medium



B. YPD medium

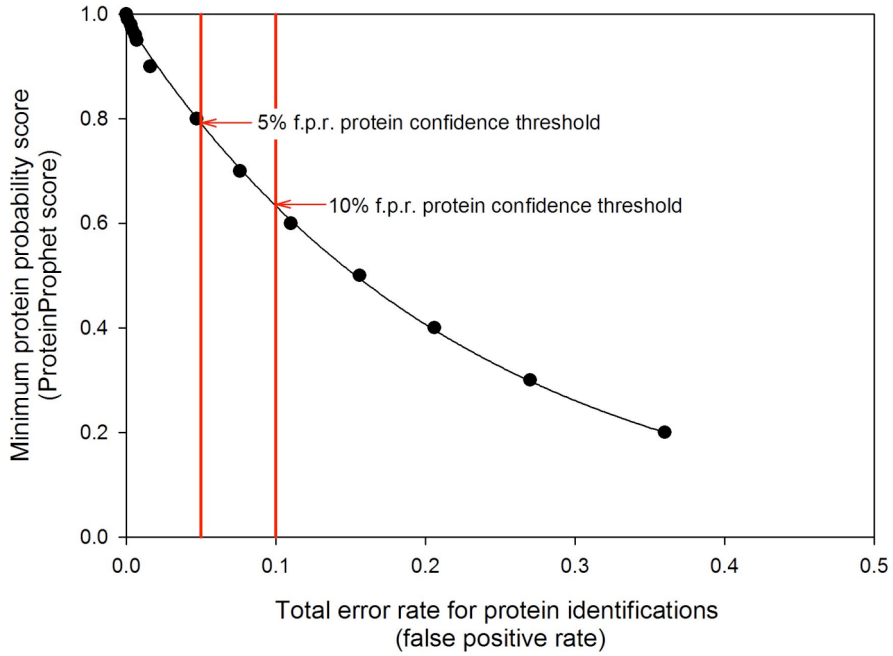
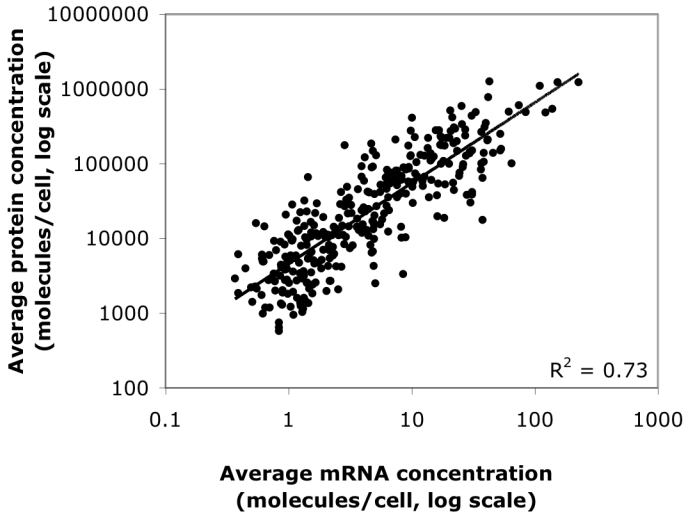
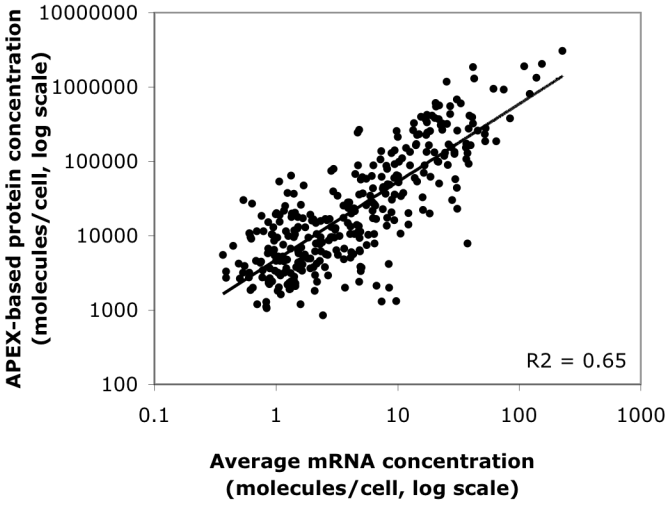


Figure S1. Yeast. Error model for protein identification in minimal (A, YMD) and rich (B, YPD) medium using ProteinProphet [Nesvizhskii, *Anal Chem* 2003].

A.



B.



C.

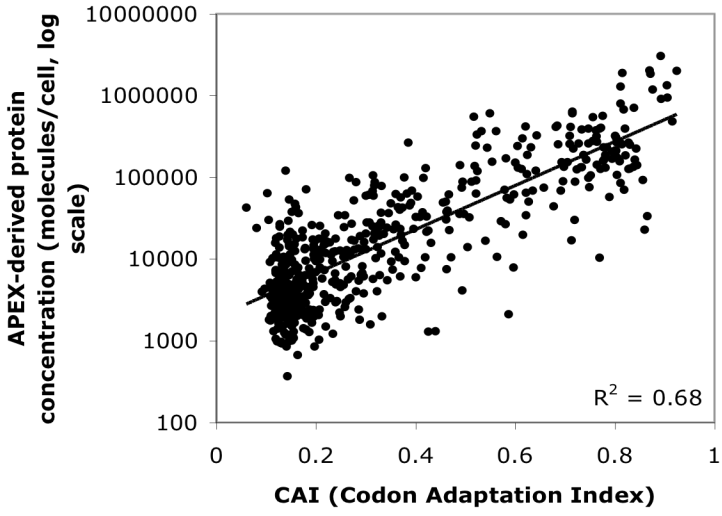
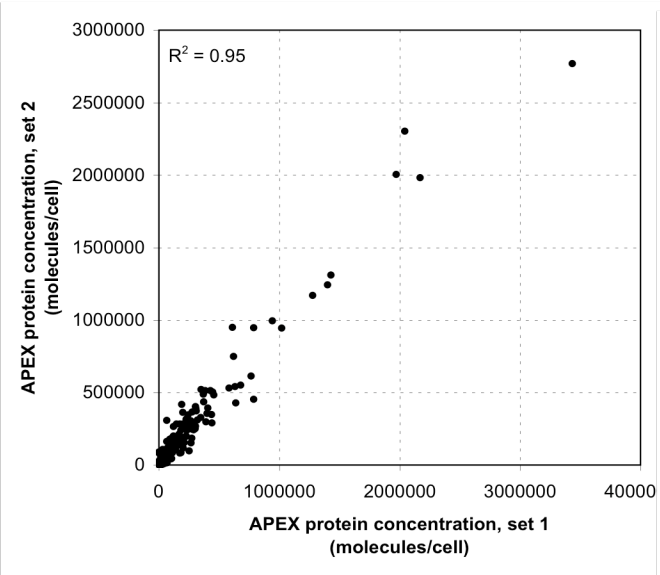
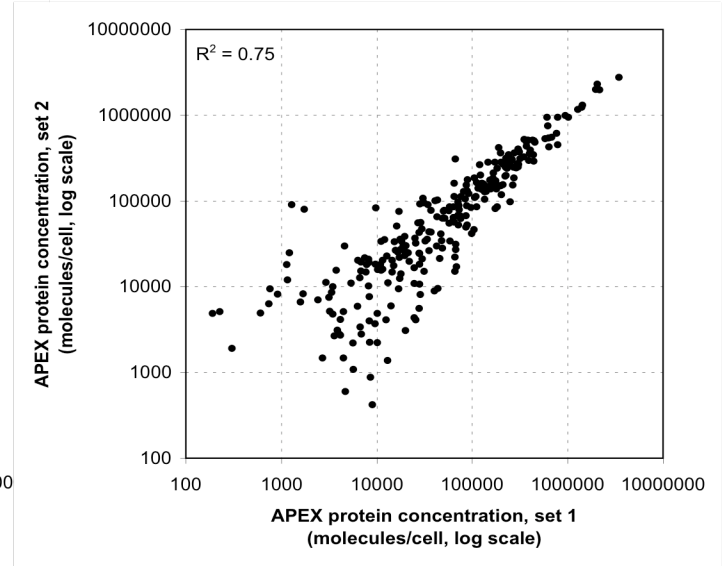


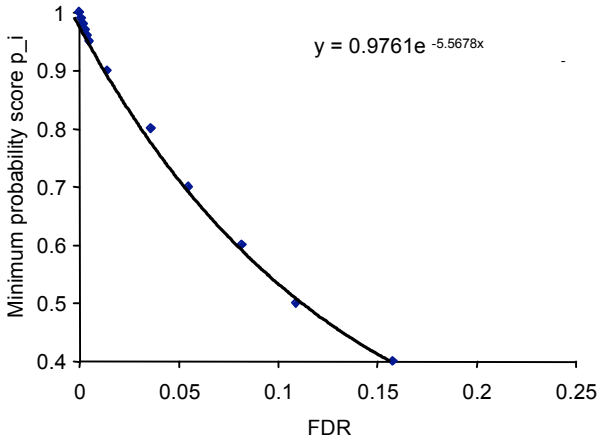
Figure S2. Yeast. Analysis of 555 proteins identified using 10% FDR for yeast growing in rich medium (YPD).





set 1 - three pooled injections

set 2 - two pooled injections



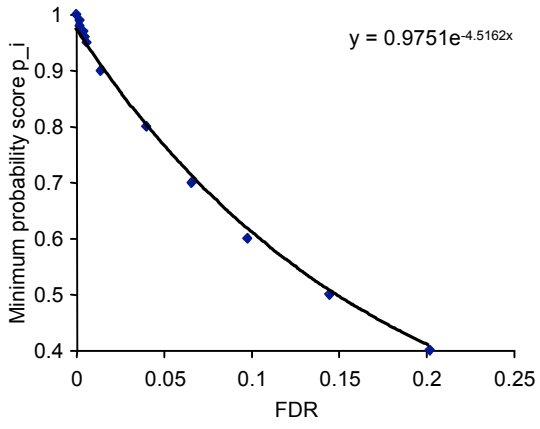


Figure S3. Yeast. Reproducibility of MS-based protein abundance measurements grown in YPD medium (for YMD see main text).

