A simple mass-action model predicts protein timecourses from mRNA trajectories during a dynamic response in two strains of *S. cerevisiae*

Sidney Kuo,[1,4] Jarrett Egertson,[2] Gennifer Merrihew,[2] Michael MacCoss,[2] Daniel Pollard,[3] Scott A. Rifkin[1,*]

1. Section of Ecology, Behavior, and Evolution. Division of Biological Sciences, University of California, San Diego.

2. Department of Genome Sciences, University of Washington.

3. Biology Department, Western Washington University

4. current affiliation: Illumina, San Diego, California

*. To whom correspondence should be addressed: sarifkin@ucsd.edu

Abstract

Although mRNA is a necessary precursor to protein, several studies[1–12] have argued that the relationship between mRNA and protein levels is often weak. This claim undermines the functional relevance of conclusions based on quantitative analyses of mRNA levels, which are ubiquitous in modern biology from the single gene to the whole genome scale. Furthermore, if post-translational processes vary between strains and species, then comparative studies based on mRNA alone would miss an important driver of diversity[13,14]. However, gene expression is a dynamic process, and most studies examining relationship between mRNA and protein levels at the genome scale have analyzed single timepoints[15,16]. Here we show that protein level timecourses can be predicted from mRNA timecourses through a simple generative model for a majority of genes involved in the yeast pheromone response. We find that this generative process acts as a buffer on mRNA level differences such that many mRNA differences

between yeast strains are not reflected in the protein level. Evolution manipulates protein-specific processes to amplify or buffer transcriptional regulation.

To determine (a) whether mRNA levels are a good proxy for protein levels during the dynamic response of an organism to an environmental stimulus[17] and (b) whether differences in mRNA levels between strains translate into differences in protein levels, we profiled the transcriptomes and proteomes of two strains of *Saccharomyces cerevisiae* during their response to pheromone. mRNA levels in S288c and YJM145 have previously been shown to differ at 30 minutes after exposure to alpha factor pheromone[15]. We collected mRNA samples at 0, 0. 5, 1, 1.6, 2.4, 3.2, 4.8, 6.4, and 8 hours after exposure and protein samples at 0, 1.6, 3.2, 4.8, 6.4, and 8 hours for each strain in three biological replicates with mRNA and protein collected from the same flasks. We measured genome-wide levels of mRNA by RNA-seq and protein levels by quantitative mass spectrometry.[18]

Even during this response to a specific environmental signal, gene expression was pervasive genome-wide throughout the timecourse. We detected transcripts for 6,231 genes at one or more time points in at least one strain and 6,078 genes with transcripts at one or more timepoints in both strains of which a shared set of 5,595 genes could be analyzed through our DESeq2 pipeline, comprising approximately 85% of yeast genes (Supplementary methods). 2,345 different proteins were detected, and all but two were present at every timepoint in both strains (Supplementary methods). This is approximately 35% of the genome. Proteins that are present in low abundances

may not be detected via mass spectrometry[19,20], which may partially account for fewer proteins being present than as detected by RNA-seq.

Although large numbers of genes had detectable levels of mRNA or protein, not all were affected by pheromone induction. We tested for significant induction or repression by checking whether at least one timepoint was significantly different from time 0 (FDR < 0.05).[21] With some variation between strains around 25% of analyzed mRNA and 35% of analyzed proteins (Table 1; Figure 1), were significantly induced or repressed, with roughly 55% of these shared between strains (Table 1). The classification of genes as significantly induced or repressed and/or different between strains at 1 hour post-induction were not, in general, representative of analyses over the whole timecourse (Figure 2).

