Accounting for experimental noise reveals that transcription dominates control of steady-state protein levels in yeast

Gábor Csárdi[1], Alexander Franks[1], David S. Choi[1] Edoardo M. Airoldi[1,2] D. Allan Drummond[3*]
1 Dept. of Statistics, Harvard University, Cambridge, MA, USA
2 The Broad Institute of Harvard & MIT, Cambridge, MA, USA
3 Depts. of Biochemistry & Molecular Biology and Human Genetics, University of Chicago, Chicago, IL, USA
∗ E-mail: dadrummond@uchicago.edu, airoldi@fas.harvard.edu

Abstract

Cells respond to their environment by modulating protein levels through mRNA transcription and post-transcriptional control. Modest correlations between global steady-state mRNA and protein measurements have been interpreted as evidence that transcript levels determine roughly 40% of the variation in protein levels, indicating dominant post-transcriptional effects. However, the techniques underlying these conclusions, such as correlation and regression, yield biased results when data are noisy, missing systematically, and collinear—properties of mRNA and protein measurements—which motivated us to revisit this subject. Noise-robust analyses of 25 studies of budding yeast reveal that mRNA levels explain roughly 80% of the variation in steady-state protein levels. Post-transcriptional regulation amplifies rather than competes with the transcriptional signal. Measurements are highly reproducible within but not between studies, and are distorted in part by between-study differences in gene expression. These results substantially revise current models of protein-level regulation and introduce multiple noise-aware approaches essential for proper analysis of many biological phenomena.

Author Summary

Cells regulate protein levels in multiple ways. The weak correlations found between steady-state levels of mRNA transcripts and proteins have led to the widely held conclusion that mRNA transcription contributes less than half of the variation in protein levels. Alternatively, weak correlations may reflect biases arising from measurement noise. By using noise-aware approaches to analyze many independent studies of budding yeast, we show that mRNA levels explain roughly 80% of protein-level variation, far higher than previously appreciated. The resulting integrated quantitative data suggests substantial revisions to many common assumptions, such as that protein levels are proportional to mRNA levels. To combat biases, both gold-standard measurements and noise-aware analyses are needed.

Introduction

Cellular protein levels reflect the balance of transcript levels, protein production by translation initiation and completion, and protein removal by degradation, secretion and dilution [1, 2](Figure 1A). The standard quantitative model for protein-level regulation is

$$\frac{\partial P_i}{\partial t} = \tau_i M_i - \delta_i P_i \tag{1}$$

where P_i is the cellular protein level (molecules per cell) of gene i, M_i is the mRNA level, and τ_i and δ_i are the mRNA translation and net protein removal rates, respectively. At steady-state, protein levels will be proportional to mRNA levels with proportionality constants of τ_i/δ_i, such that if rates of translation and removal did not vary by gene, steady-state mRNA and protein levels would correlate perfectly [1]. Consequently, the mRNA–protein correlation observed in global measurements of mRNA

and protein levels has been intensely studied, and deviations from perfect correlation used to quantify the contribution of post-transcriptional processes to cellular protein levels [1–6].

The consensus emerging from these studies holds that, across organisms, transcriptional regulation explains 40–50% of the variation in steady-state protein levels, leaving half or more to be explained by posttranscriptional regulatory processes [2,4,6–9]. Higher correlations are observed, generally for subsets of less than half the genome that are biased toward high-abundance mRNA and protein expression [1,6,10]. Low observed mRNA–protein correlations have motivated the search for alternate forms of regulation capable of accounting for the majority of protein-level variability [2, 6, 8]. Recent studies have indeed uncovered wide between-gene variation in posttranscriptional mechanisms such as translation rates [11] and protein degradation rates [2].

However, as frequently noted [1, 4, 6, 12, 13], noise in measurements can cause many of the observations attributed to post-transcriptional regulation. Here, noise encompasses variability due to cell-to-cell variation, growth conditions, sample preparation and other effects due to experimental design [14], and measurement biases and error [13]. Uncorrelated noise between mRNA and protein measurements will reduce the observed mRNA–protein correlation relative to the true value, while inflating the variation in measurements of translational efficiency and other posttranscriptional processes [15, 16]. Empirically, disentangling noise effects from biological effects is critical for an accurate understanding of how cells regulate protein levels.

Rapid progress has been made in global measurement of transcript and protein levels by multiple methods, as underscored by recent high-coverage drafts of the human proteome [17, 18]. These methods were largely pioneered in budding yeast, and have been replicated many times by different groups. Motivated by the ongoing and intense interest in the contribution of mRNA levels to protein levels, we were prompted to revisit the subject in this well-studied model eukaryote.

Results

We collected 38 measurements of mRNA levels and 20 measurements of protein levels from 14 and 11 separate studies respectively, each of haploid *S. cerevisiae* growing exponentially in shaken liquid rich medium with 2% glucose between 22°C and 30°C (Table S1). These data cover varying amounts of the genome and display a wide range of correlations between studies (Figure 1B, Pearson correlations on log-transformed values with zeros and missing values omitted). Although correlations of replicates within studies are quite high [6], with median $r = 0.97$ for mRNA and 0.93 for protein levels, between-study correlations are far more modest, $r = 0.62$ for mRNA measurements and 0.57 for protein measurements. That is, data from a typical mRNA study explains 39% of the variance in another study ($r^2 = 0.39$) and a typical protein study's results explain only 32% in another study's variance, consistent with previous studies reporting wide variation between studies [9]. Strong outliers indicate high reproducibility for a two pairs of studies (Figure 1B), but each such outlier is a correlation between separate studies done by the same research group, suggesting the presence of additional variability sources between groups. The high within-study reproducibility and low between-study reproducibility indicates the presence of large systematic errors between studies.

