Comparing protein abundance and mRNA expression levels on a genomic scale

Dov Greenbaum[1], Christopher Colangelo[2,4], Kenneth Williams, [2.4, ‡] & Mark Gerstein[2,3,‡]

‡Co-corresponding authors

[1]Department of Genetics,
[2]Department of Molecular Biophysics & Biochemistry
[3]Department of Computer Science
[4]HHMI Biopolymer Laboratory and W. M. Keck Foundation
Biotechnology Resource Laboratory
Yale University
P.O. Box 208114
New Haven, CT 06520-8114, USA.

Abstract

We survey methods used in determining protein abundance levels (divided into 2D-electrophoresis and mass-spectrometry based methods) and recent attempts to correlate protein abundance and mRNA expression. We discuss the results of these comparisons, focusing on yeast. In the process we merge together much of the available yeast protein abundance data, and use this larger data set to find correlations between protein and mRNA expression data both globally, and in terms of smaller categories of proteins.
(Supplemental information is available on http://bioinfo.mbb.yale.edu/expression/mrna-v-protein/)

Text

Although some of the underlying technology for protein abundance quantification was introduced almost thirty years ago [1,2], there has recently been a significant increase in the development of new tools for determining protein abundance. Concurrently, mRNA expression analysis tools are becoming more mainstream. The quantification of both of these populations is not an exercise in redundancy; measurements taken from mRNA and protein levels are complementary and are both necessary for a complete understanding of how the cell works [3]. Additionally, since mRNA is eventually translated into protein, one might assume that there should be some sort of correlation between the level of mRNA expression and that of protein abundance. Or there may not be any correlation.

There are two commonly used high throughput methods for measuring mRNA expression, microarrays and Affymetrix chips, both have been extensively reviewed elsewhere[4-6]. There are also two basic methods for determining protein abundance: (i) those based on two dimensional electrophoresis (2DE), and (ii) mass spectrometric methods. We provide a succinct review of these technologies, and recent efforts to determine correlations between quantified protein abundances and mRNA expression.

METHODS FOR DETERMINING PROTEIN LEVELS

Two Dimensional Electrophoresis (2DE)

Determining relative protein expression by conventional 2DE requires isoelectric focusing, SDS-polyacrylamide gel electrophoresis, staining, fixing, densitometry, and careful matching of the same spots on two or more gels. Differentially expressed spots are then excised, enzymatically digested, and the resulting peptides identified using mass spectrometry. An attractiveness of this approach is the low capital equipment cost, but a high level of expertise is needed to obtain reproducible gels and 2DE is generally limited to proteins that are neither too acidic, basic, or hydrophobic, and that are between 10-200 kDa. Additionally, this approach only detects proteins expressed at relatively high levels and that have long half-lives [7,8]. Using a 40µg yeast lysate, the average protein abundance detected was 51,200 copies/cell with no proteins detected with abundances <1,000 copies/cell [8]. Since 1,500 spots were resolved on a 1.0 pH unit gel[8], several gels covering different pH ranges would be needed to resolve a whole cell lysate. Given these limitations this technology has limited potential for large scale proteome analysis[8].

2D Fluorescence difference gel electrophoresis (DIGE) utilizes mass- and charge-matched, spectrally resolvable fluorescent dyes (*e.g.,* Cy3 and Cy5) to label two different protein samples *in vitro* prior to 2DE. Its main advantage over conventional 2DE is that both the control and experimental sample are run in the *same* polyacrylamide gel. These samples are then imaged separately but can be perfectly overlaid without "warping". This substantially raises the confidence with which protein changes between samples can be detected and quantified. Changes in relative level of protein expression may be detected that are as little as 1.2-fold for large volume spots[9]. Because detection is based on fluorescence, DIGE has a large dynamic range of about 10,000, which permits differential expression analysis of relatively low copy number proteins [9]. The limit of detection of DIGE for quantifying protein expression ratios is between 0.25 - 0.95 ng protein, which is similar to that for silver staining [9,10]. In a recent study [11], the relative level of expression of ~1,050 protein spots was compared in 250,000 laser dissected normal versus esophageal carcinoma cells. This analysis identified 58 spots that were up-regulated by >3-fold and 107 that were down-regulated by >3-fold in cancer cells.

Mass Spectrometric Approaches to Protein Profiling

Peptide/protein Disease Biomarker Discovery. Current approaches involve batch chromatography, matrix assisted laser desorption ionization mass spectrometry (MALDI-MS) and statistical analysis of large numbers of disease versus normal serum or other biological samples. Most recent studies have relied on surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS)[12,13]. The SELDI approach[13] involves using a gold coated chip with eight or sixteen 2 mm spots that are modified with chromatographic surfaces (*e.g.*, anionic, cationic, hydrophobic, etc). After spotting a few microliters of serum, contaminants and salt are removed by washing with water, and the target dried by adding a MALDI matrix solution like α-cyano-4-hydroxy-cinnamic acid. In a study by Petricoin *et al* [14] SELDI-MS analysis of serum from 50 control and 50 case samples from patients with ovarian cancer resulted in identifying 5 peptide biomarkers that ranged in size from 534 to 2,465 Da. The pattern formed by these markers was then used to correctly classify all 50 ovarian cancer samples in a masked set of serum samples from 116 patients who included 50 patients with ovarian cancer and 66 unaffected women. Similar promising results have been reported in studies of serum samples from breast and prostrate cancer patients. (Li *et al* [15] and Adam *et al*[12]) While powerful, it does not, however, provide accurate relative amounts of the control versus experimental biomarker - only the relative intensity difference