Across genes, the correlation *between strains* from mRNA is 0.49 at 1 hour post-induction and for protein is 0.48 at 1.6 hours post-induction. However, despite seemingly low correlations, if we test for differences between strains over entire mRNA or protein trajectories by evaluating the significance of strain x timepoint interactions in a likelihood ratio test, we detect 3.9% (217 genes) of mRNA trajectories being significantly different between strains (FDR<0.05), and 4.6% (108 genes) of protein trajectories being significantly different (FDR<0.05). For mRNA, expression trajectories of genes involved in translation (including ribosomal genes) tended to be more divergent than expected while protein divergences were dominated by genes involve in amino-acid synthesis (Supplementary file 4). There were only 14 genes which were different between strains for both mRNA and protein. These numbers may be conservative since the false negative rate for mRNA and protein differences is not necessarily the same,

but if we rank the p-values of the likelihood ratio tests for between strain differences, there is not a noticeable relationship between those for mRNA and protein (Supplementary figure 1). This suggests that post-transcriptional regulation plays an important role, both in maintaining protein levels while mRNA differs, and in affecting protein levels without significant changes in mRNA trajectories. Large differences in mRNA do not necessarily lead to large differences in protein between strains.

If we restrict our focus to only those genes that significantly change during the response in both RNA and protein in either strain, the mean correlation between mRNA and protein trajectories is 0.18, although significantly changing genes are enriched for higher correlations (Supplementary figure 2). The low correlation is in line with the literature[22,23] and corroborates the idea that taking only the magnitude and direction of mRNA differential expression is a poor predictor of the behavior of its corresponding protein product, presumably because it does not take into account properties like protein degradation rates. Indeed, even qualitatively, mRNA and protein are not necessarily congruent. When the mRNA changes significantly during the timecourse, only around 2/3 of proteins follow the direction of this change. (Figure 3). Even among the significantly changing protein targets of STE12 – a transcriptional activator that is known to respond to pheromone – mRNA and protein change in the same direction only for 84 of 131 targets in s288c. This suggests that (a) even qualitative conclusions drawn from dynamic changes in mRNA abundance are not always representative of the behavior in protein levels and that (b) generalizations about protein behavior based on mRNA measurements at single timepoints can be strikingly mistaken.

In order to bridge the disconnect between mRNA and protein levels, we fit a mass action model to our protein data[24] of the form $dp/dt = k_s R(t-\tau) - k_D P(t)$, where k_s is the protein synthesis rate after pheromone exposure, $R(t-\tau)$ is the measured amount of mRNA at some amount of time τ before time t, k_D is the protein degradation rate after pheromone exposure, and $P(t)$ is the amount of protein at time t. Because individual parameter estimates are often poorly constrained in systems biology models,[25,26] we estimated their joint posterior distribution for each strain separately and for a model where both strains shared the same synthesis and degradation values (Supplementary methods). This framework allowed us to compare these rates between strains along with the structure of their covariation while accounting for the uncertainties in the parameter estimates.[27]

The model fit the data well. For each gene and strain we simulated trajectories from the posterior. The mean normalized root mean squared error was less than 2% for approximately 60% of the genes and less than 10% for 98% of them. At the maximum *a posteriori* estimates, protein synthesis rates during the pheromone response range are log-normally distributed with the middle 50% ranging from 1.9 to 22.5 per minute. The middle 50% of protein half lives range from 1.9 to 10.3 hours with 13.1% (S288c) and 10.7% (YJM789) of proteins with half-lives less than one hour. The predicted protein half-lives were also approximately log-normally distributed[28], although we found an overabundance of short lived proteins and an underrepresentation of long-lived ones[22] (Figure 4; Supplementary figure 3). While this could be the result of MS bias, which tends to only pick up more abundant proteins, pheromone response induces substantial

physiological changes such as cell cycle arrest, which may lead to active protein degradation for certain proteins compared with steady-state growth conditions[29].

In order to lead to differences in protein trajectories, mutations must alter either RNA trajectories or biochemical parameters such as the translation and degradation rates we modeled here. We found that strain-specific synthesis and degradation rates gave demonstrably better protein trajectory predictions than a shared parameter model for 300 of the 2316 genes with good data for both RNA and protein based on WAIC (FDR<0.05; Figure 5).[30,31] Genes involved in small molecule metabolic processes were most likely to have divergent parameters between strains (FDR < 1E-13; Supplementary data 1), particularly those active in the mitochondria.[32] This apparent divergence in the parameters controlling mitochondrial protein dynamics is consistent with other results showing that YJM145 is divergent from S288c-like lab strains in mitochondrial expression.[33]

