Investigating RNA structure and function, transcriptome-wide

by

Silvia Rouskin

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

For my family

ACKNOWLEDGMENTS

First of all I would like to thank my parents, Roumen and Antoaneta Rouskin, who have inspired me, supported me, and most of all, always believed that I could succeed in anything. I cannot overstate their vital role in encouraging my curiosity for science and letting me follow my own path even when I was only a child, and when the path led across the Pacific Ocean. Throughout my life they have provided the essential leap of faith that made me go the extra distance.

Grad school at UCSF has been an incredible experience. I couldn't imagine a better mentor than Jonathan Weissman. In fact, it is truly impossible to express in simple words his multiple talents and various ways through which he has helped me grow as a scientist. One of my favorite things is sharing a new result with him. It is always both invigorating and insightful. I would often e-mail a result and then impatiently re-fresh my inbox for the next 10 minutes, because this is how long it takes for him to reply- even if it is 1am on a Saturday night! It has been a lot of fun working in his lab, and it will remain as one of my greatest memories. I will certainly miss his brilliant advice and insatiable interest in science.

I would also like to thank Joe DeRisi, who is the reason why I came to UCSF initially as a technician in his lab. He has been a tremendous source of excitement and inspiration. Joe taught me a lot, and I wouldn't be here without him. It was a great pleasure working with and learning from the people in his lab, particularly Charles Chiu, Anatoly Urisman, Amy Kistler, and Kael Fischer. Another instrumental mentor at the time was Don Ganem, also a brilliant scientist and a delightful collaborator.

During my grad school years, I was lucky to have a great group of committee members – Christine Guthrie, Alan Frankel, and Raul Andino, who provided key guidance and advice. I would especially like to thank Christine for being an exceptional role model as one of the first women at UCSF, and sharing with me her personal experiences, both hardships and excitements.

I was also extremely fortunate to collaborate with Carol Gross and a very talented graduate student in her lab David Burkhardt. It has been a great pleasure and a truly enriching experience to work with such an intelligent and selfless group of people. This group extends to the people in the Weissman lab, in particular Meghan Zubradt, Gene-Wei Li, Calvin Jan, Josh Dunn, Martin Kampmann, Liz Costa, Owen Chen, Clement Chu, Nick Ingolia, Gloria Brar, Martin Jonikas, Noam Stern-Ginossar, and Eugene Oh.

As my personal and scientific life go hand in hand, there are a number of remarkable scientists that are also very close friends, and have provided moral support, courage, and advice in times when the path was rough and had no clear view. This group extends to my classmates as well as two instrumental people Andrew Houk and Anna Reade.

Finally, my life would not be complete without my daughter Adriana Rouskin-Faust. Having her has taught me to love and appreciate the simple things in life, such as sleep for longer than two-hour intervals. Last, but certainly not least, I would like to thank my husband Tyler Faust. There is no one else who could do a better job at questioning and criticizing my work, from the very first time we met at the annual Granlibakken Tetrad retreat. Tyler is both a vital intellectual resource and a comedic relief at the end of a long day. Thanks to him I can never be uninspired. Tyler's contributions, both personal and scientific, are unsurpassed.

CONTRIBUTIONS

Excluding the contributions listed below, work presented in this dissertation was performed by Silvia Rouskin. Chapter II, entitled “Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*” first appeared in print on January 30, 2014 in Nature 505(7485):701-5. Meghan Zubradt performed the yeast mutations and flow cytometry experiments. Dr. Stefan Washietl and Dr. Manolis Kellis conducted the conservation analysis for RNA structures. This work was supervised entirely by Dr. Jonathan S. Weissman. Chapter III, entitled “Operon mRNAs are organized into ORF-centric structures that specify translation efficiency” is currently in preparation for publication. This work was done in a very close collaboration with David Burkhardt. Dr. Burkhardt performed computational analysis as well as generated mutant strains, pulse-labelled with ^{35}S-methionine, and collected mRNA-seq and ribosome profiling data. Dr. Gene-Wei Li performed data analysis. This work was supervised by Drs. Jonathan S. Weissman and Carol Gross. Chapter IV, entitled “ Causal signals between codon bias, mRNA structure, and efficiency of elongation and translation” is currently in press at Molecular Systems Biology. Cristina Pop performed all of the computational analysis. Lu Han and Dr. Eric Phizicky quantified the amount of tRNA amino-acylation in wild type and mutant yeast strains. Dr. Nicholas Ingolia advised the experimental set up. This work was supervised by Drs. Jonathan Weissman and Daphne Koller.

TABLE OF CONTENTS

CHAPTER ONE

Introduction

INTRODUCTION

Ribonucleic acid (RNA) is hypothesized as the primordial molecule to carry biological information and therefore be the origin of all current life forms[1]. On one hand, the liner sequence of RNA makes it a source of genetic information. On the other hand, the ability of RNA to fold into secondary structures, which are sensitive to the environment, provides both flexibility and specificity in its interactions with other molecules. At a higher complexity, RNA can assume tertiary structures that create internal environments and present binding pockets for metal ions to promote catalysis. Thus RNA can serve both as an information molecule and a direct effector of a biological task. The unsurpassed range of chemical space and function place RNA at the center of all major cellular processes and gene regulation[2].

With the discovery of nucleic acids in the 1900s, and the realization years later that the flow of genetic information goes from DNA to RNA to protein, DNA took the center stage as the blueprint for heredity, and the function of mRNA was thought to be merely a messenger between DNA and proteins. Nevertheless, biochemical studies in the following decade provided increasing evidence that specialized RNA molecules such as tRNAs have complex tertiary structures. These cloverleaf-like structures of tRNA were shown to be essential for transferring the information encoded in the nucleotides of mRNA into a specific sequence of amino acids[3]. Such results suggested that RNA folding is more highly analogous to the folding of proteins rather than to the highly repetitive folded structure of the DNA double helix. In fact, Crick made the hypothesis that RNAs can perform the function of proteins as early as 1968[4], and fifteen years later the first RNAs with catalytic activity were discovered[5,6].

In retrospect, it is not surprising that the RNA field has grown tremendously in the past decade. We now know that mRNAs can also form secondary structures that can have vital control over how the genetic information is read out – by dictating which DNA segments are transcribed[7,8,9] or spliced together[10,11], and how the message of the genetic code is localized[12], translated[13], and degraded[14]. From the discovery of small interfering RNAs[15] and long intervening non-coding RNAs[16] that have essential functions in regulating gene expression, to the stunning observation that only 2% of the human genome codes for protein[17], the world of regulatory RNAs continues to expand. For example, the first eukaryotic riboswtich, or RNA that changes structure in the presence of a small metabolite in order to affect a biological outcome, was recently found in *N. crassa*[18]. Furthermore, there is now evidence through evolutionary conservation that a significant portion (~10%) of the human genome may function through the formation of specific RNA secondary structures[19]. Thus, central to our understanding of RNA as a regulatory molecule is the ability to determine RNA structure in its native environment inside cells. Therefore, the main focus of my thesis has been the development and application of a genome-wide approach to determine the structures and function of RNA molecules in a wide range of species including bacteria, yeast, and human cells.

1. Cech, T. R. The RNA worlds in context. *Cold Spring Harb. Perspect. Biol.* **4,** a006742 (2012).
2. Sharp, P. A. The centrality of RNA. *Cell* **136,** 577–580 (2009).
3. Rich, A. & RajBhandary, U. L. Transfer RNA: Molecular Structure, Sequence, and Properties. *Annu. Rev. Biochem.* **45,** 805–860 (1976).

4. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379 (1968).

5. Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31,** 147–157 (1982).

6. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35,** 849–857 (1983).

7. Haller, A., Soulière, M. F. & Micura, R. The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Acc. Chem. Res.* **44,** 1339–1348 (2011).

8. Montange, R. K. & Batey, R. T. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **37,** 117–133 (2008).

9. Grundy, F. J., Winkler, W. C. & Henkin, T. M. tRNA-mediated transcription antitermination in vitro: Codon–anticodon pairing independent of the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11121–11126 (2002).

10. McManus, C. J. & Graveley, B. R. RNA structure and the mechanisms of alternative splicing. *Curr. Opin. Genet. Dev.* **21,** 373–379 (2011).

11. Warf, M. B. & Berglund, J. A. The role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.* **35,** 169–178 (2010).

12. Martin, K. C. & Ephrussi, A. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell* **136,** 719 (2009).

13. Ray, P. S. *et al.* A stress-responsive RNA switch regulates VEGFA expression. *Nature* **457,** 915–919 (2009).

14. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8,** 113–126 (2007).

15. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391,** 806–811 (1998).

16. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).

17. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* **24,** 344–352 (2008).

18. Cheah, M. T., Wachter, A., Sudarsan, N. & Breaker, R. R. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* **447,** 497–500 (2007).

19. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41,** 8220–8236 (2013).

CHAPTER 2

Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*

Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*

Silvi Rouskin[1], Meghan Zubradt[1], Stefan Washietl[2], Manolis Kellis[2] and Jonathan S. Weissman[1]*

[1]Department of Cellular and Molecular Pharmacology, California Institute of Quantitative Biology, Center for RNA Systems Biology, Howard Hughes Medical Institute, University of California, San Francisco, CA 94158, USA.
[2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; The Broad Institute, Cambridge, Massachusetts 02139, USA

* To whom correspondence should be addressed. E-mail: weissman@cmp.ucsf.edu

RNA plays a dual role as an informational molecule and a direct effector of biological tasks. The latter function is enabled by RNA's ability to adopt complex secondary and tertiary folds and thus has motivated extensive computational[1-2] and experimental[3-8] efforts for determining RNA structures. Existing approaches for evaluating RNA structure have been largely limited to *in vitro* systems, yet the thermodynamic forces which drive RNA folding *in vitro* may not be sufficient to predict stable RNA structures *in vivo*[5]. Indeed, the presence of RNA binding proteins and ATP-dependent helicases can influence which structures are present inside cells. Here we present an approach for globally monitoring RNA structure in native conditions *in vivo* with single nucleotide precision. This method is based on *in vivo* modification with dimethyl sulfate (DMS), which reacts with unpaired adenine and cytosine residues[9], followed by deep sequencing to monitor modifications. Our data from yeast and mammalian cells are in excellent agreement with known mRNA structures and with the high-resolution crystal structure of the *Saccharomyces cerevisiae* ribosome[10]. Comparison between *in vivo* and *in vitro* data reveals that in rapidly dividing cells there are vastly fewer structured mRNA regions *in vivo* than *in vitro*. Even thermostable RNA structures are often denatured in cells, highlighting the importance of cellular processes in regulating RNA structure. Indeed, analysis of mRNA structure under ATP-depleted conditions in yeast reveals that energy-dependent processes strongly contribute to the predominantly unfolded state of mRNAs inside cells. Our studies broadly enable the functional analysis of physiological RNA structures and reveal that, in contrast to the Anfinsen view of protein folding, thermodynamics play an incomplete role in determining mRNA structure *in vivo*.

A wide range of chemicals and enzymes have been used to monitor RNA structure[11,7]. We focused on DMS as it enters cells rapidly[9,12] and is a well-established tool for the analysis of RNA structure[13]. DMS is highly reactive with solvent accessible, unpaired residues but reliably unreactive with bases engaged in Watson-Crick interactions, thus nucleotides that are strongly protected or reactive to DMS can be inferred to be base-paired or unpaired, respectively. We coupled DMS treatment to a massively parallel sequencing readout (DMS-seq) by randomly fragmenting the pool of modified RNAs and size-selecting prior to 3' ligation with a specific adapter oligo (Fig. 1a). Since DMS modifications at adenine and cytosine residues block reverse transcription[14] (RT), we used a second size selection step to collect and sequence only the prematurely terminated cDNA fragments. Sequencing of the fragments reveals the precise site of DMS modification, with the number of reads at each position providing a measure of relative reactivity of that site. The results are highly reproducible and robust against changes in the time of modification or concentration of DMS used (Fig. 1b). The sequencing readout allowed global analysis with a high signal-to-noise ratio—in DMS treated samples, >90% of reads end with an adenine and cytosine, corresponding to false positives for A and C of 7% and 17%, respectively (Fig. 1c). For each experiment, we measured RNA structure both *in vivo* and *in vitro* (i.e. refolded RNA in the absence of proteins). We also measured DMS reactivity under denaturing conditions (95°C) as a control for intrinsic biases in reactivity, library generation or sequencing, revealing only modest variability compared to that caused by structure-dependent differences in reactivity (Fig. 2c, Extended Data Fig. 1a).

The *in vivo* DMS-seq data are in excellent agreement with known RNA structures. We examined three validated mRNA structures in *S. cerevisiae*: *HAC1*, *RPS28B*, and *ASH1*[15-17]. In each case, the DMS-seq pattern qualitatively recapitulates secondary structure with high reactivity constrained to loop regions in both the *in vivo* and the *in vitro* samples but not in the denatured (Fig. 2a-b). Recent determination of a high-resolution yeast 80S ribosome crystal structure[10] allowed us to comprehensively evaluate the DMS-seq data for rRNAs. Comparison of the 18S (Fig. 2c) and 25S (Extended Data Fig. 1b) rRNA DMS signal *in vivo* versus denatured reveals a large number of strongly protected bases *in vivo*. Based on DMS reactivity, we used a threshold to bin bases into reactive and unreactive groups, then calculated agreement with the crystal structure model as a function of the threshold. True positives were defined as both unpaired and solvent accessible bases according to the crystal structure, and true negatives defined as paired bases. A receiver operator characteristic (ROC) curve shows a range of thresholds with superb agreement between the *in vivo* DMS-seq data and the crystal structure model (Fig. 2d). For example, at a threshold of 0.2 the true positive rate, false positive rate, and accuracy are 90%, 6%, and 94% respectively. Bases that were not reactive at this threshold *in vivo* showed normal reactivity when denatured (Extended Data Fig. 1c). This argues that the small fraction (~10%) of residues that are designated as accessible, but are nonetheless strongly protected from reacting with DMS, resulted from genuine differences in the *in vivo* conformation of the ribosome and the existing crystal structures. Agreement with the crystal structure was far less good for *in vitro* refolded rRNA (as expected given the absence of ribosomal proteins) and was completely absent for denatured RNA. By contrast, probing of intact purified ribosomes gave a very similar result to that seen *in vivo*, further demonstrating that DMS-seq yields comparable results *in vitro* and *in vivo* when probing the same structure.

Qualitatively, we observed many mRNA regions where structure was apparent *in vitro* but not *in vivo*. For example, computational analysis[18] predicts a stem loop structure in *RPL33A*. The *in vitro* DMS-seq data strongly supported this predicted structure whereas this region showed little to no evidence of structure in cells (Fig. 3a). To systematically explore the relationship between mRNA structure *in vivo* and *in vitro*, we quantitated structure in a given region using two metrics: Pearson correlation coefficient (r value), which reports on the degree of similarity of the modification pattern to that of a denatured control, and the Gini index[19], which measures disparity in count distribution as would be seen between an accessible loop verses a protected stem (Fig. 3b). We then applied these metrics to windows containing a total of 50 A/C nucleotides. Globally, mRNAs are much more structured *in vitro* compared to *in vivo*: there is a strong shift towards low r values and high Gini indices for the *in vitro* data that is far less pronounced *in vivo* (Fig. 3c). Thus unlike the ribosomal RNA, we find little evidence within mRNAs for *in vivo* DMS protection beyond what we observe *in vitro*, suggesting that the DMS protection we observe *in vivo* is not due to mRNA-protein interactions. For example, using a cut-off (r value <0.55, Gini index >0.14) which captured the rRNAs and functionally validated mRNA structures, including both previously characterized and newly identified structures (see below), we found that out of 23,412 mRNA regions examined (representing 1,948 transcripts), only 3.9% are structured *in vivo* compared to 24% *in vitro* (Fig. 3c and Extended Data Fig. 2 for similar results obtained with windows of different sizes). In addition, 29% of the regions *in vivo* are indistinguishable from denatured (Fig. 3c, orange circle), whereas *in vitro* only 9% of regions were fully denatured. We also applied DMS-seq to mammalian cells (both K562 cells and human foreskin fibroblast), which revealed qualitatively very similar results to yeast—a limited number of stable structures *in vivo* compared to *in vitro* (Fig. 3d, Extended Data Fig. 3-4).

Because the pool of stable structures seen *in vivo* includes previously validated functional mRNA structures, this relatively small subset of mRNA regions provides highly promising candidates for novel functional RNA structures. To explore this, we focused on two structured 5' untranslated regions (UTRs) from *PMA1* and *SFT2* and on the structured *PRC1* 3'UTR for more detailed functional analyses. We fused these UTRs upstream or downstream, respectively, of a Venus protein reporter and quantified Venus levels by flow cytometry. Stem loop structures in these UTRs significantly increased (5'*SFT2*) or decreased (5'*PMA1* and 3'*PRC1*) protein levels upon disruption of their predicted base pairing interactions, and Venus protein levels were rescued by compensatory mutations (Extended Data Fig. 5-6, Extended Data Table 1). Phylogenetic analysis revealed the 5'UTR *PMA1* stem is under positive evolutionarily selection (Extended Data Fig. 5c), lending additional support for a physiological function. A list of 189 structured regions, along with a model of their secondary structures that are similarly supported by phylogenetic analysis of compensatory mutations, is hosted on an online database (http://weissmanlab.ucsf.edu/yeaststructures/index.html). In addition, we mutated predicted stems in three 3'UTRs with evidence of strong ordered structures *in vitro* but not in cells, and these mutations resulted in minimal expression changes (Extended Data Fig. 6d). Nonetheless, it remains possible that transient, heterogenous or weakly ordered structures *in vivo* have biological roles especially if they become more ordered under different physiological conditions.

To evaluate what role *in vitro* thermodynamic stability plays in driving mRNA folding *in vivo*, we performed genome-wide structure probing experiments *in vitro* at five temperatures (30, 45, 60, 75, and 95°C). As temperature rises and structure unfolds (Fig. 4a), the DMS signal becomes more even (low Gini index) and the modification pattern resembles that of the 95°C denatured control (high r value). We defined *in vitro* temperature of unfolding (T_{unf}) as the lowest temperature where a region appeared similar to the denatured controls. Remarkably, many regions with little or no detectable structure *in vivo* show similar thermostability to highly structured regions, including structures that are functionally validated (Fig. 4a, b). For example, the regions of *RPL33A* (unfolded *in vivo*) and *RPS28B* (a functionally validated structure *in vivo*) are both highly structured *in vitro* and have T_{unf} = 60°C. Nonetheless, we find that structures present *in vivo* do have a strong propensity for high thermostability (Fig. 4b), consistent with a recent *in vitro* mRNA thermal unfolding study[8]. In addition to the role of thermostability in explaining the disparity of RNA structure between *in vivo* and *in vitro* samples, we tested the effect of Mg^{2+} concentration *in vitro.* We obtained similar structure results with 2-6mM Mg^{2+}. However, at 1 mM Mg^{2+}, we observe unfolding of most structures including the functionally validated ones (Extended Data Fig. 7a). The above observations indicate that Mg^{2+} concentration and thermodynamic stability play an important but incomplete role in determining mRNA structure *in vivo*.

A central question is what accounts for the differences between *in vivo* and *in vitro* mRNA structure. Although translation by ribosomes plays a role in unwinding structure, this is unlikely to be the dominant force for unfolding *in vivo* since the average *in vivo* structure for coding regions was not distinguishable from 5' and 3'UTRs (Extended Data Fig. 7b). Moreover, within coding regions, high ribosome occupancy of an mRNA as measured by ribosome profiling[21] was not generally associated with lower structure (Extended Data Fig. 7c). It is likely that both active mechanisms (e.g., RNA helicases) and passive mechanisms (e.g., single stranded RNA binding proteins) counteract mRNA's intrinsic propensity to form the stable structures[22] seen with *in vitro* studies[3,23] and computational approaches[18]. To investigate how energy-dependent processes contribute to unfolding mRNA *in vivo*, we performed DMS-seq on yeast depleted of ATP[24]. We observed a dramatic increase in mRNA structure *in vivo* following ATP depletion (Fig. 4c). Moreover, the structural changes seen upon ATP depletion are strongly correlated ($r = 0.54$, $p < 10^{-307}$) to the changes between *in vivo* and *in vitro* samples (Fig. 4d-e, and Extended Data Fig. 8). We also observed a large increase in mRNA structure at 10°C *in vivo* (Extended Data Fig. 9a), but these changes are not as strongly correlated with those seen upon ATP depletion (Extended Data Fig. 9b). Thus the mRNA structures present in a cell are impacted by a range of factors, underscoring the value of DMS-seq in defining the RNA structures present in a specific physiological condition or perturbation.

In summary, DMS-seq provides the first comprehensive exploration of RNA structure in a cellular environment and reveals that in rapidly dividing cells, mRNAs *in vivo* are far less structured than *in vitro*. This scarcity of structure is well suited for the primary role of mRNA as an informational molecule providing a uniform substrate for translating ribosomes. Nonetheless, we identify hundreds of specific mRNA regions that are highly structured *in vivo,* and we show for three examples that these structures impact protein expression. Our studies provide an excellent set of candidate regions, among the truly enormous number of structured regions seen *in vitro*, for exploring the regulatory role of structured mRNAs. The DMS-seq approach is readily extendable to other organisms, including human-derived samples as we show here, and to the analysis of the wide range of functional RNA molecules present in a cell. Thus DMS-seq broadly enables the analysis of structure-function relationships for both informational and functional RNAs. Among the many potential applications, attractive candidates include the analysis of long noncoding RNAs[25,26] , the relationship between mRNA structure and microRNA/RNAi targeting[27], and functional identification and analysis of ribozymes[28], riboswitches[29], and thermal sensors[30].

Method Summary

DMS Modification

For *in vivo* DMS modification, 15 ml of exponentially growing yeast (strain BY4741) at 30°C were incubated with 300-600 µl DMS for 2-4 min (which results in multiple modifications per mRNA molecule). DMS was quenched with the addition of 30 ml stop solution (30% BME, 25% Isoamyl Alcohol). Total RNA was purified using hot acid phenol (Ambion). PolyA(+) mRNA was obtained using magnetic poly(A)+ Dynal beads (Invitrogen).

Library Generation

Sequencing libraries were prepared as outlined in Fig. 1. Specifically, DMS treated mRNA samples were denature at 95°C and fragmented in 1X RNA fragmentation buffer (Ambion). Fragments of 60-80 nucleotides were gel purified and ligated to microRNA cloning linker-1 (IDT) and reverse transcribed using SuperscriptIII (Invitrogen). Truncated RT products were gel purified and circularized using circ ligase (Epicenter). Illumina sequencing adapters were introduced by 8-10 cycles of PCR.

Sequencing and sequence alignment

Raw sequences obtained from Hiseq2000 (Illumina) were aligned against Saccharomyces cerevisiae assembly R62 (UCSC: sacCer2). Aligned reads were filtered so that no mismatches were allowed and alignments were required to be unique.

Online Resources:

For secondary structure models that are supported by DMS seq and have evidence for phylogenetic conservation, visit http://weissmanlab.ucsf.edu/yeaststructures/index.html

Acknowledgments: We thank R. Andino, M. Bassik, J. Doudna, J. Dunn, T. Faust, N. Stern-Ginossar, C. Gross, C. Guthrie, N. Ingolia, C. Jan, M. Kampmann, D. Koller, G.W. Li, S. Mortimer, E. Oh, C. Pop, and members of the Weissman lab for discussions; J. Stewart-Ornstein and O. Brandman for plasmids; C. Chu, N. Ingolia, and J. Lund for sequencing help. This research was supported by the Center for RNA Systems Biology (J.S.W.), the Howard Hughes Medical Insitute (J.S.W.), and the National Science Foundation (M. Z.).

Author Contributions: S.R., M.Z. and J.S.W designed the experiments. S.R. and M.Z. performed the experiments, and S.R. analyzed the data. S.W. and M.K. completed the phylogenetic analysis. S.R., M.Z., and J.S.W. drafted and revised the manuscript.

1. Gruber, A. R., Neuböck, R., Hofacker, I. L. & Washietl, S. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.* **35,** W335–338 (2007).
2. Ouyang, Z., Snyder, M. P. & Chang, H. Y. SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.* **23,** 377–387 (2013).
3. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
4. Underwood, J. G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* **7,** 995–1001 (2010).
5. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9,** 18–20 (2013).
6. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.* **108,** 11063–11068 (2011).
7. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 97–102 (2009).
8. Wan, Y. *et al.* Genome-wide measurement of RNA folding energies. *Mol. Cell* **48,** 169–181 (2012).

9. Wells, S. E., Hughes, J. M., Igel, A. H. & Ares, M., Jr. Use of dimethyl sulfate to probe RNA structure in vivo. *Meth. Enzymol.* **318,** 479–493 (2000).

10. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334,** 1524–1529 (2011).