Correlations are modest even between studies using similar methods (e.g., $r = 0.81$ between two RNA-Seq datasets using Illumina instruments [11, 19]). Comparing mRNA studies performed using similar or different methods on a shared set of 4,595 genes revealed little difference in reproducibility whether similar or different methods were used (Figure 1C, no t-test $P < 0.05$ for differences in correlation when comparing studies employing shared methods versus independent methods after false discovery rate correction).

Between-study correlations quantify the studies' mean ratio of true variance to total variance, termed the reliability [20, 21] (see *Methods*). In turn, setting aside sampling error, the maximum observable correlation between any two datasets is equal to the geometric mean of their reliabilities. Because virtually

all reported global mRNA–protein correlations involve mRNA and protein levels measured in separate studies, between-study reliabilities are the relevant quantity. The modest reliability values—setting aside those of the same group reporting two studies, which we exclude from this analysis—sharply limit the maximum observable mRNA–protein correlations. This limit has startling consequences: if steady-state mRNA and protein levels actually correlated perfectly (true $r = 1.0$), then given the median observed between-study correlations in Figure 1B, we would expect to observe mRNA–protein correlations of only $r = \sqrt{0.57 \times 0.62} = 0.60$.

The data reveal a wide range of modest mRNA–protein correlations with a median of $r = 0.54$ (Figure 1C) quantified either by the Pearson correlation between log-transformed measurements or the nonparametric Spearman rank correlation (Figure S1; both measures produce similar results and we employ the former throughout). Coverage of the 5,887 verified protein-coding genes in yeast [22] also varies widely. The largest pair of datasets covers 4,367 genes and shows an mRNA–protein correlation of $r = 0.618$ ($r^2 = 0.38$, 38% of protein-level variance explained by mRNA levels), close to consensus values [6].

Reduction of correlations by noise can be corrected using information from repeated measurements [16, 21]. Quantitative corrections for correlation attenuation were first introduced more than a century ago by Spearman [16], are widely used in the social sciences [21,23,24], and have found recent applications in biology [20,25–27]. Given two measurements each of variables X and Y, each with uncorrelated errors, the true correlation can be estimated using only correlations between the four measurements X_1, X_2, Y_1, Y_2 (see *SI Materials and Methods*):

$$\hat{r}_{XY}^{\text{true}} = \frac{\sqrt[4]{r_{X_1Y_1} r_{X_2Y_2} r_{X_1Y_2} r_{X_2Y_1}}}{\sqrt{r_{X_1X_2} r_{Y_1Y_2}}} \tag{2}$$

The correction reflects a simple intuition: the denominator quantifies the reliabilities of the measurements, which determine the maximum observable correlation, and the numerator quantifies the observed correlation using a geometric mean of four estimates and is divided by this maximum value to yield an estimate for the true value. In simulated data, this noise-corrected estimate accurately ascertains true correlations in the presence of noise far exceeding that apparent in most mRNA and protein data (Figure S2). The estimate is not itself a correlation coefficient, and may take values outside (−1,1) due to sampling error [21] (cf. Figure S2B,C).

Using Spearman's correction, we estimated mRNA–protein correlations for pairs of studies, obtaining a median corrected correlation of 0.92. Variability due to sampling error was large for small datasets as expected (cf. Figure S2, and decreased with as size increased, with estimates stabilizing for large datasets (> 3000 genes) at a mean of $r = 0.88 \pm 0.02$ (Figure 1C). This value is echoed by consideration of the largest dataset with two mRNA [19, 28] and two protein [29, 30] measurements each. For these data, the four observed mRNA–protein correlations are $r = 0.60$, 0.63, 0.62 and 0.64, and the correlation between mRNA and protein measurements are $r_{\text{mRNA}} = 0.86$ and $r_{\text{protein}} = 0.57$ respectively, yielding the corrected estimate $\hat{r}^{\text{true}} = \frac{\sqrt[4]{0.60 \times 0.63 \times 0.62 \times 0.64}}{\sqrt{0.85 \times 0.57}} = 0.89$.

Extending these estimates to the full genome requires a more sophisticated approach. Measurements vary widely in coverage, are quantified on a range of scales arising from use of a diverse array of techniques, and cannot be assumed to have equal levels of noise. Even seemingly simple approaches to reduce noise, such as averaging measurements normalized to the same scale, are unworkable: only 16 proteins are detected by all 11 protein quantification studies, and these proteins are all highly abundant. Throwing out smaller datasets discards potentially valuable measurements, and it is unclear when to stop, since all datasets are incomplete to some degree.

To address these challenges, we adapted structural equation modeling to admit nonrandomly missing data (see *Methods*). We introduce a structured covariance model (SCM) that explicitly accounts for structured noise arising from replicates and use of shared measurement techniques, explicitly estimates noise at multiple levels, and allows inferences of latent covariance relationships with imputation of missing

data. The SCM (Fig. S3) recovers true correlations in simulated data when substantial data are missing nonrandomly (Fig. S2), and satisfies posterior predictive checks using real data (Fig. S4). Fitting the SCM yields estimates of mRNA and protein levels integrating all data (Figure 2A) and estimates a whole-genome steady-state mRNA–protein correlation of $r = 0.91$ across all 5,854 genes for which an mRNA transcript has been detected in at least one of the 38 mRNA quantitation experiments (Figure1C). We emphasize that this method does not involve any attempt to maximize the mRNA–protein correlation or any assumptions about the strength of the correlation.