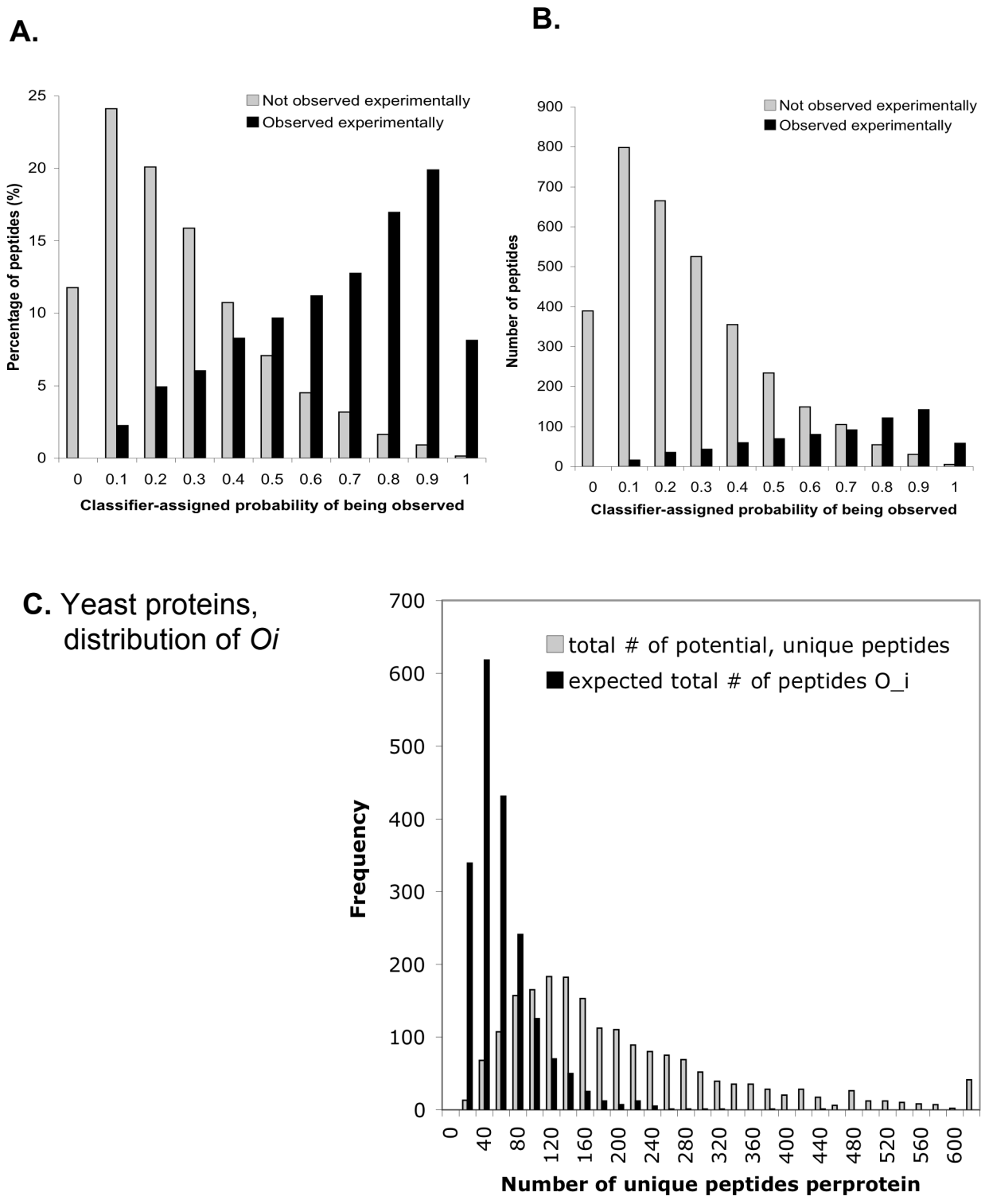
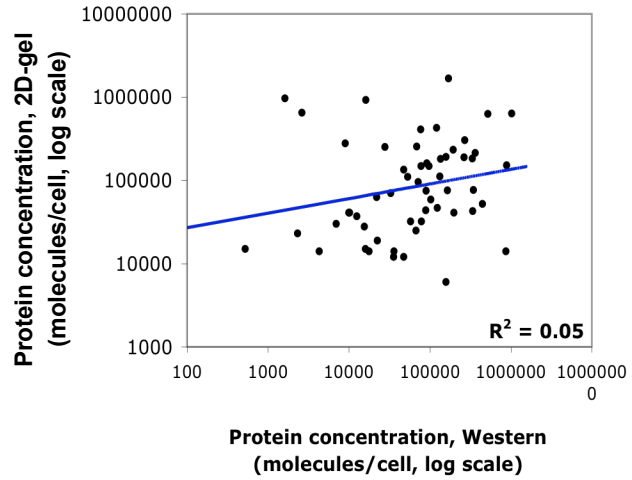
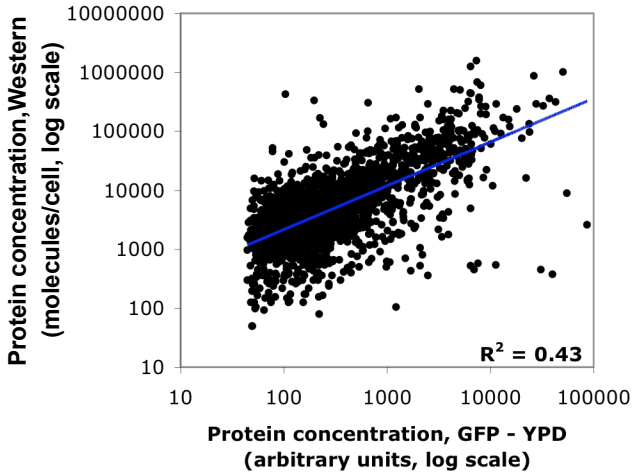


Figure S4. Classifier training. Most of the peptides are not observed. For those peptides that are observed, and those that are not observed, the classifier predicts their occurrence with 86% accuracy (A, B). Distribution of *O_i* values and number of unique expected peptides for all yeast proteins (C).





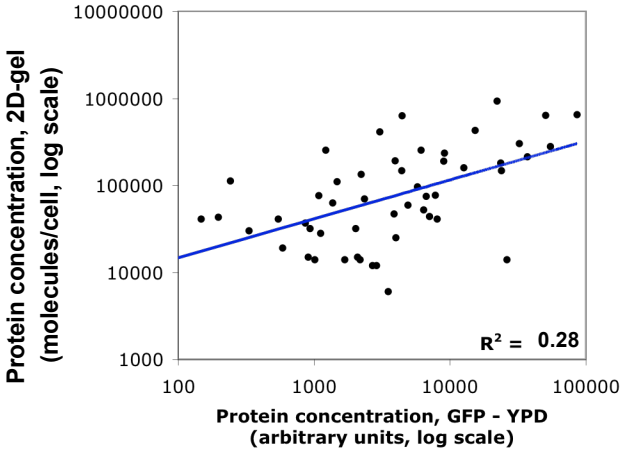


Figure S5. Yeast. APEX-based protein abundance for cells grown in rich medium (YPD). APEX correlates better with other measurements of protein abundance than these data sets correlate with each other.

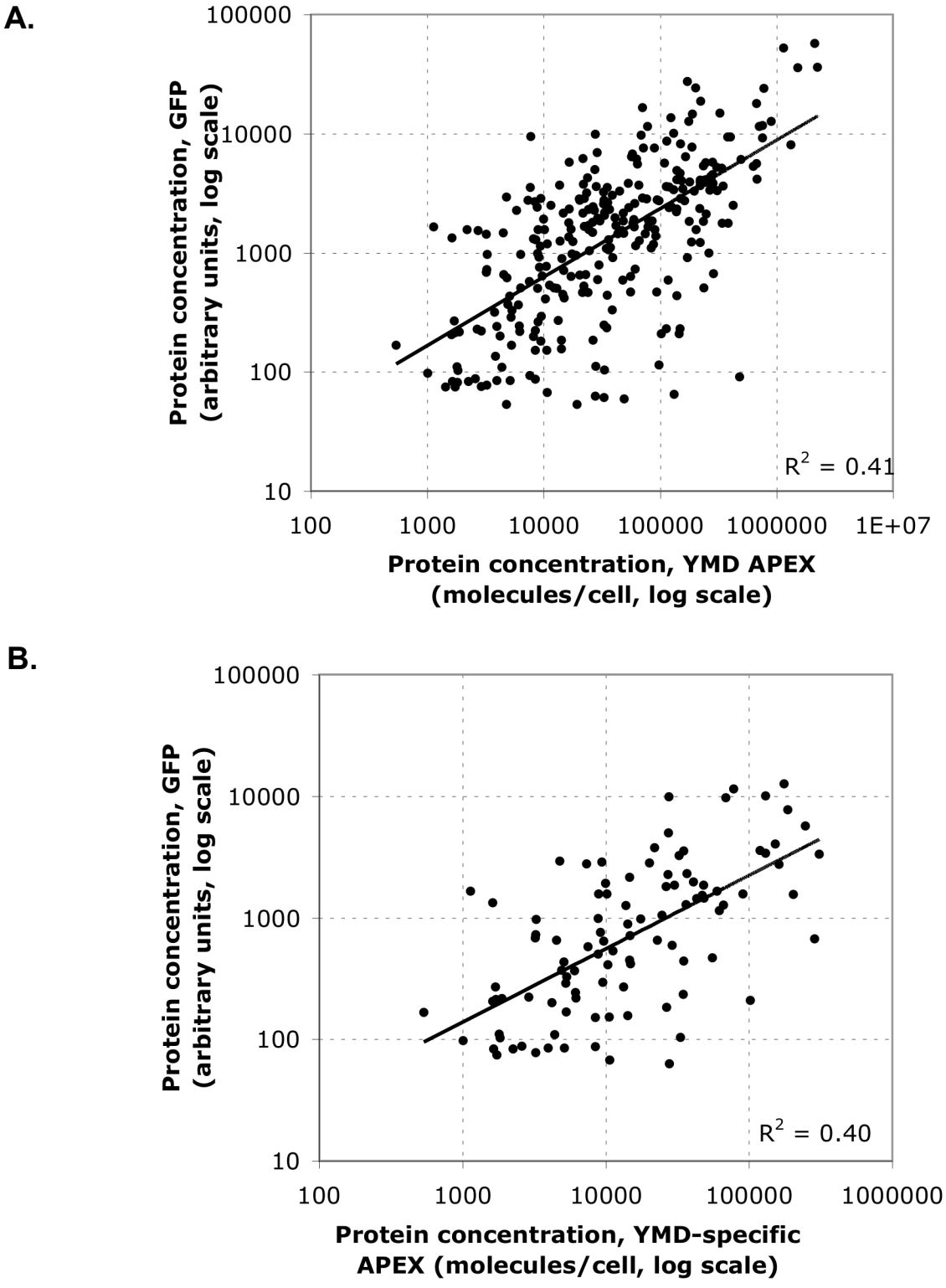


Figure S6. Yeast. APEX-based protein abundance for cells grown in minimal medium (YMD). (A) all proteins in YMD, N=290; (B) YMD-specific proteins (not observed in YPD), N=102. APEX versus GFP-labeled protein abundance: Rs=0.64 for (A) and (B).

A. Yeast, YMD
High abundance proteins

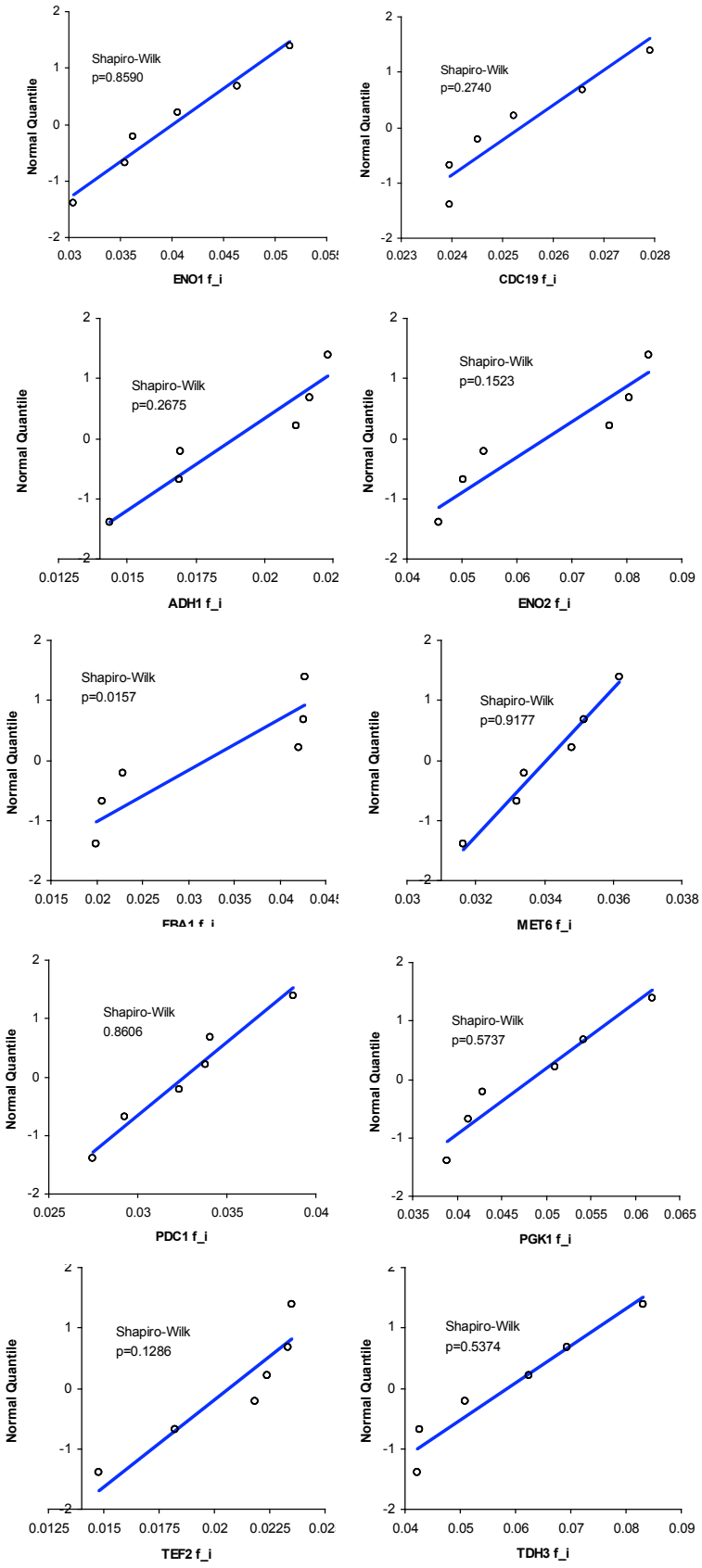


Figure S7A. Yeast. Normality test for redundant peptide counts *n_i*. High abundance proteins:Each set contains 6 *obs* of *fi* for a given protein from 6 YMD datasets. Distributions with p>0.05 are normal under the Shapiro-Wilk test (95% confidence)