11. Ziehler, W. A. & Engelke, D. R. Probing RNA structure with chemical reagents and enzymes. *Curr Protoc Nucleic Acid Chem* **Chapter 6,** Unit 6.1 (2001).

12. Zaug, A. J. & Cech, T. R. Analysis of the structure of Tetrahymena nuclear RNAs in vivo: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA* **1,** 363–374 (1995).

13. Cordero, P., Kladwang, W., VanLang, C. C. & Das, R. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* **51,** 7037–7039 (2012).

14. Inoue, T. & Cech, T. R. Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc. Natl. Acad. Sci. U.S.A.* **82,** 648–652 (1985).

15. Gonzalez, T. N., Sidrauski, C., Dörfler, S. & Walter, P. Mechanism of non-spliceosomal mRNA splicing in the unfolded protein response pathway. *EMBO J.* **18,** 3119–3132 (1999).

16. Badis, G., Saveanu, C., Fromont-Racine, M. & Jacquier, A. Targeted mRNA degradation by deadenylation-independent decapping. *Mol. Cell* **15,** 5–15 (2004).

17. Chartrand, P., Meng, X. H., Singer, R. H. & Long, R. M. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Curr. Biol.* **9,** 333–336 (1999).

18. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

19. Wittebolle, L. *et al.* Initial community evenness favours functionality under selective stress. *Nature* **458,** 623–626 (2009).

20. Rüegsegger, U., Leber, J. H. & Walter, P. Block of HAC1 mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. *Cell* **107,** 103–114 (2001).

21. 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).

22. Herschlag, D. RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **270,** 20871–20874 (1995).

23. Li, F. *et al.* Global analysis of RNA secondary structure in two metazoans. *Cell Rep* **1,** 69–82 (2012).

24. Stade, K. *et al.* Exportin 1 (Crm1p) is an essential nuclear export factor**.** *Cell* **90,** 1041–1050 (1997).

25. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493,** 231–235 (2013).

26. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* (2013). doi:10.1038/nature11928

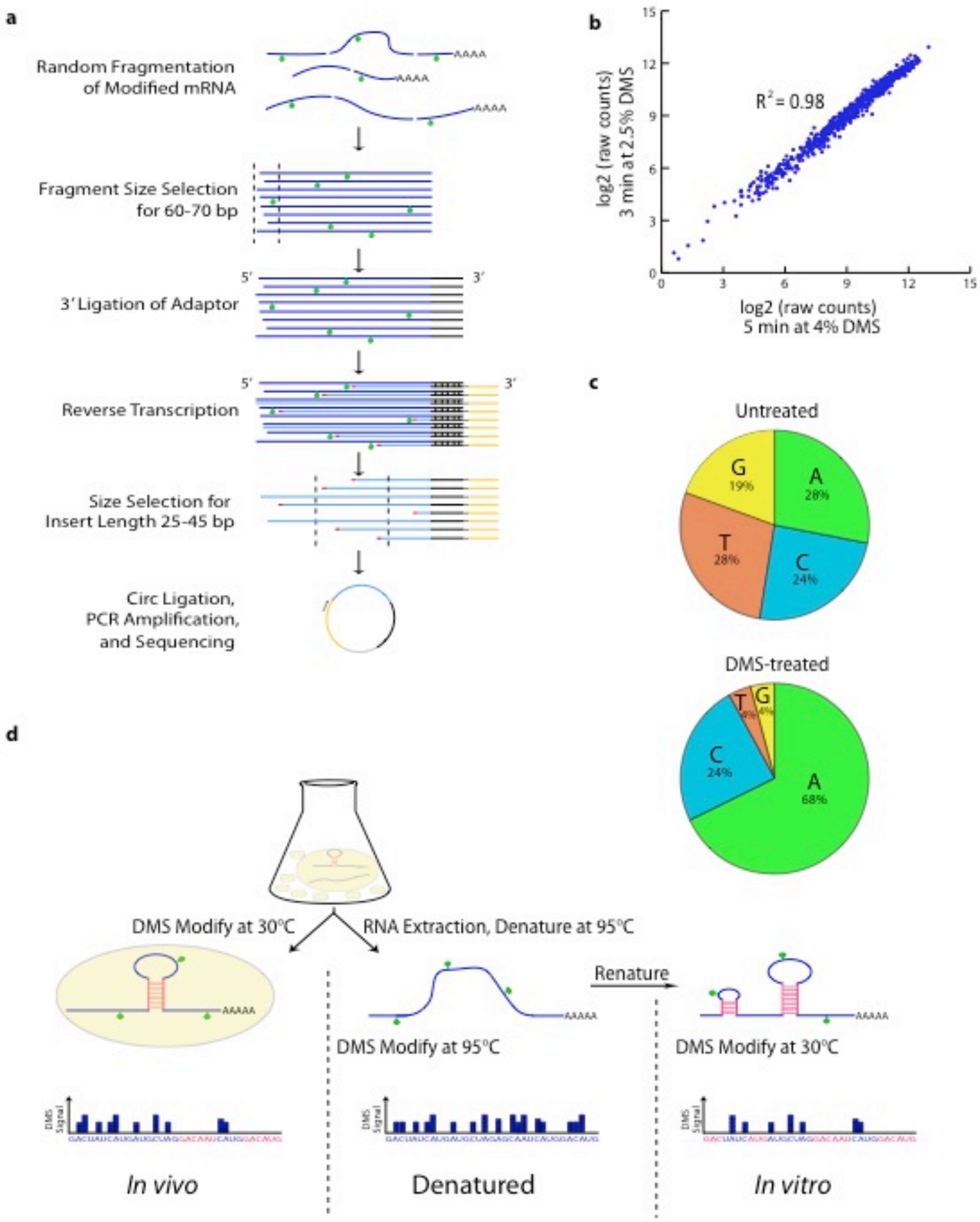
27. Tan, X. *et al.* Tiling genomes of pathogenic viruses identifies potent antiviral shRNAs and reveals a role for secondary structure in shRNA efficacy. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 869–874 (2012).

28. Tang, J. & Breaker, R. R. Structural diversity of self-cleaving ribozymes. *Proc Natl Acad Sci U S A* **97,** 5784–5789 (2000).

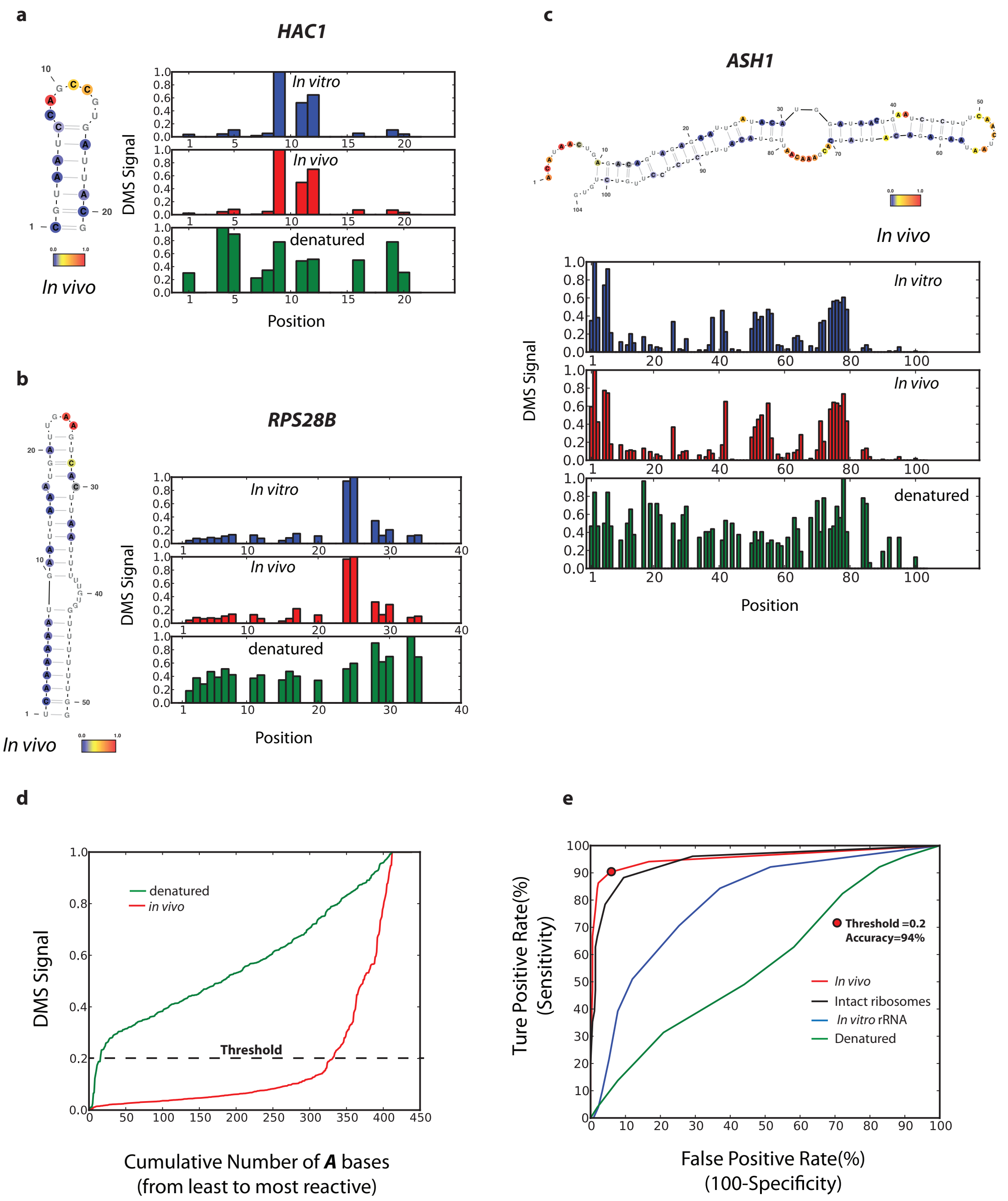
29. Li, S. & Breaker, R. R. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic Acids Res.* **41,** 3022–3031 (2013).

30. Meyer, M., Plass, M., Pérez-Valle, J., Eyras, E. & Vilardell, J. Deciphering 3′ss Selection in the Yeast Genome Reveals an RNA Thermosensor that Mediates Alternative Splicing. *Molecular Cell* **43,** 1033–1039 (2011).

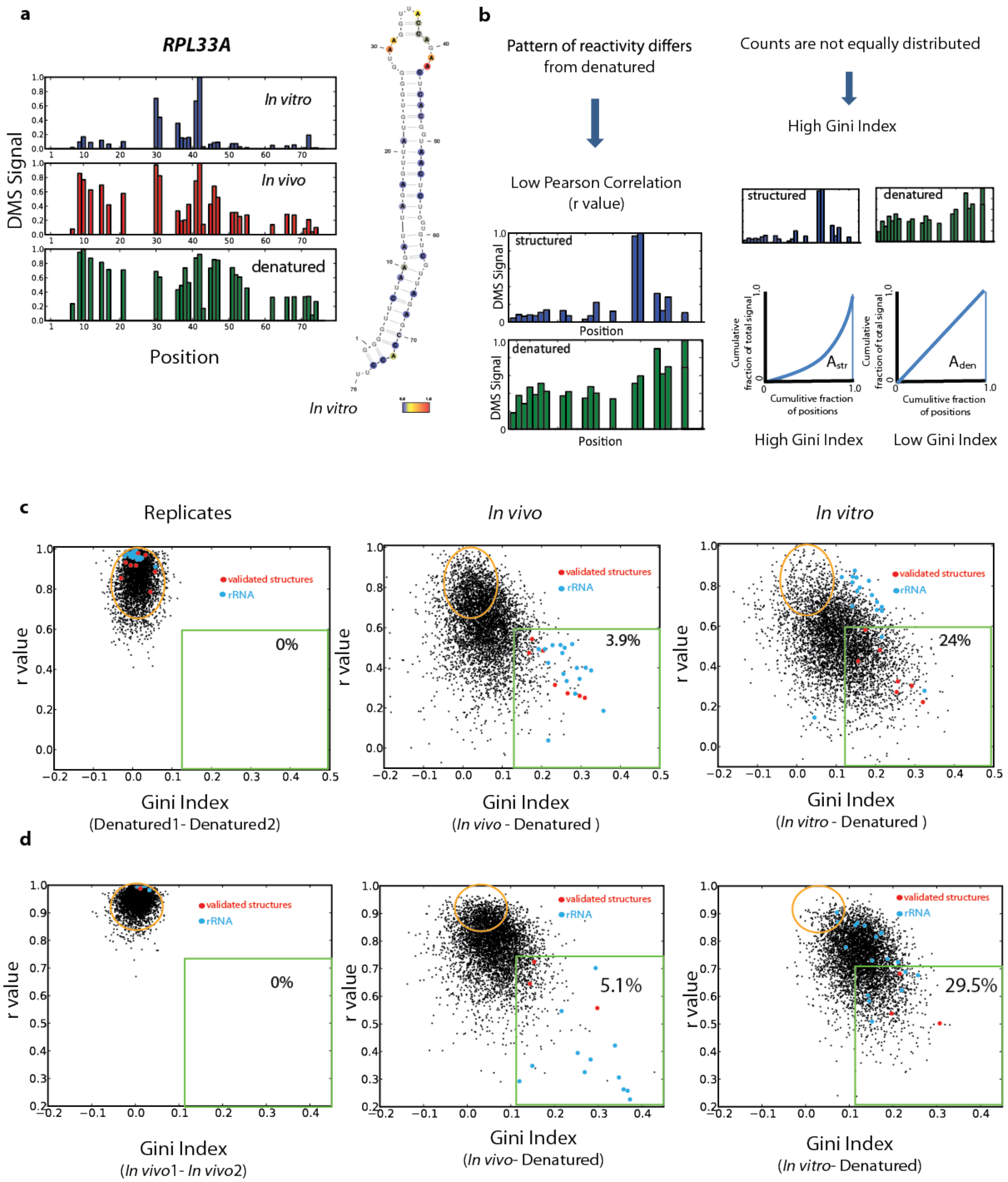
Rouskin et al., Figure 1



Rouskin et al., Figure 2



Rouskin et al., Figure 3



Rouskin et al., Figure 4

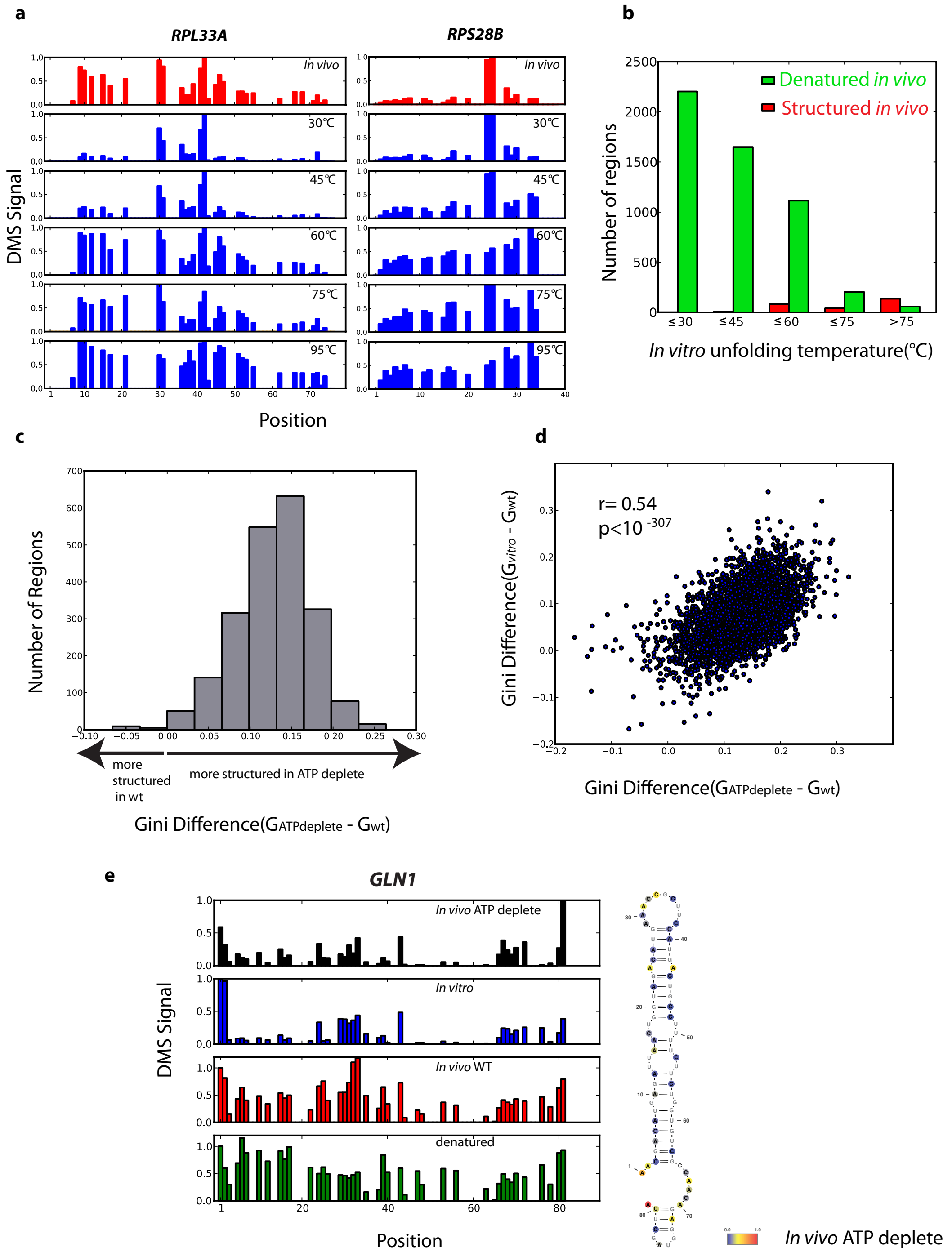


Figure 1 | Utilizing DMS for RNA structure probing by deep sequencing. a, Schematic of strategy for library preparation with DMS-modified RNAs. **b**, DMS-seq data is highly reproducible and robust against changes in time and DMS concentration in different biological replicates. **c**, *In vivo* DMS treatment dramatically enriches for sequencing reads mapping to A/C bases compared to untreated control. **d**, DMS-seq was completed for *in vivo*, denatured, and *in vitro* samples. The denatured sample served as an 'unstructured' control.

Figure 2 | Comparison of DMS-seq data to known RNA structures. a-b, DMS signal in (a) *HAC1* (position 1 corresponds to chrVI:75828) (b) *ASH1* (position 1 corresponds to chrXI:96245). Number of reads per position was normalized to the highest number of reads in the inspected region, which is set to 1.0. Also shown are the known secondary structures with nucleotides color-coded reflecting DMS-seq signal *in vivo*. **c,** DMS signal on 18S rRNA A bases plotted from least to most reactive. **d,** ROC curve on the DMS signal for A/C bases from the 18S rRNA. Threshold at 94% accuracy corresponds to 0.2 for the A bases.

Figure 3 | Identification of structured mRNA regions reveals far less structure *in vivo* than *in vitro*. a, DMS signal in *RPL33A* mRNA, position 1 corresponds to chrXVI:282824. *In vitro* DMS signal color-coded proportional to intensity and plotted onto the Mfold structure prediction. **b,** Schematic representation of the two metrics used to define structured regions within mRNAs. **c-d,** Scatter plots of Gini index versus r value from biological replicates or *in vivo* and *in vitro* relative to denatured samples for non-overlapping mRNA regions of 50 A/C nucleotides for (c) yeast and (d) K562 cells. 5,000 randomly selected regions are shown. Red dots represent regions spanning validated mRNA structures and blue dots are regions from rRNA. Evaluated regions have a minimum of 15 reads per A/C on average and their total number for *in vivo* data is (c) 23,412 and (d)17,242.

Figure 4 | Factors affecting the difference between mRNA structure *in vivo* and *in vitro.* **a,** Example of DMS signal changes for *RPL33A* and *RPS28B in vivo* and *in vitro* with increasing temperature. **b,** Histogram of *in vitro* unfolding temperature (T_{unf}) for denatured (green bars) or structured (red bars) regions *in vivo*. **c**, Histogram of Gini index difference between ATP-depleted and wildtype yeast samples. **d**, Gini index differences in ATP-depleted yeast or *in vitro* refolded mRNAs relative to wildtype yeast, calculated over 50 A/C nucleotides. **e,** Example of *in vivo* structure changes during ATP depletion. Position 1 corresponds to chrXVI:643,069.

METHODS

Media and Growth Conditions

Yeast strain BY4741 was grown in YPD at 30°C. Saturated cultures were diluted to OD_{600} of roughly 0.09 and grown to a final OD_{600} of 0.7 to 0.8 in YPD at the time of DMS treatment or mRNA harvesting. For ATP depletion experiments, cells were incubated for 1h in 10mM sodium azide and 10mM deoxyglucose prior to DMS treatment[31]. For 10ºC experiments, cells were grown to exponential phase and shifted to 10ºC by diluting the 30ºC media with 4ºC media. Mammalian cells were grown and treated with DMS in log phase (K562 cells) or at ~80% confluency for adherent cells (human foreskin fibroblast).

DMS Modification

For *in vivo* DMS modification, 15 ml of exponentially growing yeast at 30°C were incubated with 300-600 µl DMS for 2-4 min (which results in multiple modifications per mRNA molecule). Cells at 10ºC were incubated with 400 µl DMS for 40min to achieve similar modification levels as cells grown at 30ºC. DMS was quenched by adding 30 ml stop solution (30% BME, 25% Isoamyl Alcohol) after which cells were quickly put on ice, collected by centrifugation at 3000g and 4°C for 3 min, and washed with 15 ml 30% BME solution. Cell were then re-suspended in total RNA lysis buffer (10 mM EDTA, 50 mM NaOAc pH 5.5), and total RNA was purified with hot acid phenol (Ambion). PolyA(+) mRNA was obtained using magnetic poly(A)+ Dynal beads (Invitrogen). For *in vitro* and denatured DMS modifications, mRNA was collected in the same way as described above but from yeast that were not treated with DMS or quench solution. 4 µg of mRNA was denatured at 95°C for 2 min and either incubated in 0.2% DMS for 1 min (denatured control sample) or cooled on ice and re-folded in RNA folding buffer (10 mM Tris pH 8.0, 100 mM NaCl, 6 mM MgCl2) at 30°C for 30 min then incubated in 3-5% DMS for 2-5 min (*in vitro* sample). For intact ribosomes, polysomes were isolated on a sucrose gradient and treated with 4% DMS at 10° for 40 min in polysome gradient buffer (20mM Tris pH 8.0, 150mM KCl, 0.5mM DTT, 5mM MgCl2). DMS amounts/times were chosen to give a similar overall level of modification for the *in vivo*, *in vitro* and denatured sample. For *in vitro* probing at different temperatures, the RNA was re-folded at 45°C, 60°C, or 70°C. The DMS was quenched using 30% BME, 0.3M NaOAc, 2 µl GlycoBlue solution and precipitated with 1X volume of 100% Isopropanol. For K562 cells, 15ml of cells were treated with 300µl (in-vivo replicate 1) or 400µl (in vivo replicate 2) DMS and modified for 4 minutes. DMS was quenched by adding 30ml of 30% BME solution after which cells were quickly put on ice, collected by centrifugation at 1000g at 4°C for 3 min, and washed twice with 15 ml 30% BME solution. For fibroblast cells, 15cm3^2 plates with 15ml of media were treated with 300µl DMS for 4 min. The DMS was decanted and the plates were washed twice in 30% BME stop solution. Both K562 cells and fibroblasts were resuspended in Trizol Reagent and total RNA was isolated. PolyA(+) mRNA was obtained using oligotex resin (Qiagen).

Library Generation

Sequencing libraries were prepared as outlined in Fig. 1 with a modified version of the protocol used for ribosome profiling[32]. Specifically, DMS treated mRNA samples were denatured for 2 min at 95°C and fragmented at 95°C for 2 min in 1X RNA fragmentation buffer (Zn^{2+} based, Ambion). The reaction was stopped by adding 1/10 volume of 10X Stop solution (Ambion) and quickly placed on ice. The fragmented RNA was run on a 10% TBU (Tris Borate Urea) gel for 60 min. Fragments of 60-80 nucleotides in size were visualized by blue light (Invitrogen) and excised. Gel extraction was performed by crushing the purified gel piece and incubating in 300 µl DEPC treated water at 70°C for 10 min with vigorous shaking. The RNA was then precipitated by adding 33 µl NaOAc, 2 µl GlycoBlue (Invitrogen), and 900 µl 100% EtOH, incubating on dry ice for 20 min and spinning for 30 min at 4°C. The samples were then re-suspended in 7 µl 1X PNK buffer

(NEB) and the 3'phospates left after random fragmentation were resolved by adding 2 µl T4 PNK (NEB), 1 µl of Superase Inhibitor (Ambion) and incubating at 37°C for 1h. The samples were then directly ligated to 1 µg of microRNA cloning linker-1, /5rApp/CTGTAGGCACCATCAAT/3ddC/ (IDT DNA) by adding 2µl T4 RNA ligase2, truncated K227Q (NEB), 1 µl 0.1M DTT, 6 µl 50%PEG, 1 µl 10X ligase2 buffer, and incubating at room temperature for 1.5 hr. Ligated products were run on a 10% TBU gel for 40 min, visualized by blue light, and separated from unligated excess linker-1 by gel extraction as described above. Reverse transcription (RT) was performed in 20 µl volume at 52°C using SuperscriptIII (Invitrogen), and truncated RT products of 25-50 nucleotides (above the size of the RT primer) were extracted by gel purification. The samples were then circularized using circ ligase (Epicenter), and Illumina sequencing adapters were introduced by 8-10 cycles of PCR.