To evaluate accuracy of the SCM estimates, we scaled them to molecules per haploid cell using high-quality published values. Estimates of the number of mRNA molecules per cell range from 15,000 to 60,000 molecules per cell ([31, 32]). A more recent study argued that the earlier, lower estimate resulted from misestimation of mRNA mass per cell and average mRNA length, with 36,000 molecules per cell as a revised estimate also supported by independent measurements [33]. The higher estimate resulted from rescaling the lower estimate to match expression of five genes measured by single-molecule fluorescence *in situ* hybridization (FISH) [32]. We adopted the 36,169 mRNA molecules per cell measurement [33]. 4.95pg total protein per haploid yeast cell [34]—and compared the results to small-scale gold-standard independent measurements of absolute mRNA and protein levels not used in our analysis. (No gold-standard genome-scale measurements of mRNA or protein levels exist for yeast or any other organism.) SCM estimates of absolute mRNA levels matched FISH measurements well [32] (average difference of 1.2-fold between estimated and measured levels (Figure 2B, with one outlier estimate overshooting the FISH value by 1.7-fold). Notably, these results demonstrate that the FISH estimates are compatible with roughly 36,000 mRNA molecules per cell during exponential growth, and do not require the almost two-fold higher number advanced in the FISH study. Absolute protein levels for a set of 21 proteins differing up to 25,000-fold in cellular abundance have been measured using single-reaction monitoring (SRM) with spiked-in stable-isotope standards [35]. SCM estimates correlate better with these absolute levels ($r = 0.93$ between log-transformed values) than does any individual dataset, including the only study [30] which reports levels for all 21 proteins ($r = 0.90$) (Figure 2C, average difference of 1.2-fold between SRM measurement and SCM estimate). Relative protein levels estimated by integrating multiple datasets using an alternative approach in which noise is not modeled [9] correlate with absolute levels less well ($r = 0.88$). The structured covariance modeling approach thus estimates steady-state cellular mRNA and protein levels with an unmatched combination of completeness, precision, and accuracy.

To evaluate imputation of missing data, we focused on the 813 genes with a detected mRNA transcript but no protein detected in any of the 11 studies. Some of these genes encode well-studied proteins such as the proteasomal regulator Rpn4p and the cyclin Cln3p, indicating clear false negatives. Ribosome profiling [11] provides an estimate of mRNA translation rate, a contributor to steady-state protein level. At least one of two independent studies [11, 36] detects ribosomes in the coding sequence of 542 of these 813 genes, suggesting active translation, and translation rate correlates with the imputed protein levels (Figure 2D, $r = 0.39$ and 0.41 with the two studies). Because the missing protein data correspond to genes at the detection limit of these ribosome profiling data (Figure 2D), we predict that many of the remaining genes will be found to produce proteins at low levels in exponential phase.

The structured covariance model provides direct estimates of dataset-specific noise levelsl, which allow us to inquire about the main sources of noise. Cell-to-cell variability and non-systematic instrument error cannot be dominant contributors, because the very high replicate correlations within studies, the vast majority of which are biological replicates, restrict the possible noise from these sources to less than 4% of the variance in mRNA levels and 6% for protein levels on average. We therefore examined the data for signs of systematic differences.

Because growth conditions perturb cell physiology, differences in cell culturing and harvesting may also contribute to noise. The 25 experiments in our dataset report culturing yeast cells to an optical density (OD, absorbance at 600nm) of 0.36–1.0 or, when cell density was reported, from $0.3–4 \times 10^7$ cells/mL. Budding yeast cells begin to deplete glucose and enter the diauxic shift at similar densities. Depletion

of nutrients induces a stereotypic response in which instantaneous growth rate slows and, concomitantly, ribosomal protein gene expression is strongly repressed [37, 38]. We reasoned that any differences arising from such transcriptional responses would introduce unintended variation—i.e., noise. This, in turn, would reduce the observed between-study mRNA–protein correlation.

To test for systematic gene regulatory responses as a cause of noise, we treated noise as if it were an experimental perturbation, and analyzed how gene expression depended upon the noise level. We calculated the slope in each gene's transcript level as a function of decreasing dataset noise quantified by the SCM-estimated signal-to-noise ratio. Many genes showed systematic increases and decreases in level with increasing noise (Figure 3A). GO process analysis on the top 100 genes by slope yielded "translation" and "cytoplasmic translation" as enriched terms ($P < 10^{-6}$), and ribosomal genes show systematically higher mRNA values in less-noisy datasets (Wilcoxon signed-rank test $P < 10^{-16}$) (Figure 3B). Because ribosomal proteins are highly abundant, we were concerned that some systematic regression toward the mean or other abundance-related effect might influence these results. As a control, we examined mRNA levels of genes encoding glycolytic enzymes, which have comparable abundance in yeast, but whose levels are not strongly responsive to cellular stress [38]. Glycolytic genes, exemplified by *CDC19*, showed no significant slope differences ($P > 0.05$). These results suggest systematic determinants of variability between experiments, consistent with nutrient depletion, which occurs under conditions virtually identical to those used to generate many of the analyzed samples.

Our results indicate that the true correlation between steady-state mRNA and protein levels in budding yeast is far higher than previously recognized, which might be taken as evidence that post-transcriptional regulation plays a minor role. Yet positive evidence exists for strong contributions from posttranscriptional regulatory processes, most prominently substantial per-gene variation in translational efficiency [11], prompting us to re-examine these results.

We focused first on the recent report that translation rates estimated by ribosome profiling explained more than twice the protein-level variation than did measured mRNA levels [11]. We wondered whether these findings might reflect noisier mRNA measurements than translation-rate measurements. Consistent with this, correlations using SCM-integrated protein levels are substantially higher for both mRNA and translation rate (Figure4A). Noise-corrected correlations indicate no significant difference in the predictive power of either measure for protein levels—both correlate with roughly $r = 0.9$ (Figure 4A).