Isotope coded affinity tag (ICAT)-based protein profiling. While both MALDI-MS based disease biomarker discovery and DIGE comparatively profile the *naturally* occurring forms of peptides and proteins, ICAT analysis profiles the relative amounts of cysteine-containing peptides derived from tryptic digests of protein extracts. Since only a single tryptic peptide is needed to quantify the expression of the corresponding parent protein, the ICAT reagent utilizes a thiol protein reactive group that attaches both a biotin tag and either nine C12 (light) or nine C13 (heavy) atoms to each cysteine residue. Following derivatization of the control protein extract with [^{12}C]-ICAT reagent and the experimental extract with the [^{13}C]-ICAT reagent, the pooled samples are subjected to trypsin digestion followed by cation exchange chromatography. LC/MS/MS analysis is then used to identify ICAT peptide pairs and to quantify the relative [12C]/[13C] ratios. It is important to note that the ICAT approach provides the relative expression ratios of *individual* proteins. It does not provide absolute protein concentrations nor does it provide the ratio of the concentrations of one protein to another. A nice feature of this approach is that the *in vitro* incorporation of a stable isotope into one of the two samples being compared obviates the need to analyze by MS the control and experimental samples separately. While a tryptic digest of a whole cell human protein extract might produce >500,000 peptides, less than 100,000 of these might be expected to contain cysteine. Based on a search of the Swiss Database, <5% of human proteins lack cysteine and would be missed.

The resulting ICAT data is analogous to that obtained by the use of two different fluorescent dyes in DNA microarray analysis of mRNA or DIGE analysis of protein expression. The largest number of proteins profiled by this approach from a single sample are the 491 proteins contained in microsomal fractions of naïve and *in vitro* differentiated human myeloid leukemia cells[16].

Multidimensional protein identification technology (MudPit) is similar to ICAT in that it utilizes cation exchange prefractionation followed by reverse phase (RP) HPLC separation and MS/MS analysis[17]. In contrast to the ICAT approach, MudPit technology analyzes the *entire* mixture of tryptically digested proteins and utilizes tandemly coupled (cation exchange followed by reverse phase) columns. A specific sub-set of peptides is eluted from the cation exchange column using a step gradient of increasing salt concentration onto the front of the RP column. Peptides are then eluted from the RP column and enter the mass spectrometer for analysis. After the RP gradient is complete, the next step of the salt gradient releases another sub-set of peptides from the cation exchange column onto the RP column and the process repeats itself. Using this approach on the yeast proteome, Wolters et al[11,18] identified 5,540 unique peptides from 1,484 proteins and demonstrated a dynamic range of detection of 10,000. This method has been extended to comparative protein profiling by using *in vivo* N^{14}/N^{15} metabolic labeling[18,19] In Washburn et. al[18], *S. cerevisiae* was grown in both ^{14}N and ^{15}N minimal media and then 2,264 peptides and 872 proteins were uniquely identified. Also, accurate ^{14}N/^{15}N quantitation was determined for each peptide with an average standard deviation of 30%.

Comparison of mRNA and Protein Levels

Even with these significant developments in the technologies used to quantify protein abundance over the past couple years, protein identification and quantification still lags behind the high throughput experimental techniques used to determine mRNA expression values. Yet, while mRNA expression values have shown their usefulness in a broad range of applications, including diagnosis and classification of cancers [20,21], these results are almost certainly only correlative, rather than causative; in the end it is, most probably, the concentration of proteins and their interactions that are the true causative forces in the cell, and it’s the corresponding protein quantities that we ought to be looking at.

Primarily due to the limited ability to measure protein abundances, researchers have tried to find correlations between mRNA and protein expression in the hope that they could determine protein abundance levels from the more copious mRNA experiments. Alternatively, if there is definitively no correlation between mRNA and protein data, both quantities can be used as independent sources of information in machine learning algorithms.

To date, there have been only a handful of efforts to find correlations between mRNA and protein expression levels, most notably in human cancers and yeast cells; for the most part, they have reported only minimal and limited correlations.

One of the earliest analyses on correlation looked at 19 proteins in the human liver. Anderson et al[22] found a somewhat positive correlation of 0.48. Another limited analysis of three genes MMp-2, MNP-9 and TIMP-1 in human prostate cancers showed no significant relationship [23]. An additional cancer study [24] showed a significant correlation in only a small subset of the proteins studied . Conversely, Orntoft et al [25] found highly significant correlations in human carcinomas when looking at changes in mRNA and protein expression levels

Protein and mRNA correlations in Yeast

Many of the present efforts in correlating mRNA and protein expression have been conducted in yeast using two dimensional electrophoresis techniques, in particular:
2DE-1: Gygi et al [7] found that even similar mRNA expression levels could have a wide range (up to 20 fold difference) of protein abundance levels and vice versa.
2DE-2: These results contrast with Futcher et al's [26] relatively high levels of correlations (r = 0.76) after transforming the data to normal distributions.
Merged data set-1: In a previous analysis, we merged the data from both of these 2DE datasets, comparing this new larger protein abundance set with a comprehensive mRNA expression data set. This mRNA expression reference set was constructed through iteratively combining, in a non-trivial fashion, three Affymetrix sets and a SAGE dataset [27]. Using these new reference data sets, we were able to do an all-against-all comparison of mRNA and protein expression levels in addition to a number of analyses comparing protein and mRNA expression using smaller, but broad categories [27,28].

Given the difficult, laborious, and limiting nature of 2DE analysis, much of the newer protein abundance determinations have been done using the MudPit and derivative technologies. One caveat: Mudpit data on its own, is semi-quantitative in that the number of peptides determined is relative to the actual protein abundance within the cell[29].

MudPit-1: Washburn et al[29] used MUDPIT to analyze and detect 1484 arbitrary proteins- i.e. they were able to detect a somewhat random sampling of proteins independent of their abundance, localization, size or hydrophobicity. In a further experiment the authors, comparing expression ratios for both proteins and mRNA levels, found that although they could not find correlations for individual loci, they could find overall correlations when looking at pathways and complexes of proteins that functioned together [19].
MudPit-2: Peng et al [30] analyzed 1504 yeast proteins with a false positive rate - misidentification of a protein- of less than 1%. In their analysis they contrasted their methodology with that of Washburn et al with which there was significant overlap.