Of these 2316 genes, 406 significantly diverged between strains (FDR<0.05) in either RNA, protein, or parameters or a combination of factors (Figure 5). Over half of these only had detectable differences in parameter values. For 71 genes, the divergence was in RNA alone, only 1 gene showed divergence in RNA and protein trajectories without a concomitant change in parameters, and 60 genes had significance divergence in protein trajectories and parameters. Because these tests have different statistical power, these numbers cannot be directly compared, but they make it clear both that studies focusing on the mRNA level alone will miss a large fraction of potentially functional divergence and that parameters of these dynamic biological processes can vary between strains without leading to substantial differences in protein

expression. These phenotypically neutral parameter shifts in our study place cells from the two strains in different regions of the parameter space and, we hypothesize, mean that the strains have different potentials to respond to new physiological or evolutionary pressures.

Our results show that, for any given gene, mRNA levels are not necessarily reliable predictors of protein levels even during an active response to a stimulus and that differences in mRNA abundances between strains do not always translate into differences at the protein level. However, protein trajectories can be predicted from mRNA by a simple mass-action model that uses physiologically plausible parameter values to represent the post-transcriptional processes of protein synthesis and degradation. These parameter values can differ between strains, and they may depend upon cellular conditions[13,14,34,35]. These three knobs – mRNA levels, synthesis rates, and degradation – can each be changed dynamically and independently of each other and give evolution the ability to finely tune protein trajectories to meet specific exigencies.

Methods

Strains and experimental conditions

The two *S. cerevisiae* strains used in this study were S288c and YJM145. Both are mating type **a** with the following genes knocked out via the loxP system[36]: *URA3, HO, AMN1,* and *BAR1***.** In both strains, *FIG1* had been tagged with yECitrine using the Ca.URA3 marker for other studies. Strains were grown in 50 mL YPD overnight as a starter culture, with the experimental culture being in 1L YPD in a 3L Fernbach flask to provide proper aeration. Experimental cultures were diluted to an OD of ~0.1 and incubated for 2 hours before inducing with pheromone. Pheromone was added to achieve a final concentration of 50 nM in the media. Samples taken for the mRNA were at hours 0, 0.5, 1, 1.6, 2.4, 3.2, 4.8, 6.4, 8 where hour 0 is taken immediately before induction. Samples for the protein MS were taken from hours 0, 1.6, 3.2, 4.8, 6.4, 8. Fifty mL of each sample was taken from the flask at each time point, spun down, and flash frozen in liquid nitrogen. Experiments were performed on three separate days with the RNA and protein handled together afterwards.

Total RNA extraction of all the samples was done using the Ambion RiboPure kit. Multiplexed RNA-Seq libraries were generated from the 3' end of transcripts according to a published protocol.[37] The protocol produces mRNA reads proportional to the number of transcripts independent of transcript length. Sequencing was done on an Illumina HiSeq4000 using single reads 50 base pairs in length. Quantitative MS was performed by the MacCoss laboratory including processing of raw MS data (see Supplementary methods 1).

Processing and statistical methods

Processing of RNA-Seq data was done on Galaxy, using BWA to align reads to the *S. cerevisiae* reference genome (R64-2-1, 2015-01-13) and htseq-count to extract read counts for each transcript[38–40]. Estimation of gene counts was performed using DESeq2,[41] with a generalized linear model parameterized as *count ~ timepoint + strain + timepoint:strain*, with *timepoint* being a factor variable[42]. Likelihood ratio tests between the strains use the form *count ~ timepoint + strain* as the reduced model, which would not include differences in the pre-induction time point between the strains.

Pre-processing of MS intensity values was done using Skyline, followed by sample normalization and log transformation via MSStats[43,44] (Supplmentary file 2). A general linear model using log intensities for each protein that follows the same form as the RNA-seq model was fit to the data, and the likelihood ratio test was performed the same way.

All analysis and modeling was performed in R.[45] Although each of our RNA replicates corresponded to a protein replicate from the same flask, including this pairing in the modeling did not improve the fit in general. We estimated posterior densities for each gene for a model where the synthesis and degradation parameters were shared between strains and one where they were allowed to differ using a weakly-informative prior on the parameters (Supplementary file 1). All fits were on a log scale as errors are thought to be symmetrical for protein levels on a log scale.