Major contributions to protein levels from mechanisms other than mRNA level become obvious upon inspection of the data. The dynamic range of protein expression is much wider than that of mRNA levels [30]; in the SCM estimates, consistent with previous studies, the range of mRNA expression between genes at the 1st and the 99th percentile is 1,044-fold whereas the range of protein expression is 1,039,000-fold, a thousand times broader. A surprising consequence of the relative dynamic ranges of mRNA and protein expression, coupled with the strong correlation between mRNA and protein levels, is that absolute protein levels cannot be proportional to absolute mRNA levels at the genome scale. Equation 1 predicts that, given equal rates of translation and degradation, a gene with a thousand-fold higher mRNA level should have a thousand-fold higher protein level, but the data show that this estimate is too low by three orders of magnitude, indicating that rates of translation, degradation, or both must differ profoundly and systematically between genes.

This simple analysis illustrates a fundamental asymmetry: although absence of posttranscriptional regulatory processes would produce a perfect mRNA–protein correlation [1], a perfect mRNA–protein correlation would not indicate a negligible posttranscriptional contribution to relative protein levels. In fact, contrary to the assumptions of some influential analyses, it is possible for mRNA levels and (for example) translation rates to each explain more than 50% of protein-level variation—all that is required is that these contributions not be independent.

As an example of a non-independent contribution, posttranscriptional processes can shape the the dynamic range of protein levels compared to mRNA levels. Such a contribution can be quantified by the slope of the linear relationship between log-transformed protein and mRNA levels, which is the exponent

relating the untransformed absolute levels.

Previous work has reported this slope to be roughly unity for smaller datasets using ordinary least squares (OLS) linear regression [10], a result we confirmed (Figure 4B). However, OLS regression assumes the independent variable is error-free [39, 40] and thus it is improper to apply OLS regression to these data when the objective is to determine the functional relationship between variables [40]. As with correlations, error causes systematic underestimation of slopes, a phenomenon called regression dilution bias [39]. Indeed, the million-fold protein-level variation, compared to the thousand-fold mRNA-level variation, provides strong guidance that the actual slope is closer to 2 (protein levels are proportional to squared mRNA levels) than 1. Use of a noise-tolerant technique, ranged major-axis (RMA) regression [40], yielded substantially steeper slopes, with more-complete datasets producing larger slopes (Figure 4B).

Also as with correlations, non-randomly missing data can also cause underestimation of regression slopes due to restriction of range. We looked for this effect by analyzing datasets constructed using data from two of the largest studies [11, 29], but only computing the RMA slope using genes with proteins detected in each of the smaller studies. Smaller artificial datasets yielded sharply reduced slopes (Figure 4C), confirming that missing data suffices to cause severe understimation of the nonlinear relationship between mRNA and protein levels.

The SCM approach, which accounts for both noise and missing data, yields an estimated slope of 2.2 (Figure 4B), consistent with the expectation derived from simple examination of the relative dynamic ranges. Residual noise unaccounted for by the model will tend to inflate this value, but all pairwise estimates exceed 1.0. Steady-state protein levels therefore reflect a dramatic amplification of the transcriptional signal: rather than competing with transcriptional regulation as often assumed, posttranscriptional regulation cooperates.

If translational regulation drives much of this cooperative amplification, as anticipated, then translation rate (the number of mRNAs multiplied by the translation rate per mRNA) must rise nonlinearly with mRNA level. This is visually clear from examination of the linear fit (slope = 1) compared to the RMA regression line (slope = 1.65, Figure 4D). Data from an independent study using a similar methodology shows a slope of 1.70 (Figure 4E). Thus, most of the superlinear relationship between mRNA and protein levels can be attributed to translational regulation, likely at the level of translation initiation.

Discussion

Our results demonstrate that the widely accepted consensus that steady-state mRNA levels explain less than half (40%) of the variation in protein levels is a significant underestimate; the true value, taking into account the reduction in correlation due to experimental noise, is closer to 80%.

Our study is restricted to a single well-studied growth condition for a single well-studied organism. The principles of accounting for noise, but not precise results, can and should be extrapolated to regulatory contributions in other settings and other organisms. An influential study on mouse fibroblasts measured mRNA and protein levels and degradation rates for thousands of genes [2], concluding that mRNA levels explained 41% of the variation in protein levels. However, a recent follow-up study concluded that, once effects of error and missing data were accounted for, mRNA levels explain 75% or more of the protein-level variation in these data [13]. Although translation rates were inferred to cause most protein-level variation in the original study, measured translation-rate variation is insufficient to explain the observed protein-level variation [13]. Our results support similar conclusions.

The strong correlation between steady-state mRNA and protein levels may seem to validate the use of mRNA levels as relatively faithful proxies of protein levels. We urge caution, as a tempting conclusion—that mRNA changes serve as faithful proxies for protein changes—does not follow. Attempts to infer the correlation between transcript and protein changes from steady-state mRNA–protein correlations confuse two distinct and complex phenomena. The genome-scale relationship between mRNA levels and protein levels is an evolved property of the organism, reflecting natural selection's tuning of each gene's

transcriptional and posttranscriptional controls, not merely an input-output relationship between mRNA and protein. Two genes with steady-state mRNA levels differing by 10-fold may have 100-fold differences in protein levels due to evolved differences in their posttranscriptional regulation. This information does not indicate how the protein level for a gene will change if its transcript level is induced 10-fold in a cell, because no regulatory evolution is possible at this timescale.