B. Yeast, YMD
Low abundance proteins

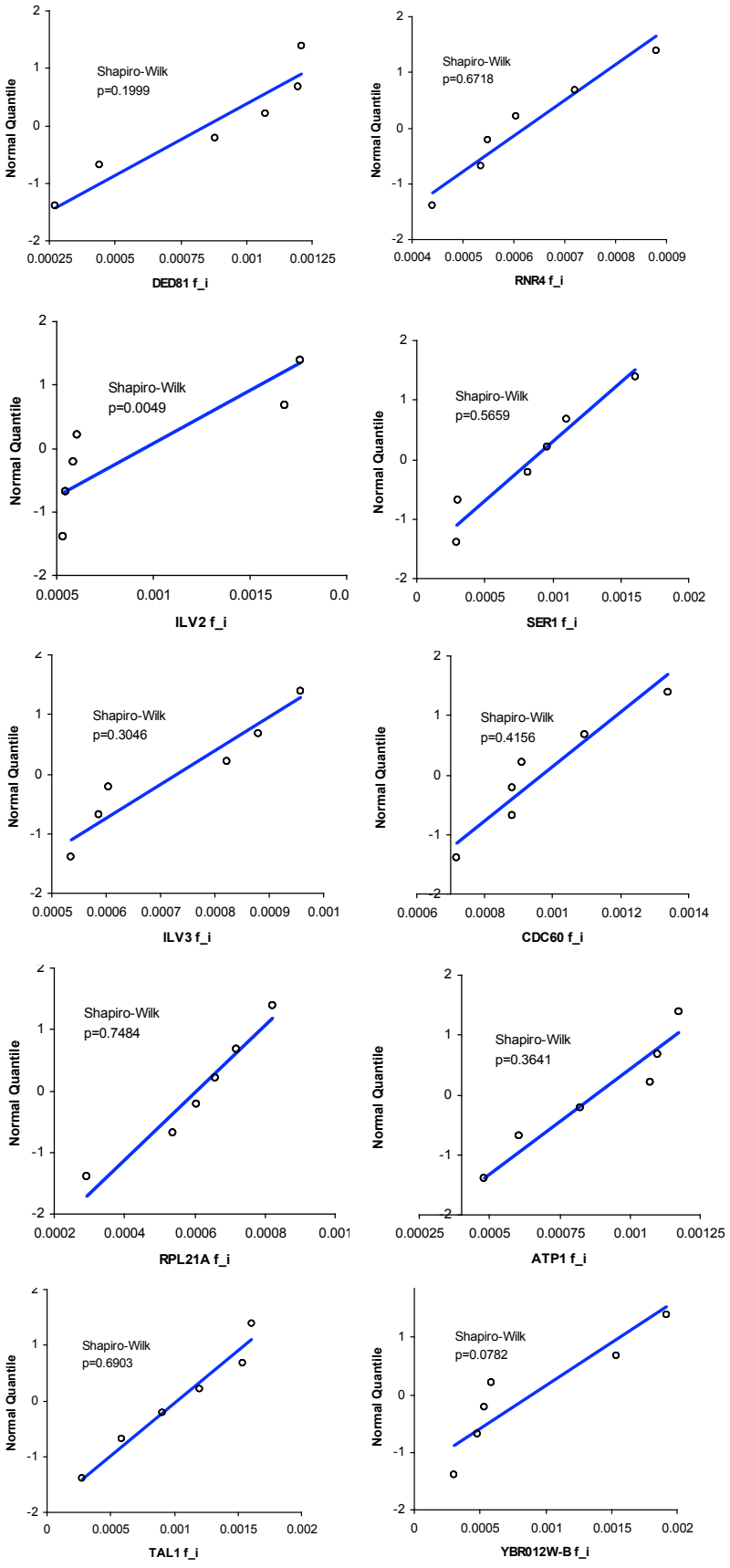
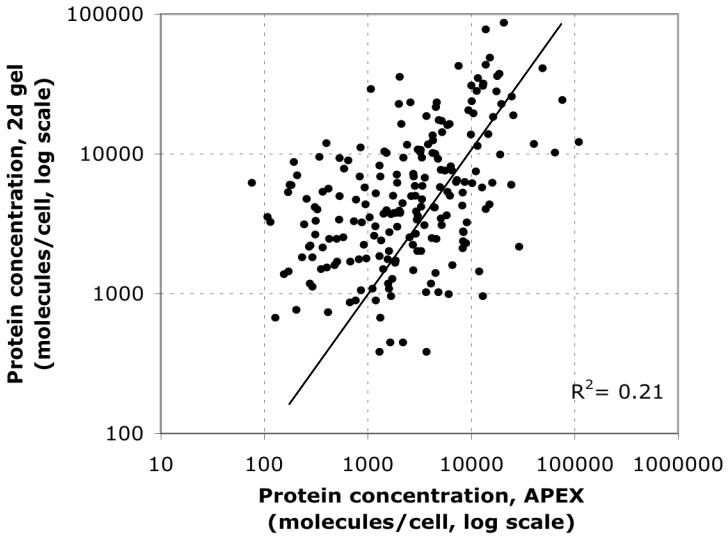
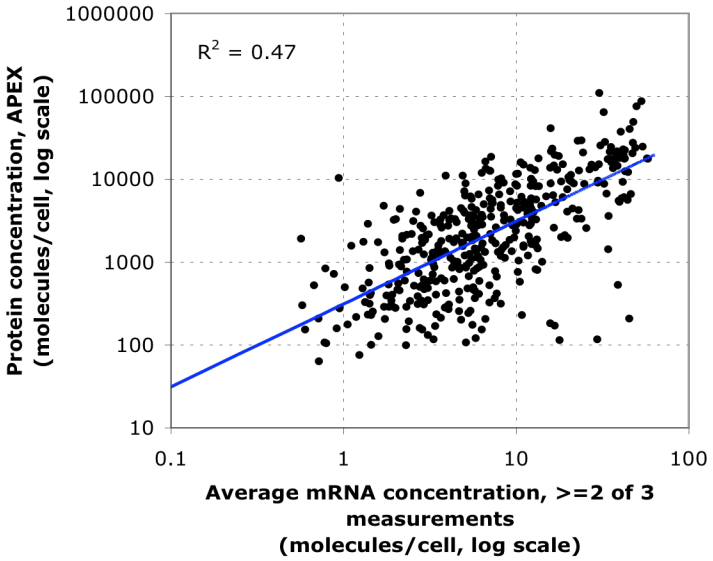


Figure S7B. Yeast. Normality test for redundant peptide counts *n_i.* Low abundance proteins:Each set contains 6 *obs* of *fi* for a given protein from 6 YMD datasets. Distributions with p>0.05 are normal under the Shapiro-Wilk test (95% confidence)

A.



B.



C.

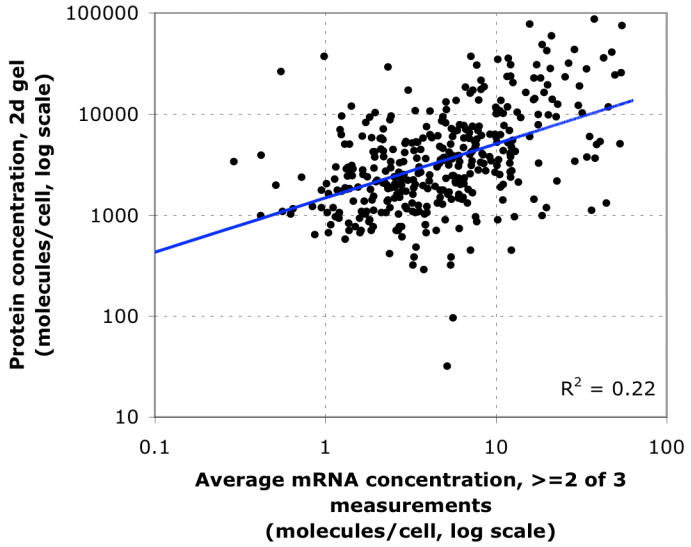


Figure S8. *E.coli*. APEX-based protein abundance vs 2D-gel based protein abundance (A). The black line indicates the diagonal. For comparison, APEX-based protein abundance and 2Dgel based protein abundance vs. average of at least 2 of 3 measurements of mRNA (B, C)

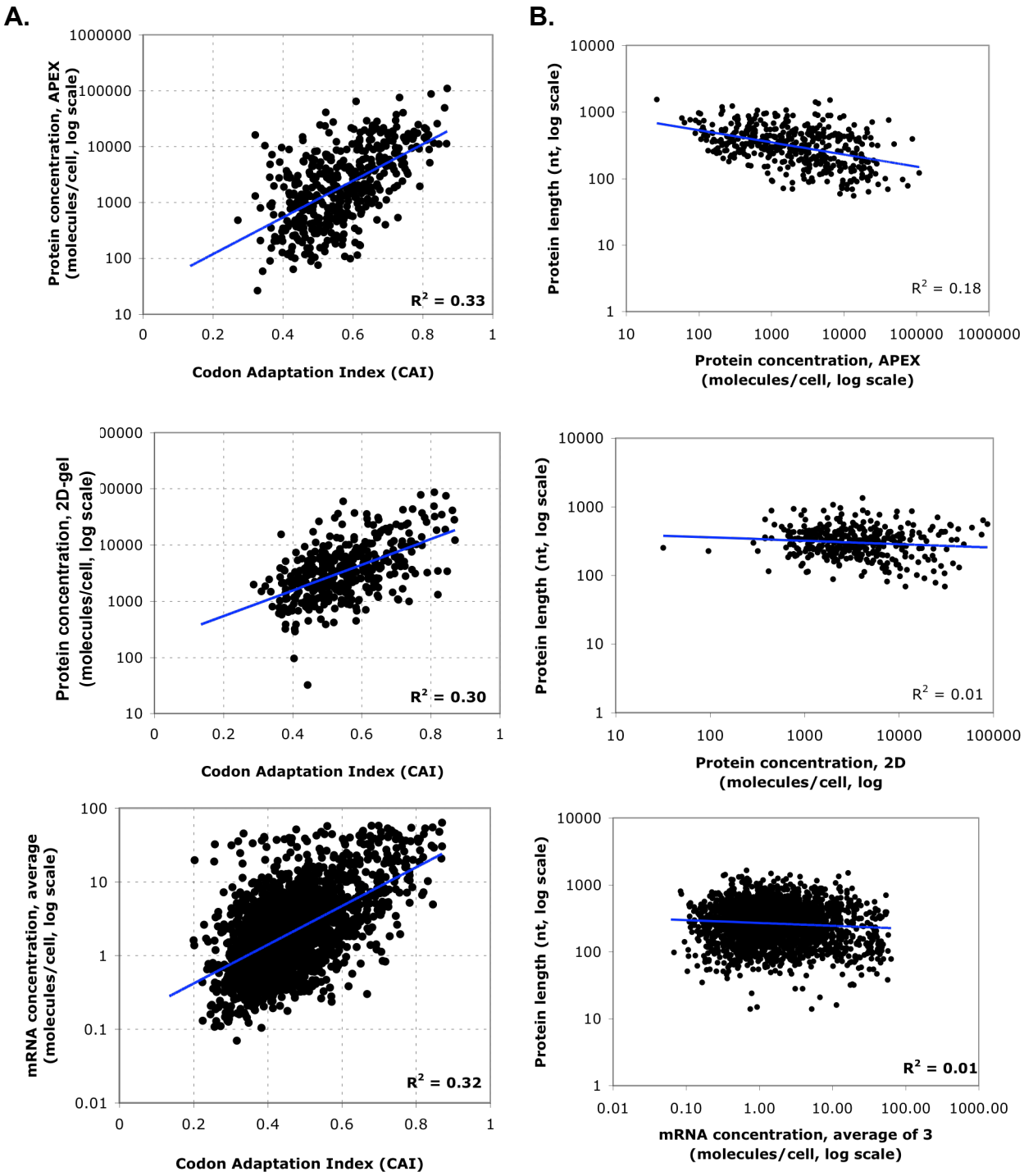


Figure S9. *E.coli.* APEX-based protein abundance, 2D-gel derived protein abundance , and mRNA abundance vs CAI (A) and gene length (B).