Sequencing and sequence alignment

Raw sequences obtained from Hiseq2000 (Illumina) corresponding to the DNA sequence from the RT termination products were aligned as described[33], against Saccharomyces cerevisiae assembly R62 (UCSC: sacCer2) downloaded from the Saccharomyces Genome Database on October 11, 2009 (SGD, http://www.yeastgenome.org/). Aligned reads were filtered so that no mismatches were allowed and alignments were required to be unique. Mammalian cells data was aligned to a transcript collection downloaded from RefSeq(http://www.ncbi.nlm.nih.gov/refseq/), in which each gene is represented by its longest protein-coding transcript. Aligned reads were filtered so that no more than 2 mismatches were allowed and the alignments were required to be unique. All data is deposited in Gene Expression Omnibus (series record GSE45803).

Computing the DMS signal

For the ribosomal RNA, the raw data was normalized proportionally to the most highly reactive residue after removing the outliers by 90% Winsorisation (all data above the 95th percentile is set to the 95th percentile)[34]. For the mRNA, the raw data was normalized proportionally to the most highly reactive base within the given structured window. Normalization of DMS data in windows of 50-200nt counteracts artifacts caused by mRNA fragmentation before polyA selection, which can lead to increased overall signal towards the 3'end of longer messages (since any 5' end that was broken off before the polyA(+) selection would be lost after the polyA(+) selection).

Computing the agreement with ribosomal RNA:

The secondary structure models for yeast ribosomal RNAs were downloaded from Comparative RNA Website and Project database (www.rna.icmb.utexas.edu/DAT/3C/Structure/index.php). The crystal structure model was downloaded from Protein Data Bank (PDB) (DOI:10.2210/pdb3u5b/pdb). The solvent accessible surface area[35] was calculated in Pymol, and DMS was modeled as a sphere with 3 Å radius (representing a conservative estimate for accessibility since DMS is a flat molecule). Accessible residues were defined as residues with solvent accessibility area of greater than 2 Å^2. True positive bases were defined as bases that are both unpaired in the secondary structure model and solvent accessible in the crystal structure model. True negative bases were defined as bases than are paired (A-U or C-G specifically) in the secondary structure model. The DMS data was normalized as described above.

Accuracy was calculated as the number of true positive bases plus the number of true negative bases divided by all tested bases.

Secondary structure models

Secondary structure models were generated using mfold[36]. Color coding by DMS signal was done using VARNA (http://varna.lri.fr/)

***In vivo* and *in vitro* DMS analysis**

Saccharomyces cerevisiae transcriptome coordinates were taken from Nagalakshimi et al.[37]. In total we collected between 140-200 million reads that uniquely aligned to the yeast genome per each sample (*in vivo*, *in vitro*, and denatured). Raw data was filtered for messages that have at least 15 reads on average per A or C position. The full yeast dataset is comprised of two biological and two technical *in vivo* replicates, two biological and one technical *in vitro* replicates, and two biological and one technical denatured replicates. For mammalian K562 dataset we collected two biological *in vivo* replicates (at 2%DMS and 2.7%DMS), one in vitro, and one denatured samples. For mammalian fibroblast data we collected of one *in vivo*, one *in vitro*, and one denatured samples. Sliding non-overlapping windows spanning a specified number of As and Cs starting at the 5'UTR were used to parse each message into a number of regions. Regions with matching length were taking from the 18S ribosomal RNA. A Gini Index and r value relative to that of a denatured control was calculated for each region. Highly structured regions in windows of 50 A or C nucleotides were defined with r value <0.55 and Gini Index >0.14 to encompass the *in vivo* regions containing validated structures and ribosomal RNA. Regions that are denatured *in vivo* were defined with r value >0.70 and Gini Index <0.08. Melting temperature ($T_{unf)}$ was defined as the lowest temperature where the DMS signal for a given mRNA region resembles denatured and was estimated based on the temperature at which a region reached r >= 0.70 or Gini Index <= 0.11. This represents relaxed criteria for unfolding to avoid bias towards overestimating thermostablity of regions due to sample variability caused by sequencing depth of *in vitro* temperature samples (which have 5-10 fold less coverage than the *in vivo*, *in vitro* (30°C), and denatured samples). For metagene analyses, the DMS signal was normalized in windows of 200 A or C nucleotides (relative to the top five most reactive residues), and the *in vivo* data was normalized by the denatured. Translation efficiency (TE) per message was calculated as number of ribosome footprints divided by the number of mRNA fragments.

Conservation Analysis

For a list of regions as well as secondary structure models supported by DMS data and conservation analysis visit:

http://weissmanlab.ucsf.edu/yeaststructures/index.html

Multiple sequence alignments generated by MultiZ[38] were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/sacCer2/multiz7way. Small (50 As or Cs) and large (100 As or Cs) overlapping regions with evidence for structure from the DMS probing experiment were inspected by the phylogenetic conservation analysis. The consensus secondary structure prediction was compared to normalized DMS data. The DMS values were separated in two groups for paired and unpaired bases, respectively. The median of both groups and the p-value from a one-sided Wilcoxon rank sum test is reported, testing the hypothesis that unpaired bases have higher DMS values. Both distributions are shown as box plots for each region on the website. For each region (i) a consensus secondary structure was predicted and (ii) the consensus structure was assessed for features typical of a functional RNA. The consensus secondary structure (i) was done using RNAalifold[39] which extends the classical thermodynamic folding for single sequences in two ways: it averages over the sequences while evaluating the energy for a given fold and it adds "pseudoenergies" to account for consistent or inconsistent mutations. The goal is to find the structure of the minimum free energy in this extended energy model. RNAalifold readily predicts a consensus structure even if there is no selection pressure for a conserved RNA structure. RNAz[40,41] was used to address the question if a predicted structure is likely to be a functional structure that is evolutionarily conserved. RNAz calculates two metrics typical for functional RNAs: (i) thermodynamic stability and (ii) evolutionary conservation. RNAz calculates a z-score indicating how much more stable a structure is compared to a random background of sequences of the same dinucleotide content. By convention, negative z-score indicates more stable structures and all reported z-scores are the average of all sequences in the alignment. RNAz calculates a metric known as structure conservation index (SCI). The SCI takes values between 0 and 1.0 means there is no structure conserved at all, 1 means the structure is perfectly conserved. The SCI is not normalized with respect to sequences conservation, so an alignment with sequences 100% conserved has by definition SCI = 1.0. RNAz evaluates z-score, SCI and sequence diversity of the alignment and provides an overall classification score that is based on a support vector machine classifier. It ranges from very negative values with little evidence for a functional RNA, over 0 which means undecided to high positive values with good evidence for a functional RNA. For convenience, this score is mapped to a probability of being a functional RNA which is reported in the results (the higher the better). A total of 189 structures with RNAz significance value > 0.5 and a correlation p-value between the predicted structure and the DMS signal of < 0.01 are displayed on the aforementioned website.

Functional UTR Cloning

A fluorescent Venus reporter driven by a Nop8 promoter (chrXV:52262-53096) and *C. albicans ADH1* terminator was genomically integrated into yeast strain BY4741 at the *TRP1* locus (chrIV:461320-462280). Plasmids containing kanamycin resistance and the untranslated region (UTR) of interest were made in a pUC18 plasmid backbone (Thermo Scientific). For the *PMA1* 5'UTR, the entire 1kb promoter region and 5'UTR (chrVII:482672-483671) was used. The pNop8 promoter was retained for the *SFT2* 5'UTR investigation, with only the Nop8 5'UTR replaced by the *SFT2* 5'UTR. All 3'UTRs were cloned to include >100bp after evidence of transcription ends (see Extended Data Fig. 5-6 and Extended Data Table 1 for sequence of *PMA1*, *SFT2,* and *PRC1* structures). BY4741-Venus yeast were transformed using the standard technique of homologous recombination from a plasmid PCR product containing either a wildtype, mutant, or compensated UTR. Successfully transformed yeast were identified by check PCR and subsequently sequenced to confirm the presence of only the desired mutations. Mutagenesis in the endogenous PMA1 locus was done via the strategy described above for the PMA1 5'UTR, except homologous recombination was targeted to the endogenous PMA1 locus and surrounding genomic region rather than to Venus. After sequencing to confirm the presence of only the desired mutations, PMA1 was C-terminally tagged with Venus via PCR product from the pFA6a-link-yEVenus-SpHIS5 plasmid[42].

Flow Cytometry

A saturated yeast culture was diluted 1:200 fold in minimal media and grown at 30°C for 6-8 hr before flow cytometry using a LSRII flow cytometer (Becton Dickinson) and 530/30 filter. 10°C cultures were grown for 72 hr. Venus signal from each cell was normalized to cell size (Venus/sidescatter) using Matlab 7.8.0 (Mathworks)[43], and once normalized, all events (~20,000 per experiment) were averaged for a final Venus/sidescatter value.

References:

31. Kortmann, J., Sczodrok, S., Rinnenthal, J., Schwalbe, H. & Narberhaus, F. Translation on demand by a simple RNA-based thermosensor. *Nucleic Acids Res.* **39,** 2855–2868 (2011).
32. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7,** 1534–1550 (2012).
33. 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).
34. Hastings, C., Mosteller, F., Tukey, J. W. & Winsor, C. P. Low Moments for Small Samples: A Comparative Study of Order Statistics. *The Annals of Mathematical Statistics* **18,** 413–426 (1947).
35. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55,** 379–400 (1971).

36. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

37. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).

38. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).

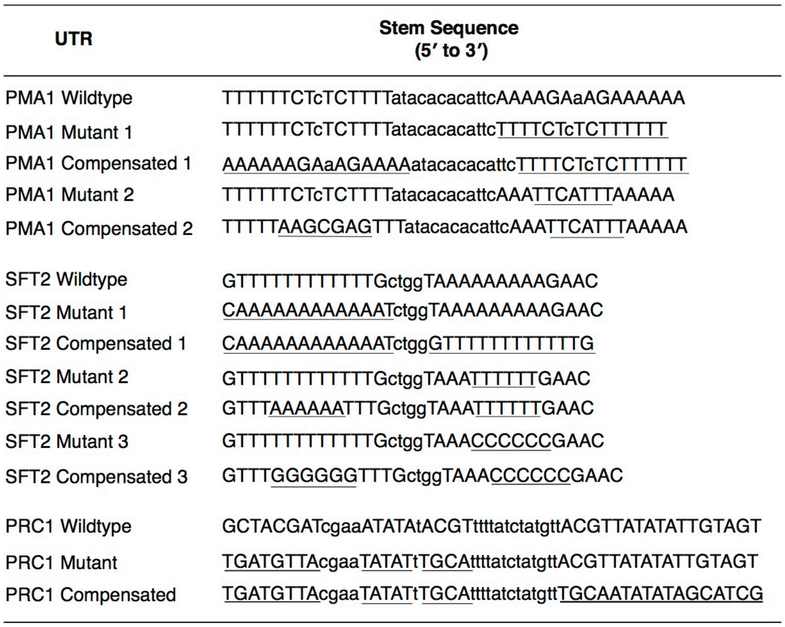
39. Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9,** 474 (2008).

40. Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 2454–2459 (2005).

41. Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. Rnaz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput* 69–79 (2010).

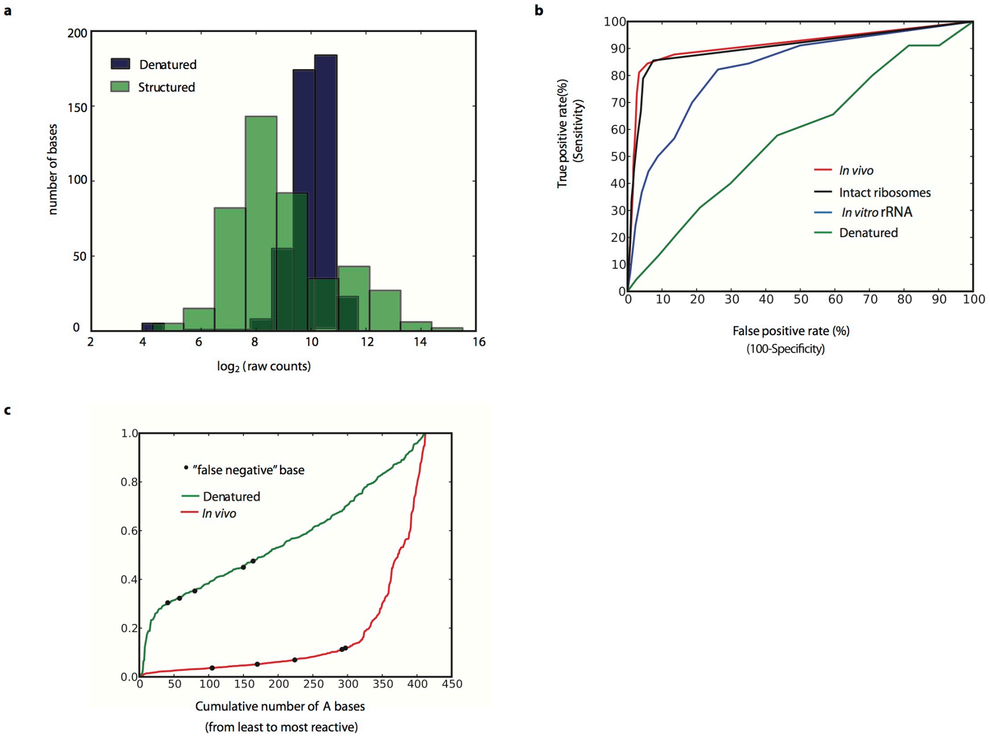
42. Sheff, M. A. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in Saccharomyces cerevisiae. *Yeast* **21,** 661–670 (2004).

43. Brandman, O. *et al.* A Ribosome-Bound Quality Control Complex Triggers Degradation of Nascent Peptides and Signals Translation Stress. *Cell* **151,** 1042–1054 (2012).

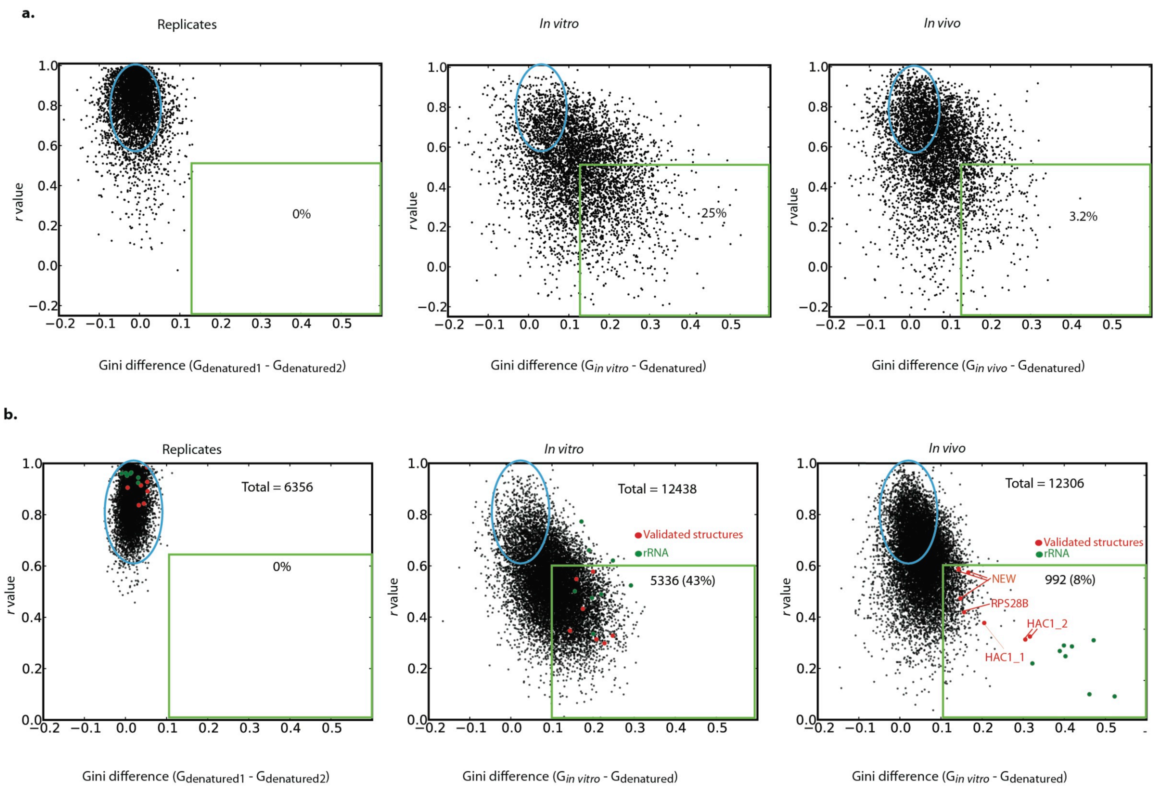


UTR	Stem Sequence (5′ to 3′)
PMA1 Wildtype	TTTTTTCTcTCTTTTatacacacattcAAAAGAaAGAAAAAA
PMA1 Mutant 1	TTTTTTCTcTCTTTTatacacacattcTTTTCTcTCTTTTTT
PMA1 Compensated 1	AAAAAAGAaAGAAAAatacacacattcTTTTCTcTCTTTTTT
PMA1 Mutant 2	TTTTTTCTcTCTTTTatacacacattcAAATTCATTTAAAAA
PMA1 Compensated 2	TTTTTAAGCGAGTTTatacacacattcAAATTCATTTAAAAA
SFT2 Wildtype	GTTTTTTTTTTTTGctggTAAAAAAAAAGAAC
SFT2 Mutant 1	CAAAAAAAAAAAATctggTAAAAAAAAAGAAC
SFT2 Compensated 1	CAAAAAAAAAAAATctggGTTTTTTTTTTTTG
SFT2 Mutant 2	GTTTTTTTTTTTTGctggTAAATTTTTTGAAC
SFT2 Compensated 2	GTTTAAAAAATTTGctggTAAATTTTTTGAAC
SFT2 Mutant 3	GTTTTTTTTTTTTGctggTAAACCCCCCGAAC
SFT2 Compensated 3	GTTTGGGGGGTTTGctggTAAACCCCCCGAAC
PRC1 Wildtype	GCTACGATcgaaATATAtACGTttttatctatgttACGTTATATATTGTAGT
PRC1 Mutant	TGATGTTAcgaaTATATtTGCAttttatctatgttACGTTATATATTGTAGT
PRC1 Compensated	TGATGTTAcgaaTATATtTGCAttttatctatgttTGCAATATATAGCATCG

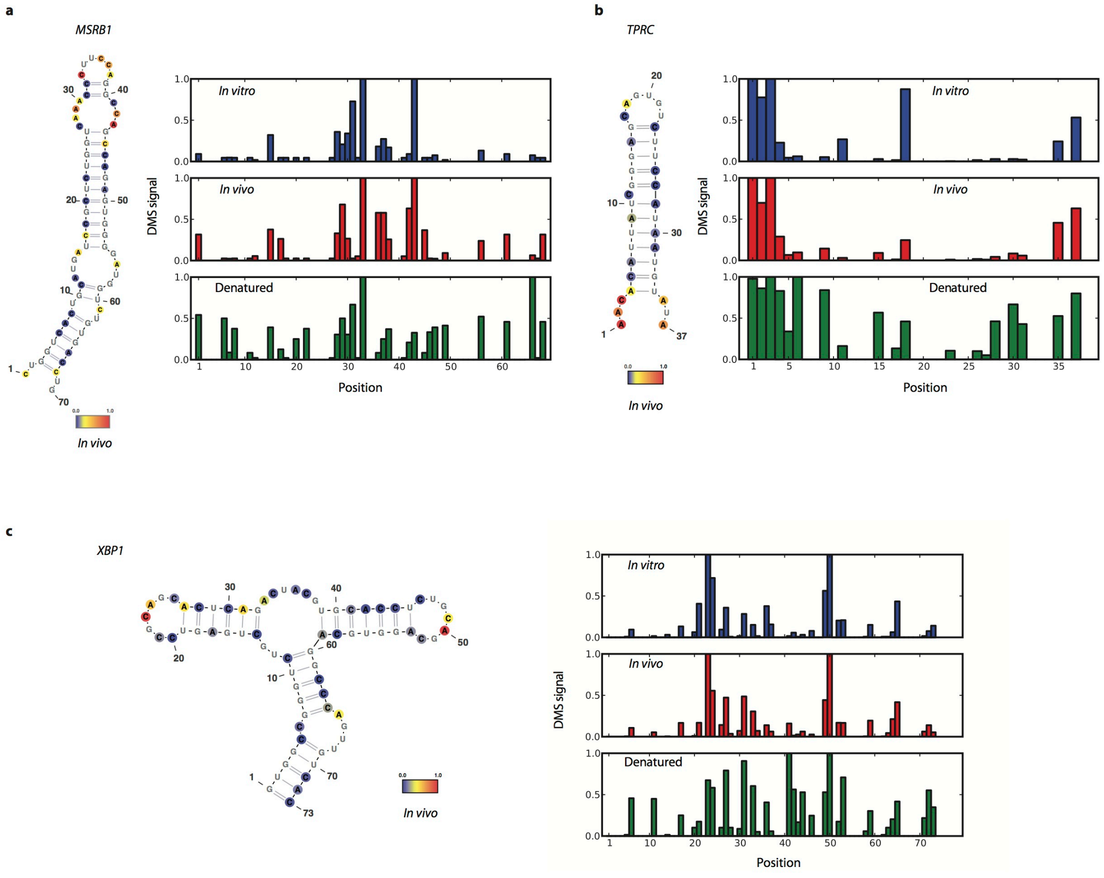
Extended Data Table 1 | Sequences of functional structure mutations. 5'-3' mRNA structure sequences are listed. Lowercase letters correspond to non-paired bases, found in bulges or loops within the stem. Mutated bases are underlined.



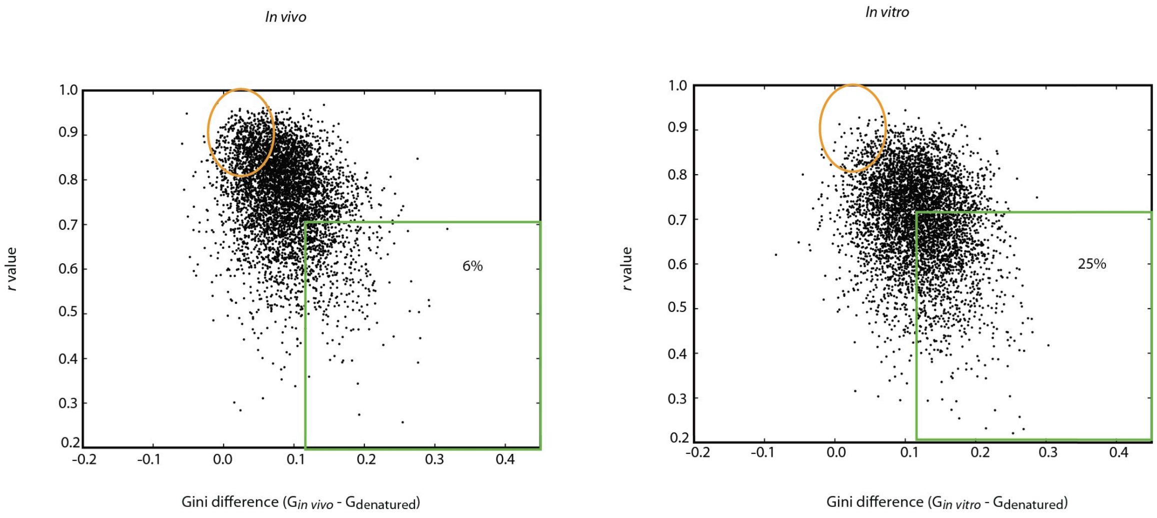
Extended Data Figure 1 | Ribosomal RNA analysis. **a**, Histogram of raw counts distribution for denatured and structured 18S rRNA. Log2 (raw counts) for A bases plotted for in vivo and denatured samples. **b**, ROC curve on the DMS signal for A and C bases from the 25S rRNA in denatured, in vitro, intact ribosomes, and in vivo samples. True positives are defined as bases that are both unpaired and solvent accessible, and true negatives are defined as bases that are paired. **c**, DMS signal on all of the 18S rRNA A bases plotted from least to most reactive in the denatured or in vivo samples. The A bases that are false negatives relative to the crystal structure are colored as black dots on both the denatured and in vivo samples.