New Merged Dataset: Merged data set-2

Expanding upon our previous merged data set, we constructed a new merged data set using the two 2DE and two Mudpit data sets presented above. Succinctly (more information is available on our web site: http://bioinfo.mbb.yale.edu/expression/mrna-v-protein/), we transformed each of the protein abundance data sets into more quantitative data via fitting them individually onto the reference mRNA expression data set. The Mudpit-1 dataset was also fit onto the more finely-grained Mudpit-2 dataset. Each of the new, fitted datasets was then inversely transformed back into protein space. These datasets were then combined into a larger reference data set; when we had more than one abundance value for an ORF, we chose the value from the dataset according to a proscribed quality-ranking (see figure 1 caption). The resulting set contained protein abundance for ~2000 ORFs. (Although some may argue that the less quantitative nature of some of the MudPit data should not be used to compare with the mRNA data, we feel that the merging process creates a more quantitative and representative data set.) Using this data we could compare, globally, mRNA expression and protein abundance (Figure 1a) as well as looking at smaller, broad, categories -i.e. functional categories or localization (See Figure 1b,c). In particular, we show that some localization categories, e.g. the nucleolus, have significantly higher correlations than the global correlation. Other localizations may present less of a correlation between mRNA and protein data, e.g. the mitochondria, possibly reflecting the heterogeneous nature and function of the organelle. In terms of MIPS functional categories, we show that while some categories, such as cell rescue, show a lower correlation than the whole merged set, other functional categories, such as cell cycle, show a significant increase in correlation. Logically, this increased correlation reflects the co-regulated nature of the proteins in this functional category.

Reasons for the absence of correlation

There are presumably at least three reasons, for the poor correlations generally reported in the literature for the level of mRNA versus protein expression – they may not be mutually exclusive: (i) there are many complicated and varied post translational mechanisms that are involved in turning mRNA into protein that are not yet sufficiently well defined to be able to compute protein concentrations from mRNA; (ii) proteins may differ substantially in their *in vivo* half lives; and/or (iii) there is a significant amount of error and noise in both protein and mRNA experiments that limit our ability to get a clear picture [31,32].

Examining the first option that there are a number of complex steps between transcription and translation, we looked at correlations between mRNA and protein abundance for those ORFs that had varied or steady levels of mRNA expression over the course of the cell cycle [33]. To normalize for the varied degrees of expression for different ORFs, we took the standard deviation divided by the average expression level as representative of the variation of each ORF over the course of the yeast cell cycle. (Figure 2)

Broadly, the cell can control the levels of protein at the transcriptional level and/or at the translational level. Logically, we would assume that those ORFs that show a large degree of variation in their expression are controlled at the transcriptional level – the variability of the mRNA expression is indicative of the cell controlling mRNA expression at different points of the cell cycle to achieve the resulting and desired protein levels. Thus we would expect, and we found (r= 0.89), a high degree of correlation between the mRNA and protein levels for these particular ORFs; the cell has already put significant energy into dictating the final level of protein through tightly controlling the mRNA expression, we assume that there would then be minimal control at the protein level. In contrast, those genes that show minimal variation in their mRNA expression throughout the cell cycle are more likely to have little if no correlation with the final protein level; the cell would be controlling these ORFs at the translational and/or post-translational level, with the mRNA levels being somewhat independent of the final protein concentration. We found only minimal correlation between protein and mRNA expression for these ORFs (r = 0.2)

Further, we found that those ORFs that have higher than average levels of occupancy, that is that a large percentage of their cellular mRNA concentration is associated with ribosomes (i.e. being translated), have well correlated mRNA and protein expression levels. (Figure 2) These cases probably represent a situation wherein the cell, having significantly controlled the mRNA expression to produce a specific level of protein, will probably not employ mechanisms to control the translation. Alternatively, those proteins that have very low occupancy rates have uncorrelated mRNA and protein expression; thus, given that the cell has not controlled the mRNA expression, it will dictate the resulting protein levels through rigorous controls on the translation (i.e. through tight limits on occupancy) of these genes[34].

A second option responsible for a general lack of correlation between mRNA and protein abundance may be the result of varied protein synthesis and degradation. Protein turnover can vary significantly depending on a number of different conditions[35]; the cell can control the protein level in the cell through the rates of degradation or synthesis for a given protein. There is significant heterogeneity even within similar functioning proteins[36].

Recent efforts have been made to computationally measure these rates[37]. Simplistically it can be presumed that the change in a protein's concentration over time will be equal to the rate of translation minus the rate of degradation. Similar to concepts in chemical kinetics, we can approximate this equation: $dP(i,t)/dt = SE(i,t) - DP(i,t)$, where P is protein abundance i at time t, E is the mRNA expression level of protein P, S is a general rate of protein synthesis per mRNA, and D is a general rate of protein degradation per protein[37]. Additionally there are some experimental methods that can also be used to measure turnover and translational control of protein levels[36-39].

Due to degenerate nature of the genetic code, there are many synonymous codons, i.e. they translate into the same amino acid. Given that the cell is biased in its usage of synonymous codons (i.e. the usage of a subset of codons results in a higher level of mRNA expression possibly due to cellular tRNA levels[40]), the Codon Adaptation Index (CAI), a measurement of codon usage, can be used to predict the expression of a gene[41] (new parameters for this model were recently calculated with some improvement in predictive strength [42]). It is thought that the CAI will correlate differently with mRNA levels than with protein abundance levels due, partially, to protein turnover rates [43]. Ranking the ORFs in terms of their CAI, we found that those ORFs in that ranked the highest in terms of CAI, while not showing a very strong correlation between mRNA and protein levels, still showed a significantly higher correlation than the ORFs that were ranked as having the lower CAI values ($r = 0.48$ v 0.02). The low correlations reflect the fact that CAI will correlate differently for protein and mRNA values because of the additional cellular controls on protein translation, i.e. the affect of protein turnover rates. Still, the sizable difference in correlations between the two groups of high and low ranking CAI values (Figure 2) shows that there is some relationship between mRNA and protein values, possibly indicating that highly expressed genes tend to result in a more correlated level of protein abundance than more lowly expressed ones.