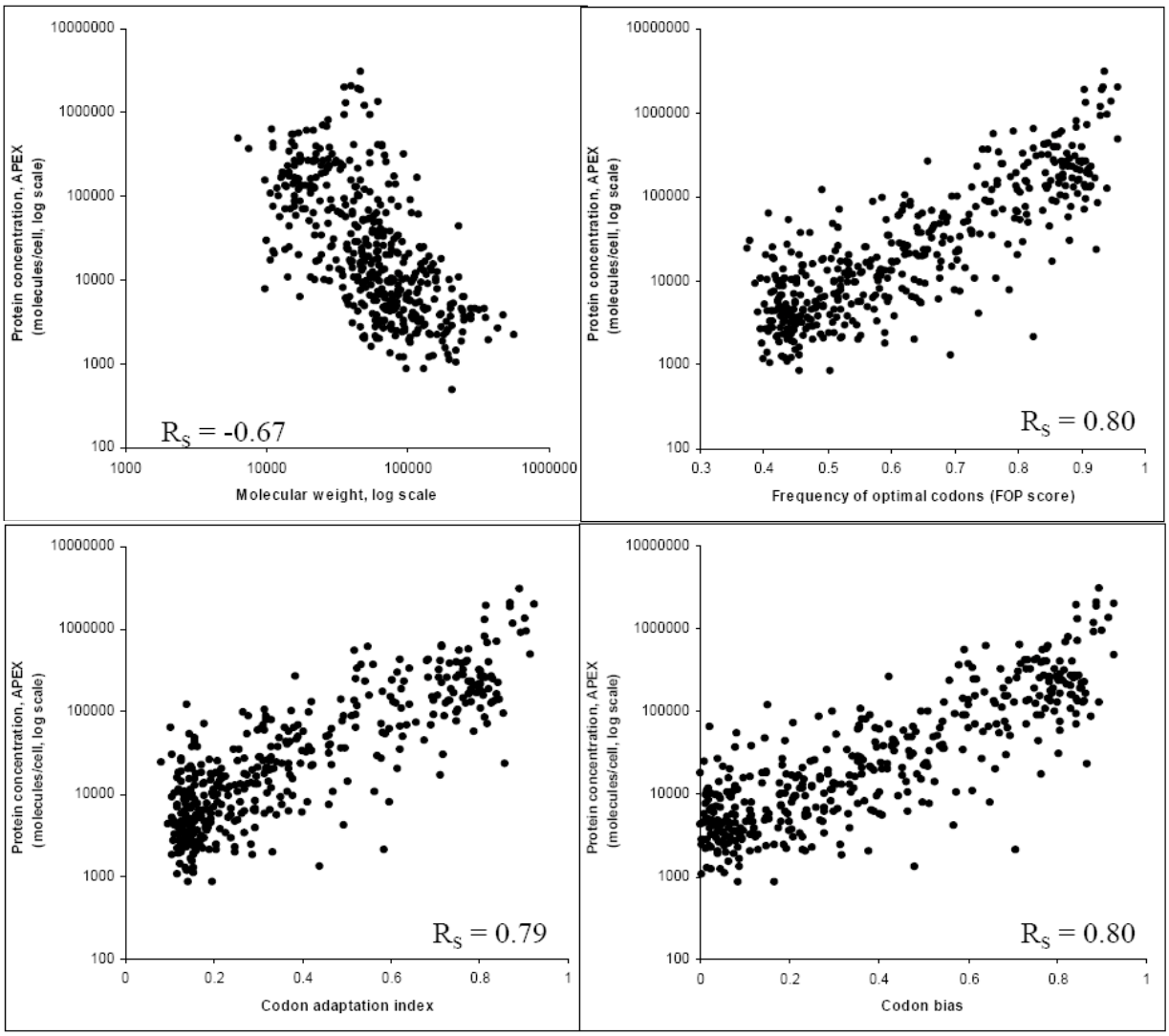


Figure S10. Yeast. APEX-based protein abundances vs. molecular weight, FOP, CAI, codon bias.

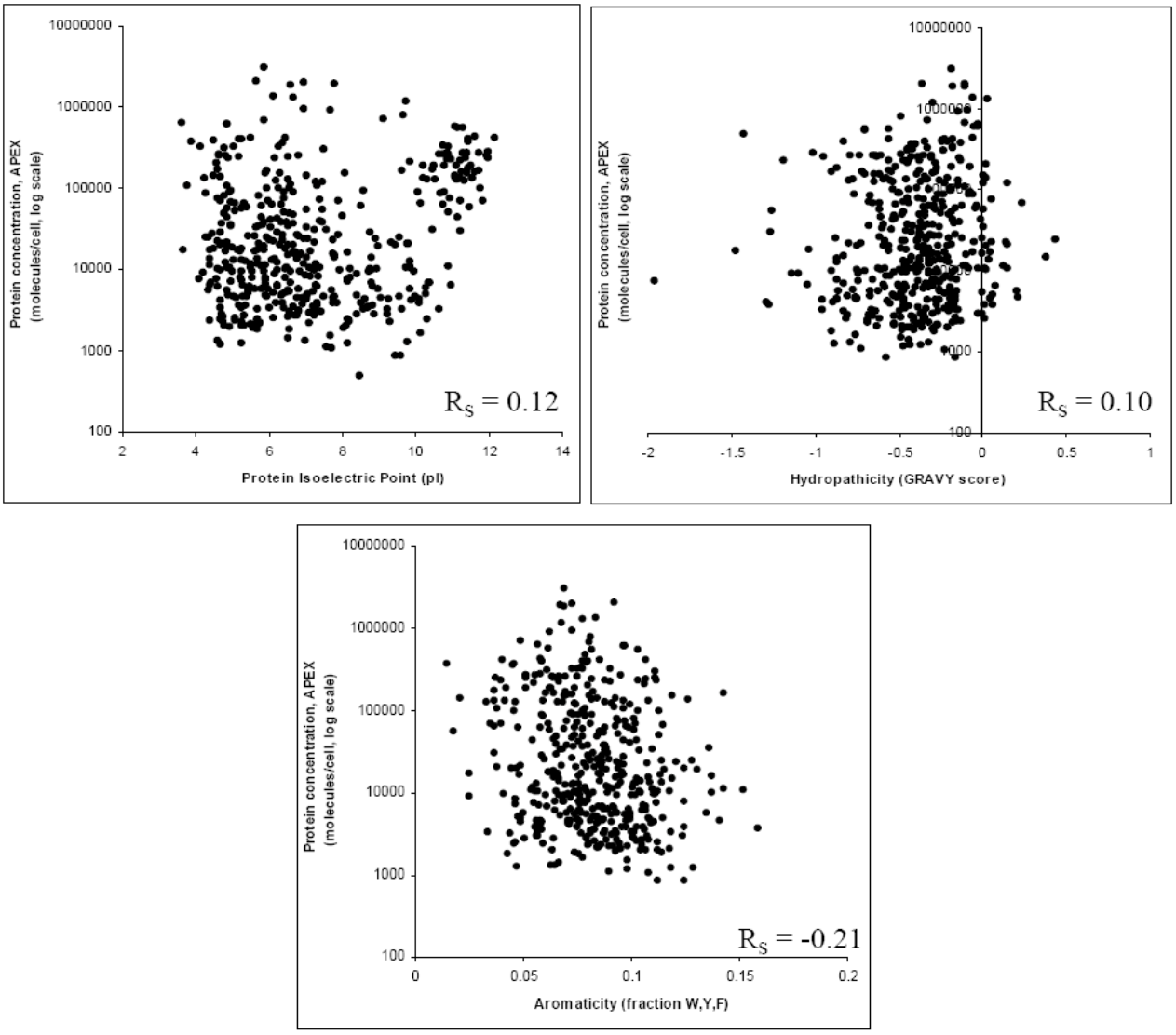


Figure S11. Yeast. APEX-based protein abundances vs. pI, hydrophobicity, aromaticity.

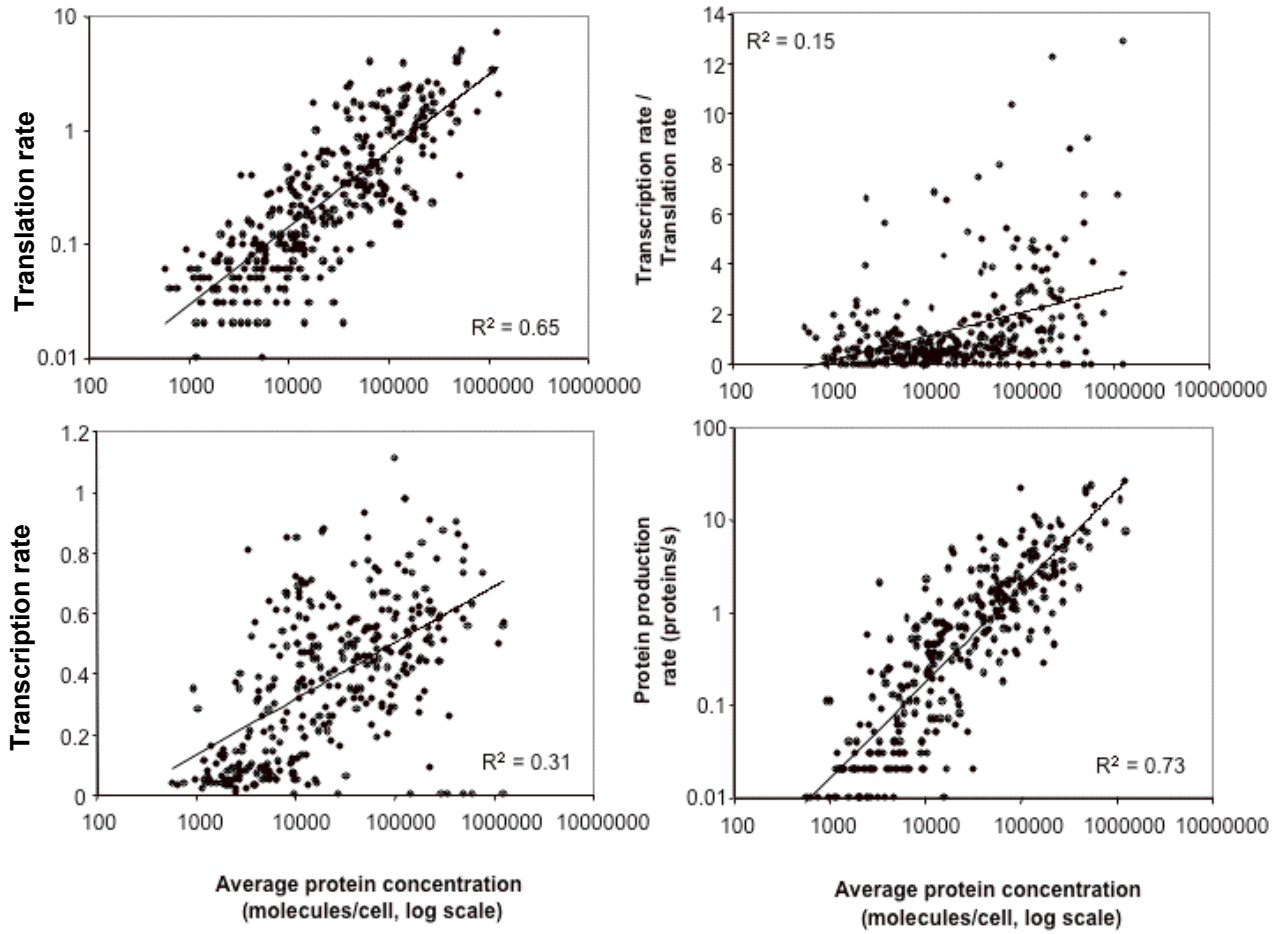
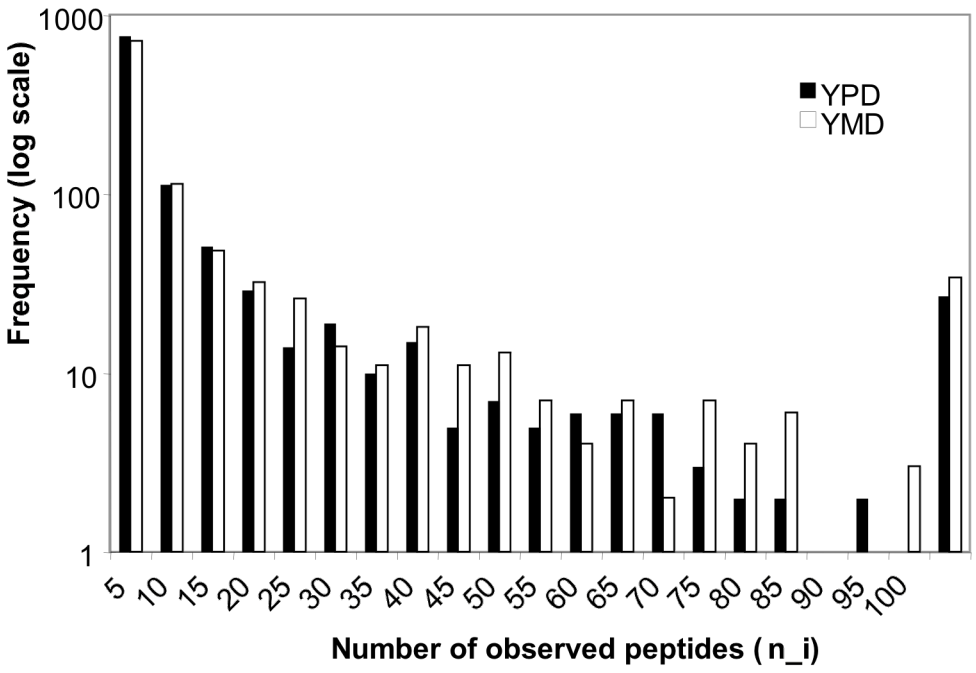


Figure S12. Yeast. Average protein abundance vs. parameters of transcriptional and translational regulation (transcription, translation and protein production rates).