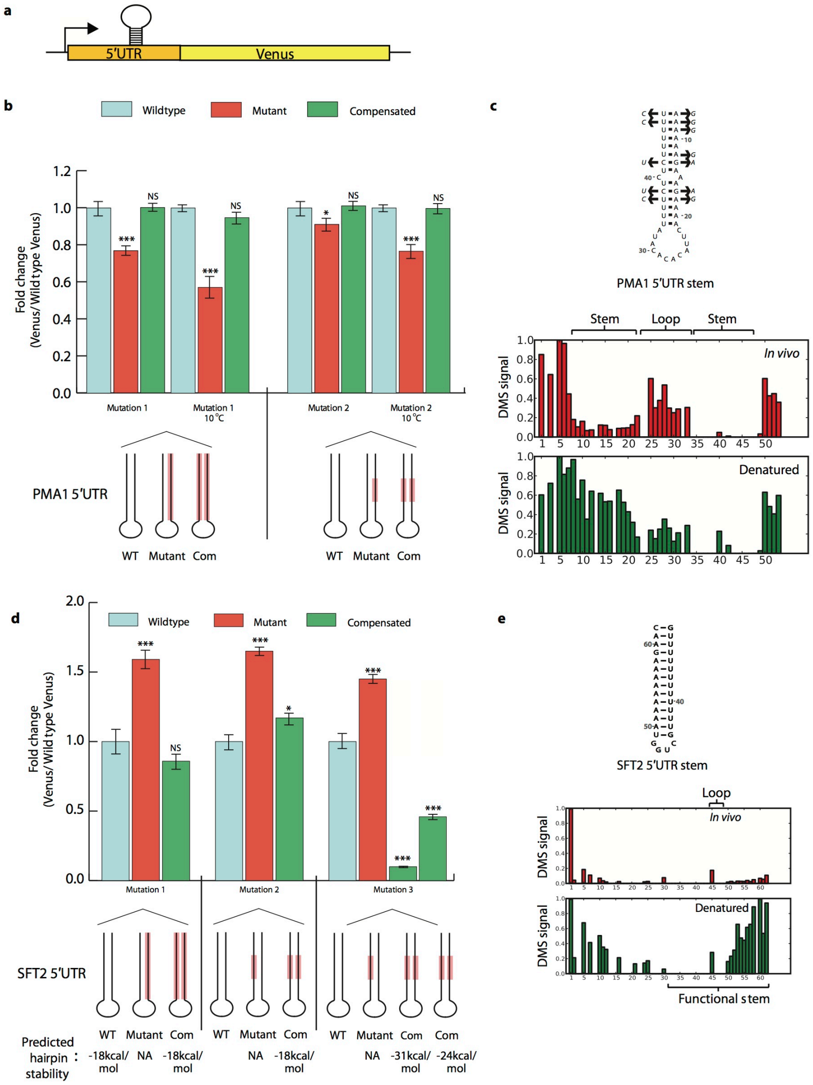
Extended Data Figure 2 | mRNA structure analysis with different window sizes. Scatter plots of Gini Index verses r values in replicate samples and for in vivo or in vitro samples relative to denatured sample for mRNA regions with an average of at least 15 counts per position, spanning the sequence of (**a**) 25 A/C nucleotides (5000 randomly selected regions are shown) or (**b**) 100A/C nucleotides. Shown are regions spanning validated secondary structures (red dots) and regions from the 18S rRNA (green dots).



Extended Data Figure 3 | Agreement of DMS-seq with validated structures in mammalian K562 cells. Raw DMS counts were normalized to the most reactive nucleotide in the given region. A and C bases were normalized separately. The DMS signal is color coded proportional to intensity and plotted onto the secondary structure model of (**a**) MSRB1 selenocysteine insertion element, nucleotide 1 corresponds to nucleotide 966 of the transcript, (**b**) TPRC iron recognision element, nucleotide 1 corresponds to nucleotide 3901 of the transcript, (**c**) XBP1 conserved non cannonical intron recognized by Ire1, nucleotide 1 corresponds to nucleotide 520 of the transcript.

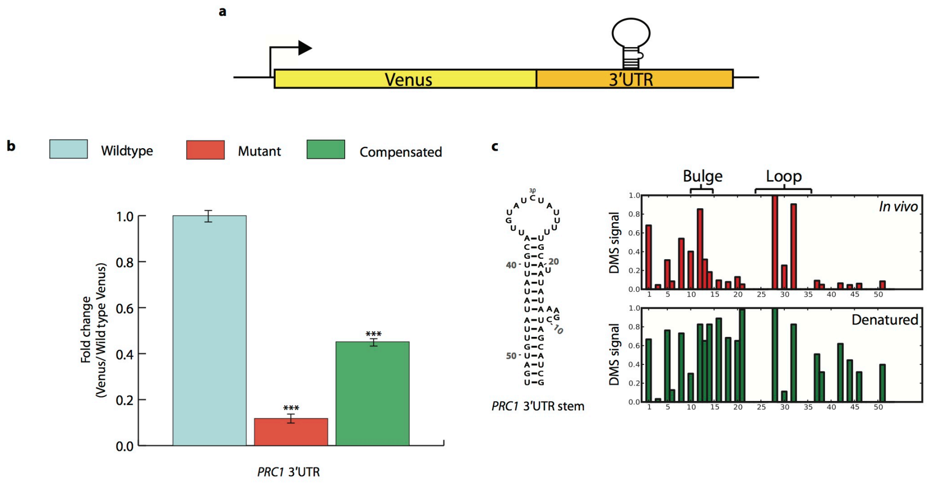


Extended Data Figure 4 | Global mRNA analysis of human foreskin fibroblast cells. Scatter plots of Gini Index verses r values for in vivo or in vitro samples relative to denatured sample for mRNA regions spanning the sequence of 50 A/C nucleotides.



Extended Data Figure 5 | Functional verification of novel 5’UTR structures in vivo.

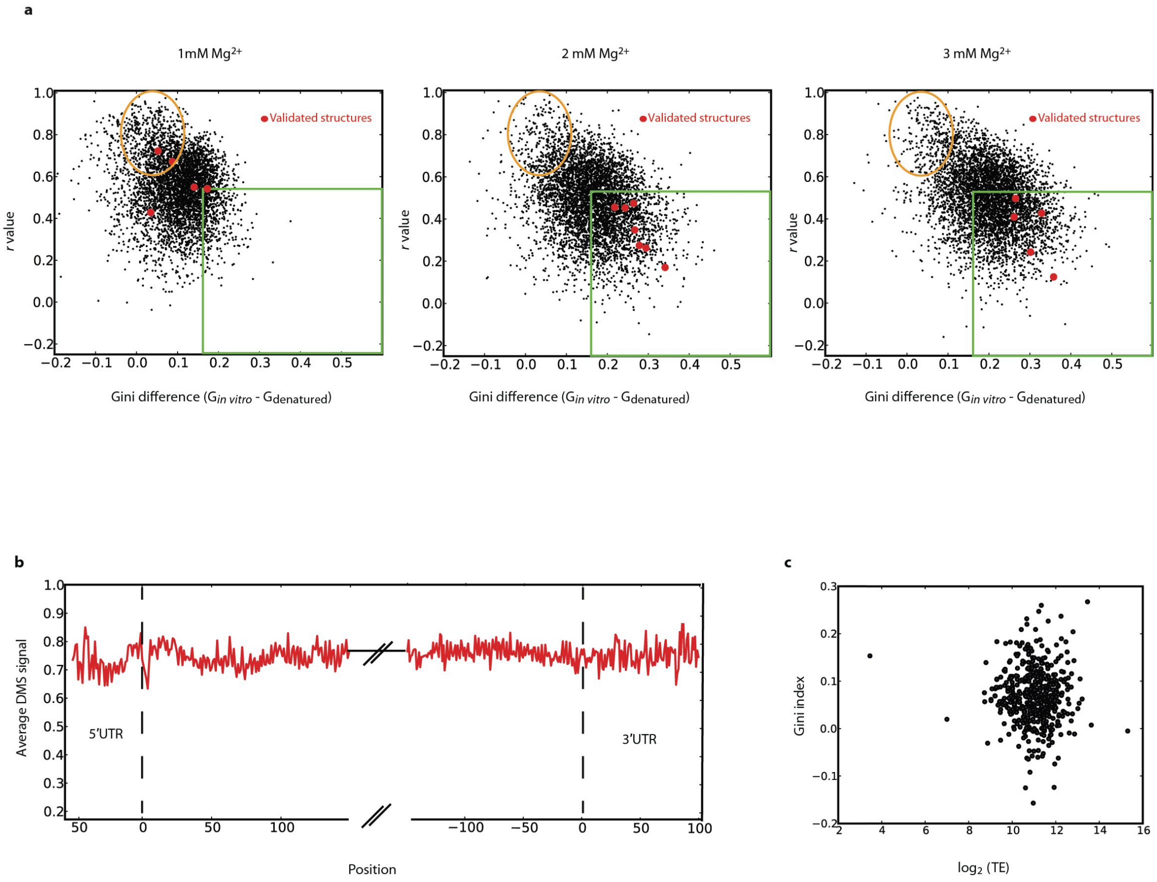
a, Putative 5’UTR stems were manipulated in the context of a Venus reporter in vivo. **b**, PMA1 5’UTR structure was mutated and compensated twice with Venus reporter, differing in number and character of bases mutated. Mutation location shown in red on schematic. Reported p-values relative to wildtype Venus levels, calculated by two-sided t-test (p < .01, .001, and .0001 represent *, **, and *** respectively). For all graphs, Venus signal normalized to cell size before calculating fold change and data presented is from two biological and two technical replicates. Error bars represent SEM. **c**, Secondary structure of functional PMA1 5’UTR stem, with compensatory mutations (arrows) found in S. paradoxus, S. mikatae, S. kudriavzevii, and S. bayanus. Raw DMS signal shown below (position 1 = chrVII:482745). **d**, SFT2 5’UTR structure was mutated and compensated three times in Venus reporter system, differing in number, character, and location of bases mutated. Mutation location shown in red on schematic. Stem stability as predicated by mfold. Reported p-values relative to wild type Venus levels, also by two-sided t -test (p < .01, .001, and .0001 represent *, **, and *** respectively). Error bars represent standard deviation c, Secondary structure of functional SFT2 5’UTR stem. Position 1 = chrII:24023.



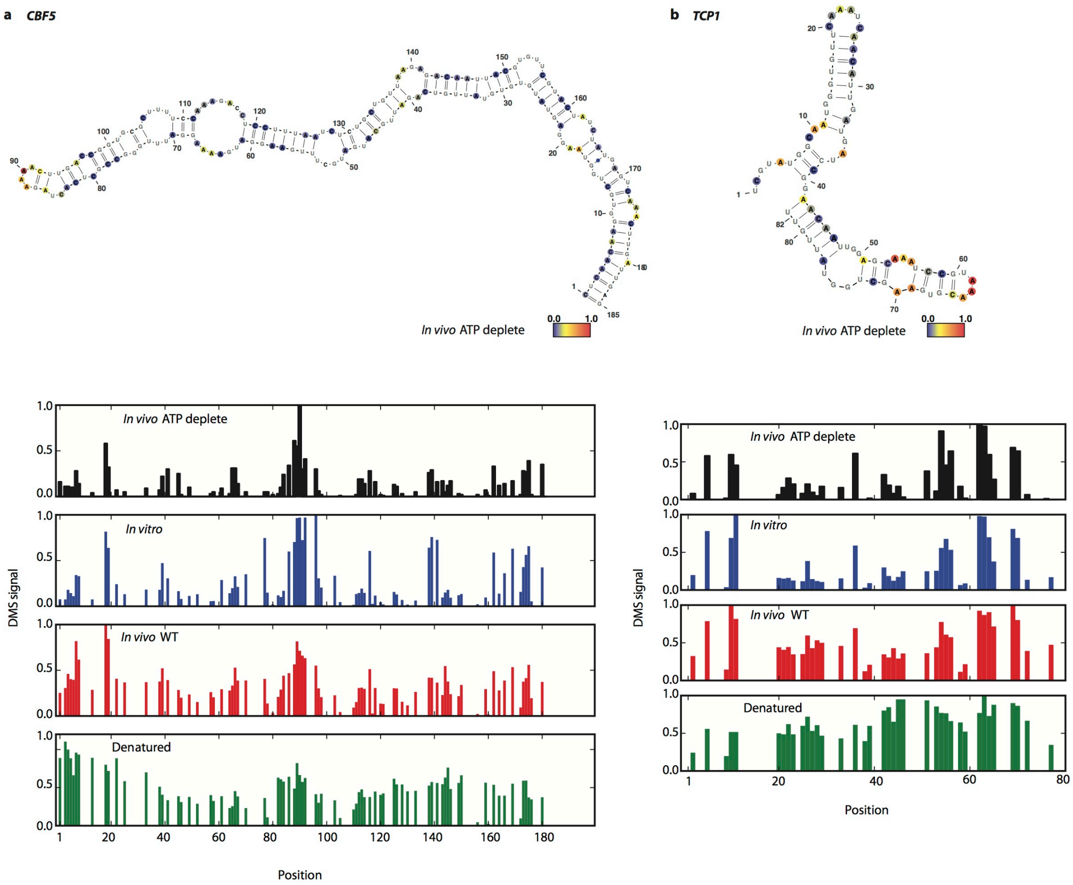
d

3'UTR	Chromosome and position	Mutant fold change	Compensated fold change
SAM2	chr IV: 1,453,207–1,453,250	1.15 (p = 0.02)	1.46 (p = 1.3 x 10^{-5})
YDL085C-A	chr IV: 302,416–302,440	1.03 (p = 0.48)	0.99 (p = 0.79)
TSC3	chr II: 356,285–356,245	1.02 (p = 0.57)	0.94 (p = 0.05)

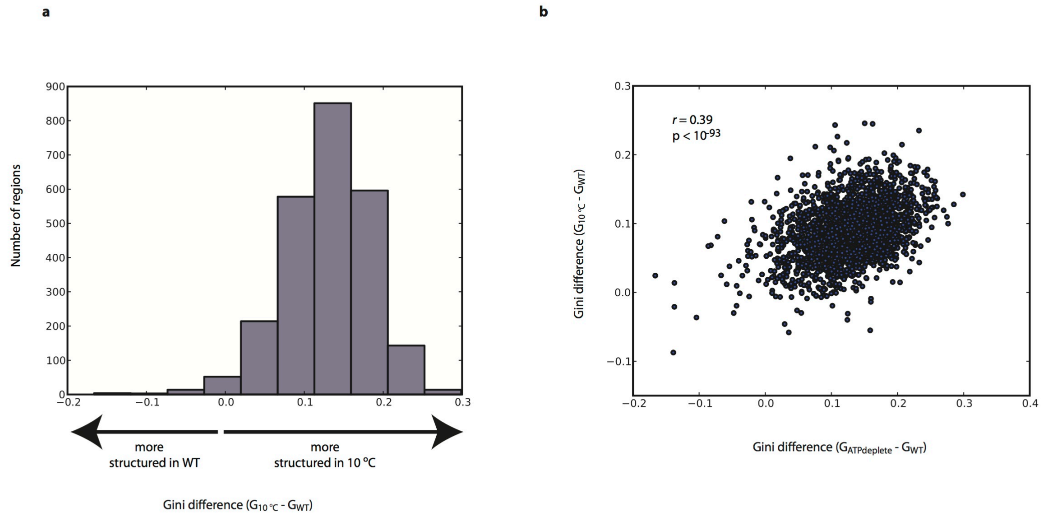
Extended Data Figure 6 | Functional verification of novel PRC1 3'UTR structure in vivo. a, Putative 3'UTR stems were manipulated in the context of a Venus reporter in vivo, followed by Venus quantitation with flow cytometry. **b**, PRC1 3'UTR structure was mutated and compensated in Venus reporter system. For all data, reported p-values relative to wildtype Venus levels, calculated by two-sided t-test (p < .01, .001, and .0001represent *, **, and *** respectively). Venus signal was normalized to cell size with fold change reported relative to Venus levels seen with the wild type stem. All results shown are derived from four measurements: two biological and two technical replicates. Error bars show standard deviation. **c**, Secondary structure of functional PRC1 3'UTR stem, shown with raw DMS signal for in vivo and denatured samples. Position 1 = chrXIII:863554. **d**,Weakly structured 3'UTRs in vivo were tested for function as in (**b**) but reveal little effectwhen mutated and no evidence for compensation.



Extended Data Figure 7 | Global analysis of mRNA structure. a, In vitro DMS-seq on RNA re-folded in different Mg^{+2} concentrations. **b**, Metagene plot of the average DMS signal (normalized to denatured control) over 5'UTR, coding, and 3'UTR regions. **c**, Scatter plot of Gini Index (calculated over the first 100 A/C bases) of in vivo messages (relative to denatured) verses translation efficiency.



Extended Data Figure 8 | In vivo structures forming in ATP depleted conditions. Raw DMS counts were normalized to the most reactive nucleotide in a given region. The DMS signal is color coded proportional to intensity and plotted onto the mfold predicted secondary structure model of (**a**) CBF5, nucleotide 1 corresponds to chrXII position 506,479 (**b**) TCP1, nucleotide 1 corresponds to chrIV position 887,991.



Extended Data Figure 9 | Analysis of mRNA structure at 10ºC. a, Histogram of gini index difference (calculated over 100A or Cs) between 10ºC and wt (30ºC) samples. **b**, Scatter plot of the gini index differences in ATP depleted or 10ºC yeast relative to wt yeast calculated over 50 As or Cs.

CHAPTER 3

Operon mRNAs are organized into ORF-centric structures that specify translation efficiency

Operon mRNAs are organized into ORF-centric structures that specify translation efficiency

David H. Burkhardt[1,2,5]*, Silvi Rouskin[4-6]*, Gene-Wei Li[4-6†], Jonathan S. Weissman[4-6†],
Carol A. Gross[2,3,5†]

[1]Graduate Group in Biophysics,

[2]Department of Microbiology and Immunology,

[3]Department of Cell and Tissue Biology,

[4]Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute,

[5]California Institute of Quantitative Biology,

[6]Center for RNA Systems Biology,

University of California, San Francisco, CA 94158, USA.

*These authors contributed equally to this work.

† To whom correspondence should be addressed. E-mail: gene-wei.li@ucsf.edu, weissman@cmp.ucsf.edu, cgrossucsf@gmail.com

Summary

Prokaryotic mRNAs are organized into operons consisting of discrete open reading frames (ORFs) that are differentially translated by as much as 100-fold. To understand the mRNA features instructing these differing translation efficiencies, we monitored the relationship between mRNA structure and translation on endogenous messages genome-wide *in vivo*. We find that operon mRNAs are organized into structural domains divided by ORF boundaries. This modular mRNA structure, rather than Shine-Dalgarno strength, specifies ORF translation efficiency. Upon cold shock, mRNA structure increases and translation decreases, but both are restored by massive induction of the Cold Shock Proteins (Csps). Csps modulate global mRNA structure and autoregulate their expression via an RNA element cued to the cellular environment, enabling mRNA structure surveillance both at cold and normal growth temperatures. Operons and Csps are present in all bacteria, suggesting that the organization of operonic mRNA structure and its surveillance system we describe are universally used to set and maintain translation.

Introduction

Protein synthesis is the most energetically costly process in bacteria, consuming ~ 50% of cellular energy (Russell and Cook, 1995). To optimize cellular efficiency, the rate of synthesis of each protein is carefully controlled. The bacterial strategy to achieve this control entails organizing open reading frames (ORFs) into operons so that mRNA level for genes with related functions are co-regulated (Jacob and Monod, 1961). Fine-grained control of protein synthesis rate is then achieved by tuning translation efficiency (TE) of each ORF, with efficiency of adjacent ORFs varying by as much as 100-fold (Li et al., 2014). Thus, optimal energy utilization depends on the ability to reliably drive ORF-specific translation efficiencies. Understanding the rules that govern how mRNA sequence features drive these specific translation efficiencies is important for decoding genomes and for designing synthetic ORFs.

The role of *cis* elements proximal to the ribosome-binding site in setting and maintaining translation efficiencies on *E. coli* ORFs has been extensively studied. Translation initiation minimally requires an accessible Shine-Dalgarno (SD) sequence upstream from the initiation codon (Steitz and Jakes, 1975). Consequently, highly stable structures in direct proximity to the initiation codon diminish translation efficiency (de Smit and van Duin, 1990; Hall et al., 1982; Lodish, 1970). Rare codons that disfavor structure are enriched in positions immediately following the translation start site (Bentele et al., 2013; Eyre-Walker and Bulmer, 1993; Scharff et al., 2011), and mutational analysis of these early codons in synthetic reporters has shown that changes in protein expression can be explained by changes to predicted RNA structure at the translation start (Goodman et al., 2013; Kudla et al., 2009; Salis et al., 2009). However, biophysical models based on structural prediction around the start codon are only weakly predictive of relative translation efficiencies of messages that differ in sequence beyond the early coding region (Kosuri et al., 2013) or on endogenous messages (Li et al., 2014).

mRNA structural elements extending past the ribosome binding site into ORF bodies (Wikström et al., 1992) or into 5' untranslated leaders (Borujeni et al., 2013; Marzi et al., 2007) can both inhibit and enable translation initiation, raising the possibility that *cis* features in mRNA sequence beyond the ribosome binding site may play a role in setting the translational efficiency of each ORF. Using recently developed global technologies (Ingolia et al., 2009; Li et al., 2014; Oh et al., 2011; Rouskin et al., 2014), we simultaneously probed the *in vivo* structure and translation of endogenous messages in *E. coli*. We find that mRNA structure of operons is organized around open reading frames, and is strongly correlated with translation efficiency.

We then used cold temperature stress, anticipated to drive an increase in RNA structure, to determine whether *E. coli* can sense and repair changes to mRNA structure. We find that cold shock drives a global increase in mRNA structure with concomitant defects in translation initiation and that the immediate cold recovery program alters the structure of each mRNA in a gene-specific manner. We find that this program is dependent on induction of the Csp RNA binding proteins (Goldstein et al., 1990; Jiang et al., 1997) to modulate mRNA structure globally, and RNase R to degrade stabilized mRNA. Finally, Csps autoregulate their expression by modulating their 5' UTRs structure, and this structural transition is cued to the global structure in the cell, enabling appropriate transcript structure in all conditions.

Results

Development of global structure determination in E. coli

New genomic technologies enable the determination of RNA structure *in vivo* on a global scale (Ding et al., 2014; Rouskin et al., 2014; Wan et al., 2014). We monitored global *in vivo* RNA structure with DMS (dimethyl sulfate)-seq (Rouskin et al., 2014), which uses next generation sequencing to determine chemical accessibility of RNA to DMS, a reagent that reacts with unpaired adenosine and cytosine nucleotides (Inoue and Cech, 1985) (Figure 1A). DMS-seq, adapted here to *E. coli*, is highly reproducible (Figure S1A) and in strong agreement both with the *E. coli* ribosome crystal structure (Figure 1B), and a mutationally-verified *E. coli* mRNA structure (Figure 1C) (Wikström et al., 1992). We quantified the degree of secondary structure on each open reading frame using the Gini index metric, which measures the variability in reactivity of residues in the region being examined (Rouskin et al., 2014). A low Gini index indicates an even distribution of DMS-seq reads, and occurs when a region of the mRNA is unstructured. A high Gini index occurs when a subset of residues is strongly protected from DMS reactivity, and indicates a high degree of structure (Figure S1B-D). We found that the degree of RNA secondary structure varied greatly between ORFs: some are nearly as structured as rRNA, whereas some are close to the denatured state (Figure 1D).

mRNA structure is organized around open reading frames and specifies TE

Despite the variability in the degree of secondary structure among ORFs, the degree of structure within a given ORF is well correlated (Figure 2A). This relationship holds even when controlling for GC content (Figure S2A). Structural correlation does not extend to adjacent ORFs on the same mRNA (Figure 2B), suggesting that the structures are a property of the ORFs rather than of the polycistronic transcript.

We next asked whether structure is correlated with translation efficiency, which we quantified by combining ribosome density obtained from ribosome profiling with total mRNA measured by mRNAseq (Ingolia et al., 2009; Li et al., 2014; Oh et al., 2011). Indeed, better-translated ORFs have lower structure, and the difference in the degree of folding between adjacent ORFs is highly predictive of their relative levels of translation (Figure 2B, S1D). Notably, ORF pairs with overlapping stop and start codons show as much variability in their relative translation as non-overlapping ORF pairs (Figure 2C). We next expanded our analysis beyond operons to all ORFs, and found that structure is strongly correlated to TE on all endogenous open reading frames ($\rho = 0.76$, Figure 2E). These results indicate that ORF-specific RNA structure specifies differential translation between genes in the same operon.