Correlations *have* been found between the mRNA expression of different protein subunits within protein complexes[44]. This implies that there should be, in practice, a correlation between mRNA and protein abundance, as these subunits also have to be available in stoichiometric amounts of proteins for the complexes to function. Thus, we believe that a major limitation to finding correlations is the degree of natural and manufactured systematic noise in mRNA and protein expression experiments. There is a continued effort to both describe and reduce this noise [45]. Meanwhile, in an attempt to get around the noise one could look at broad categories of proteins (e.g. groups defined by function, structure, or localization) such that the background noise is cancelled out to some degree, to discover correlations[27].

While proteomics is still in its infancy - given the pace of technological advancement in protein quantification, mRNA expression analysis and noise reduction- more comprehensive correlation studies will soon be feasible. This will allow for more robust analyses of the relationship between mRNA expression and protein abundance values. Finally, to be fully able to understand the relationship between mRNA and protein abundances, the dynamic processes of synthesis and degradation of transcripts has to be better understood; is the protein level changing because of a change in transcript syntheses, degradation, or protein turnover? These questions need to be looked into further before we can appreciate the relationship between mRNA and protein.

Figure 1a shows an XY plot comparing our mRNA reference expression set[27] with a newly compiled protein abundance data set. This data set is the result of iteratively fitting 2 MudPit data sets, (MudPit 1[30] and MudPit 2[29]) and two 2 Dimensional Electrophoresis data sets (2DE1[7] and 2DE2[26]). Given the semi quantitative nature of the Mudpit data[29], we transformed the data into a more quantitative set via fitting each set individually onto our reference mRNA expression data set. In addition, we fit the Mudpit-1 dataset onto the more finely-grained MudPit-2 data set. Each of the datasets was then moved back into 'protein space' using an inverse transformation. This inverse transformation was derived from the 2DE-1 set, as this set has the most precise values. These data sets were then combined into the new reference abundance data set. In cases where there were overlapping values for a given ORF we used the data set in accord with the following ordering: 2DE-1, 2DE-2, Mudpit-2, Mudpit-1. The resulting reference protein abundance set (N = 2044) had a correlation of 0.66 with the mRNA reference data set.

Figure 1B,C Additionally, we show that when looking at specific subsets, (i.e. subcellular localization[46] or functional groups[47]) we can find both higher and lower correlations amongst these groups. The lower correlations are generally reflective of a more heterogeneous category. This analysis indicates that while correlations may be weak when looking at the global data, we tend to find higher correlations when looking at smaller well-defined subsets of ORFs.

Figure 2 shows the differences in correlation between mRNA and protein expression values using novel categories. In particular, we see significant differences when looking at the highest and lowest ranking of groups of ORFs in the following categories: Occupancy, CAI and Variability. Occupancy refers to the percentage of transcripts associated with ribosomes; we compared the correlation between the top 100 ORFs in terms of occupancy and the bottom 100. (0.78 vs. 0.30) For the CAI, the codon adaptation index, we compared the correlation between mRNA and protein for those ORFs with the highest CAI and those with the lowest (0.48 vs. 0.02). Variability refers to the normalized standard deviation (i.e. standard deviation divided by average expression level) for all the ORFs as measured by the Cho et al cell cycle expression data set [33]. Here we compared the correlations between protein abundance and mRNA expression for the most variable compared with the least variable proteins (0.89 vs. 0.20). We found significant differences between the correlations of mRNA and protein levels for the top and bottom ranking populations for each of the comparisons.

Acknowledgments. This project has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28186.

Table 1: Overview of Selected Protein Profiling Technologies

Technology	Type of Labeling Required?	Ability to Detect Many Post-translational Modifications	Biomolecules that are Optimally Quantified	Approximate Dynamic Range	Number of Proteins/Spots Quantified
2D Gel Electrophoresis	silver staining	Yes	*naturally* occurring forms of >10 kD proteins	10[9]	1,500[8]
Differential 2D Fluorescence Gel Electrophoresis (DIGE)	*in vitro* with Cy-2,3 or 5 fluorophores at primary amines	yes		10,000[9]	1,100[48]
SELDI or MALDI-MS Disease Biomarker Discovery	None	yes	*naturally* occurring forms of <10 kD proteins	25	not applicable
Isotope Coded Affinity Tag (ICAT) - LC/MS	*in vitro* with H^1/D or C^{12}/C^{13} ICAT reagent at cysteine	no	cysteine-containing tryptic peptides from digests of protein extracts	10,000[a]	496[16]
N^{14}/N^{15} - LC/MS	*in vivo* at nitrogens in amino acids	yes	tryptic peptides from digests of protein extracts	10,000[17]	872[18]

[a] Assumed to be similar to that for multidimensional protein identification.

Literature Cited

1. O'Farrell, PH: **High resolution two-dimensional electrophoresis of proteins**. *J Biol Chem* 1975, **250:**4007-21.
2. Klose, J: **Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals**. *Humangenetik* 1975, **26:**231-43.
3. Hatzimanikatis, V, Choe, LH & Lee, KH: **Proteomics: theoretical and experimental considerations**. *Biotechnol Prog* 1999, **15:**312-8.
4. Schena, M, Heller, RA, Theriault, TP, Konrad, K, Lachenmeier, E & Davis, RW: **Microarrays: biotechnology's discovery platform for functional genomics**. *Trends Biotechnol* 1998, **16:**301-6.
5. McGall, GH & Christians, FC: **High-density genechip oligonucleotide probe arrays**. *Adv Biochem Eng Biotechnol* 2002, **77:**21-42.
6. Brown, PO & Botstein, D: **Exploring the new world of the genome with DNA microarrays**. *Nat Genet* 1999, **21:**33-7.
7. Gygi, SP, Rochon, Y, Franza, BR & Aebersold, R: **Correlation between protein and mRNA abundance in yeast**. *Mol Cell Biol* 1999, **19:**1720-30.
8. Gygi, SP, Corthals, GL, Zhang, Y, Rochon, Y & Aebersold, R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology**. *Proc Natl Acad Sci U S A* 2000, **97:**9390-5.
9. Tonge, R, Shaw, J, Middleton, B, Rowlinson, R, Rayner, S, Young, J, Pognan, F, Hawkins, E, Currie, I & Davison, M: **Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology**. *Proteomics* 2001, **1:**377-96.
10. Gharbi, S, Gaffney, P, Yang, A, Zvelebil, MJ, Cramer, R, Waterfield, MD & Timms, JF: **Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system**. *Mol Cell Proteomics* 2002, **1:**91-8.
11. Zhou, G, Li, H, DeCamp, D, Chen, S, Shu, H, Gong, Y, Flaig, M, Gillespie, JW, Hu, N, Taylor, PR *et al.*: **2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers**. *Mol Cell Proteomics* 2002, **1:**117-24.
12. Adam, BL, Vlahou, A, Semmes, OJ & Wright, GL, Jr.: **Proteomic approaches to biomarker discovery in prostate and bladder cancers**. *Proteomics* 2001, **1:**1264-70.