Data and analysis script availability

Upon publication, raw RNA-seq reads will be found on GEO under the accession number GSE131211 and the mass spectrometry data will be found on Panorama public

at the URL https://panoramaweb.org/KuoRifkin.url. The complete processed data, R code, and outputs from the analyses will be available from the UCSD library digital collection archive at https://library.ucsd.edu/dc/

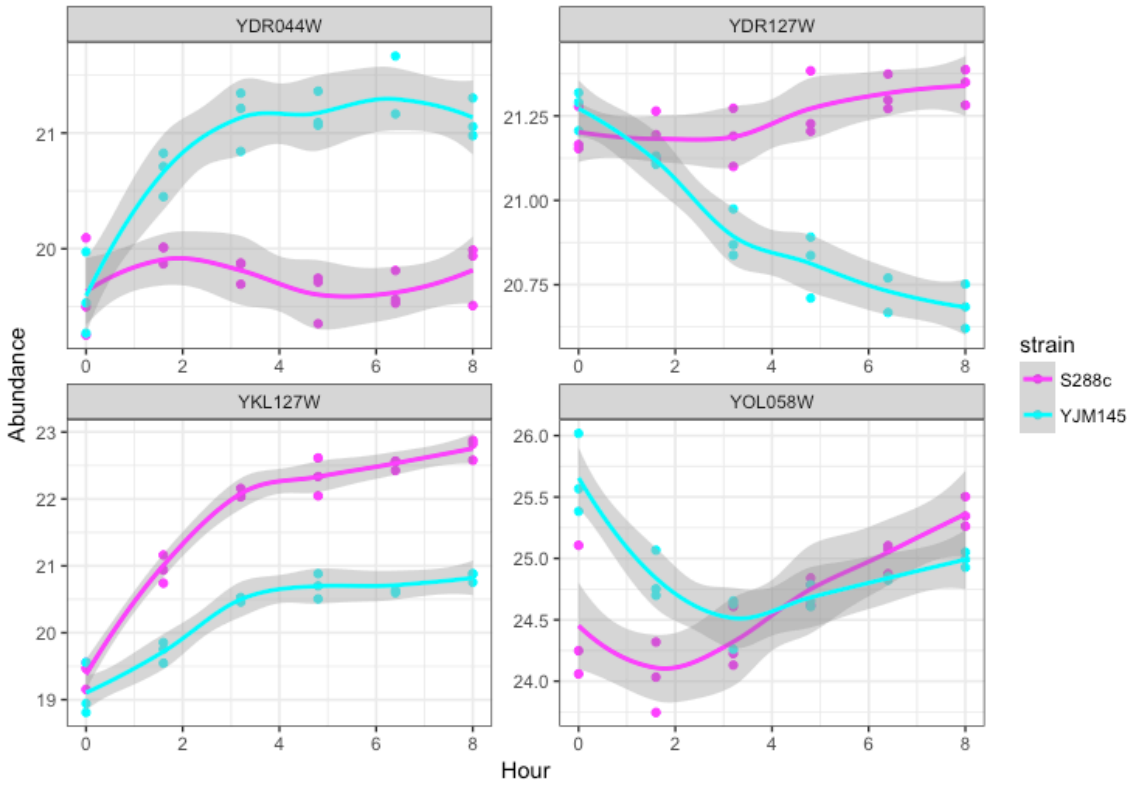

Figure 1: Four proteins that were detected to be significantly different according to the likelihood ratio test. Protein abundances were plotted over time, with the color denoting the strain. Dots represent actual experimentally measured time points, whereas the curves are the loess-smoothed fits of the linear model with shaded 95% confidence intervals, which represent our best estimate of average abundance at that time point. Top-left: YDR044W (HEM13) is involved in the heme biosynthetic pathway. Top-right: YDR127W (ARO1) is involved in chromisate biosynthesis, which is the precursor to aromatic amino acids. Bottom-left: YKL127W (PGM1) codes for phosphoglucomutase, which converts glucose-1-phosphate to glucose-6-phosphate, a process necessary for hexose metabolism. Bottom-right: YOL058W (ARG1) is an enzyme in the arginine biosynthesis pathway.