Bacterial operons are densely packed with ORFs, and the majority of adjacent ORFs (62%) are separated by only 25nt or less (Figure 2D). It is therefore important to examine how structure changes at ORF boundaries. At translational start sites, the local degree of folding correlates with the TE only downstream from the start site and rapidly diminishes upstream of the start site, whereas structure upstream of the start site is correlated with the TE of the upstream ORF (Figure 2F). Thus, structure undergoes a sharp transition at ORF boundaries and polycistronic mRNAs consist of distinct structural domains.

mRNA sequence drives the organization of mRNA structure around open reading frames

We evaluated whether the ORF-centric structures of mRNAs *in vivo* arises as an intrinsic property of sequence by using DMS-seq to determine the structure of mRNAs refolded *in vitro* at 37°C. *In vitro* RNA structure was correlated with *in vivo* TE (ρ = 0.48, Figure 2G, S2B-C), whereas control samples without DMS modification was not correlated (ρ = .05, Figure S2D). This correlation persists through windows that do not include the translation start site (Figure 2H). The correlation between *in vivo* TE and structure was also maintained after addition of the translation initiation inhibitor kasugamycin at 10°C, where longer mRNA half-life permits this measurement (see below). Importantly, *in vitro* refolded mRNAs possess a sharp structural boundary between adjacent ORFs similar to that observed on *in vivo* mRNA (Figure 2H). Computationally predicted ORF-length structures also retained a strong correlation to translation efficiency (ρ = 0.48, Fig S2E-F). As the correlation of *in vivo* mRNA structures with TE is stronger than the correlation to structures determined *in vitro* or computationally, the contribution of the ribosome to mRNA folding, as well as differences in folding environment (e.g. salt, molecular crowding) and pathway (lack of vectorial folding) contribute to the eventual *in vivo* structure. Notably, the strength of the Shine-Dalgarno sequence does not have predictive power on TE, even after controlling for structure as measured by Gini (Figure S2G). *In toto*, these analyses indicate that the linear sequences of bacterial mRNAs encode not only open reading frames, but also the blueprint for ORF-wide secondary structures that specifies levels of translation.

Whereas ORFs are marked by start and stop codons, the signatures that define structural domains have remained elusive. To understand how structural boundaries might be set up by the linear sequence, we computationally predicted the structure of mRNA extending -250 to +250 nt from the translation start at the boundary of adjacent ORF pairs. Because folding algorithms often predict a large ensemble of possible folds for a long stretch of RNA, we used the DMS-seq data (both *in vivo* and *in vitro*) to constrain the predictions by forcing positions that were highly DMS-modified to be unpaired in the predicted structures. We then examined the propensity for each position to interact with each other position. Consistent with previous studies, we found a lack of structure in the immediate vicinity of the start sites for most ORFs (Figure S2H). Downstream from this structure-free zone (25-50 nt), endogenous mRNA has a high propensity to base pair with regions further downstream, i.e. pairing within the same ORF (Figure 2I). Conversely, nucleotides located 25-50nt upstream of the start site have a strong preference to interact with regions further upstream in the preceding ORF (Figure 2I). Importantly, these results are similar for both *in vivo* and *in vitro* probed RNA. Therefore, a sequence-driven sharp transition in the directionality of pairing around start sites can provide a mechanism for organizing structure around ORFs.

Cold shock increases mRNA structure and drives a global ribosome run-off

Given the importance of structure in setting translational efficiency, we asked whether the cell is able to monitor and repair the structure of its mRNAs. Cold shock (shift to 10˚C) is expected to increase mRNA structure, and therefore provides an avenue to determine whether such a system exists.

Upon shift to cold, protein synthesis dramatically decreases and cell growth stops, resuming after a ~6 hr lag (Friedman et al., 1971; Ng et al., 1962). Existing evidence suggested that cold shock inhibits translation initiation, as polysomes decrease and monosomes increase (Friedman et al., 1969; Jones and Inouye, 1996). Additionally, at 5°C, a temperature at which ribosomes dissociate, it was observed *in vivo* that ribosomes on a specific RNA phage-encoded transcript complete one round of translation following cold shock but do not initiate new rounds (Friedman et al., 1971).

With ribosome profiling experiments, we identified an immediate global reduction in translation initiation after shift to 10°C, as ribosome density is depleted from the 5' end of all genes (Figure 3A). Gradual run-off of ribosomes that had initiated translation at 37˚C is reflected in a gradual decrease in ^{35}S-methionine incorporation, plateauing at 30 min after cold shock when the run-off observed by ribosome profiling is presumably complete. At that point, ^{35}S-methionine incorporation indicates a 200-fold reduction in translation initiation (Figure 3B). Concomitant DMS-seq measurements indicated a large, gene-specific increase in mRNA structure across the transcriptome relative to 37°C (Figure 3C), with structure remaining correlated with TE (Figure S3A). Similar to 37°C, the mRNA structure probed *in vitro* is correlated with TE *in vivo* at 10°C (Figure S3D). Furthermore, we removed the contribution of translation on structure *in vivo* by treating cells with the translation initiation inhibitor kasugamycin, and observed the same trend (Figure S3C). Taken together, these results indicate that cold shock induces a global and sequence-dependent increase in mRNA structure that leads to reduction in protein synthesis.

After the initial shock, total protein synthesis increases ~4-fold during cold recovery prior to resumption of growth (Figure 3B). We tested whether remodeling mRNA structures drives this increase by comparing global mRNA structure and TE at 6 hr vs. 30 min after cold shock. Notably, this enabled comparison of TE and structure for the same set of mRNAs in the same environmental condition, revealing the effect of internal changes within the cell. Structure and TE remain correlated (Figure S3B), and their dramatic global changes are also correlated (Figure 3D). This result indicates a recovery program that drives a decrease in the mRNA structure of specific genes to permit their TE to increase.

RNase R and Csps mediate initial recovery by restoring mRNA degradation and structure

A number of proteins are induced by cold shock (Goldstein et al., 1990; Jones et al., 1987), including most prominently 4 of the 9 structurally homologous Cold shock proteins (CspA-I) (Wang et al., 1999) that have been implicated in modulating mRNA structure (Jiang et al., 1997; Phadtare et al., 2004). However, there was limited understanding of which factors drive recovery of protein synthesis during the 6 hrs following cold shock. We identified actuators of the recovery circuit by examining gene deletion phenotypes of the 53 proteins whose measured synthesis rates indicate a copy number increase of ≥ 2-fold during the 6 hr recovery period (Extended Data Table 1). Only single gene deletions of *rnr* (RNase R), an exonuclease that degrades damaged rRNA (Basturea et al., 2011; Cheng and Deutscher, 2003) and processes tmRNA (Awano et al., 2010; Cairrão et al., 2003), and *cspA* reduced protein synthesis during recovery

(Figure 4A). Together, Csps and RNase R constitute 40% of total protein synthesis at 3 hours after cold shock (Figure 4B), supporting their dominant role in initial recovery.

We determined the RNA targets of RNase R by sequencing total RNA immediately prior to and 2hr after addition of the transcriptional inhibitor rifampicin at 10°C in WT and Δ*rnr* strains. In a WT strain, mRNA decreases from 5.2% to 2.3% of total RNA during this 2hr window, indicating a half-life of ~2 hr at 10°C (Figure 4C) but a Δ*rnr* strain exhibits a minimal decrease in mRNA level. Moreover, mRNA accumulates to 9.8% of total RNA at 8hr after cold shock in Δ*rnr* cells, whereas WT cells maintain mRNA as 4.2% of total RNA (Figure 4D). Thus, mRNA degradation requires RNase R during cold recovery.

We next examined the role of CspA and its homologues in early recovery. Csps promote read-through of a transcriptional terminator in the metY-rpsO operon through its nucleic acid binding activity (Bae et al., 2000; Phadtare, 2002), and a quadruple deletion Csp strain (*cspA* and its homologues *cspB*, *cspG*, and *cspE*) is unable to grow at low temperature (Xia et al., 2001). However, the role of Csps in facilitating growth at cold temperature has remained elusive. We found that the quadruple Csp mutant (Δ*cspABEG*) did not recover protein synthesis during the 6-hr immediate recovery period (Figure 4A). Because Csps bind and melt nucleic acid structures *in vitro* (Jiang et al., 1997; Phadtare and Severinov, 2005), we tested whether they promote translation recovery via direct, genome-wide modulation of mRNA structure. Indeed, Δ*cspABEG* cells remained trapped in the state observed immediately following cold shock in which all mRNAs were highly-structured and poorly-translated (Figure S4A). The most structured mRNAs in a Δ*cspABEG* strain had the greatest defect in recovery of TE relative to their TE's in the

WT strain (Figure 4E, S4), indicating that Csps drive the alteration of mRNA secondary structure and translation efficiency that accompanies cold recovery.

The Csps are well-expressed at 37°C (Brandi et al., 1999; Li et al., 2014; Taniguchi et al., 2010), and we therefore tested whether they also play a role in maintaining TE at normal growth temperature. A quintuple Δ*cspABCEG* strain (additionally deleted for *cspC*, the homologue that is well-expressed at 37°C), has a 10% growth defect at 37°C indicating that Csps are required for optimal growth. TE measurements in the Δ*cspABCEG* strain indicate that the TEs of the best-translated ORFs in WT (≥ top 10%), which requires less structure, exhibited an ~10% decrease in TE without Csps, whereas the remainder are only marginally influenced (Figure 4F). Thus, Csp expression is crucial for achieving high TEs at 37˚C, just as it is at 10˚C.

Cold recovery is regulated by Csp autoregulation of their own mRNA structures

Csp expression increases dramatically upon cold shock, and then declines during cold recovery. Cold induction is known to involve *csp* message stabilization, with *cspA* mRNA shifting from a rapidly degraded state at 37°C ($T_{½}$= 10-20”) (Fang et al., 1997) to a stable state at 10°C (Giuliodori et al., 2010; Hankins et al., 2007; Yamanaka et al., 1999). *CspA* message stability is regulated through its long 5’UTR, a thermosensor that was shown to undergo a change in structural conformation when shifted from 37˚C to 10˚C (Giuliodori et al., 2010). A conserved element at the 5’ end of the *cspA* UTR, the "cold box," is especially critical to regulation of message stability (Xia et al., 2002). At 37°C, the cold box forms a helix at the 5' end of the message, whereas at 10°C it pairs with a downstream region within the UTR, an interaction that presumably stabilizes the message (Giuliodori et al., 2010). Using a standard minimal free energy structural prediction constrained by DMS-seq data (Hofacker, 2003) to model the structure changes upon cold shock, we validated that cold box interactions are altered on *cspA* upon cold shock *in vivo* (Figure S5).

During cold recovery, *csp* message is destabilized in a process dependent on Csp protein activity (Bae et al., 1997). The mechanistic basis for this destabilization was not known. We found that the long 5' UTRs of Csps were among the most dramatically changing mRNA structures during recovery (Figure S6A-B), suggesting that changes in the UTR structure might be responsible for the *csp* message destabilization. Indeed, during recovery, the 5' UTR shifts to a structure in which the cold box is in a helix with the 5' end of the message, similar to the structure observed for the 37˚C state, as illustrated for *cspB* (Figure 5A-B). The ability of the Csp transcript structure to shift as a function of time at 10°C indicates that the Csp UTR structure senses the state of the cell in addition to sensing temperature. Importantly, these structural transitions do not occur in a Δ*cspABEG* strain, which lacks the Csp ORFs but retains their 5'UTRs (Figure 5C-D), but the CspB 5'UTR does change structure in a Δ*cspBG* strain, where the CspB ORF is deleted and recovery is driven by CspA expression (Figure S6C). These results indicate that the structural change of the 5'UTR during recovery is not dependent on the sequence of the ORF but requires Csp protein expression at cold temperature. Since the Csps are known to interact with their 5' UTRs (Jiang et al., 1997), we propose that Csps remodel their own 5' UTRs, thereby tying their own regulation to their role of structure surveillance in the cell.

Discussion

By determining the relationship between mRNA structure and translation efficiency at a genome scale, we discovered that operons are comprised of ORF-centric mRNA structures that contribute to translation efficiency both under steady state conditions and following perturbation. We consider the implications of these findings for operon function (Figure 6A) and then discuss the self-regulating structure surveillance system that maintains appropriate mRNA structure (Figure 6B).

Operons are the fundamental unit of bacterial gene expression. They enable common transcriptional control of genes with related functions while achieving appropriate protein expression by regulating translational efficiency. We show here that bacteria regulate TE with ORF–centric structures that both drive and insulate the TE of each protein. A blueprint for ORF-centric mRNA structures is encoded in the mRNA sequence itself, including the propensity for in-ORF basepairing, but is likely reinforced by the activity of ribosomes and Csps.

The necessity for achieving discrete TEs for close-packed ORFs may have driven the evolution of this strategy. The translation termination codon of most ORFs is separated by less than 25nt of untranslated mRNA from the start site of the downstream ORF, yet the TE's of these adjacent ORFs can vary as much as 100-fold. If an mRNA structure were to span the boundary between a highly translated and a poorly translated ORF, the abundant ribosomes of the highly translated ORF would have potential to transiently open the structure of the poorly-translated ORF and increase the accessibility of its start site. ORF-centric mRNA structures with predominantly intra-ORF base pairing may prevent the upstream ORF from influencing the downstream ORF's structure and translation efficiency, effectively insulating each ORF from its neighbors. RNA polymerase pausing is enriched at translation start sites (Larson et al., 2014) and may reinforce ORF-centric structural insulation by allowing ORFs to fold independently. For ~15% of operonic ORFs, this insulation is broken as the stop codon of the upstream ORF overlaps the downstream ORF. These ORFs have been hypothesized to be "translationally coupled" through diffusion of the upstream ribosome to the downstream start site (Aksoy et al., 1984; Oppenheim and Yanofsky, 1980; Schümperli et al., 1982; Yates and Nomura, 1981). As the TE's of such ORF pairs vary as much as other ORF pairs, overlap does not cause all ribosomes to reinitiate on the downstream ORF, but may enable upstream ribosomes to influence downstream ORF translation by unwinding mRNA structure.

We find Shine-Dalgarno (SD) strength to be unpredictive of translation efficiency, even after removing the contribution of mRNA structure to TE. This observation is in contrast with the common belief that a stronger SD site indicates stronger translation. Although the presence of SD sites is critical for translation initiation in *E. coli* (Steitz and Jakes, 1975), the role of their quantitative strength for endogenous transcripts has not been defined prior to this work. Large-scale studies using synthetic libraries noted the difficulty in predicting TE from SD strength, which can be mitigated by actively reducing RNA structures (Mutalik et al., 2013). Our results suggest that cells face the same challenge and rely on RNA structure rather than SD sites to tune the level of translation. This conclusion favors the 'standby model' of translation initiation in which the 30S subunit quickly binds to regions near the initiation site and waits for the opening of the SD and start codon (Adhin and van Duin, 1990; de Smit and van Duin, 2003). In this scenario, the major role of the SD is to capture ribosomes diffusing from standby sites and to ensure that the correct start codon is selected rather than to set translation efficiency.

The length-scale of the relationship between mRNA structure and TE is also in line with the standby model of translation initiation. High translation efficiency may require an open structure over long distances to capture a large pool of non-specifically bound ribosomes, whereas the stable structures of poorly translated ORFs may form inhibitory structures that prevent this binding over a large region thereby inhibiting translation. Poorly-translated ORF structures may additionally be necessary to protect the ORFs from premature endonucleolytic cleavage on the frequent occasions when they are bare of ribosomes.

When the cell is subjected to cold shock, mRNA structure increases with a concomitant decrease in translation initiation. Cold recovery consists of a highly correlated ORF-specific decrease in mRNA structure and recovery of translation. Only the Csps and RNaseR are necessary for this recovery. Notably, other proteins important for long term growth at 10˚C (e.g. DeaD [alias CsdA] (Jones et al., 1996), RbfA (Jones and Inouye, 1996) and PNPase (Luttinger et al., 1996)), do not affect initial recovery of protein synthesis. Thus, the cell has an initial emergency system to restore mRNA structure and degradation, comprised of only two proteins, and a long-term program to sustain growth in the cold.

Our data, together with existing data, support a model in which Csps perform mRNA structure surveillance (Figure 6B). The Csps are RNA binding proteins that also bind their own 5'UTRs (Jiang et al., 1997). Their peak abundance is estimated at ~ 2 *

10^6/cell, (Xia et al., 2001), which is consistent with Csp-mRNA interaction over the entire length of open reading frames (~10^7 nt of total mRNA / cell). We suggest that at early times, after cold shock, Csps are predominantly engaged in interacting with cellular mRNA, and do not perturb the long range pairing of the cold box element in the Csp 5'UTR triggered by cold temperature. As recovery proceeds and Csp concentration increases, Csps bind their 5'UTRs, triggering the switch in pairing of the cold box element to the 5' helix and promoting message degradation. In this circuit, the cell sets Csp expression by monitoring the free level of Csps, determined by the extent to which Csps are required to globally remodel mRNA structure. This circuit explains why RNase R deletion, which increases the amount of mRNA to be remodeled, delays recovery as more Csps must be produced to attain the appropriate Csp/mRNA level required for resumption of the 10˚C translational program. This regulatory system closely resembles that of the bacteriophage T4 Gene32 protein (gp-32) autoregulatory circuit. Gp-32 is a single-strand DNA binding protein, and its production is translationally controlled to maintain a constant amount of free gp-32 in the face of changing amounts of ss-DNA (von Hippel et al., 1982; McPheeters et al., 1988; Shamoo et al., 1993).

The Csp RNA surveillance system is likely utilized in a wide variety of conditions and in most bacteria. We show here that Csps are important for growth and proper TE at 37˚C. Other perturbations, including stationary phase and sublethal antibiotic exposure modulate Csp expression (Brandi et al., 1999; VanBogelen and Neidhardt, 1990), suggesting that many environmental changes drive changes in mRNA structure and hence Csp expression. Csps span the gram-negative/positive divide, and Csps in *B. subtilis* exhibit strikingly similar properties to those in *E. coli*—high abundance during normal growth (Eymann et al., 2004) and induction during cold shock (Willimsky et al., 1992). Thus, it is likely that the Csp RNA surveillance system arose early in evolution and was then maintained throughout the bacterial world. Csp orthologues have been identified in all domains (Graumann et al., 1997; Karlson et al., 2002), and ectopic expression of bacterial Csps in maize enhances growth at cold and in water-limited conditions (Castiglioni et al., 2008), indicating that the mechanism through which they modulate protein synthesis is likely broadly relevant.

The relationship between structure and translation efficiency that we identify is a constraint on mRNA sequence beyond codon adaptation (Sharp and Li, 1987). Further work will identify the relative contributions of these considerations to codon choice, but there are immediate implications for synthetic construct design, as optimal codon selection must reflect both message abundance and translation efficiency. This presents both a challenge and an opportunity to efforts to synthesize synthetic operons: synthetic designs must be carried out with cognizance of the entire open reading frame sequence. However, design approaches that incorporate appropriate mRNA structures should have the potential to produce finely- tuned synthesis rates as are observed on natural operons.

Experimental Procedures

E. coli K-12 MG1655 was used as the wild-type strain. All culture experiments were performed in MOPS complete medium with all amino acids except methionine. All samples were captured at OD_{420} = 0.4. Multiple deletion strains were generated by transduction of FRT-flanked deletion alleles from the Keio collection followed by marker excision by Flp recombinase. Ribosome profiling, mRNAseq, and DMS-seq were performed as previously described (Ingolia et al., 2009; Li et al., 2014; Rouskin et al., 2014). Gini indices were calculated on all A and C residues for the designated windows. For all Gini index calculations, analysis was limited to genes with greater than an average of 15 reads per nucleotide across the gene body. Mean ribosome and mRNA densities, used to compute translation efficiencies, were calculated as described (Li et al., 2014). Adjacent ORFs within operons were identified by analyzing mRNA-seq data: pairs of adjacent ORFs with constant mRNA-seq signal were considered to be in the same operon. Secondary structure models were constructed using a minimum free energy prediction generated by the RNAfold function of ViennaRNA (Hofacker, 2003). These RNA structure predictions were constrained with DMS-seq measurements where indicated. Ribosome run-off was determined as described (Ingolia et al., 2011).

Author Contributions

D.H.B. S.R., G.W.L., J.S.W., and C.A.G. designed the experiments. D.H.B. and S.R. performed the experiments, D.H.B. S.R. and G.W.L. analyzed the data, D.H.B. S.R., G.W.L., J.S.W., and C.A.G. drafted and revised the manuscript.

Acknowledgements

This research was supported by the Center for RNA Systems Biology (J.S.W.), the Howard Hughes Medical Institute (J.S.W.), the Helen Hay Whitney Foundation (G.W.L.) and the National Institutes of Health (C.A.G. and D.H.B., T32GM8284 and T32 EB009383, G.W.L., K99GM105913).

References

Adhin, M.R., and van Duin, J. (1990). Scanning model for translational reinitiation in eubacteria. Journal of Molecular Biology *213*, 811–818.

Aksoy, S., Squires, C.L., and Squires, C. (1984). Translational coupling of the trpB and trpA genes in the Escherichia coli tryptophan operon. Journal of Bacteriology *157*, 363–367.

Awano, N., Rajagopal, V., Arbing, M., Patel, S., Hunt, J., Inouye, M., and Phadtare, S. (2010). Escherichia coli RNase R Has Dual Activities, Helicase and RNase. Journal of Bacteriology *192*, 1344–1352.

Bae, W., Jones, P.G., and Inouye, M. (1997). CspA, the major cold shock protein of Escherichia coli, negatively regulates its own gene expression. Journal of Bacteriology *179*, 7081–7088.

Bae, W., Xia, B., Inouye, M., and Severinov, K. (2000). Escherichia coli CspA-family RNA chaperones are transcription antiterminators. Proceedings of the National Academy of Sciences *97*, 7784–7789.

Basturea, G.N., Zundel, M.A., and Deutscher, M.P. (2011). Degradation of ribosomal RNA during starvation: Comparison to quality control during steady-state growth and a role for RNase PH. Rna *17*, 338–345.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Blüthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. Molecular Systems Biology *9*, 1–10.

Borujeni, A.E., Channarasappa, A.S., and Salis, H.M. (2013). Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Research, *42*, 2646-2659.

Brandi, A., Spurio, R., Gualerzi, C.O., and Pon, C.L. (1999). Massive presence of the Escherichia coli “major cold-shock protein” CspA under non-stress conditions. The EMBO Journal *18*, 1653–1659.

Cairrão, F., Cruz, A., Mori, H., and Arraiano, C.M. (2003). Cold shock induction of RNase R and its role in the maturation of the quality control mediator SsrA/tmRNA. Mol Microbiol *50*, 1349–1360.

Castiglioni, P., Warner, D., Bensen, R.J., Anstrom, D.C., Harrison, J., Stoecker, M., Abad, M., Kumar, G., Salvador, S., D'Ordine, R., et al. (2008). Bacterial RNA Chaperones Confer Abiotic Stress Tolerance in Plants and Improved Grain Yield in Maize under Water-Limited Conditions. Plant Physiology *147*, 446–455.

Cheng, Z.-F., and Deutscher, M.P. (2003). Quality control of ribosomal RNA mediated by polynucleotide phosphorylase and RNase R. Proceedings of the National Academy of Sciences *100*, 6388–6393.

de Smit, M.H., and van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proceedings of the National Academy of Sciences *87*, 7668–7672.

de Smit, M.H., and van Duin, J. (2003). Translational Standby Sites: How Ribosomes May Deal with the Rapid Folding Kinetics of mRNA. Journal of Molecular Biology *331*, 737–743.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696-700.

Eymann, C., Dreisbach, A., Albrecht, D., Bernhardt, J., Becher, D., Gentner, S., Tam, L.T., Büttner, K., Buurman, G., Scharf, C., et al. (2004). A comprehensive proteome map of growingBacillus subtilis cells. Proteomics *4*, 2849–2876.

Eyre-Walker, A., and Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Research *21*, 4599–4603.

Fang, L., Jiang, W., Bae, W., and Inouye, M. (1997). Promoter-independent cold-shock induction of cspA and its derepression at 37 degrees C by mRNA stabilization. Mol Microbiol *23*, 355–364.

Friedman, H., Lu, P., and Rich, A. (1969). An In Vivo Block in the Initiation of Protein Synthesis. Cold Spring Harbor Symposia on Quantitative Biology *34*, 255–260.

Friedman, H., Lu, P., and Rich, A. (1971). Temperature control of initiation of protein synthesis in Escherichia coli. Journal of Molecular Biology *61*, 105–121.