A.



B.

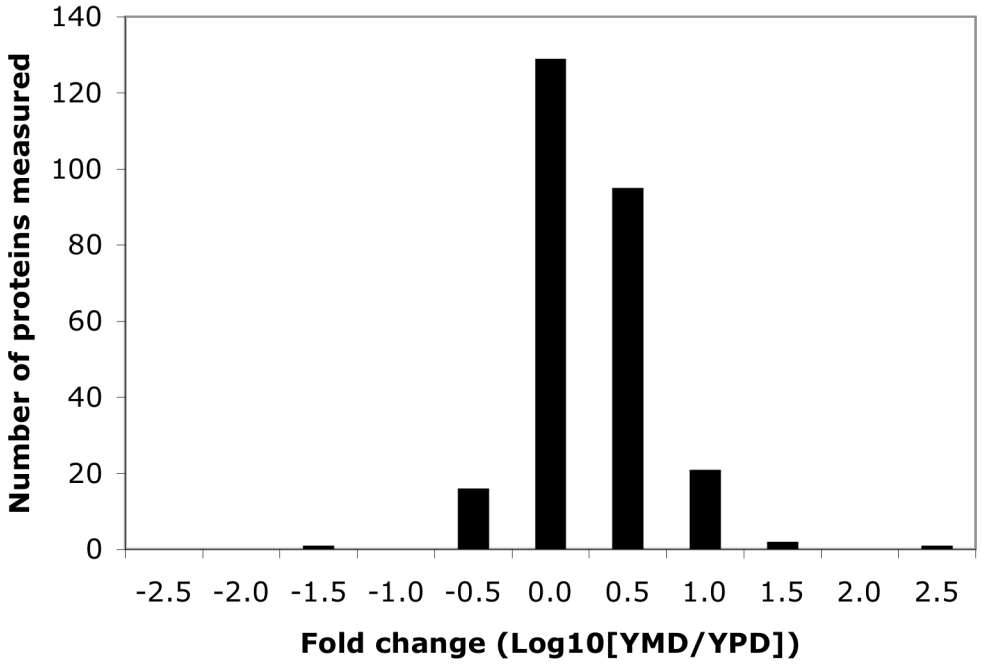
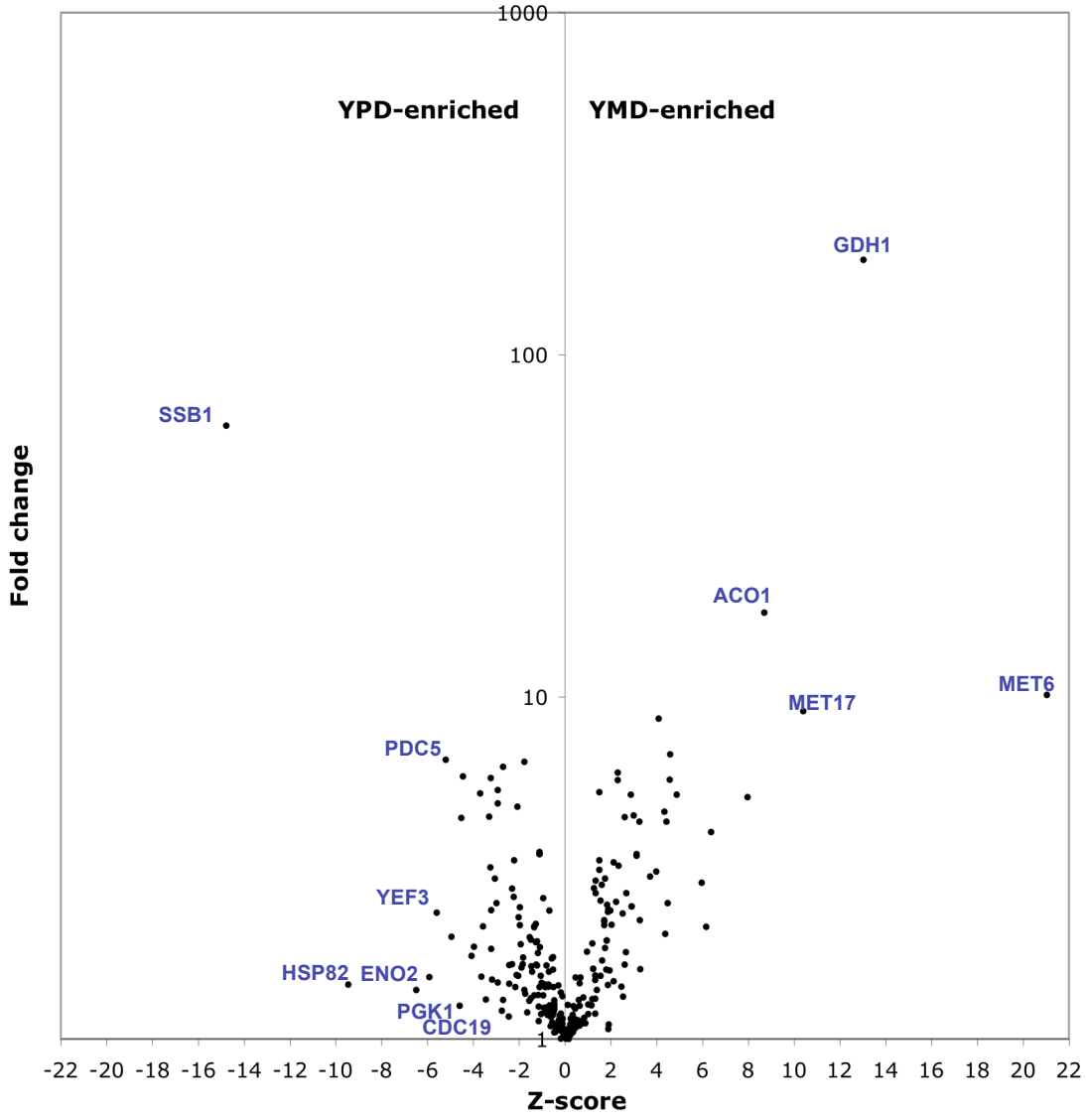


Figure S13A,B. Yeast. Peptide count distribution in YMD and YPD (A). Distribution of fold change between YPD and YMD medium (B).

C.



D.

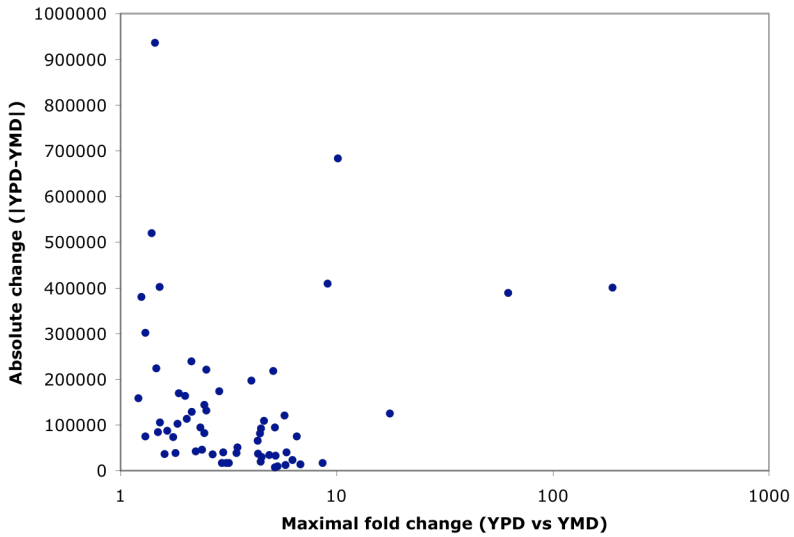


Figure S13C,D. Yeast. (C) Significance (Z-score) versus fold change (YPD vs. YMD). (D) Fold change versus absolute change, only significant (99% confidence) values and only those with observations both in YPD and YMD medium are plotted. |Z|>2.58 denotes 99% confidence.

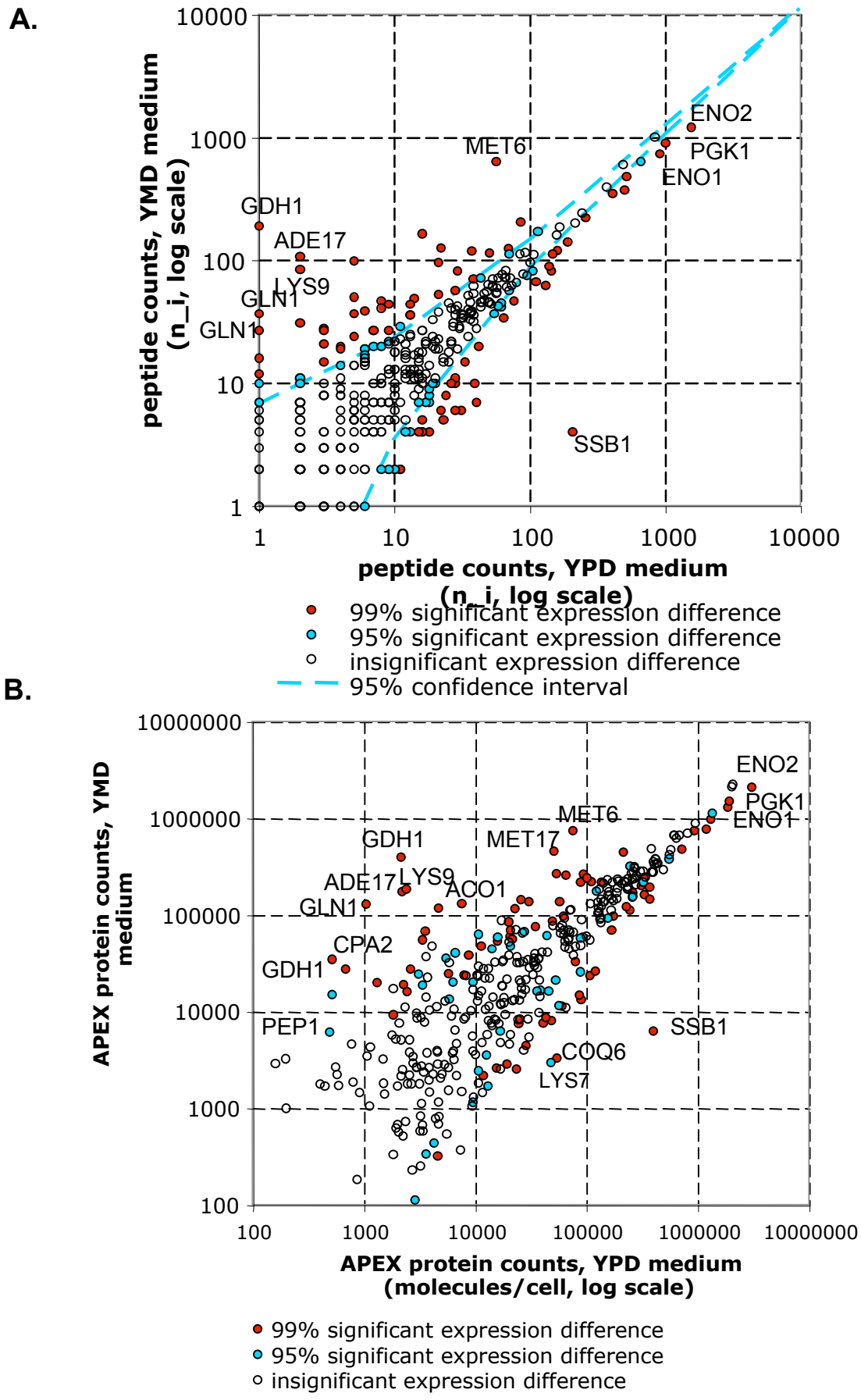
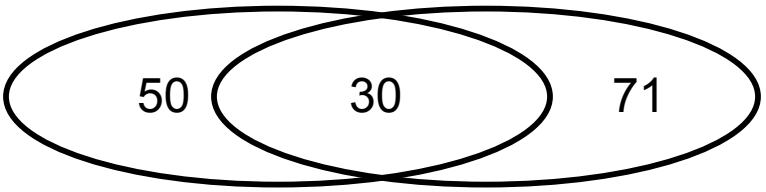


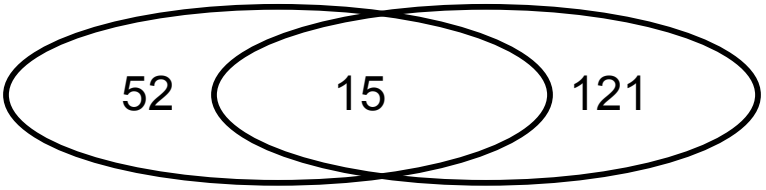
Figure S14. Yeast. Differential expression analysis. Z-score test (based on peptides) - significant data points plotted with respect to peptide and protein counts.