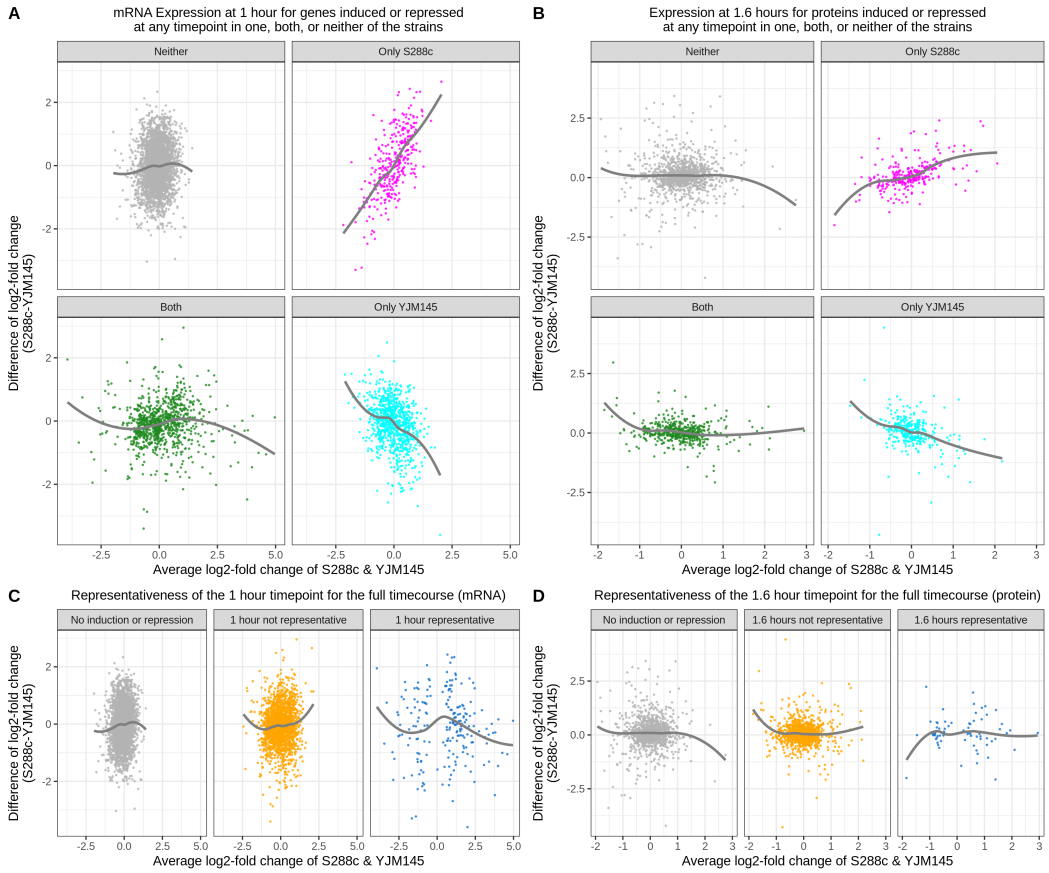

Figure 2. Studies comparing one timepoint to a baseline misclassify differential expression during a dynamic response. **(A)** MA plots of mRNA expression for the two strains at 1 hour after pheromone exposure separated into whether the mRNA is significantly induced or repressed during the timecourse in one, both, or neither strain. Grey lines are loess fits through the data as guides to the eye. **(B)** Like (A) except for protein expression at 1.6 hours after induction. **(C)** Like (A) except separated by whether the induction at 1 hour was reprsentative of the whole timecouse. **(D)** Like (C) excpect for protein at 1.6 hours.