Giuliodori, A.M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C.O., and Pon, C.L. (2010). The cspA mRNA Is a Thermosensor that Modulates Translation of the Cold-Shock Protein CspA. Molecular Cell *37*, 21–33.

Goldstein, J., Pollitt, N.S., and Inouye, M. (1990). Major cold shock protein of Escherichia coli. Proceedings of the National Academy of Sciences *87*, 283–287.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. Science *342*, 475–479.

Graumann, P., Wendrich, T.M., Weber, M.H., Schröder, K., and Marahiel, M.A. (1997). A family of cold shock proteins in Bacillus subtilis is essential for cellular growth and for efficient protein synthesis at optimal and low temperatures. Mol Microbiol *25*, 741–756.

Hall, M.N., Gabay, J., Débarbouillé, M., and Schwartz, M. (1982). A role for mRNA secondary structure in the control of translation initiation. Nature *295*, 616–618.

Hankins, J.S., Zappavigna, C., Prud'homme-Genereux, A., and Mackie, G.A. (2007). Role of RNA Structure and Susceptibility to RNase E in Regulation of a Cold Shock mRNA, cspA mRNA. Journal of Bacteriology *189*, 4353–4358.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Research *31*, 3429–3431.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell *147*, 789–802.

Inoue, T., and Cech, T.R. (1985). Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. Proceedings of the National Academy of Sciences *82*, 648–652.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology *3*, 318–356.

Jiang, W., Hou, Y., and Inouye, M. (1997). CspA, the major cold-shock protein of Escherichia coli, is an RNA chaperone. J. Biol. Chem. *272*, 196–202.

Jones, P.G., and Inouye, M. (1996). RbfA, a 30S ribosomal binding factor, is a cold-shock protein whose absence triggers the cold-shock response. Mol Microbiol *21*, 1207–1218.

Jones, P.G., Mitta, M., Kim, Y., Jiang, W., and Inouye, M. (1996). Cold shock induces a major ribosomal-associated protein that unwinds double-stranded RNA in Escherichia coli. Proceedings of the National Academy of Sciences *93*, 76–80.

Jones, P.G., VanBogelen, R.A., and Neidhardt, F.C. (1987). Induction of proteins in response to low temperature in Escherichia coli. Journal of Bacteriology *169*, 2092–2095.

Karlson, D., Nakaminami, K., Toyomasu, T., and Imai, R. (2002). A Cold-regulated Nucleic Acid-binding Protein of Winter Wheat Shares a Domain with Bacterial Cold Shock Proteins. Journal of Biological Chemistry *277*, 35248–35256.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proceedings of the National Academy of Sciences *110*, 14024–14029.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science *324*, 255–258.

Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., and Weissman, J.S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science *344*, 1042–1047.

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell *157*, 624–635.

Lodish, H.F. (1970). Secondary structure of bacteriophage f2 ribonucleic acid and the initiation of in vitro protein biosynthesis. Journal of Molecular Biology *50*, 689–702.

Luttinger, A., Hahn, J., and Dubnau, D. (1996). Polynucleotide phosphorylase is necessary for competence development in Bacillus subtilis. Mol Microbiol *19*, 343–356.

Marzi, S., Myasnikov, A.G., Serganov, A., Ehresmann, C., Romby, P., Yusupov, M., and Klaholz, B.P. (2007). Structured mRNAs Regulate Translation Initiation by Binding to the Platform of the Ribosome. Cell *130*, 1019–1031.

McPheeters, D.S., Stormo, G.D., and Gold, L. (1988). Autogenous regulatory site on the bacteriophage T4 gene 32 messenger RNA. Journal of Molecular Biology *201*, 517–535.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.-A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P., et al. (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. Nature Methods *10*, 354–360.

Ng, H., Ingraham, J.L., and Marr, A.G. (1962). Damage and derepression in Escherichia coli resulting from growth at low temperatures. Journal of Bacteriology *84*, 331–339.

Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R.J., Typas, A., Gross, C.A., Kramer, G., et al. (2011). Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo. Cell *147*, 1295–1308.

Oppenheim, D.S., and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of Escherichia coli. Genetics *95*, 785–795.

Phadtare, S. (2002). Three Amino Acids in Escherichia coli CspE Surface-exposed Aromatic Patch Are Critical for Nucleic Acid Melting Activity Leading to Transcription Antitermination and Cold Acclimation of Cells. Journal of Biological Chemistry *277*, 46706–46711.

Phadtare, S., and Severinov, K. (2005). Nucleic acid melting by Escherichia coli CspE. Nucleic Acids Research *33*, 5583–5590.

Phadtare, S., Inouye, M., and Severinov, K. (2004). The Mechanism of Nucleic Acid Melting by a CspA Family Protein. Journal of Molecular Biology *337*, 147–155.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505,* 701-705.

Russell, J.B., and Cook, G.M. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. Microbiol. Rev. *59*, 48–62.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nature Biotechnology *27,* 946-950.

Scharff, L.B., Childs, L., Walther, D., and Bock, R. (2011). Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites. PLoS Genetics *7*, e1002155.

Schümperli, D., McKenney, K., Sobieski, D.A., and Rosenberg, M. (1982). Translational coupling at an intercistronic boundary of the Escherichia coli galactose operon. Cell *30*, 865–871.

Shamoo, Y., Tam, A., Konigsberg, W.H., and Williams, K.R. (1993). Translational repression by the bacteriophage T4 gene 32 protein involves specific recognition of an RNA pseudoknot structure. Journal of Molecular Biology *232*, 89–104.

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

Steitz, J.A., and Jakes, K. (1975). How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. Proceedings of the National Academy of Sciences *72*, 4734–4738.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science *329*, 533–538.

VanBogelen, R.A., and Neidhardt, F.C. (1990). Ribosomes as sensors of heat and cold shock in Escherichia coli. Proceedings of the National Academy of Sciences *87*, 5589–5593.

von Hippel, P.H., Kowalczykowski, S.C., Lonberg, N., Newport, J.W., Paul, L.S., Stormo, G.D., and Gold, L. (1982). Autoregulation of gene expression. Quantitative evaluation of the expression and function of the bacteriophage T4 gene 32 (single-stranded DNA binding) protein system. Journal of Molecular Biology *162*, 795–818.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706–709.

Wang, N., Yamanaka, K., and Inouye, M. (1999). CspI, the ninth member of the CspA family of Escherichia coli, is induced upon cold shock. Journal of Bacteriology *181*, 1603–1609.

Wikström, P.M., Lind, L.K., Berg, D.E., and Björk, G.R. (1992). Importance of mRNA folding and start codon accessibility in the expression of genes in a ribosomal protein operon of Escherichia coli. Journal of Molecular Biology *224*, 949–966.

Willimsky, G., Bang, H., Fischer, G., and Marahiel, M.A. (1992). Characterization of cspB, a Bacillus subtilis inducible cold shock gene affecting cell viability at low temperatures. Journal of Bacteriology *174*, 6326–6335.

Xia, B., Ke, H., and Inouye, M. (2001). Acquirement of cold sensitivity by quadruple deletion of the cspA family and its suppression by PNPase S1 domain in Escherichia coli. Mol Microbiol *40*, 179–188.

Xia, B., Ke, H., Jiang, W., and Inouye, M. (2002). The Cold Box Stem-loop Proximal to the 5'-End of the Escherichia coli cspA Gene Stabilizes Its mRNA at Low Temperature. Journal of Biological Chemistry *277*, 6005–6011.

Yamanaka, K., Mitta, M., and Inouye, M. (1999). Mutation analysis of the 5' untranslated region of the cold shock cspA mRNA of Escherichia coli. Journal of Bacteriology *181*, 6284–6291.

Yates, J.L., and Nomura, M. (1981). Feedback regulation of ribosomal protein synthesis in E. coli: localization of the mRNA target sites for repressor action of ribosomal protein L1. Cell *24*, 243–249.

Zhang, W., Dunkle, J.A., and Cate, J.H.D. (2009). Structures of the Ribosome in Intermediate States of Ratcheting. Science *325*, 1014–1017.

Figures

Figure 1

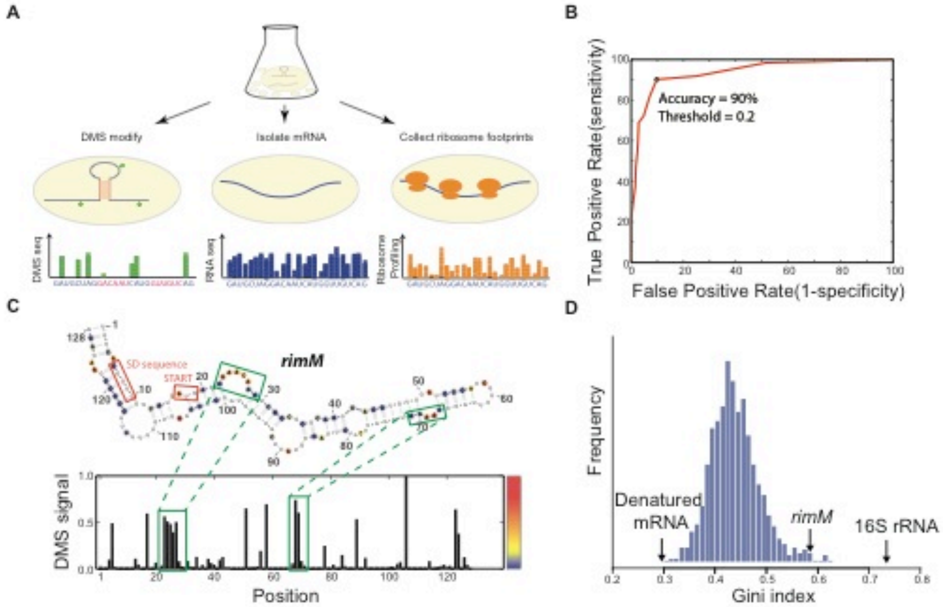


Figure 2

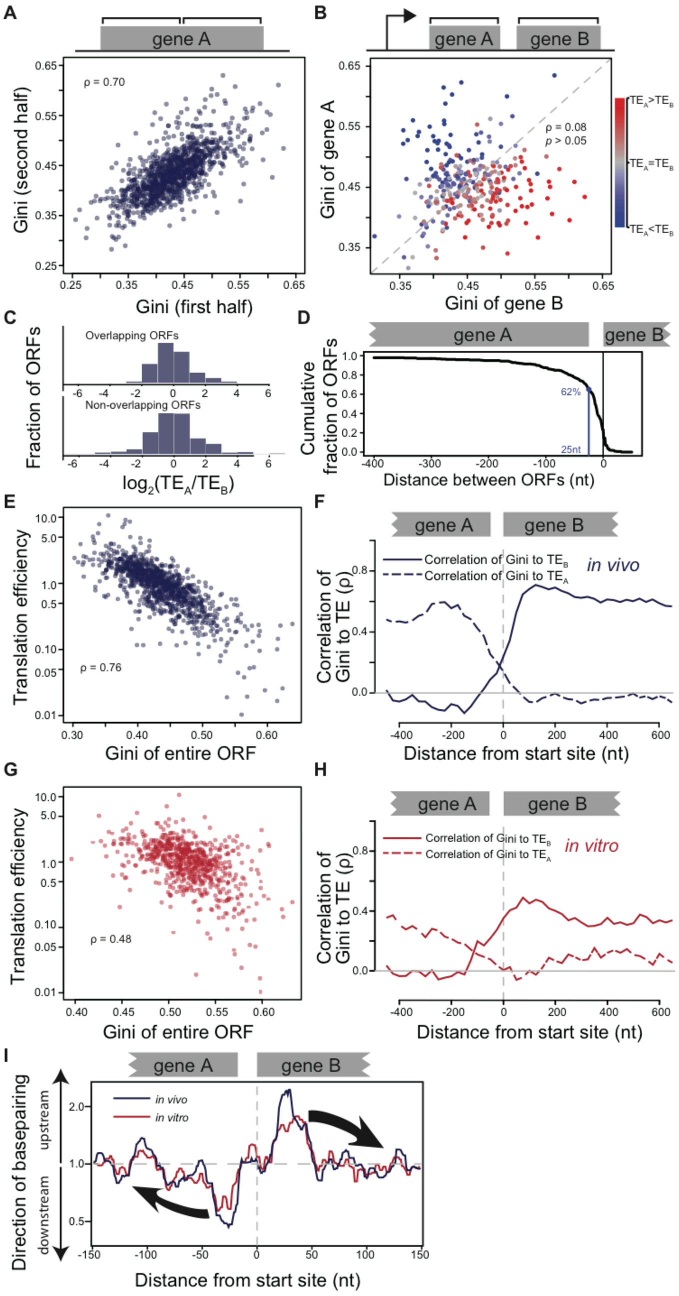

A
gene A
ρ = 0.70
Gini (second half)
Gini (first half)
0.25
0.35
0.45
0.55
0.65
B
gene A
gene B
Gini of gene A
Gini of gene B
ρ = 0.08
$p > 0.05$
$TE_A>TE_B$
$TE_A=TE_B$
$TE_A<TE_B$
C
Fraction of ORFs
Overlapping ORFs
Non-overlapping ORFs
-6
-4
-2
0
2
4
6
$\log_2(TE_A/TE_B)$
D
gene A
gene B
Cumulative fraction of ORFs
62%
25nt
1.0
0.8
0.6
0.4
0.2
0.0
-400
-300
-200
-100
0
Distance between ORFs (nt)
E
Translation efficiency
10.0
5.0
1.0
0.5
0.10
0.05
0.01
ρ = 0.76
0.30
0.40
0.50
0.60
Gini of entire ORF
F
gene A
gene B
Correlation of Gini to TE_B
Correlation of Gini to TE_A
in vivo
Correlation of Gini to TE (ρ)
0.8
0.4
0.0
-400
-200
0
200
400
600
Distance from start site (nt)
G
Translation efficiency
10.0
5.0
1.0
0.5
0.10
0.05
0.01
ρ = 0.48
0.40
0.45
0.50
0.55
0.60
Gini of entire ORF
H
gene A
gene B
Correlation of Gini to TE_B
Correlation of Gini to TE_A
in vitro
Correlation of Gini to TE (ρ)
0.8
0.4
0.0
-400
-200
0
200
400
600
Distance from start site (nt)
I
gene A
gene B
in vivo
in vitro
Direction of basepairing
upstream
downstream
2.0
1.0
0.5
-150
-100
-50
0
50
100
150
Distance from start site (nt)

Figure 3

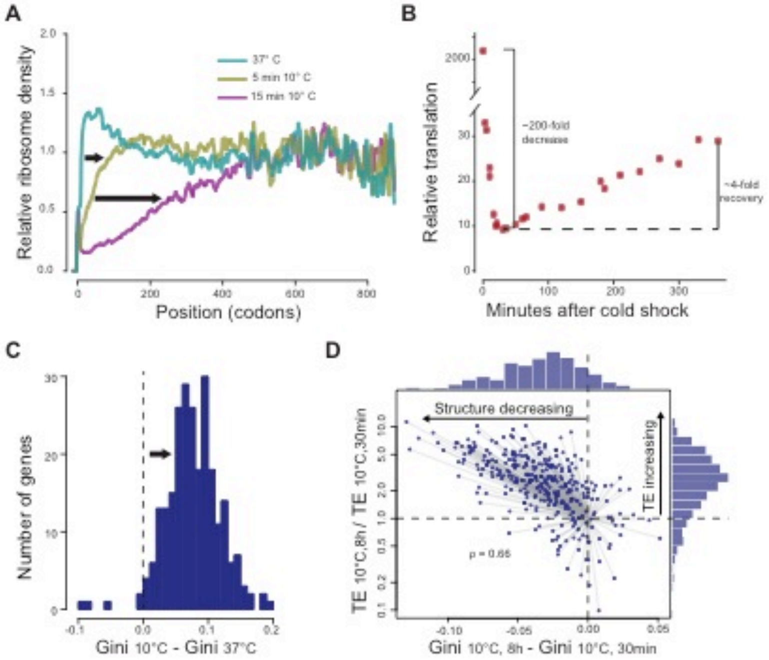

A
Relative ribosome density
2.0
1.5
1.0
0.5
0.0
37° C
5 min 10° C
15 min 10° C
0
200
400
600
800
Position (codons)
B
Relative translation
2000
30
20
10
0
~200-fold decrease
~4-fold recovery
0
100
200
300
Minutes after cold shock
C
Number of genes
30
20
10
0
-0.1
0.0
0.1
0.2
Gini 10°C - Gini 37°C
D
Structure decreasing
TE increasing
TE 10°C,8h / TE 10°C,30min
10.0
5.0
2.0
1.0
0.5
0.2
0.1
ρ = 0.66
-0.10
-0.05
0.00
0.05
Gini 10°C, 8h - Gini 10°C, 30min

Figure 4

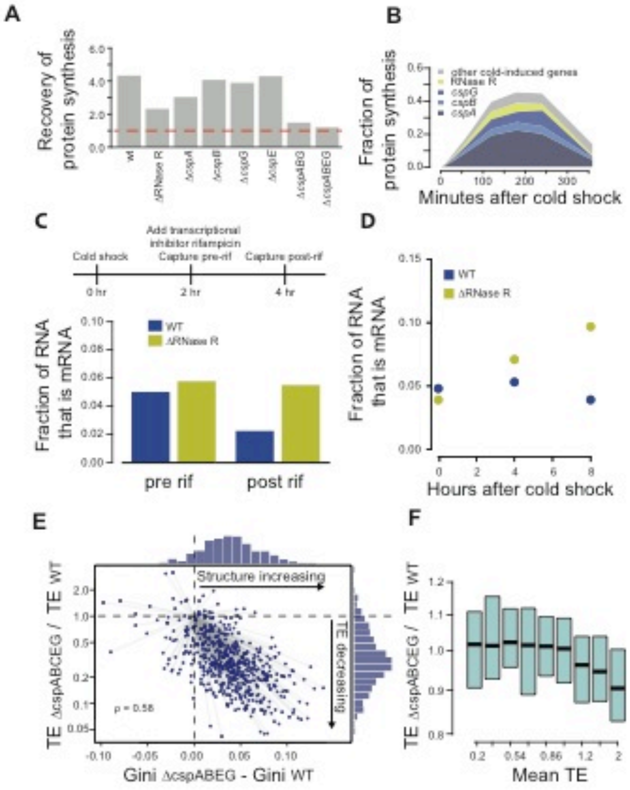

A
Recovery of protein synthesis
6.0
4.0
2.0
0.0
wt
ΔRNase R
ΔcspA
ΔcspB
ΔcspG
ΔcspE
ΔcspABG
ΔcspABEG
B
Fraction of protein synthesis
0.6
0.4
0.2
0.0
other cold-induced genes
RNase R
cspG
cspB
cspA
0
100
200
300
Minutes after cold shock
C
Cold shock
Add transcriptional inhibitor rifampicin
Capture pre-rif
Capture post-rif
0 hr
2 hr
4 hr
Fraction of RNA that is mRNA
0.10
0.08
0.06
0.04
0.02
0.00
WT
ΔRNase R
pre rif
post rif
D
Fraction of RNA that is mRNA
0.15
0.10
0.05
0.00
WT
ΔRNase R
0
2
4
6
8
Hours after cold shock
E
TE ΔcspABCEG / TE WT
2.0
1.0
0.5
0.2
0.1
0.05
Structure increasing
TE decreasing
ρ = 0.58
-0.10
-0.05
0.00
0.05
0.10
Gini ΔcspABEG - Gini WT
F
TE ΔcspABCEG / TE WT
1.2
1.1
1.0
0.9
0.8
0.2
0.54
0.86
1.2
2
Mean TE

Figure 5

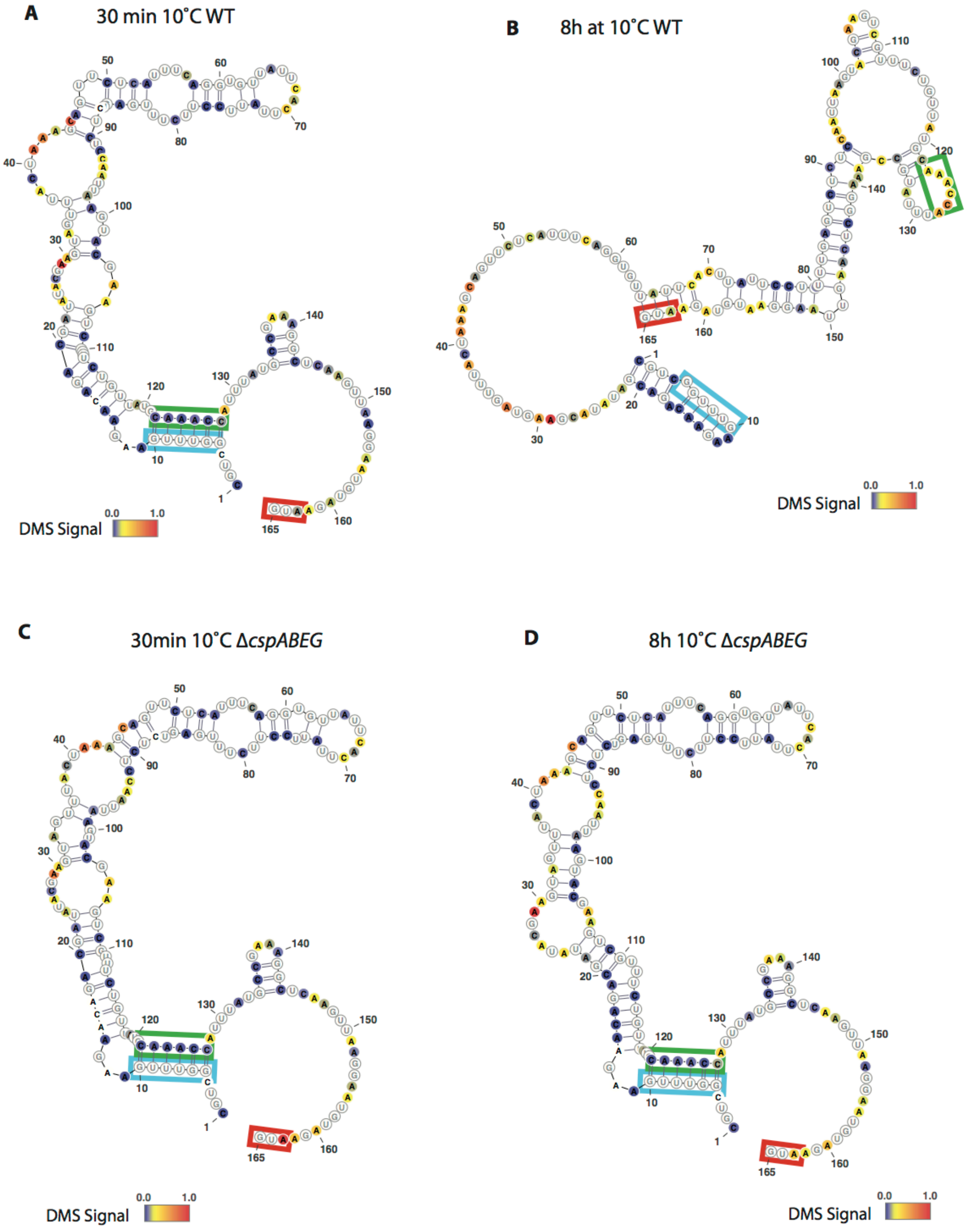

A
30 min 10˚C WT
B
8h at 10˚C WT
C
30min 10˚C Δ*cspABEG*
D
8h 10˚C Δ*cspABEG*
DMS Signal
0.0
1.0
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
165

Figure 6

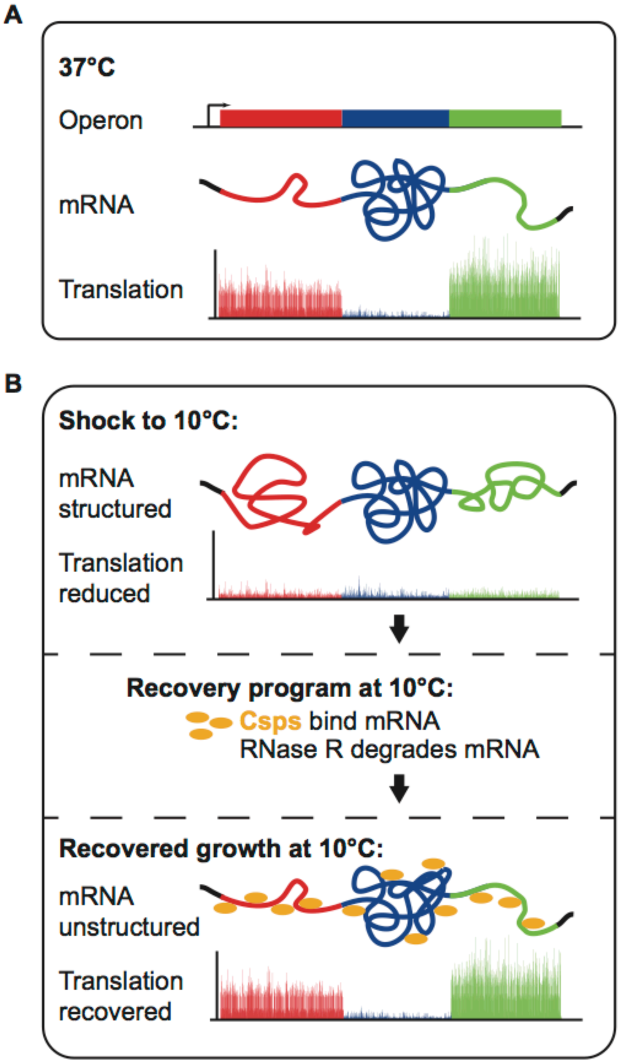

A
37°C
Operon
mRNA
Translation
B
Shock to 10°C:
mRNA
structured
Translation
reduced
Recovery program at 10°C:
Csps bind mRNA
RNase R degrades mRNA
Recovered growth at 10°C:
mRNA
unstructured
Translation
recovered

Figure Legends

Figure 1: DMS-seq effectively probes RNA structures in E. coli

(A) Schematic for obtaining mRNA structures and translation efficiency using DMS-seq, mRNA-seq, and ribosome profiling.