A related consequence is that the number of proteins per mRNA, often treated as roughly constant, increases steeply with gene expression level. The increased density of ribosomes on high-expression transcripts suggests increased rates of translation initiation as a major contributor to this evolved nonlinearity. Consistent with this, recent work has shown that in yeast and a wide range of other organisms, the stability of mRNA structures in the 5' region weakens as expression level increases, favoring more efficient translation initiation [41].

Our results underscore the urgent need for genome-scale gold-standard measurements of absolute mRNA and protein levels to enable identification and correction of systematic errors in widely used gene-expression measurement techniques. That different groups have, as yet, been unable to reliably reproduce these bread-and-butter measurements using different methods implies that advantages can be gained in improved accuracy, rather than mere precision.

Materials and Methods

Reliability

We wish to measure latent variables ϕ and ψ but, due to noise, actually observe variables $X = \phi + \epsilon$ and $Y = \psi + \delta$ where the random noise variables ϵ and δ have zero mean and are uncorrelated with the latent variables and with each other. The reliability

$$\alpha_X = \frac{\mathrm{Var}(\phi)}{\mathrm{Var}(X)} = \frac{\mathrm{Var}(\phi)}{\mathrm{Var}(\phi) + \mathrm{Var}(\epsilon)} \tag{3}$$

quantifies the ratio of latent-variable variance to total (latent plus noise) variance in X. Given two random variables X_1 and X_2 representing replicate measurements of ϕ, the latent (true) variance can be estimated by $\mathrm{Cov}(X_1, X_2) = cov(\phi + \epsilon_1, \phi + \epsilon_2) = \mathrm{Cov}(\phi, \phi) = \mathrm{Var}(\phi)$, where the error terms vanish because they are uncorrelated. Thus, the expected correlation between replicates is

$$\begin{aligned} r_{X_1,X_2} &= \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}} = \frac{\mathrm{Cov}(\phi, \phi)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}} \\ &= \sqrt{\frac{\mathrm{Var}(\phi)}{\mathrm{Var}(X_1)}\frac{\mathrm{Var}(\phi)}{\mathrm{Var}(X_2)}} = \sqrt{\alpha_{X_1}\alpha_{X_2}}, \end{aligned} \tag{4}$$

which is the geometric mean of the reliabilities of the two measurements.

Spearman's correction

We wish to measure the Pearson correlation coefficient between latent variables $r_{\phi,\psi} = \frac{\mathrm{Cov}(\phi,\psi)}{\sqrt{\mathrm{Var}(\phi)\,\mathrm{Var}(\psi)}}$ but, due to noise, actually observe

$$\begin{aligned} r_{X,Y} &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \\ &= \frac{\mathrm{Cov}(\phi,\psi)}{\sqrt{(\mathrm{Var}(\phi)+\mathrm{Var}(\epsilon))(\mathrm{Var}(\psi)+\mathrm{Var}(\delta))}} \\ &\leq r_{\phi\psi}. \end{aligned} \tag{5}$$

Uncorrelated noise has no average effect on the numerator because errors cancel (see above), but the error terms in the denominator do not cancel. This effect additively inflates the variances in the denominator, biasing the observed correlations downward relative to the truth.

Given the reliabilities α_X and α_Y, Spearman's correction is given by

$$\hat{r}_{\phi\psi} = \frac{r_{XY}}{\sqrt{\alpha_X \alpha_Y}} = r_{\phi\psi} \tag{6}$$

with equality in expectation.

Given two measurements each of X and Y, all with different unknown reliabilities, the true correlation can then be estimated using only correlations between measurements:

$$\begin{aligned} \sqrt{\frac{r_{X_1Y_1} r_{X_2Y_2}}{r_{X_1X_2} r_{Y_1Y_2}}} &= \sqrt{\frac{r_{X_1Y_1} r_{X_2Y_2}}{\sqrt{\alpha_{X_1}\alpha_{X_2}}\sqrt{\alpha_{Y_1}\alpha_{Y_2}}}} \\ &= \sqrt{\frac{r_{X_1Y_1} r_{X_2Y_2}}{\sqrt{\alpha_{X_1}\alpha_{Y_1}}\sqrt{\alpha_{X_2}\alpha_{Y_2}}}} = r_{\phi\psi} \end{aligned} \tag{7}$$

We extend this estimate to

$$\hat{r}_{\phi\psi} = \sqrt[4]{\frac{r_{X_1Y_1} r_{X_2Y_2} r_{X_1Y_2} r_{X_2Y_1}}{r_{X_1X_2} r_{Y_1Y_2}}}$$

which again has expected value $r_{\phi\psi}$ and has the further desirable properties of exploiting all pairwise correlations and being independent of the choice of indices. In practice, each of correlations contributing to Spearman's correction are replaced with correlations estimated from the data, such that the result is also an estimate of the true correlation.

Data collection

We gathered 16 data sets that measure mRNA expression and 11 that measure protein concentrations, mostly published, yielding a total of 58 high-throughput measurements of mRNA and protein levels from 5,854 genes in budding yeast. The measurements were taken using different technologies including custom and commercial microarrays, high-throughput sequencing and mass spectrometry. All yeast cultures were growing in rich media and sampled during the exponential growth phase. Details of the data sets are summarized in Table 1.

We gathered 16 data sets that measure mRNA expression and 11 that measure protein concentrations, mostly published, yielding a total of 58 high-throughput measurements of mRNA and protein levels from 5,854 genes in budding yeast. The measurements were taken using different technologies including custom and commercial microarrays, high-throughput sequencing and mass spectrometry. All yeast cultures were

growing in rich media and sampled during the exponential growth phase. Details of the data sets are summarized in Table 1.