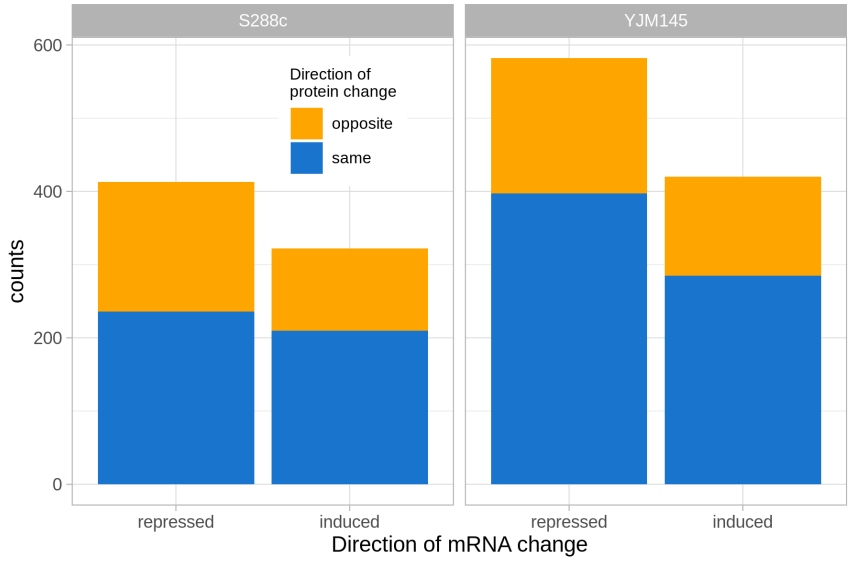

Figure 3. Concordance of induction or repression between mRNA and protein.

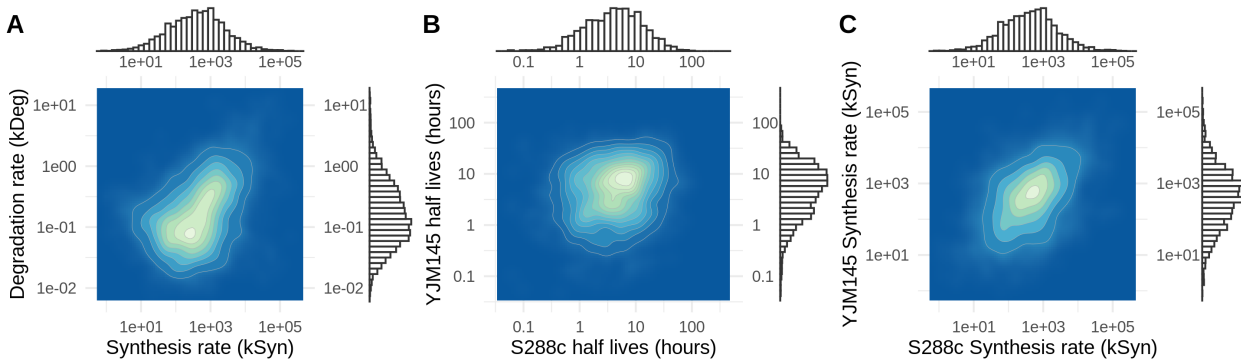

Figure 4: Maximum *a posteriori* parameter estimates. (**A**) Density plot and histograms of synthesis rate and degradation rate pairs for both strains together. (**B**) Density plot and histograms of protein half-lives. (**C**) Density plot and histograms of estimated protein synthesis rates during pheromone response.