A. up-regulated in YMD (minimal) medium (Z>2.58)



GCN4-targets	15	14	14
	4	12	8
BAS1-targets	5	3	0
HAP1-targets	3	0	4

B. up-regulated YPD (rich) medium (Z<-2.58)



GCN4-targets	4	0	8
	4	0	1
BAS1-targets	0	0	5
HAP1-targets	1	1	4

Figure S15. Yeast. Differential expression analysis. Comparison with Zybailov et al.'s data on differential expression in YPD vs. YMD. |Z|>2.58.

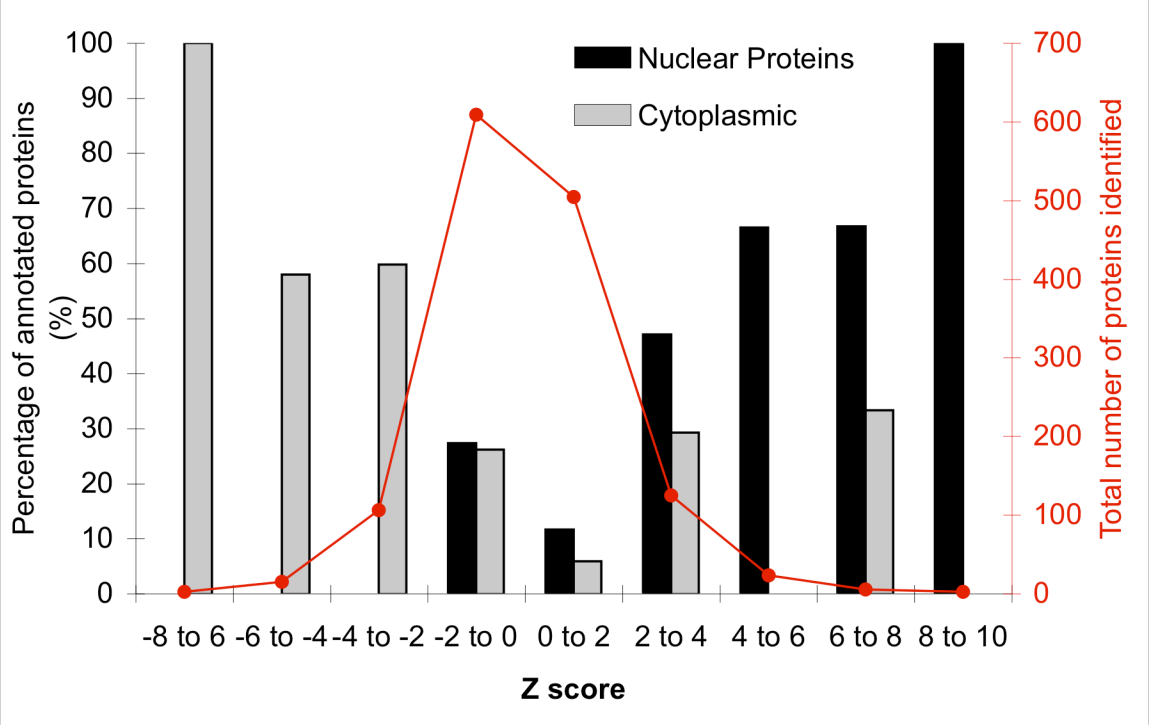


Figure S16. Mouse. Differential protein expression analysis in mouse T-lymphocytes. Known nuclear and cytoplasmic proteins are enriched in the respective sub-cellular fractions (measured by Z-scores)(1391 proteins).

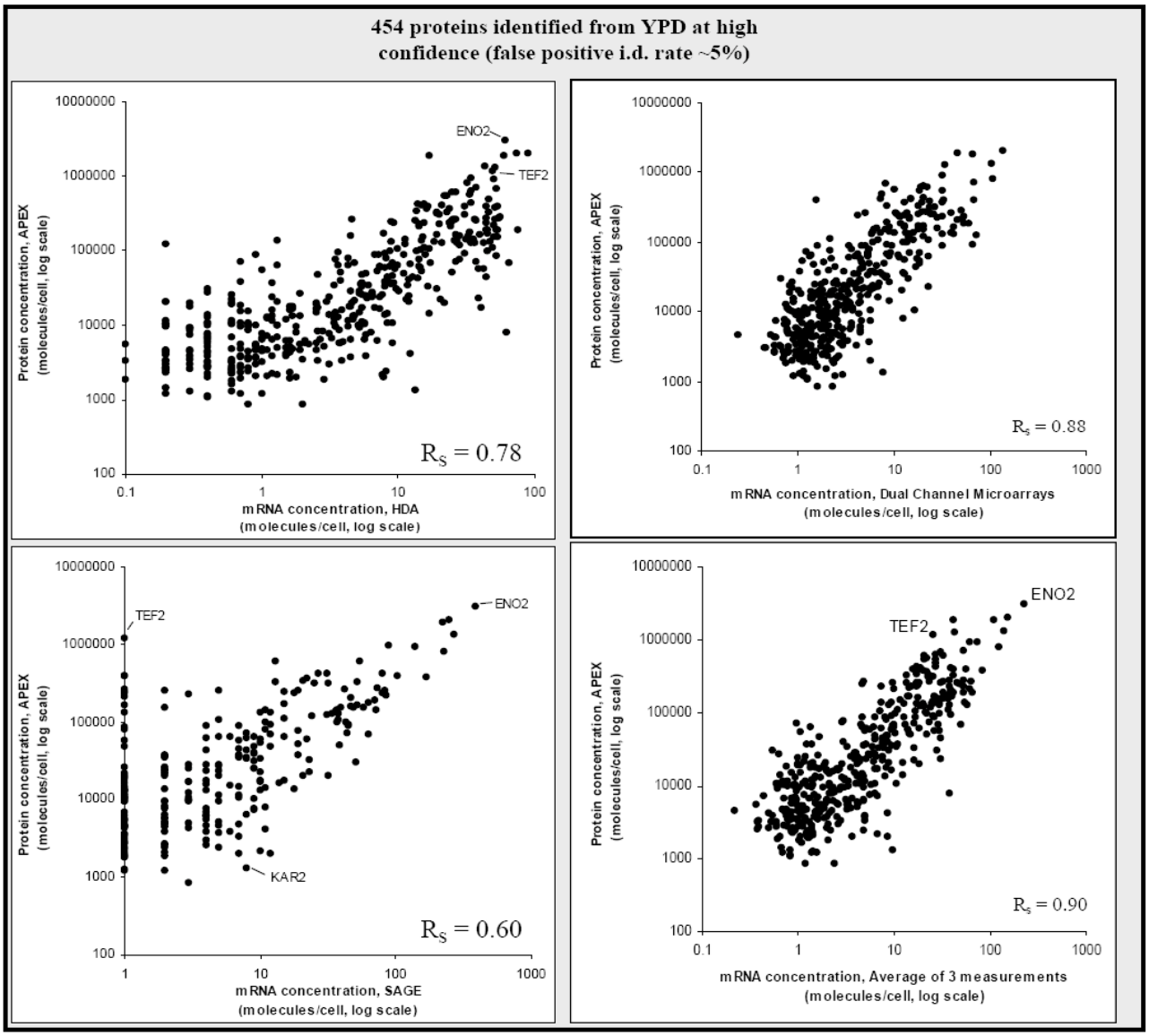


Figure S17. Yeast . APEX-based protein abundances vs. three individual mRNA measurements and their average.

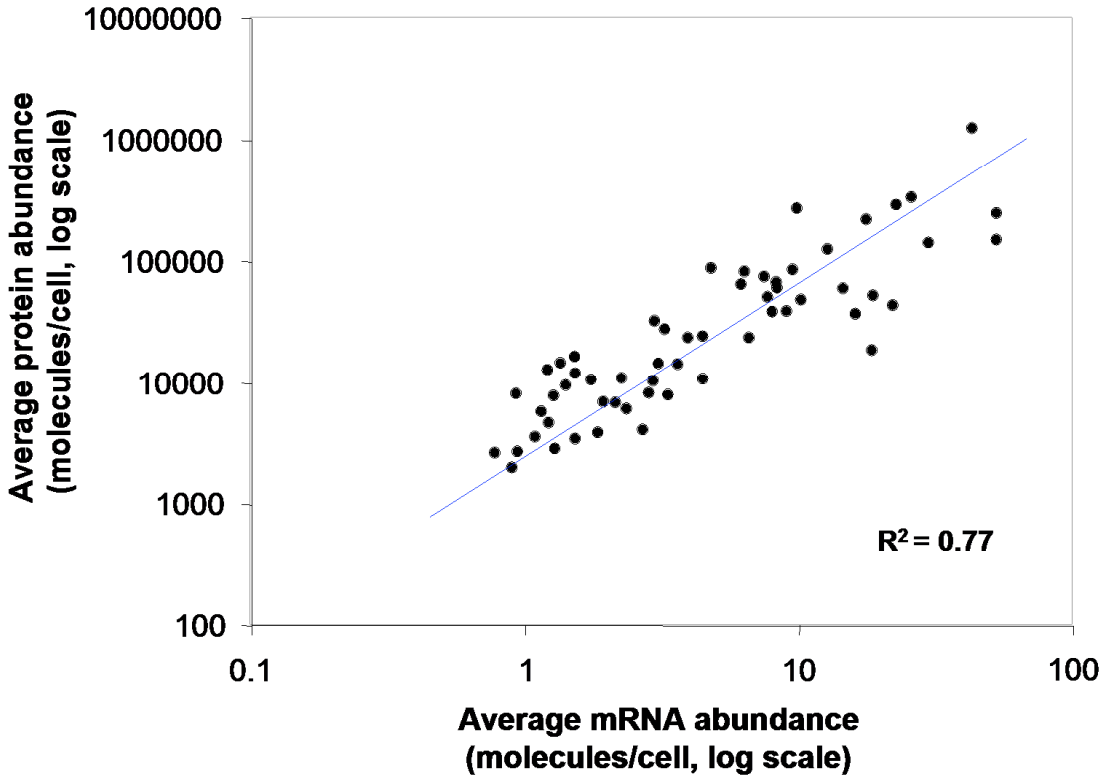
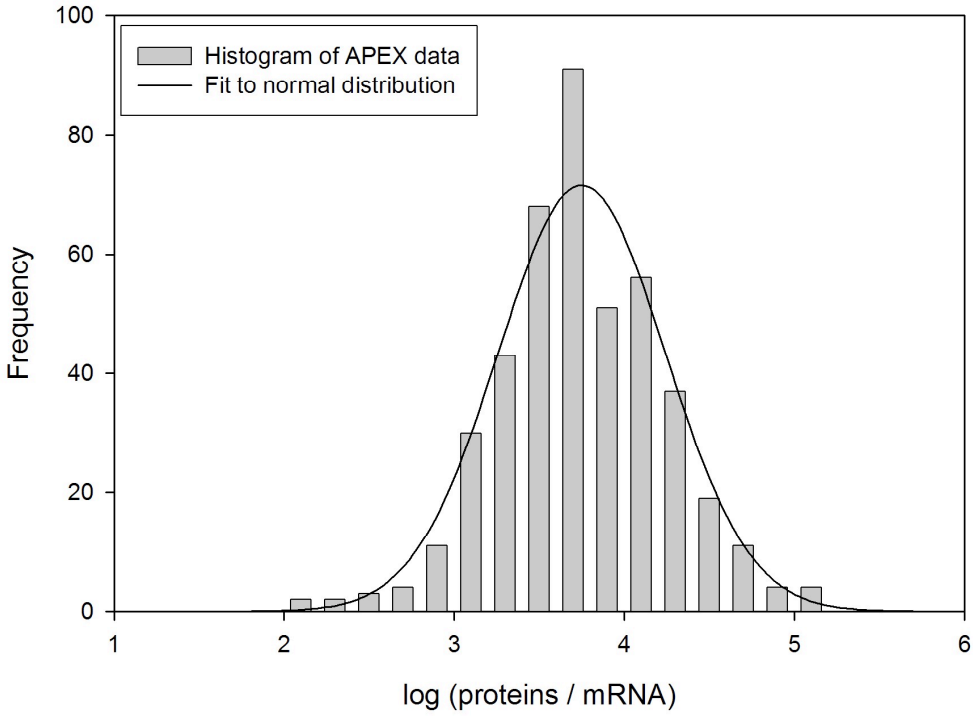


Figure S18. Yeast. High-confidence protein data set. The correlation between protein and mRNA abundance is conserved for a high-confidence data set (N=58).