Raw data (with missing values), data normalized and imputed using the SCM, and merged molecules-per-cell estimates are archived in Dryad (http://datadryad.org) with DOI doi:10.5061/dryad.rg367.

The structured covariance model (SCM)

The model has two components: an observation model $p(I_{i,j}|X_{i,j})$, which provides the probability of observing a value for mRNA/protein i in replicate j, given the underlying mRNA/protein level, and a hierarchical model $p(X_{i,j}|\dots)$ for the underlying mRNA/protein levels themselves. The full model is specified as

$$\begin{aligned} X_{i,j} &= L_{i,l[j]}G_{k[j]} + T_{i,t[j]} \\ &\quad + E_{i,k[j]} + R_{i,j} + \nu_j && (8) \\ \mathbf{L}_i &\sim \mathcal{N}_2(0, \mathbf{\Psi}) && (9) \\ T_{i,t} &\sim \mathcal{N}_{N_T}(0, \tau_t) && (10) \\ E_{i,k} &\sim \mathcal{N}(0, \xi_k) && (11) \\ R_{i,j} &\sim \mathcal{N}(0, \theta_j) && (12) \\ p(I_{i,j} = 0|X_{i,j} = x) &= \frac{1}{1 + \exp(-\eta^0_{k[j]} - \eta^1_{k[j]}X_{i,j})}. && (13) \end{aligned}$$

Random variables $L_{i,l}$ correspond to the true denoised protein ($l = 1$) and mRNA ($l = 2$) levels, for mRNAs and proteins $i = 1, \dots, N$, and $\mathbf{L}_i = [L_{i,1}, L_{i,2}]'$. The random variables $T_{i,t}$ and $E_{i,k}$ capture common technological variation and batch effects, respectively, $t = 1, \dots, N_t$, $k = 1, \dots, N_E$. $R_{i,j}$ are measurement noise for replicate $j = 1, \dots, N_R$.

Both technology effects and batch effects between experiments are assumed to be independent, $\mathrm{Cov}(T_{i_1,t_1}, T_{i_2,t_2}) = 0$ if $t_1 \neq t_2$, and $\mathrm{Cov}(E_{i_1,k_1}, E_{i_2,k_2}) = 0$ if $k_1 \neq k_2$. Measurement noise is independent between replicates, $\mathrm{Cov}(R_{i_1,j_1}, R_{i_2,j_2}) = 0$ if $j_1 \neq j_2$.

The parameter ν_j reflects replicate specific bias common to all mRNAs/proteins. The coefficient G_k is an experiment specific scaling factor for the true underlying expression and abundance, and reflects the amount of post-transcriptional amplication.

Missing data model

Equation 13 models the probability that measurement $X_{i,j}$ is missing, $p(I_{i,j} = 0)$, as a logistic function of the value of the measurement. The parameters of the missing data mechanism, η^0_k and η^1_k, are shared by all replicates within an experiment; they uniquely determine the probability that measurements are observed, conditional on $X_{i,j}$.

Prior specifications

To complete the model specifications we place priors on $\mathbf{\Psi}$, τ_t, ξ_k, θ_j, η^0_k and η^1_k. We use either flat, or weakly informative priors on all parameters so as to bias the inference as little as possible. For the parameters η^0_k and η^1_k of the logistic observation model we use a Cauchy prior with mean zero and scale 2.5 as suggested by [42]. We assume flat priors on the scaling factors, G_k, and the measurement bias parameters ν_j. For the replicate and experiment variances θ_j and ξ_k we use independent conjugate Inv-Gamma(3/2, 3/10) prior. Finally, for the estimand of interest, we assume $\mathbf{\Psi}$ is a priori drawn from the set of correlation matrices with marginally uniform correlations [43].

Acknowledgments

We thank S. Nesterko for early work, and E.W.J. Wallace, K. Cook, and many other colleagues for thoughtful discussions. This work was supported by NIH grants GM088344 and GM096193, and by two Alfred P. Sloan Research Fellowships. D.A.D. is a Pew Scholar in the Biomedical Sciences.

References

1. de Sousa Abreu R, Penalva L, Marcotte E, Vogel C (2009) Global signatures of protein and mRNA expression levels. Mol Biosyst 5: 1512–1526.

2. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2011) Global quantification of mammalian gene expression control. Nature 473: 337–342.

3. Gygi S, Rochon Y, Franza B, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720–1730.

4. Maier T, Guell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. FEBS Lett 583: 3966–3973.

5. Siwiak M, Zielenkiewicz P (2010) A comprehensive, quantitative, and genome-wide model of translation. PLoS Comput Biol 6: e1000865.

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

7. Brockmann R, Beyer A, Heinisch J, Wilhelm T (2007) Posttranscriptional expression regulation: what determines translation rates? PLoS Comput Biol 3: e57.

8. Castrillo J, Zeef L, Hoyle D, Zhang N, Hayes A, et al. (2007) Growth control of the eukaryote cell: a systems biology study in yeast. J Biol 6: 4.

9. Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf S, et al. (2012) Paxdb, a database of protein abundance averages across all three domains of life. Molecular & Cellular Proteomics 11: 492–500.

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

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

12. Futcher B, Latter G, Monardo P, McLaughlin C, Garrels J (1999) A sampling of the yeast proteome. Mol Cell Biol 19: 7357–7368.

13. Li J, Bickel P, Biggin M (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2: e270.

14. Leek J, Scharpf R, Bravo H, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in highthroughput data. Nat Rev Genet 11: 733–739.