(B) ROC curve on the DMS signal for A and C bases in the 16S rRNA from *in vivo* ribosomes using the *E.coli* ribosome crystal structure (Zhang et al., 2009) as a model. True positives are defined as bases that are both unpaired and solvent-accessible, and true negatives are bases that are paired. The total number of evaluated bases is 438 As or Cs. Signal threshold of 0.2 has 90% agreement with the crystal structure.

(C) Structural prediction for *rimM*. The predicted *rimM* structure is based on a minimum free energy prediction constrained by our DMS-seq measurements, using the same 0.2 threshold used for the 16S rRNA in (B). The DMS signal across *rimM* is shown below the structure. The color bar indicates the intensity of the DMS-seq signal at each position.

(D) Histogram of Gini indices on *E. coli* open reading frames from DMS-seq data obtained *in vivo* at 37°C. Gini index calculated on 16S rRNA and mean of Gini indices calculated on mRNAs heat-denatured at 95°C are indicated.

Figure 2: mRNA structure is organized around open reading frames

(A) Plot of Gini index calculated on the first half of ORF body vs. the Gini index calculated on the second half of ORF body. Spearman's rank order correlation (ρ) of Gini indices is indicated.

(B) Plot of Gini index calculated on adjacent ORFs in operons. ρ indicates the correlation between Gini indices of adjacent ORFs. Coloring indicates ratio of the translation efficiency (TE, ribosome footprint density / mRNA-seq density) of the adjacent ORFs. Correlation of Gini and TE is indicated by clustering of red (lower) and blue (upper) dots.

(C) Histogram of TE ratios for overlapping and non-overlapping open reading frames. Overlapping ORFs are ORF pairs for which the annotated stop codon of the upstream ORF overlaps or is 3' of the start codon of the downstream ORF.

(D) Plot of Gini index of *in vivo* DMS-modified mRNA calculated across the entire ORF body against *in vivo* TE for well-expressed ORFs. TEs are plotted on a log scale.

(E) Correlation (Spearman's ρ) between *in vivo* mRNA structure quantified by Gini index and *in vivo* TE of well-expressed ORFs. Gini index was calculated for 300 nt windows that scan gene bodies, using genes that extend through the 300 nt window being examined. The correlation to TE is plotted at the center of each 300nt window.

(F) Plot of Gini index of *in vitro* DMS-modified mRNA calculated across the entire ORF body against *in vivo* TE for well-expressed ORFs.

(G) Correlation (Spearman's ρ) between *in vitro* mRNA structure quantified by Gini index and *in vivo* TE of well-expressed ORFs, as in (E).

(H) Cumulative distribution of spacing between ORFs within operons.

(I) Plot of directionality of RNA interactions. mRNA structure at each operonic ORF boundary was predicted by calculating either the *in vivo* or the *in vitro* DMS-constrained minimum free energy structure for a region extending from -250 nt to

+250 nt relative to translation start site. At each position, the probability of interaction with each other position was calculated for each ORF examined. The average sum probability of interacting with any nucleotide in a 100 nt window upstream and in a 100 nt downstream was calculated. The ratio of the downstream interaction probability to the upstream interaction probability is plotted at each position.

Figure 3: Cold induces a defect in translation instigated by an increase in mRNA structure

(A) Meta-gene analysis of ribosome run-off after cold shock. Ribosome read density at each position in the gene was averaged across well-expressed genes for samples prepared at the indicated times. Analysis at each position is limited to ORFs that are at least of that length.

(B) Total translation during cold recovery. Total translation was measured by pulse-labeling with ^{35}S-methionine at 37°C and at timepoints following cold shock.

(C) Histogram of change in Gini index following cold shock. mRNA was probed with DMS at 37°C and 25 min after shift to 10°C. Gini index was calculated for all genes that were well-expressed in both conditions. The difference in the Gini index of each gene at 10°C vs. its Gini index at 37°C is plotted.

(D) Plot of the change in Gini index between 30min and 8hr following cold shock against change in translation efficiency during this same time window. Histograms above each axis indicate the distribution of changes in structure and translation efficiency. During recovery, the large majority of genes fall in the upper left quadrant, indicating that their structure is decreasing while their translation efficiency is increasing.

Figure 4: RNase R and Csps facilitate cold recovery

(A) Deleting RNase R and the Csps inhibit cold recovery. Ratio of total translation (^{35}S-methionine pulse labeling) at 8 hrs versus 30 min following cold shock for WT cells, Δ*rnr* and single or multiple *csp* deletion strains.

(B) Fraction of ribosome footprint reads that map to cold-induced genes during cold recovery.

(C) RNA content of cells prior to and following rifampicin treatment at 10°C. Total RNA was purified and sequenced immediately prior to and 2hr after rifampicin treatment of WT and Δ*rnr* cells. The fraction of all sequencing reads that map to mRNA are plotted.

(D) RNA content of cells following cold shock. Total RNA was purified at the indicated timepoints following shock to 10°C in WT and in Δ*rnr* cells. The fraction of all sequencing reads that map to mRNA at each timepoint are plotted.

(E) Comparison of the change in Gini index and translation efficiency of well-expressed mRNAs in a cold recovery-inhibited strain (Δ*cspABEG)* vs. a WT strain at 6 hr following recovery. Histograms above each axis indicate the distribution of changes in structure and translation efficiency. The large majority of genes fall in the lower right quadrant, indicating that mRNA structure is higher and translation efficiency lower in the *csp* deletion strain relative to the WT strain.

(F) Distribution of change in translation efficiency in Δ*cspABCEG* at 37˚C. Genes were binned into 9 groups based on the TE in WT cells, and the distribution of changes in TE in the Δ*cspABCEG* strain was calculated for each bin. For each bin, box center and limits indicate the median change and the 25th and 75th percentile changes.

Figure 5: Csp expression is controlled by an auto-regulatory feedback loop

Change in structure of the *cspB* 5' UTR during cold recovery. The predicted structure of the *cspB* 5' UTR was generated by constraining a minimum free-energy prediction with our DMS-seq measurements in WT (A, B) and Δ*cspABEG* (C, D) strains at 30 min and 8hr after cold shock. The cold box element is highlighted in the blue box and the long range interaction regions is highlighted in a green box. A color bar indicates the intensity of the DMS-seq signal at each position. DMS reactive bases (based on the ribosomal ROC derived threshold) are in yellow to red.

Figure 6: Model of operon structural organization and surveillance

(A) Operon mRNAs are organized into ORF-centric structures that specify translation efficiency of each ORF.

(B) Cold shock induces a genome-wide increase in mRNA structures and reduction in translation efficiency. A recovery system consisting of Csps and RNase R facilitate recovery by unstructuring and degrading structured mRNAs.

Supplementary Figures

Figure S1

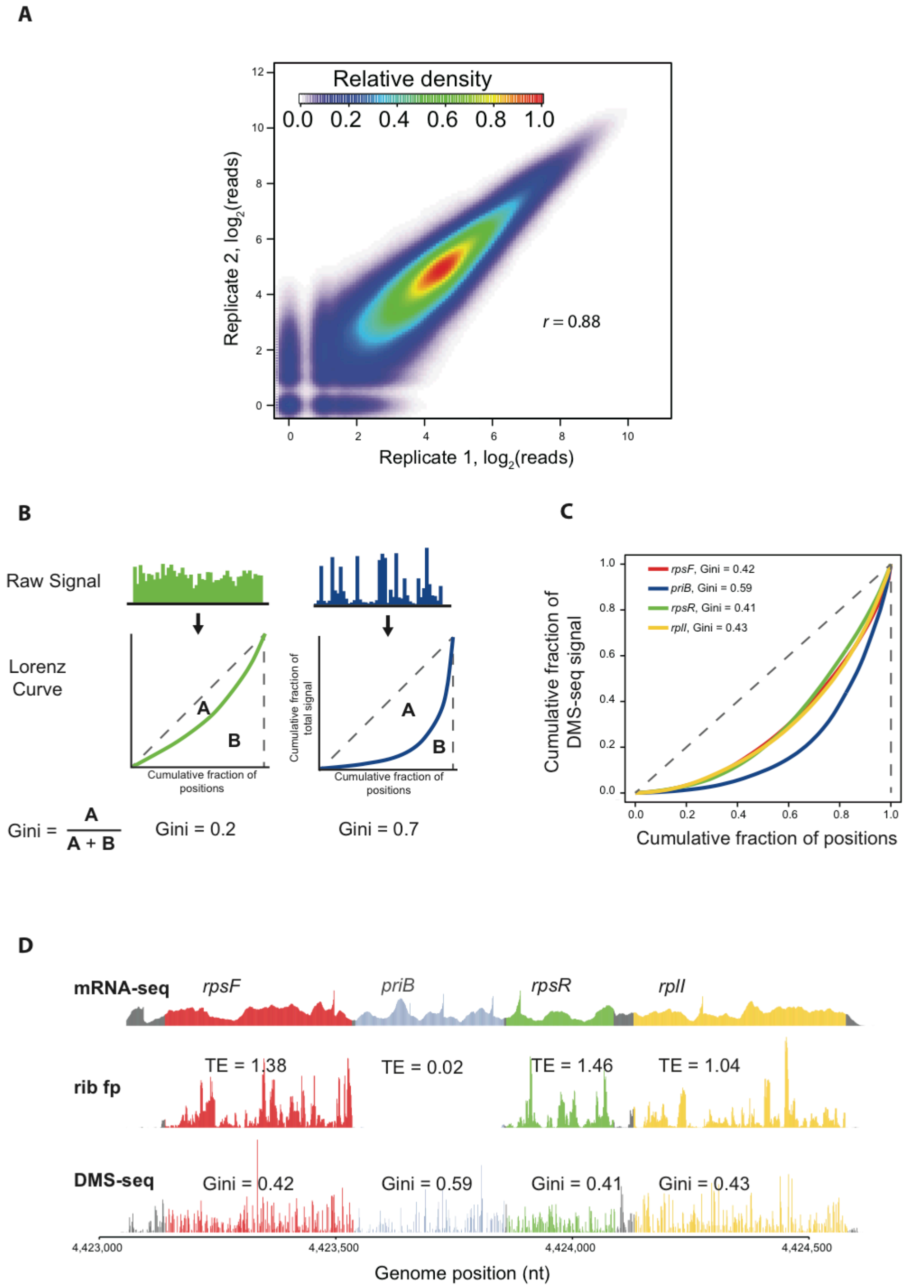


Figure S2

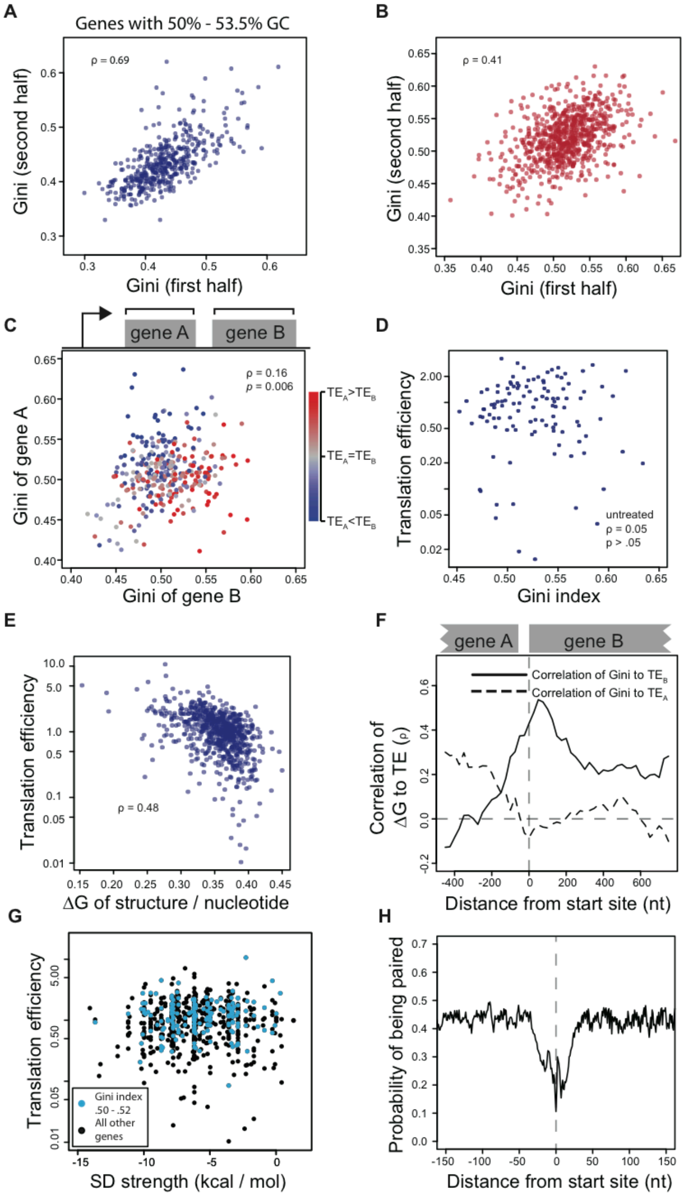

A
Genes with 50% - 53.5% GC
ρ = 0.69
Gini (second half)
0.3
0.4
0.5
0.6
Gini (first half)
B
ρ = 0.41
Gini (second half)
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Gini (first half)
C
gene A
gene B
ρ = 0.16
p = 0.006
$TE_A>TE_B$
$TE_A=TE_B$
$TE_A<TE_B$
Gini of gene A
Gini of gene B
0.40
0.45
0.50
0.55
0.60
0.65
D
Translation efficiency
2.00
0.50
0.20
0.05
0.02
untreated
ρ = 0.05
p > .05
Gini index
E
Translation efficiency
10.0
5.0
1.0
0.5
0.1
0.05
0.01
ρ = 0.48
ΔG of structure / nucleotide
0.15
0.20
0.25
0.30
0.35
0.40
0.45
F
gene A
gene B
Correlation of Gini to TE_B
Correlation of Gini to TE_A
Correlation of ΔG to TE (ρ)
-0.2
0.0
0.2
0.4
0.6
Distance from start site (nt)
-400
-200
0
200
400
600
G
Translation efficiency
5.00
0.50
0.05
0.01
Gini index .50 - .52
All other genes
SD strength (kcal / mol)
-15
-10
-5
0
H
Probability of being paired
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Distance from start site (nt)
-150
-100
-50
0
50
100
150

Figure S3

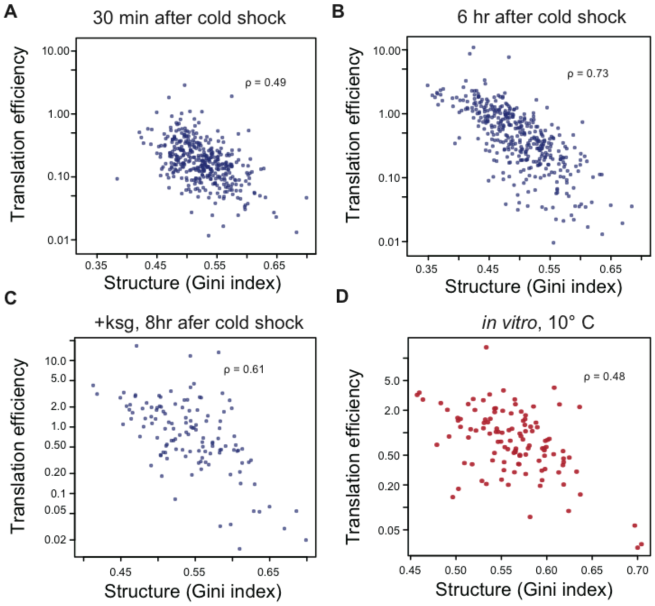


Figure S4

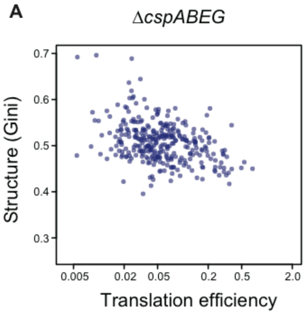


Figure S5

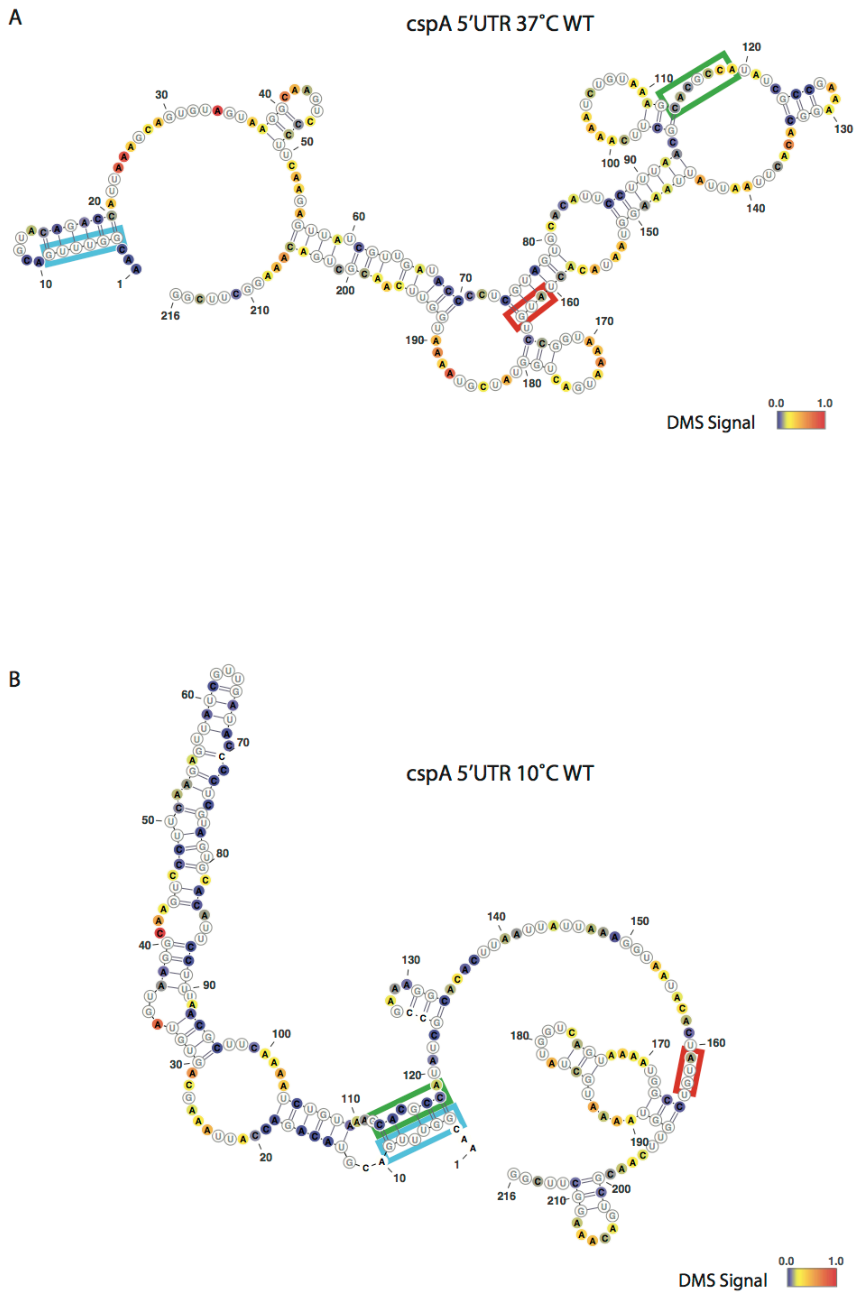
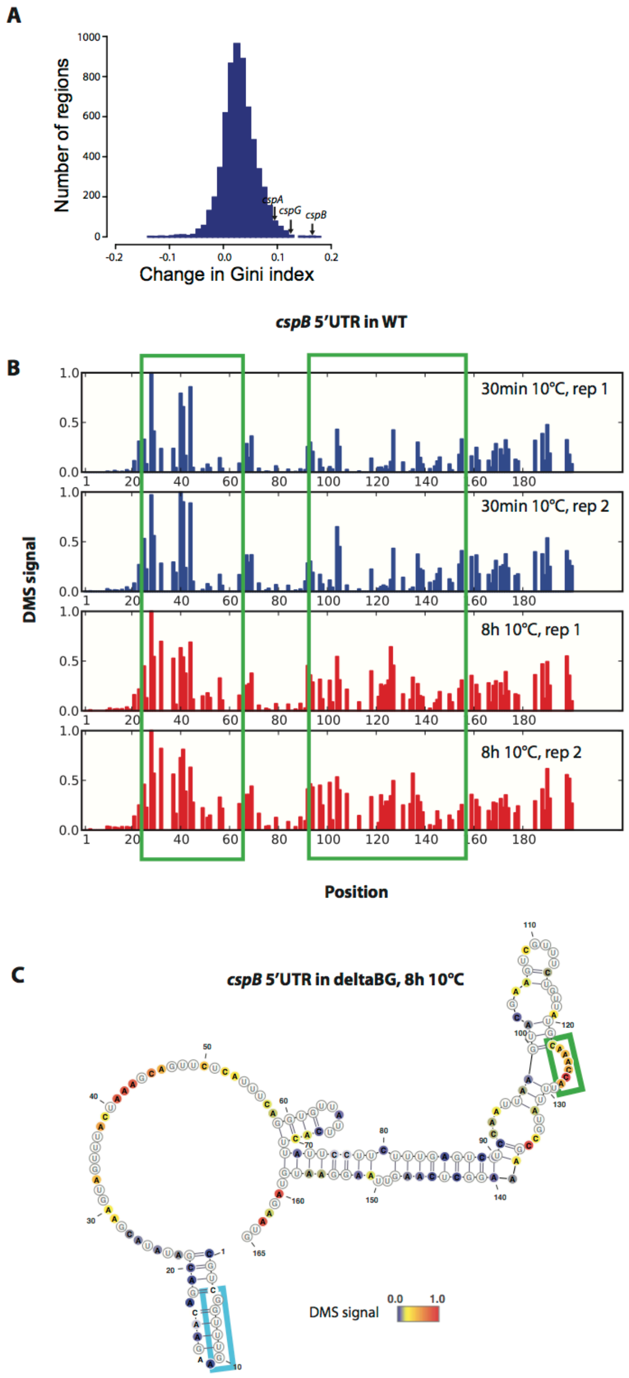

A
cspA 5'UTR 37°C WT
DMS Signal
0.0
1.0
B
cspA 5'UTR 10°C WT
DMS Signal
0.0
1.0

Figure S6


A
Number of regions
1000
800
600
400
200
0
-0.2
-0.1
0.0
0.1
0.2
Change in Gini index
cspA
cspG
cspB
cspB 5'UTR in WT
B
DMS signal
30min 10°C, rep 1
30min 10°C, rep 2
8h 10°C, rep 1
8h 10°C, rep 2
1.0
0.5
0.0
1
20
40
60
80
100
120
140
160
180
Position
C
cspB 5'UTR in deltaBG, 8h 10°C
DMS signal
0.0
1.0
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
165

Supplementary Figures Legends

Figure S1: DMS-seq effectively probes RNA structures in E. coli, related to Figure 1

(A) DMS signal at all positions within well-expressed mRNAs in two biological replicates.

(B) Schematic representation of the Gini index calculation

(C) Lorenz curve of DMS-seq data of each gene in the operon represented in D

(D) mRNA-seq, ribosome profiling, and DMS-seq data for a single operon.

Figure S2: mRNA structure is organized around open reading frames, related to Figure 2

(A) Plot of *in vivo* Gini index calculated on the first half of ORF body against the *in vivo* Gini index calculated on the second half of ORF body restricted to genes with GC content between 50% - 53.5%.