A.



B.

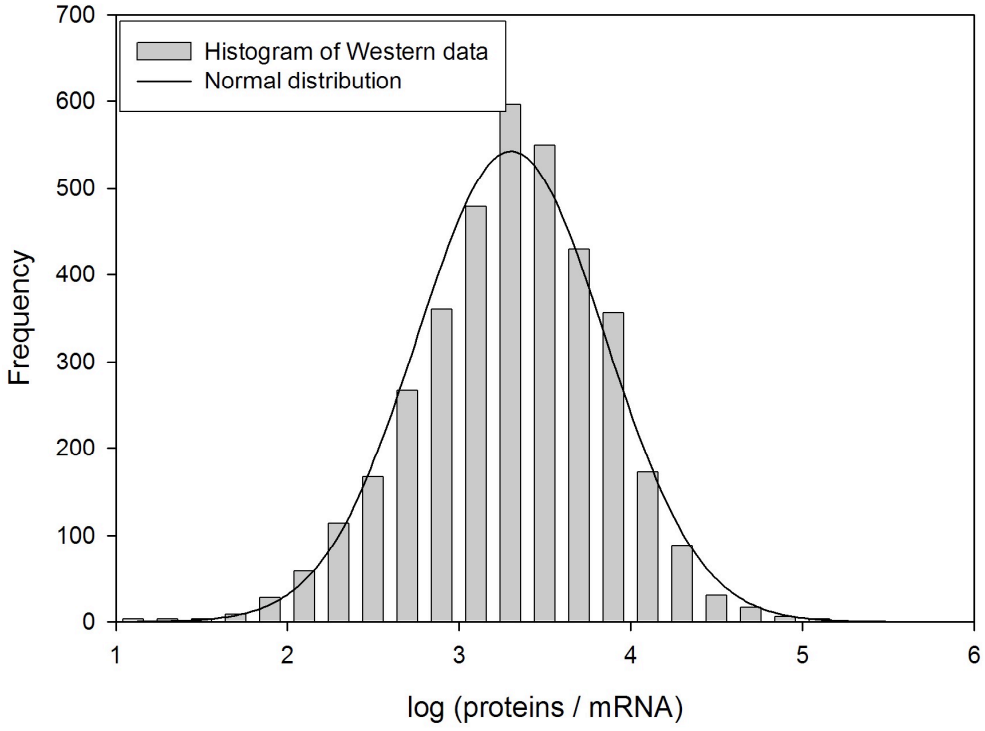


Figure S19.Yeast. Protein per mRNA ratios are log-normally distributed, even for individual data sets.

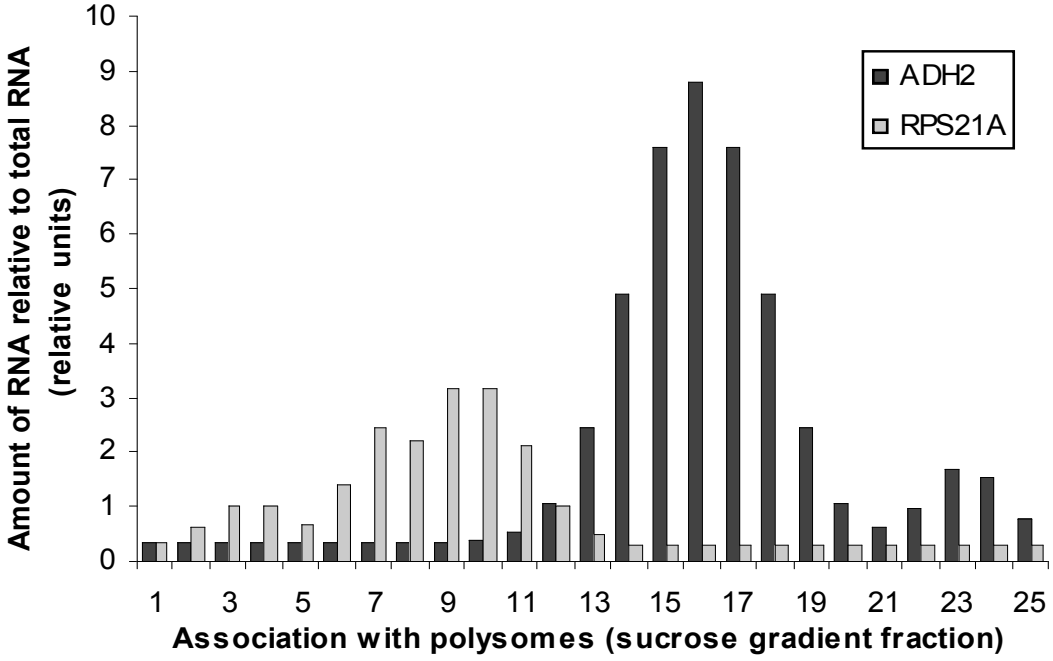


Figure S20. Yeast. Ribosomal occupancy for two proteins with extreme protein per mRNA ratios [Arava *PNAS* 2003; Arava *Nucl. Acids Res.* 2005].

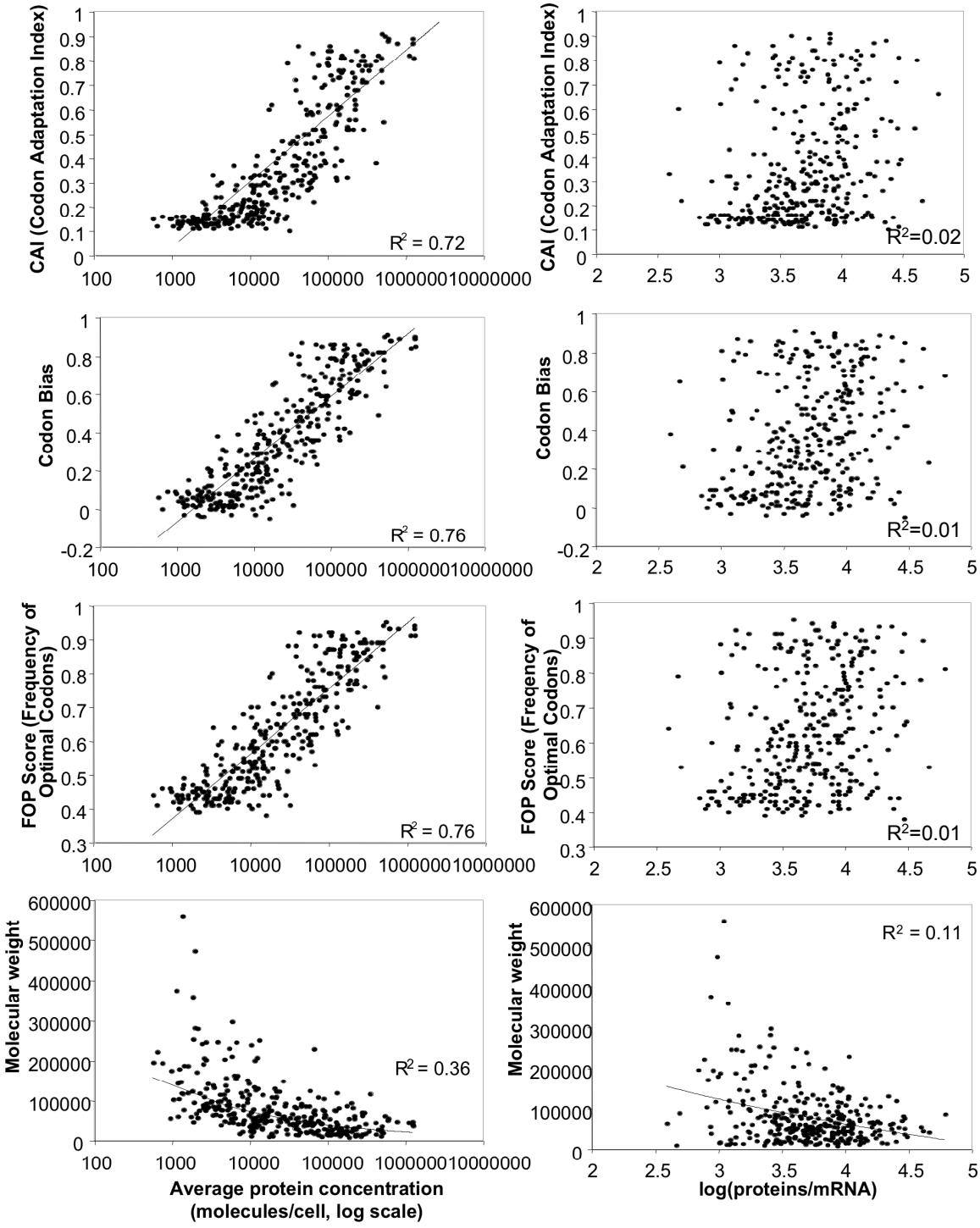


Figure S21. Yeast. Absolute protein abundance (left) correlates with some protein characteristics, while this is not the case for the protein per mRNA ratio (right), except for molecular weight.

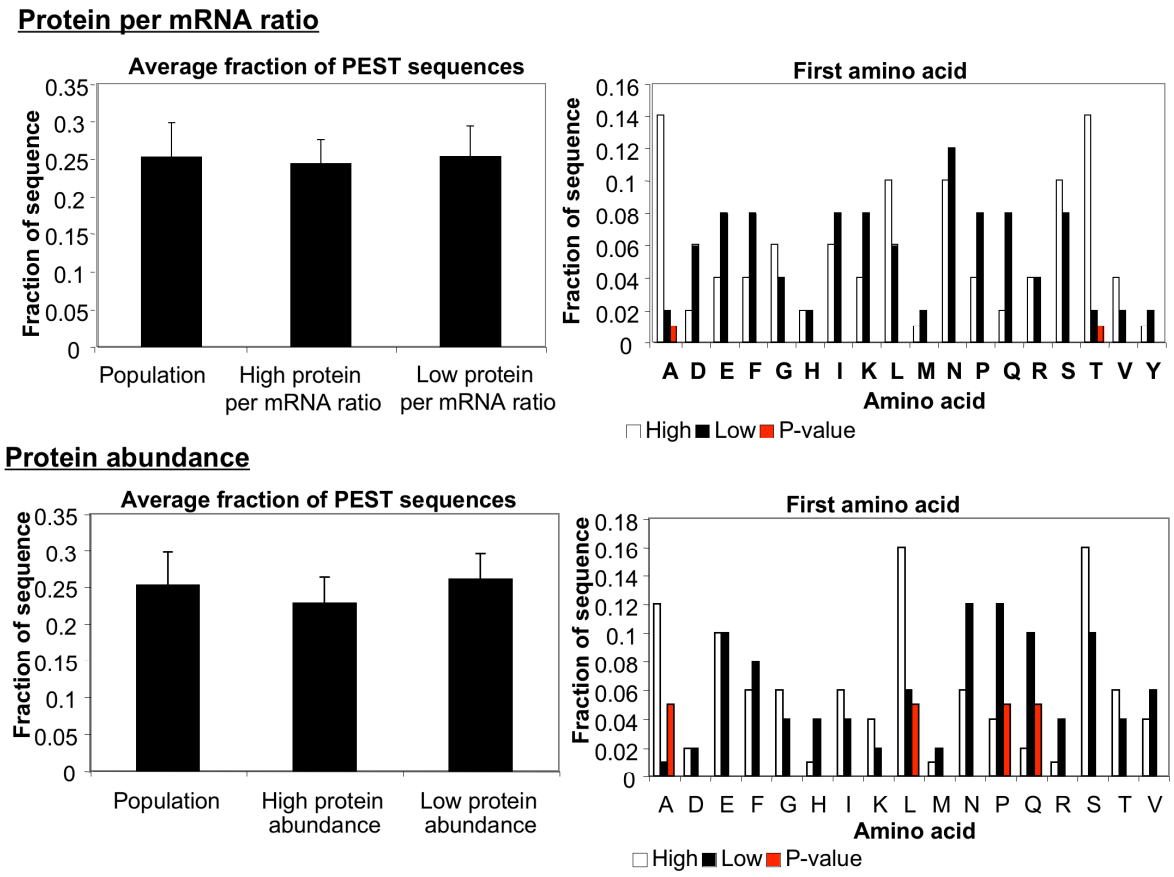


Figure S22. Yeast. Sequence analysis. proteins of high abundance and/or high protein per mRNA ratios have a slightly biased amino acid composition.