13. Issaq, HJ, Veenstra, TD, Conrads, TP & Felschow, D: **The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification**. *Biochem Biophys Res Commun* 2002, **292:**587-92.
14. Petricoin, EF, Ardekani, AM, Hitt, BA, Levine, PJ, Fusaro, VA, Steinberg, SM, Mills, GB, Simone, C, Fishman, DA, Kohn, EC *et al.*: **Use of proteomic patterns in serum to identify ovarian cancer**. *Lancet* 2002, **359:**572-7.
15. Li, J, Zhang, Z, Rosenzweig, J, Wang, YY & Chan, DW: **Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer**. *Clin Chem* 2002, **48:**1296-304.
16. Han, DK, Eng, J, Zhou, H & Aebersold, R: **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry**. *Nat Biotechnol* 2001, **19:**946-51.
17. Wolters, DA, Washburn, MP & Yates, JR, 3rd: **An automated multidimensional protein identification technology for shotgun proteomics**. *Anal Chem* 2001, **73:**5683-90.
18. Washburn, MP, Ulaszek, R, Deciu, C, Schieltz, DM & Yates, JR, 3rd: **Analysis of quantitative proteomic data generated via multidimensional protein identification technology**. *Anal Chem* 2002, **74:**1650-7.
19. Washburn, MP, Koller, A, Oshiro, G, Ulaszek, RR, Plouffe, D, Deciu, C, Winzeler, E & Yates, JR, 3rd: **Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 2003, **100:**3107-12.
20. Golub, TR, Slonim, DK, Tamayo, P, Huard, C, Gaasenbeek, M, Mesirov, JP, Coller, H, Loh, ML, Downing, JR, Caligiuri, MA *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286:**531-7.
21. Macgregor, PF & Squire, JA: **Application of microarrays to the analysis of gene expression in cancer**. *Clin Chem* 2002, **48:**1170-7.
22. Anderson, L & Seilhamer, J: **A comparison of selected mRNA and protein abundances in human liver**. *Electrophoresis* 1997, **18:**533-7.
23. Lichtinghagen, R, Musholt, PB, Lein, M, Romer, A, Rudolph, B, Kristiansen, G, Hauptmann, S, Schnorr, D, Loening, SA & Jung, K: **Different mRNA and protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinases 1 in benign and malignant prostate tissue**. *Eur Urol* 2002, **42:**398-406.
24. Chen, G, Gharib, TG, Huang, CC, Taylor, JM, Misek, DE, Kardia, SL, Giordano, TJ, Iannettoni, MD, Orringer, MB, Hanash, SM *et al.*: **Discordant protein and mRNA expression in lung adenocarcinomas**. *Mol Cell Proteomics* 2002, **1:**304-13.
25. Orntoft, TF, Thykjaer, T, Waldman, FM, Wolf, H & Celis, JE: **Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas**. *Mol Cell Proteomics* 2002, **1:**37-45.
26. Futcher, B, Latter, GI, Monardo, P, McLaughlin, CS & Garrels, JI: **A sampling of the yeast proteome**. *Mol Cell Biol* 1999, **19:**7357-68.
27. Greenbaum, D, Jansen, R & Gerstein, M: **Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts**. *Bioinformatics* 2002, **18:**585-96.
28. Greenbaum, D, Luscombe, NM, Jansen, R, Qian, J & Gerstein, M: **Interrelating different types of genomic data, from proteome to secretome: 'oming in on function**. *Genome Res* 2001, **11:**1463-8.

29. Washburn, MP, Wolters, D & Yates, JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology**. *Nat Biotechnol* 2001, **19:**242-7.

30. Peng, J, Elias, JE, Thoreen, CC, Licklider, LJ & Gygi, SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome**. *J Proteome Res* 2003, **2:**43-50.

31. Baldi, P & Long, AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17:**509-19.

32. Szallasi, Z: **Genetic network analysis in light of massively parallel biological data acquisition**. *Pac Symp Biocomput* 1999, 5-16.

33. Cho, RJ, Campbell, MJ, Winzeler, EA, Steinmetz, L, Conway, A, Wodicka, L, Wolfsberg, TG, Gabrielian, AE, Landsman, D, Lockhart, DJ *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle**. *Mol Cell* 1998, **2:**65-73.

34. Arava, Y, Wang, Y, Storey, JD, Liu, CL, Brown, PO & Herschlag, D: **Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 2003, **100:**3889-94.

35. Glickman, MH & Ciechanover, A: **The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction**. *Physiol Rev* 2002, **82:**373-428.

36. Pratt, JM, Petty, J, Riba-Garcia, I, Robertson, DH, Gaskell, SJ, Oliver, SG & Beynon, RJ: **Dynamics of protein turnover, a missing dimension in proteomics**. *Mol Cell Proteomics* 2002, **1:**579-91.

37. Lian, Z, Kluger, Y, Greenbaum, DS, Tuck, D, Gerstein, M, Berliner, N, Weissman, SM & Newburger, PE: **Genomic and proteomic analysis of the myeloid differentiation program: global analysis of gene expression during induced differentiation in the MPRO cell line**. *Blood* 2002, **100:**3209-20.

38. Gerner, C, Vejda, S, Gelbmann, D, Bayer, E, Gotzmann, J, Schulte-Hermann, R & Mikulits, W: **Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling**. *Mol Cell Proteomics* 2002, **1:**528-37.