15. Franks A, Csardi G, Choi D, Drummond D, Airoldi E (2014) Estimating a structured covariance matrix from multi-lab measurements in high-throughput biology. Technical report, Harvard University, Dept of Statistics.

16. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15: 72–101.

17. Wilhelm M, Schlegl J, Hahne H, Gholami A, Lieberenz M, et al. (2014) Mass-spectrometry-based draft of the human proteome. Nature 509: 582–587.

18. Kim MS, Pinto S, Getnet D, Nirujogi R, Manda S, et al. (2014) A draft map of the human proteome. Nature 509: 575–581.

19. Yassour M, Kaplan T, Fraser H, Levin J, Pfiffner J, et al. (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci USA 106: 3264–3269.

20. Archer K, Dumur C, Taylor G, Chaplin M, Guiseppi-Elie, et al. (2008) A disattenuated correlation estimate when variables are measured with error: illustration estimating cross-platform correlations. Stat med 27: 1026–1039.

21. Schmidt F, Hunter J (1999) Theory testing and measurement error. Intelligence 27: 183198.

22. Cherry J, Hong E, Amundsen C, Balakrishnan R, Binkley G, et al. (2012) Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res 40: D700–705.

23. Muchinsky P (1996) The correction for attenuation. Educational and psychological measurement 56: 63–75.

24. Zimmerman D, Williams R (1997) Properties of the spearman correction for attenuation for normal and realistic non-normal distributions. Applied Psychological Measurement 21: 253270.

25. Adolph S, Hardin J (2007) Estimating phenotypic correlations: correcting for bias due to intraindividual variability. Functional Ecology 21: 178–184.

26. Archer K, Dumur C, Taylor G, Chaplin M, Guiseppi-Elie A, et al. (2007) Application of a correlation correction factor in a microarray crossplatform reproducibility study. BMC Bioinformatics 8: 447.

27. Behseta S, Berdyyeva T, Olson C, Kass R (2009) Bayesian correction for attenuation of correlation in multi-trial spike count data. J neurophysiol 101: 2186–2193.

28. Lipson D, Raz T, Kieu A, Jones D, Giladi E, et al. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol 27: 652–658.

29. de Godoy L, Olsen J, Cox J, Nielsen M, Hubner N, et al. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455: 1251–1254.

30. Ghaemmaghami S, Huh W, Bower K, Howson R, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737–741.

31. Holstege F, Jennings E, Wyrick J, Lee T, Hengartner C, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–728.

32. Zenklusen D, Larson D, Singer R (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. Nat Struct Mol Biol 15: 1263–1271.

33. Miura F, Kawaguchi N, Yoshida M, Uematsu C, Kito K, et al. (2008) Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. BMC Genomics 9: 574.

34. von der Haar T, McCarthy J (2002) Intracellular translation initiation factor levels in Saccharomyces cerevisiae and their role in cap-complex function. Mol Microbiol 46: 531–544.

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

36. Gerashchenko M, Lobanov A, Gladyshev V (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc Natl Acad Sci USA 109: 17394–17399.

37. Brauer M, Huttenhower C, Airoldi E, Rosenstein R, Matese J, et al. (2008) Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. Mol Biol Cell 19: 352–367.

38. Airoldi E, Huttenhower C, Gresham D, Lu C, Caudy A, et al. (2009) Predicting cellular growth from gene expression signatures. PLoS Computational Biology 5: e1000257.

39. Hutcheon J, Chiolero A, Hanley J (2010) Random measurement error and regression dilution bias. BMJ 340: c2289.

40. Legendre P, Legendre L, Legendre L, Legendre L (1998) Numerical ecology. Amsterdam, New York: Elsevier, 2nd English edition.

41. Gu W, Zhou T, Wilke C (2010) A universal trend of reduced mrna stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Computational Biology 6: e1000664.

42. Gelman A, Jakulin A, Pittau M, Su Y (2008) A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics 2: 1360–1383.

43. Barnard J, McCulloch R, Meng X (2000) Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica 10: 1281–1311.

44. Causton H, Ren B, Koh S, Harbison C, Kanin E, et al. (2001) Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell 12: 323–337.

45. Dudley A, Aach J, Steffen M, Church G (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci USA 99: 7554–7559.

46. Garcia-Martinez J, Aranda A, Perez-Ortin J (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell 15: 303–313.

47. MacKay V, Li X, Flory M, Turcott E, Law G, et al. (2004) Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. Mol Cell Proteomics 3: 478–489.

48. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.

49. Pelechano V, Pérez-Ortín J (2010) There is a steady-state transcriptome in exponentially growing yeast cells. Yeast 27: 413–422.

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

51. Velculescu V, Zhang L, Zhou W, Vogelstein J, Basrai M, et al. (1997) Characterization of the yeast transcriptome. Cell 88: 243–251.

52. Lee M, Topper S, Hubler S, Hose J, Wenger C, et al. (2011) A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Mol Syst Biol 7: 514.

53. Nagaraj N, Kulak N, Cox J, Neuhauser N, Mayr K, et al. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol Cell Proteomics 11: M111.013722.

54. Newman J, Ghaemmaghami S, Ihmels J, Breslow D, Noble M, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441: 840–846.

55. Peng J, Elias J, Thoreen C, Licklider L, Gygi S (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2: 43–50.

56. Thakur S, Geiger T, Chatterjee B, Bandilla P, Frohlich F, et al. (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol Cell Proteomics 10: M110.003699.