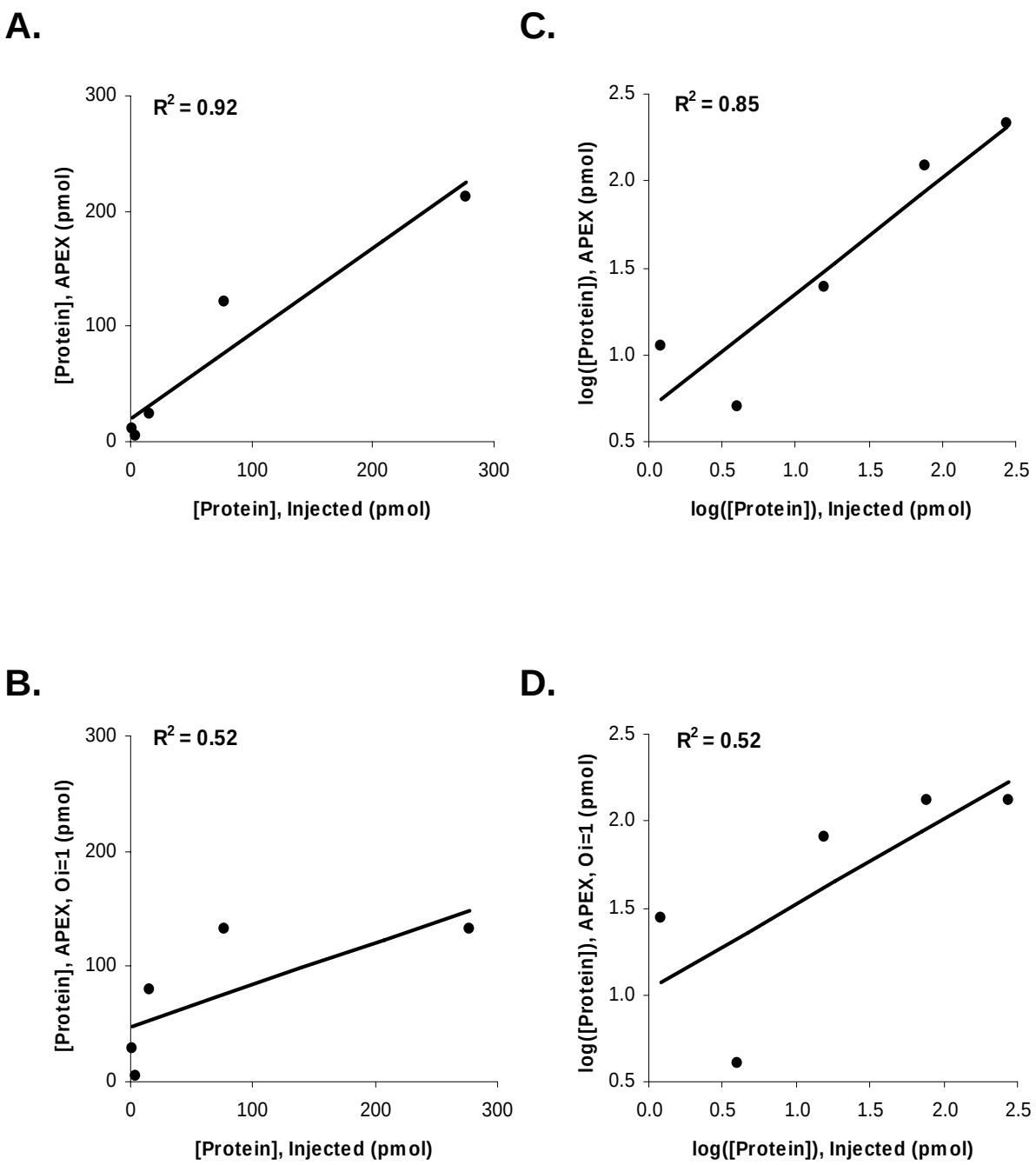


Figure S23. Synthetic 5 protein mixture. APEX protein concentrations vs. known abundances (A,C) and APEX calculated with flat priors ($O_i = 1$) vs. known abundances (B,D), plotted as linear (A,B) and log-transformed (C,D) abundances.

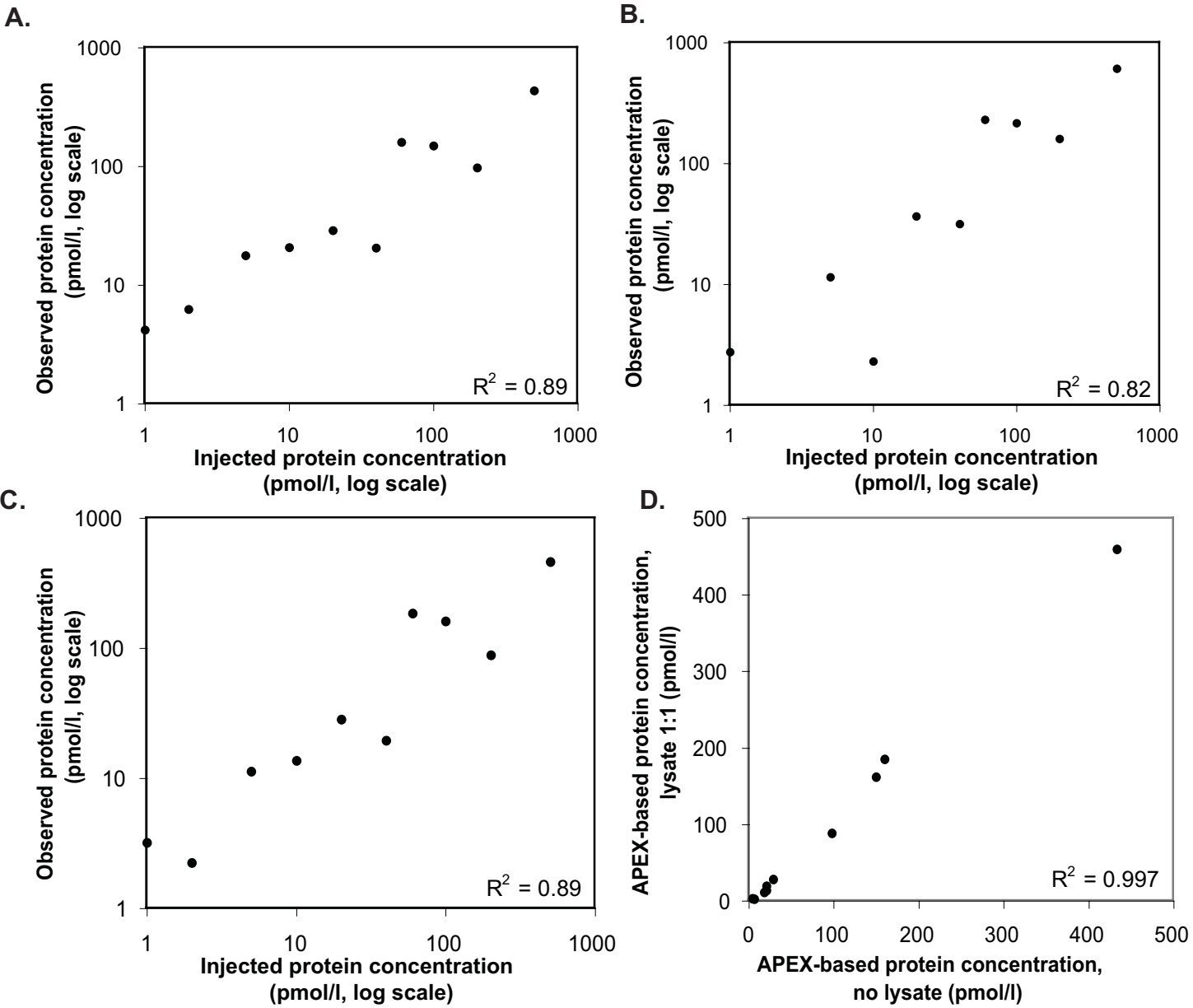


Figure S24. Analysis of synthetic mixture of 10 proteins of known concentrations, without lysate (A.) or with yeast cell lysate added in ratios 1:10 (B.) or 1:1 (C.). Graphs A.-C. compare the injected concentrations with APEX-based estimates, graph D. compares APEX-based estimates from A. (no lysate) with C. (lysate 1:1).

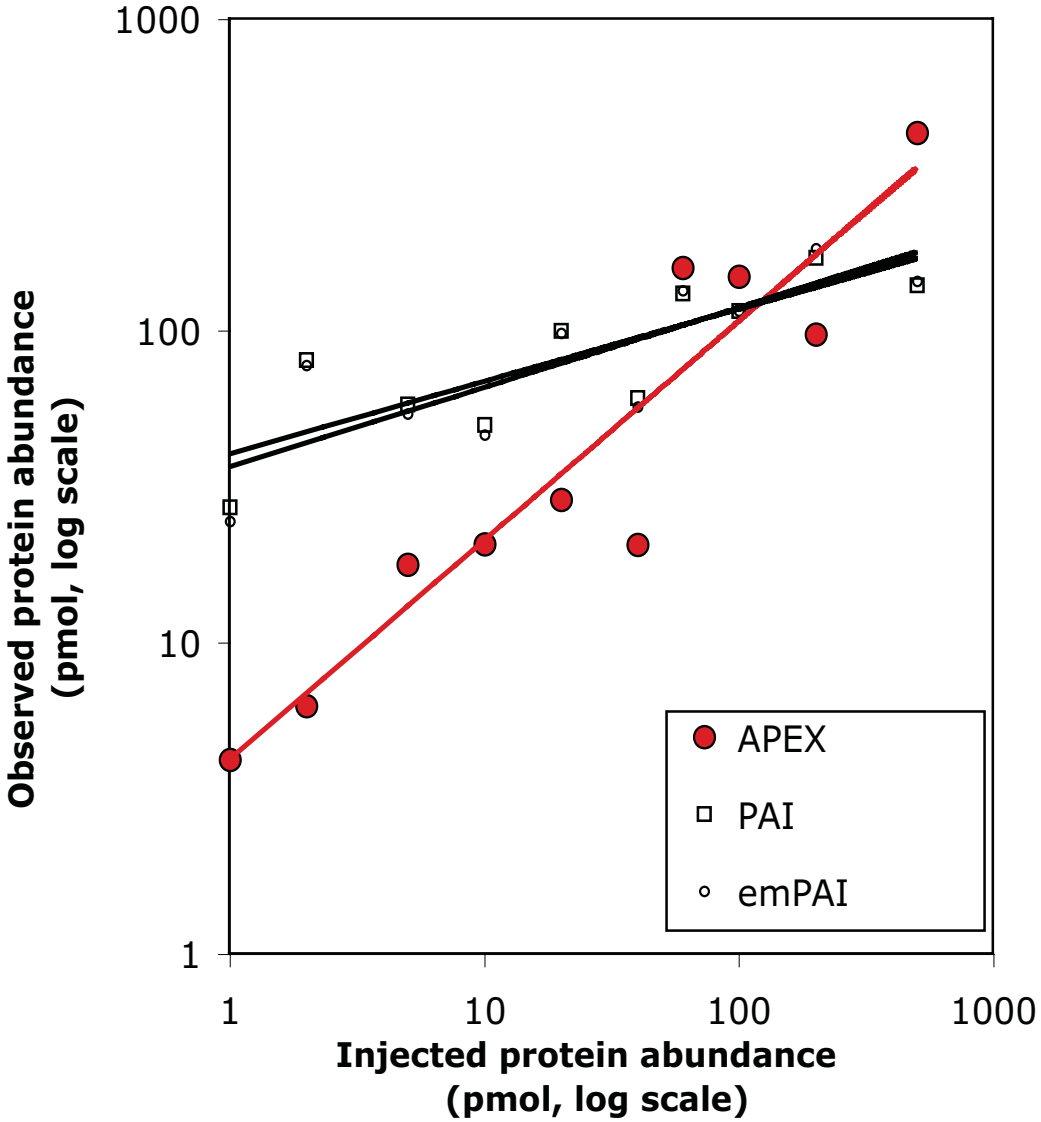


Figure S25. Analysis of synthetic mixture of 10 proteins of known concentrations, without lysate; comparison of APEX-based estimates with other methods.