39. Serikawa, KA, Xu, XL, MacKay, VL, Law, GL, Zong, Q, Zhao, LP, Bumgarner, R & Morris, DR: **The Transcriptome and Its Translation during Recovery from Cell Cycle Arrest in Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2003, **2:**191-204.

40. Bennetzen, JL & Hall, BD: **Codon selection in yeast**. *J Biol Chem* 1982, **257:**3026-31.

41. Sharp, PM & Li, WH: **The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic Acids Res* 1987, **15:**1281-95.

42. Jansen, R, Bussemaker, HJ & Gerstein, M: **Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models**. *Nucleic Acids Res* 2003, **31:**2242-51.

43. Coghlan, A & Wolfe, KH: **Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae**. *Yeast* 2000, **16:**1131-45.

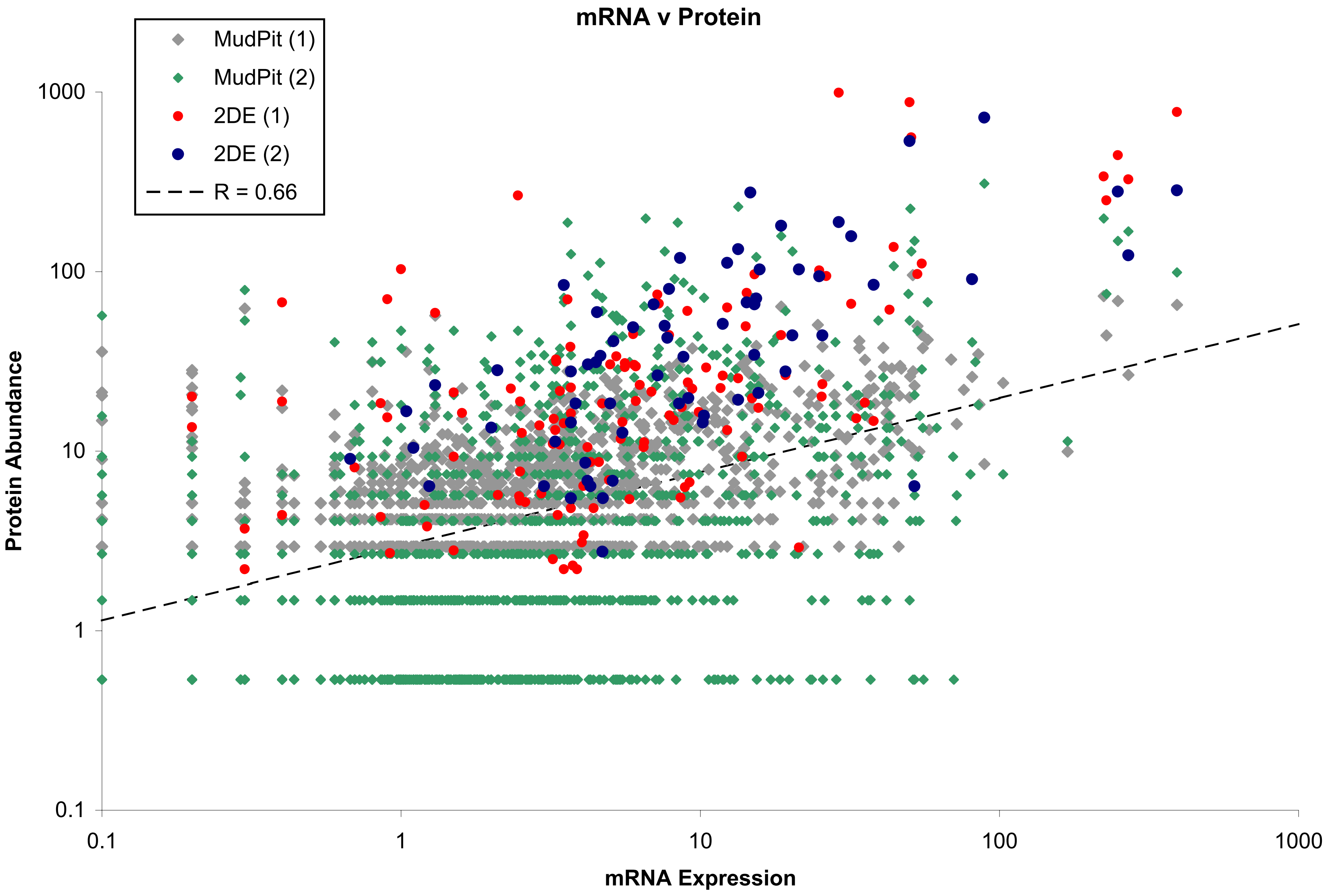
44. Jansen, R, Greenbaum, D & Gerstein, M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **12:**37-46.