57. Washburn M, Wolters D, Yates 3rd J (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19: 242–247.

Figure Legends

Figure 1. Quantification and consequences of noise on the correlation between measurements of steady-state mRNA and protein levels. **A**, Steady-state protein levels reflect the balance of mRNA translation and protein removal. **B**, Global measurements of mRNA and protein levels vary widely in reproducibility and coverage. Each point represents a pair of studies. Dots show between-study correlations (median shown by dashed line), a measure of reliability. Dotted line, median of within-study correlations. Blue dots show pairs of studies from the same research group. **C**, Correlations between studies sharing the same quantification method or different methods (dark and light gray bars, respectively), using mRNA datasets with $\geq$ 5000 genes (4,595 genes quantified by all datasets). For example, the second column from the left shows the 18 correlations between each of three commercial microarray studies and six studies using custom microarrays or RNA-Seq. **D**, Large-scale datasets vary widely in coverage of 5,887 yeast coding sequences and in resulting estimates of the mRNA–protein correlation. Shown are all pairwise correlations between 14 mRNA and 11 protein datasets, with within-study replicates averaged if present. Correlations are shown between mRNA and protein levels reported without correction (dots); using Spearman's correction on pairs of datasets (binned, boxes show mean and bars indicate standard deviation); using Spearman's correction on the largest set of paired measurements (red box); and as estimated by structured covariance modeling for 5,854 genes with a detected mRNA or protein (red diamond). **E**, Correlations obtained for the largest set of paired measurements, two of mRNA and two of protein levels (N=3,418).

Figure 2. Integrated estimates of mRNA and protein levels using structured covariance modeling (SCM). **A**, Integrated estimates across 58 global measurements reveal a strong genome-wide dependence between steady-state protein and mRNA levels ($r = 0.91$). Light gray points and marginal density indicate genes with detected mRNA but no detected protein. **B**, Absolute mRNA level estimates versus single-molecule fluorescence *in situ* hybridization counts [32]. **C**, Absolute protein level estimates versus stable-isotope-standardized single reaction monitoring measurements [35]. Dotted lines in **B** and **C** show perfect agreement. **D**, Evidence for active translation of undetected proteins inferred from ribosome profiling; data from one [36] of two [11] studies. Dashed line shows ranged major-axis regression best fit. Marginal densities show ribosome density for all detected transcripts (medium gray), all transcripts with a detected protein (dark gray), and transcripts with no detected protein (light gray).

Figure 3. Cellular responses linked to growth are apparent in gene expression data. A, Gene expression varies systematically with noise; shown are normalized mRNA levels for genes encoding large ribosomal protein 23A (*RPL23A*), the glycolytic enzyme pyruvate kinase 1 (*CDC19*), and a proteasome lid subunit (*RPN6*). Lines show linear fits; slopes for *RPL23A* and *RPN6* are significantly nonzero with $P < 0.05$. **B**, Expression of classes of genes changes systematically with noise. Box and whisker plots show all genes with at least 25 measurements (N=5,326), 133 ribosomal proteins, and 20 glycolytic enzymes. Wilcoxon signed-rank tests, ***, $P < 10^{-16}$; n.s., $P > 0.05$.

Figure 4. Transcriptional and translational regulation act coherently to set protein levels. **A**, The correlation of mRNA (light gray) and rates of translation (dark gray) reported in the original ribosome-profiling study, using averaged mRNA and protein levels, and corrected for noise using Spearman's correction on the same set of genes (N=3,266). Diamond shows whole-proteome SCM estimate. **B**, The exponent relating protein and mRNA concentrations estimated by noise-blind (ordinary least squares) and noise-aware (ranged major-axis) regression analyses. Gray points, all pairs of datasets; black points, pairs of datasets with > 3500 measurements. Dotted line shows perfect agreement; dashed line marks SCM estimate. **C**, Missing data leads to underestimation of the mRNA–protein exponent. The exponent from two large mRNA and protein studies was computed after limiting analysis to only genes with proteins detected in each of the 11 protein studies. **D**, Ribosome density depends nonlinearly on mRNA level. Dashed line shows linear (slope = 1) fit. Solid gray line shows RMA regression fit. **E**, mRNA–ribosome-density exponents estimated from independent studies [11, 36].

Tables

Data set, reference	Technology, replicates	% miss
CAUSTON [44]	Commercial microarray, 5	19–22
DUDLEY [45]	Custom microarray, 4	5
GARCIA [46]	Custom microarray, 1	1
HOLSTEGE [31]	Commercial microarray, 1	12
INGOLIA [11]	RNA-Seq, 6	4–10
LIPSON [28]	RNA-Seq, 6	1
LIPSON.ma [28]	Commercial microarray, 1	4
MACKAY [47]	Custom microarray, 1	28
MIURA [33]	Competitive PCR, 4	26–29
NAGALAKSHMI [48]	RNA-Seq, 1	22
PELECHANO [49]	Custom microarray, 1	14
ROTH [50]	Commercial microarray, 2	59–70
VELCULESCU [51]	SAGE, 1	58
YASSOUR [19]	RNA-Seq, 4	5
FUTCHER [12]	2D gel, 1	99
GHAEMMAGHAMI [30]	Western blot, 1	34
DEGODOY [29]	LC MS/MS, 1	25
GYGI [3]	2D gel, 1	98
LEE [52]	LC MS/MS, 3	67–76
LU [10]	LC MS/MS, 1	83
NAGARAJ [53]	LC MS/MS, 6	31
NEWMAN [54]	GFP/flow cytometry, 1	60
PENG [55]	LC MS/MS, 1	74
THAKUR [56]	LC MS/MS, 3	84–85
WASHBURN [57]	LC MS/MS, 1	77

Table 1. List of mRNA data sets (above the midline) and protein concentration data sets (below the midline). The number of replicates in each data set is given after the technology name.