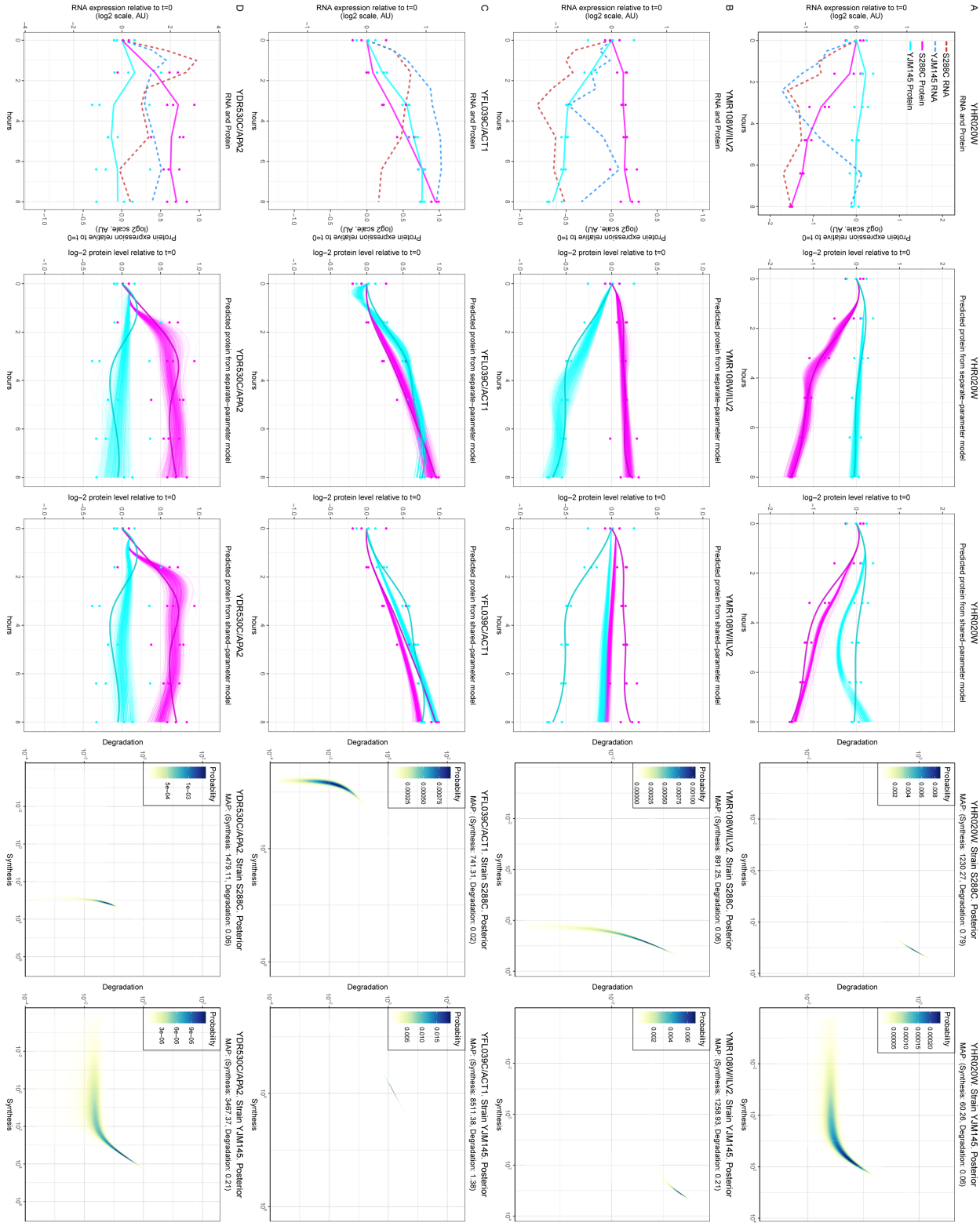

Figure 5: RNA, protein, and parameter variation can take several different forms. The figure is described rotated into landscape orientation. From left to right, the 5 panels in

each row are: (1) the mean mRNA and protein trajectories with protein data plotted as filled circles. Dotted lines are RNA (red: S288C; blue: YJM145) and solid lines are protein (magenta: S288C; cyan: YJM145). The values are on a log scale relative to the value at t=0. (2) 500 trajectories simulated from the posterior of the separate-parameter model. Spline fits to the actual protein data are shaded in a slightly darker color than the simulations. Protein data is plotted as filled circles. As in (1), the values are on a log scale relative to the value at t=0. (3) Same as (2) except simulated trajectories are generated from the shared-parameter model. (4) Heat map of the posterior density estimate for S288C. (5) Heat map of the posterior density estimate for YJM145. (**A**): YHR020W has significantly different mRNA, protein, and parameters; (**B**) YMR108W (ILV2) has significantly different protein and parameters but not mRNA trajectories; (**C**) YFL039C (ACT1) has significantly different mRNA trajectories and parameters but not protein trajectories; (**D**) YDR530C (APA2) has shared parameter values between strain but significantly different mRNA and protein trajectories.

Table 1: Number of genes induced or repressed via RNA-seq and protein-MS

	mRNA (Total 5595 genes detected)	Protein (Total 2345 genes detected)
S288c	1275 (22.8%)	808 (35.5%)
YJM145	1781 (31.8%)	835 (36.7%)
Both	938 (16.8%)	514 (22.6%)

Acknowledgements

This project was supported by National Science Foundation grant NSF MCB-1517482 to SAR and NSF MCB-1518314 to DP. SK was partially supported on NIH training grant T32GM007240-39A1.