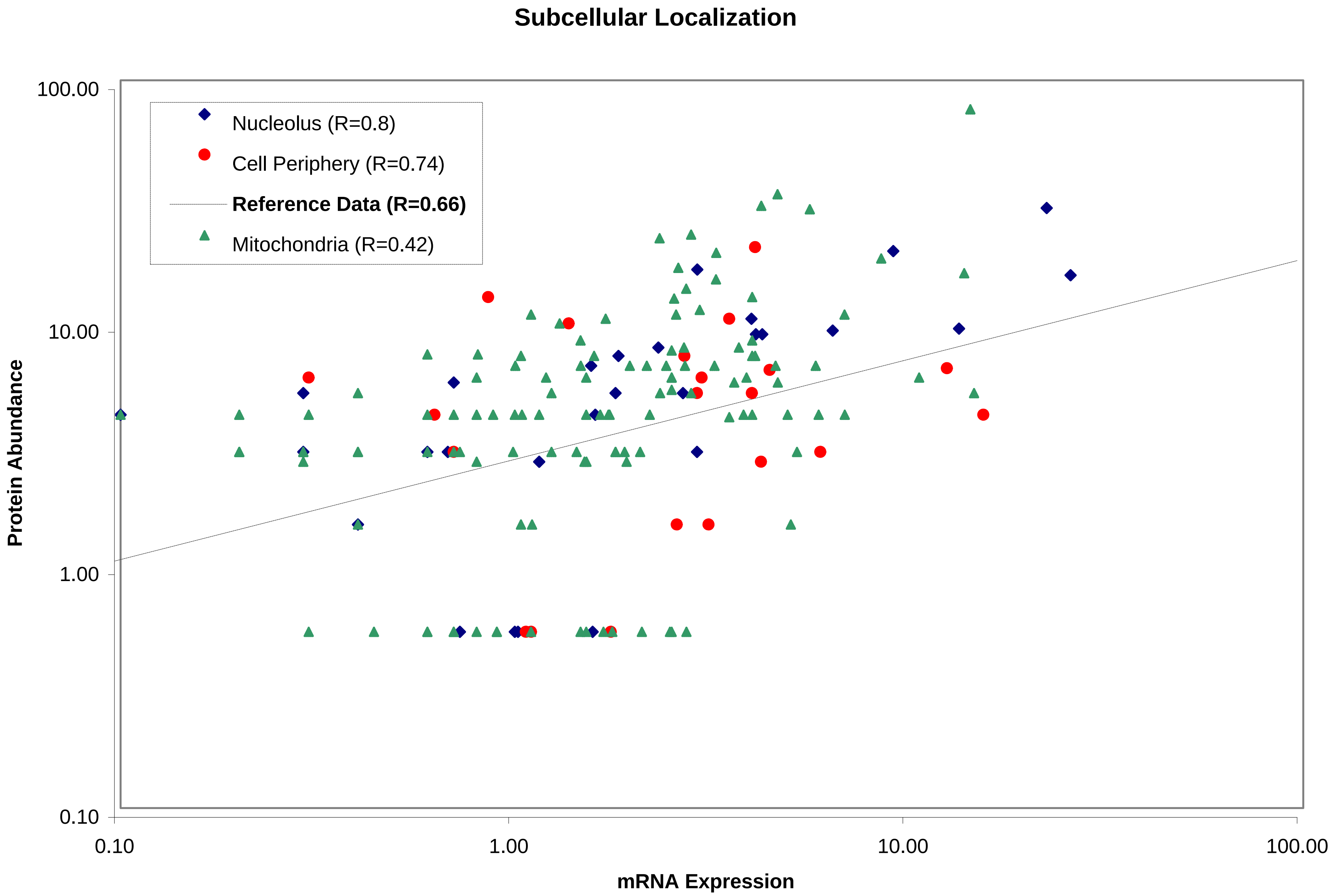
45. Qian, J, Kluger, Y, Yu, H & Gerstein, M: **Identification and correction of spurious spatial correlations in microarray data**. *BioTechniques* (In Press).

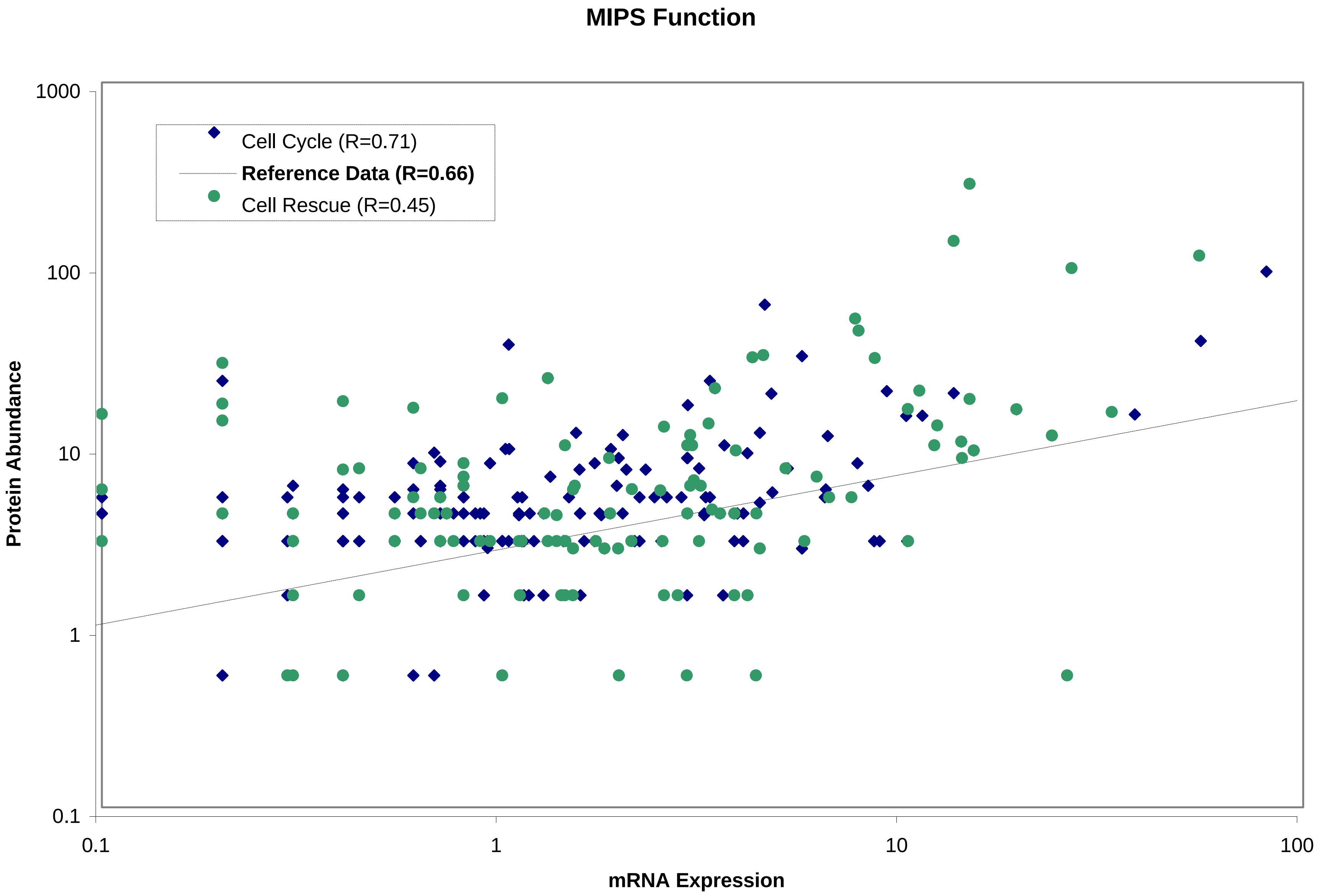
46. Kumar, A, Agarwal, S, Heyman, JA, Matson, S, Heidtman, M, Piccirillo, S, Umansky, L, Drawid, A, Jansen, R, Liu, Y *et al.*: **Subcellular localization of the yeast proteome**. *Genes Dev* 2002, **16:**707-19.

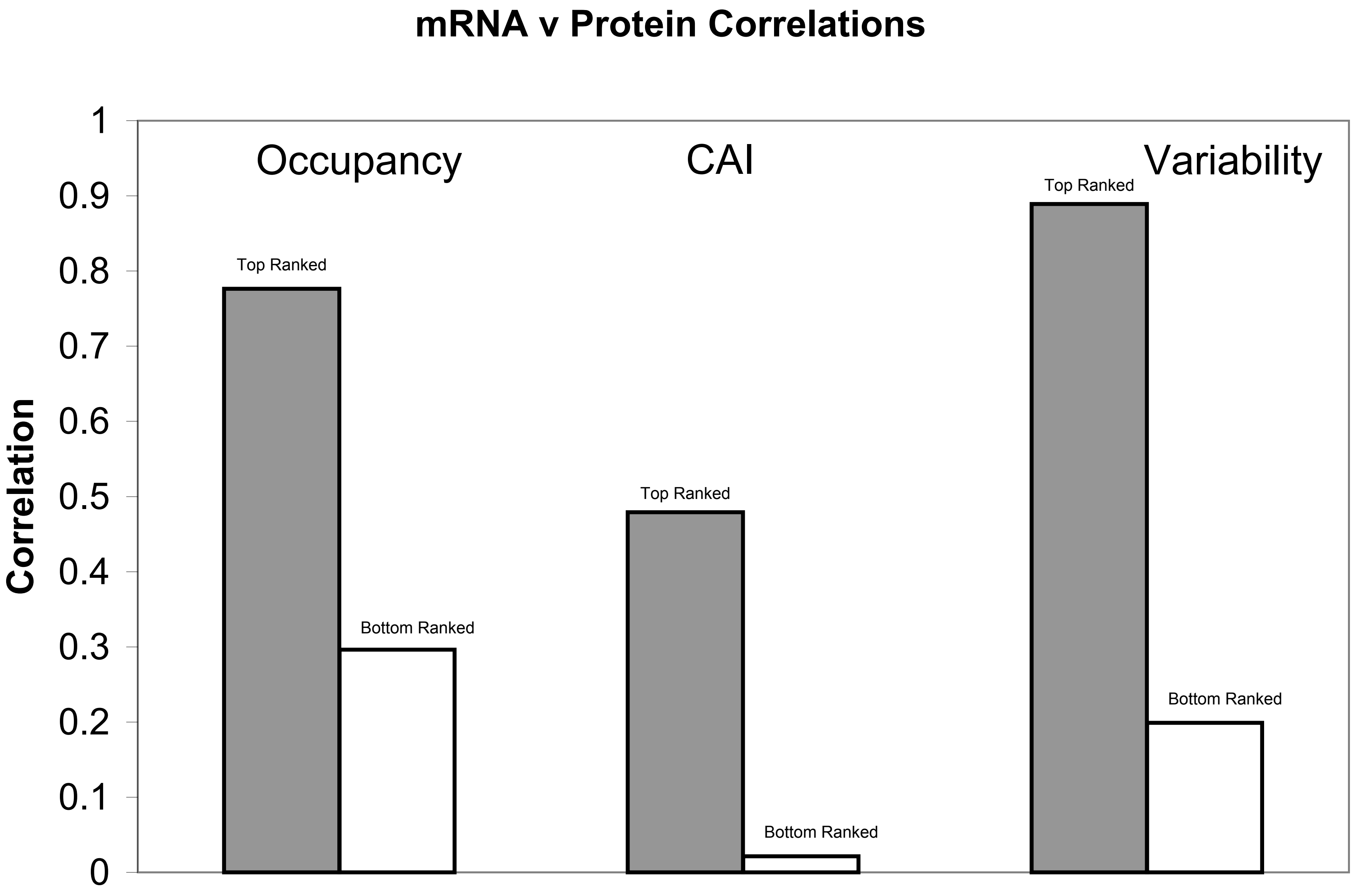
47. Mewes, HW, Frishman, D, Guldener, U, Mannhaupt, G, Mayer, K, Mokrejs, M, Morgenstern, B, Munsterkotter, M, Rudd, S & Weil, B: **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 2002, **30:**31-4.

48. Yan, JX, Devenish, AT, Wait, R, Stone, T, Lewis, S & Fowler, S: **Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of Escherichia coli**. *Proteomics* 2002, **2:**1682-98.


mRNA v Protein
MudPit (1)
MudPit (2)
2DE (1)
2DE (2)
R = 0.66
Protein Abundance
1000
100
10
1
0.1
0.1
1
10
100
1000
mRNA Expression


Subcellular Localization
Nucleolus (R=0.8)
Cell Periphery (R=0.74)
Reference Data (R=0.66)
Mitochondria (R=0.42)
Protein Abundance
100.00
10.00
1.00
0.10
0.10
1.00
10.00
100.00
mRNA Expression


MIPS Function
Cell Cycle (R=0.71)
Reference Data (R=0.66)
Cell Rescue (R=0.45)
Protein Abundance
1000
100
10
1
0.1
0.1
1
10
100
mRNA Expression


mRNA v Protein Correlations
Correlation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Occupancy
CAI
Variability
Top Ranked
Bottom Ranked
Top Ranked
Bottom Ranked
Top Ranked
Bottom Ranked