Author Contributions

SK, SAR, and DP designed the study. SK performed the experiments. JE, GM, MM performed the MS measurements. SK and SAR analyzed the data and wrote the paper.

Supplementary material

Supplementary methods

Supplementary figures

Supplementary data 1. Zip archive with the results of the GOrilla[32] GO analyses for mRNA, protein, and parameter values for Process, Function, and Component ontologies.

1. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology* **25**, 117–124 (2007).
2. Wu, G., Nie, L. & Zhang, W. Integrative Analyses of Posttranscriptional Regulation in the Yeast Saccharomyces cerevisiae Using Transcriptomic and Proteomic Data. *Current Microbiology* **57**, 18–22 (2008).
3. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).
4. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology* **19**, 1720–30 (1999).
5. Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. A sampling of the yeast proteome. *Molecular and cellular biology* **19**, 7357–68 (1999).
6. Washburn, M. P. *et al.* Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3107–12 (2003).
7. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology* **4**, 117 (2003).
8. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).

9. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Letters* **583**, 3966–3973 (2009).
10. Edfors, F. *et al.* Gene‐specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology* **12**, 883 (2016).
11. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20 (2017).
12. Wilhelm, M. *et al.* Wilhelm *et al*. reply. *Nature* **547**, E23 (2017).
13. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**, 752–755 (2002).
14. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Reviews Genetics* **7**, 862–872 (2006).
15. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187–1191 (2010).
16. Roberts, C. J. *et al.* Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles. *Science* **287**, (2000).
17. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
18. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrometry Reviews* **33**, 452–470 (2014).
19. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry* **389**, 1017–1031 (2007).

20. Peshkin, L. *et al.* On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Developmental Cell* **35**, 383–394 (2015).
21. Spies, D., Renz, P. F., Beyer, T. A. & Ciaudo, C. Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Brief Bioinform* doi:10.1093/bib/bbx115
22. Christiano, R., Nagaraj, N., Fröhlich, F. & Walther, T. C. Global Proteome Turnover Analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Reports* **9**, 1959–1965 (2014).
23. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Letters* **583**, 3966–3973 (2009).
24. Lee, M. V. *et al.* A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology* **7**, (2011).
25. Gutenkunst, R. N. *et al.* Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Comput. Biol.* **3**, e189 (2007).
26. Erguler, K. & Stumpf, M. P. H. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol. BioSyst.* **7**, 1593–1602 (2011).
27. Rifkin, S. A., Atteson, K. & Kim, J. Constraint structure analysis of gene expression. *Funct. Integr. Genomics* **1**, 174–185 (2000).
28. Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E. K. Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 13004–9 (2006).

29. Bardwell, L. A walk-through of the yeast mating pheromone response pathway. *Peptides* **26**, 339–350 (2005).

30. Watanabe, S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research* **11**, 3571–3594 (2010).

31. McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. (Chapman and Hall/CRC, 2018).

32. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

33. Sun, X., Wang, Z., Guo, X., Li, H. & Gu, Z. Coordinated Evolution of Transcriptional and Post-Transcriptional Regulation for Mitochondrial Functions in Yeast Strains. *PLOS ONE* **11**, e0153523 (2016).

34. Albert, F. W., Muzzey, D., Weissman, J. S. & Kruglyak, L. Genetic Influences on Translation in Yeast. *PLoS Genetics* **10**, e1004692 (2014).

35. MacKay, V. L. *et al.* Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Molecular & cellular proteomics : MCP* **3**, 478–89 (2004).

36. Gueldener, U., Heinisch, J., Koehler, G. J., Voss, D. & Hegemann, J. H. A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Research* **30**, 23e–223 (2002).

37. Kadoki, M. *et al.* Organism-Level Analysis of Vaccination Reveals Networks of Protection across Tissues. *Cell* **171**, 398-413.e21 (2017).

38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–9 (2015).
40. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* **44**, W3–W10 (2016).
41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
42. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
43. Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).
44. Egertson, J. D., MacLean, B., Johnson, R., Xuan, Y. & MacCoss, M. J. Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nature Protocols* **10**, 887–903 (2015).
45. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2017).