(B) Plot of Gini index calculated on the first half of ORF body against the Gini index calculated on the second half of ORF body for samples modified with DMS *in vitro*.

(C) Plot of Gini index calculated on adjacent ORFs in operons, calculated from mRNA refolded and modified with DMS *in vitro*. Coloring indicates ratio in Translation Efficiency.

(D) Plot of Gini index of unmodified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs.

(E) Plot of predicted ΔG of computationally folded mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs.

(F) *in vivo* Correlation (Spearman's ρ) between computationally predicted mRNA structure of the ORF, quantified by predicted ΔG of minimum free energy structure, and the translation efficiency of the ORF. ΔG index was calculated for 300 nt windows that scan gene bodies, using genes that extend through the 300 nt window being examined, and is plotted at the center of each window.

(G) Plot of predicted Shine-Dalgarno strength (Salis et al., 2009) against measured translation efficiency. Genes with Gini indices in a tight range (.5 - .52) are indicated in cyan.

(H) Plot of mean predicted interaction probability across all well-expressed open reading frames.

Figure S3: Structure and translation efficiency remain correlated at 10°C, related to Figure 3

(A) Plot of Gini index of *in vivo* DMS-modified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs, measured 30 min following cold shock to 10°C.

(B) Plot of Gini index of *in vivo* DMS-modified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs, measured 6hr following cold shock to 10°C.

(C) Plot of Gini index of *in vivo* DMS-modified mRNA following addition of the translation initiation inhibitor against *in vivo* translation efficiency for well-expressed ORFs. TE was measured 8hr following cold shock, while structure was measured 40 min later following addition of kasugamycin.

(D) Plot of Gini index of *in vitro* DMS-modified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs. Translation efficiency was measured 30 min following cold shock to 10°C.

Figure S4: Csp deletion increases mRNA structure and reduces translation efficiencies, related to Figure 4

Plot of Gini index of *in vivo* DMS-modified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs, measured in a Δ*cspABEG* strain at 6 hr following cold shock.

Fig S5: CspB UTR structure is modulated by cold shock, related to Figure 5

Change in structure of the *cspA* 5' UTR upon cold shock. The predicted structure of the *cspA* 5' UTR was generated by constraining a minimum free-energy prediction with our DMS-seq measurements taken at 37°C (A) and immediately after shock to 10°C (B). The cold box element is highlighted in the blue box and the long range interaction regions is highlighted in a green box. Start codon is indicated by a red box. A color bar indicates the intensity of the DMS-seq signal at each position. DMS reactive bases (based on the ribosomal ROC derived threshold) are in yellow to red.

Fig S6: CspB UTR structure is modulated during cold recovery, related to Figure 5

(A) Histogram showing change in structure on Csp UTRs during cold recovery relative to other mRNAs. Gini index was calculated for 150 nt windows tiling all expressed mRNAs 6 hr vs 30 min after cold shock. The difference in Gini index between timepoints for each window was calculated.

(B) Plot of raw DMS signal at early and late times after cold shock, scaled relative to the most reactive position in the 5'UTR of cspB. Position 1 corresponds to nucleotide 1,639,739 in *E. coli* genome. Regions with large change in DMS signal between timepoints are boxed.

(C) Structure of *cspB* UTR 8hr after cold shock in Δ*cspBG*, presented as in Figure 5. Change in structure of the *cspA* 5' UTR upon cold shock.

Extended Experimental Procedures:

Strains and growth conditions *E. coli* K-12 MG1655 was used as the wild-type strain. All culture experiments were performed in MOPS medium supplemented with 0.2% glucose, all amino acids except methionine, vitamins, bases and micronutrients (Teknova). Cells were grown in an overnight liquid culture at 37˚C, diluted to an OD_{420} = .001 in fresh medium and grown until OD_{420} reached 0.4 where samples were collected. For 10°C samples, cultures were grown to OD_{420} = 1.1 at 37˚C and cold shock was performed by mixing 70mL of 37°C culture with 130mL of 0°C media, with continued growth of the culture in a 10°C shaker. Multiple deletion strains were generated by transduction of FRT-flanked deletion alleles from the Keio collection (Baba et al., 2006) followed by marker excision by Flp recombinase (Cherepanov and Wackernagel, 1995).

Ribosome profiling sample capture The protocol for bacterial ribosome profiling with flash freezing was described (Li et al., 2014). Briefly, 200 mLs of cell culture were filtered rapidly and the resulting cell pellet was flash-frozen in liquid nitrogen and combined with 650 µl of frozen lysis buffer (10 mM $MgCl_2$, 100 mM NH_4Cl, 20 mM Tris-HCl pH 8.0, 0.1% Nonidet P40, 0.4% Triton X-100, 100 U ml^{-1} DNase I (Roche), 1 mM chloramphenicol). Cells were pulverized in 10-ml canisters pre-chilled in liquid nitrogen. Lysate containing 0.5 mg of RNA was digested for 1 h with 750 U of micrococcal nuclease (Roche) at 25°C. The ribosome-protected RNA fragments were isolated using a sucrose gradient followed by hot acid phenol extraction. Library generation was performed using the previously described strategy (Li et al., 2014) detailed below.

Total mRNA sample capture For experiments performed in parallel with ribosome profiling, total RNA was phenol extracted from the same lysate that was used for ribosome footprinting. For experiments performed independently of ribosome profilng experiments, and for total mRNA used for *in vitro* DMS-seq experiments, 4mL of OD_{420} = 0.4 culture was added to 500µL of ice-cold stop solution (475 µL of 100% EtOH and 25µL acid phenol), vortexed, and spun for 2 min at 8000rpm. Supernatant was poured off, and the cell pellet was flash frozen in liquid nitrogen. Total RNA was then hot acid phenol extracted. For mRNA-seq experiments, ribosomal RNA and small RNA were removed from the total RNA with MICROBExpress (Ambion) or Ribozero (Epicenter) and MEGAclear (Ambion), respectively, following the manufacturers' protocols. mRNA was randomly fragmented as described (Ingolia et al., 2009). For total RNA sequencing experiments, these subtractions were not performed. The fragmented mRNA sample was converted to a complementary DNA library with the same strategy as for ribosome footprints.

mRNA-seq following rifampicin addition Rifampicin was added to a final concentration of 250 µg/mL at the designated time. Total RNA-seq samples were prepared as described for mRNA-seq samples, except that tRNA and rRNA subtraction was not performed.

Library generation for ribosome profiling and mRNA seq samples The footprints and mRNA fragments were ligated to miRNA cloning linker-1 (IDT) 5rApp/CTGTAGGCACCATCAAT/3ddC/, using a recombinantly expressed truncated T4 RNA ligase 2 K227Q produced in our laboratory. The ligated RNA fragments were reverse transcribed using the primer 5'/5Phos/GATCGTCGGACTGTAGAACTCTGAACCTGTCGGTGGTCGCC GTATCATT/iSp18/CACTCA/iSp18/CAAGCAGAAGACGGCATACGAATTGATGGT GCCTACAG 3'. The resulting cDNA was circularized with CircLigase (Epicentre) and PCR amplification was done as described previously (Ingolia et al., 2009).

DMS modification For *in vivo* DMS modification, 15 ml of exponentially growing *E. coli* were incubated with 750 µl DMS. Incubation was performed for 2 min at 37°C, and for 45 min at 10°C. For kasugamycin experiments, kasugamycin was added to a final concentration of 10 mg/mL after 8 hr at 10°C for 40 min prior to DMS modification. DMS was quenched by adding 30 ml 0°C stop solution (30% ß-mercaptoethanol, 25% isoamyl alcohol) after which cells were quickly put on ice, collected by centrifugation at 8,000g and 4 °C for 2 min, and washed with 8 ml 30% BME solution. Cell were then resuspended in 450 µL total RNA lysis buffer (10 mM EDTA, 50 mM sodium acetate pH

5.5), and total RNA was purified with hot acid phenol (Ambion). For *in vitro* DMS modifications, mRNA was collected in the same way as described above but from *E. coli* that were not treated with DMS. 2μg of mRNA was denatured at 95 °C for 2 min, cooled on ice and refolded in 90 μL RNA folding buffer (10 mM Tris pH 8.0, 100 mM NaCl, 6 mM MgCl2) at 37°C or 10°C for 30 min then incubated in either .2% (95°C) or 4% (37°C and 10°C) DMS for 1 min (95°C), 5 min (37°C) or 40 min (10°C). The DMS reaction was quenched using 30% BME, 0.3 M sodium acetate pH 5.5, 2 μl GlycoBlue solution and precipitated with 1X volume of 100% isopropanol.

Library generation for DMS-seq samples Sequencing libraries were prepared as described (Rouskin et al., 2014). Specifically, DMS treated mRNA samples were denatured for 2 min at 95 °C and fragmented at 95 °C for 2 min in 1x RNA fragmentation buffer (Zn^{2+} based, Ambion). The reaction was stopped by adding 1/10 volume of 10X Stop solution (Ambion) and quickly placed on ice. The fragmented RNA was run on a 10% TBU (Tris borate urea) gel for 60 min. Fragments of 60–70 nucleotides in size were visualized by blue light (Invitrogen) and excised. Reverse transcription was performed in a 20 μl volume at 52 °C using Superscript III (Invitrogen), and truncated reverse transcription products of 25–45 nucleotides (above the size of the reverse transcription primer) were extracted by gel purification.

Measurement of total protein synthesis 1μC of Perkin Elmer EasyTag ^{35}S labeled methionine (Product # NEG709A) was mixed with 5μL 288 μmol unlabeled methionine and 24 μL media. At the time of capture, 900 μL of culture was added to methionine mix, and was labeled on a shaker for the time of capture, 1 min at 37°C and 5min at

10°C. After labeling, 100 µL of 50% trichloracetic acid on ice was added to the sample, which was vortexed and placed on ice. Samples were left on ice for at least 20 min to allow precipitation. Samples were then counted by running 100µL of sample through a 25mm APFC glass fiber filter (Millipore APFC02500) pre-wetted with 750 µL of 5% TCA on a vacuum stand, and washing three times with 750 µL 5% TCA and three times with 750 µL 100% ethanol. Filters were then placed in MP Ecolume scintillation fluid and counted.

Sequencing Sequencing was performed on an Illumina HiSeq 2000 system. Sequence alignment with Bowtie v. 0.12.0 mapped the footprint data to the reference genomes NC_000913.fna obtained from the NCBI Reference Sequence Bank. Sequencing data from mutated strains were aligned to appropriately modifed versions of the NC_000913.fna genome. For ribosome footprint and mRNA-seq samples, the center residues that were at least 12 nucleotides from either end were given a score of 1/N in which N equals the number of positions leftover after discarding the 5' and 3' ends.. For DMS-seq samples, read counts were assigned to the base immediately 5' of the 5' end of each read, which is the base that was DMS-modified.

Computational prediction of RNA structures For identification of unpaired bases, raw DMS-seq data was normalized to the most highly reactive residue after removing outliers by 95% Winsorisation (all data above the 95th percentile is set to the 95th percentile). Bases with DMS-seq signal greater that 20% of the signal on the most highly reactive residue (after Winsorisation) were called "unpaired". For determination of *rimM* mRNA structures constrained by DMS-seq data, A Viennafold (Hofacker, 2003) minimum free energy model of the indicated region was generated, constrained by bases experimentally determined to be unpaired in the indicated dataset. For *csp* structure predictions, a conservative model was made in which the 20% of bases with highest DMS modification in the window were constrained to be unpaired. Color coding by DMS signal was done using VARNA (http://varna.lri.fr/).

Computing the agreement with ribosomal RNA The secondary structure models for *E. coli* ribosomal RNAs were downloaded from Comparative RNA Website and Project database (http://www.rna.icmb.utexas.edu/DAT/3C/Structure/index.php). The crystal structure model was downloaded from Protein Data Bank (http://www.pdb.org, PDB entries 3I1M, 3I1N, 3I1O, and 3I1P). The solvent-accessible surface area was calculated in PyMOL, and DMS was modeled as a sphere with 2.5 Å radius (representing a conservative estimate for accessibility because DMS is a flat molecule). Accessible residues were defined as residues with solvent accessibility area of greater than 2 Å^2. Unpaired residues in DMS-seq data were identified as described above. True positive bases were defined as bases that are both unpaired in the secondary structure model and solvent-accessible in the crystal structure model. True negative bases were defined as bases than are paired (A-U or C-G specifically) in the secondary structure model. Accuracy was calculated as the number of true positive bases plus the number of true negative bases divided by all tested bases.

Translation efficiency calculation Data analysis was performed with custom scripts written for R version 2.15.2 and Python 2.6.6. Mean ribosome density was calculated as described (Li et al., 2014). mRNA density was calculated by calculating the mean density of mRNA reads following a Winsorization applied to trim the top and bottom 5% of reads. For comparisons of translation efficiency between timepoints and between strains at 10°C, relative translation efficiencies were normalized by relative total protein synthesis, quantified through 35S-methionine incorporation as described above.

Metagene analysis of ribosome run-off and DMS structure Metagene analysis of ribosome run-off was perfomed as done previously (Ingolia et al., 2011). Codons 600-800, which appeared undepleted in all timepoints measured, were used to normalize timepoints.

Calculation of Gini index on DMS-seq data All Gini indices were calculated using the R package "ineq" to calculate Gini over As and Cs in the region specified for each experiement. For each DMS-seq sample, Gini indices were calculated only for genes that had greater than an average of 15 reads per nucleotide (A or C) across the gene body. Genes for which mRNA-seq data was discontinuous (due to an early termination event or an internal promoter, 1% of genes) were excluded from the analysis. Specifically, Gini indices were calculated on mRNA-seq data, and a cut-off was created based on two standard deviations from the mean.

Identification of adjacent open reading frames on operons Adjacent open reading frames in annotated operons often have differing levels of mRNA-seq reads, suggesting that they are not always on the same mRNA molecule. To identify adjacent ORFs expressed as a single operon, we assessed mRNA-seq data for equivalent mean message level, and for signal continuity, as described below. Equivalent mean message level was assessed by first determining the variability in mean mRNA-seq read density within individual ORFs. There is a single transcript that extends over the entire body of the large majority of ORFs, and so the variability in mean read density level in the first half of each ORF was compared to mean read density in the second half of each ORF, and the variability in this distribution was used to define a cut-off for ORFs on a single message. Adjacent ORFs that fell within a 2σ cut-off in mean level (calculated to be a 1.5-fold difference in mRNA level) were determined to have equivalent mRNA level, and were then assessed for signal continuity. Signal continuity was assessed by first determining the distribution of read density in windows within messages. Gini index of mRNA signal were calculated for all 50nt windows within ORF bodies, and the variability in this distribution was again used to define a cut-off for continuous mRNA regions. Gini index were then calculated for 50nt windows tiling the region between adjacent open reading frames. Gene pairs that fell within a 2σ cut-off defined by the intra-ORF distribution, were considered to be a pair of adjacent ORFs on a single message.

Directionality of interaction predictions For the determination of directionality of interaction at ORF boundaries, sequence from -250 to +250 nt relative to the translation start site was extracted for each adjacent pair of ORFs. A Viennafold (Hofacker, 2003) minimum free energy model of each 500nt sequence was then generated, constrained by DMS-seq dataset indicated, using DMS constraints as described above. The predicted probability of each base interacting with each other base in each mRNA structure model was then extracted from the Viennafold output. The mean probability of each position interacting with each other position across all analyzed messages was then calculated, generating a square matrix of interaction probability between all positions in the analyzed region. For each position between -150 to +150 nt relative to the translation start site, the summed probability of that position interacting with any of the previous 100 upstream positions was then calculated. The same calculation was performed for the 100 downstream positions. The ratio between sum upstream interaction and sum downstream interaction probability was then calculated for each position.

Identification of cold-induced open reading frames Cold-induced ORFs were identified by calculating synthesis rates through integrating ribosome profiling with 35S-methionine total protein synthesis measurements. At 37°C and at all timepoints following cold shock, the relative synthesis rate of each ORF was determined by multiplying total protein synthesis, measured by 35S-methionine total incorporation (see above) by the fraction of ribosome footprints mapping to that open reading frame. To calculate 37°C synthesis, the 37°C doubling time (26 min) was multiplied by 37°C synthesis rate. To calculate 10°C synthesis, the accumulated protein at each timepoint was multiplied by the window between that and the subsequent timepoint to estimate total synthesis within each window between timepoints. The total synthesis during all windows spanning the growth arrest period was then summed, and the ratio of 10°C synthesis to 37°C synthesis was calculated. For the large majority of genes, this ratio was << 1, as the absolute total protein synthesis rate was down > 100-fold relative to 37°C.

SD strength calculation For each open reading frame, SD strength was determined using the model established by (Salis et al., 2009). We used the RBS Calculator established by Salis et al downloaded from http://www.github.com/hsalis/Ribosome-Binding-Site-Calculator-v1.0.

Supplemental references

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology *2*.

Cherepanov, P.P., and Wackernagel, W. (1995). Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene *158*, 9–14.

CHAPTER 4

Causal signals between codon bias, mRNA structure, and efficiency of elongation and translation

Causal signals between codon bias, mRNA structure, and efficiency of elongation and translation

Cristina Pop[1], Silvi Rouskin[2], Nicholas T. Ingolia[3], Lu Han[1], Eric M. Phizicky[4], Jonathan S. Weissman[2], Daphne Koller[1]

Abstract: Ribosome profiling data reports on the distribution of translating ribosomes, at steady-state, with codon-level resolution. We present a robust method to extract codon translation rates and protein synthesis rates from these data, and identify causal features associated with elongation and translation efficiency in physiological conditions in yeast. We show that neither elongation rate nor translational efficiency is improved by experimental manipulation of the abundance or body sequence of the rare AGG tRNA. Deletion of three of the four copies of the heavily used ACA tRNA shows a modest efficiency decrease that could be explained by other rate-reducing signals at gene start. This suggests that correlation between codon bias and efficiency arises as selection for codons to utilize translation machinery efficiently in highly translated genes. We also show a correlation between efficiency and RNA structure calculated both computationally and from recent structure probing data, as well as the Kozac initiation motif, which may comprise a mechanism to regulate initiation.

[1] Computer Science Department, Stanford University, Stanford, California 94305, USA
[2] Department of Cellular and Molecular Pharmacology, California Institute of Quantitative Biology, Center for RNA Systems Biology, Howard Hughes Medical Institute, University of California, San Francisco, California 94158, USA
[3] Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California 94720, USA
[1] School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, New York, 14642, USA

Introduction

The translation of RNA into protein is the nexus of decoding genetic information into functional polypeptides and also a central biosynthetic process consuming a substantial fraction of the cell's resources. Although apparently redundant nucleotide sequences encode each protein, usage of different synonymous codons is highly biased (Plotkin & Kudla, 2011). These preferences are strongest in highly-expressed genes throughout diverse organisms (Man & Pilpel, 2007; Hershberg & Petrov, 2008), suggesting selective pressure for the efficient use of the translational apparatus during the synthesis of abundant proteins. At the same time, less common codons may be used in order to modulate translation, or may arise due to competing sequence constraints such as mRNA secondary structure. While the evolutionary signature of codon bias is clear, its biochemical basis remains unsettled.

Ribosome profiling (Ingolia et al, 2009) is an emerging technique for profiling translation *in vivo* that is well suited to provide insights into the factors controlling the speed of translation as well as the amounts of each protein produced by the cell. Ribosome profiling data comprise a set of ribosome-protected fragments (*footprints*) marking ribosome density along mRNA transcripts with codon resolution. We can therefore extract from these data both the yield of each protein (*protein synthesis rate*) and the rate at which each codon is translated (*codon translation rate* or *elongation rate*). However, estimation of these two quantities is nontrivial, and ad-hoc approaches disregard differences in elongation rates between genes or exclude mRNAs with sparse footprint coverage. A number of studies with different analysis approaches present varying hypotheses for the mechanisms underlying variation in elongation and *translation efficiency* in yeast and other organisms (Tuller et al, 2010a; Tuller et al, 2010b; Ingolia et al, 2011; Tuller et al, 2011; Stadler & Fire, 2011; Qian et al, 2012; Charneski & Hurst, 2013; Shah et al, 2013; Woolstenhulme et al, 2013; Lareau et al, 2014; Gardin et al, 2014). These include codon effects mediated by tRNA abundance or wobble base pairing, as well as effects of mRNA structure and the nascent peptide on the ribosome.

Here, we present a rigorous statistical method that estimates, from ribosome profiling data, both elongation rates and protein synthesis levels on individual transcripts; as a byproduct, it also estimates *translation efficiency* (TE), the propensity of a transcript to generate complete protein, defined as the total amount of protein produced from an mRNA message, and calculated here as our model-derived protein synthesis rates divided by the mRNA levels. We use our robust modeling framework in conjunction with new high-resolution data from wild-type yeast, along with three tRNA mutants, to explore some of the conflicting views on the causality between codon usage and elongation rate, as well as between codon usage and TE, in physiological conditions at a genome-wide level.

We first apply our model to examine biological factors contributing to local translation kinetics. Due to differences in tRNA levels that correlate with synonymous codon bias, variability in codon translation rates observed per gene is commonly thought to be governed by the abundance of cognate tRNAs (Varenne et al, 1984; Sorensen et al, 1989). However, codon bias does not correlate with indirect measures of decoding speed, at least in bacteria (Bonekamp et al, 1989; Curran and Yarus, 1989). Similar to other observations in ribosome profiling datasets (Li et al, 2012; Qian et al, 2012; Charneski & Hurst, 2013), we find that codon usage bias is a poor predictor of elongation rate. We further test for causal influence and illustrate that experimentally manipulating tRNA abundance or body similarly does not affect the elongation rate when decoding with the manipulated tRNA. In addition, our model identifies positions where elongation is slower than expected based on codon identity and suggests that such pauses commonly occur closer to the 5' end but are unrelated to codon bias.

Finally, we use our model to disentangle the factors underlying message-specific differences in translational efficiency. In physiological conditions, initiation rather than elongation may largely determine overall protein production; initiation predominates when it is slow relative to the time needed to elongate through the width of one ribosome (~10 codons), so that translating ribosomes rarely interfere with each other, and when elongation is highly processive, so that most initiation events result in a protein (Andersson & Kurland, 1990; Bulmer, 1991; Arava et al, 2003; Lackner & Bahler, 2008). Analysis of our tRNA-perturbed mutant experiments shows that efficiency is not causally affected by improving tRNA levels, leading us to focus on initiation signals in understanding variation in translational efficiency across different messages. Several causes for slow initiation have been proposed: codon bias at the 5' end (Tuller et al, 2010a; Tuller et al, 2011), secondary structure (Kudla et al, 2009; Gu et al, 2010; Kertesz et al, 2010; Tuller et al, 2011; Keller et al, 2012; Zur & Tuller, 2012), and gene length

(Arava et al, 2005; Lackner et al, 2007; Ding et al, 2012). We find that a Kozak-like initiation motif (Kozak, 1981) and lack of structure around the start codon are predictors of TE. Overall, our experimental and analytical results provide support to a previously proposed model in which initiation is rate-limiting in physiological conditions (Bulmer, 1991), in which initiation rate is affected largely by mRNA sequence features, and where translational efficiency is not significantly affected by codon usage (Andersson & Kurland, 1990; Bulmer, 1991). In contrast with experiments in non-physiological conditions, our results endorse the resulting explanation that, in endogenous conditions, perhaps in combination with other pressures, selection for efficient use of ribosomes and associated factors in the synthesis of highly-translated proteins is a potential driver of the observed codon usage biases.

Results

Queuing Model for Elongation Process

To extract high-quality estimates of protein synthesis rates and codon translation rates from the ribosome footprint data, we model the process of ribosome flow, using gene- and codon-dependent parameters, and the physical sampling that occurs in the experimental protocol from which these data are derived. Our design choices are motivated by potential biases in the data including sparse footprint counts for low abundance genes, biases due to the position along the mRNA, and biases due to the identity of the mRNA.

Our model inputs are the set of ribosome footprint counts ***d*** at each codon in the genome, sparsely sampled (due to sequencing depth) from an (unobserved) steady-state distribution $\boldsymbol{\pi}$. In particular, d_{mk} is the observed footprint count at position k in mRNA message m and π_{mk} encodes the fraction of ribosomes at *(m,k)*. Consequently, the distribution must satisfy *flow conservation constraints*: if ribosomes do not fall off the message, then due to conservation of matter, the protein synthesis rate J_m for message m (the ribosome flow out of the stop codon) must be the same as the flow J_{mk} from any position k on m. If we define μ_{mk} as the dwell time of the ribosome at *(m,k),* flow conservation also implies that rapidly translating positions (small μ_{mk}) are occupied for a smaller fraction of time (small π_{mk}) than positions that are slow to translate. The dwell time μ_{mk} is the inverse of the rate at which the ribosome elongates off of position *(m,k)* and so intuitively depends on the amount of time the ribosome requires to perform one elongation step (recruit tRNA, form the peptide bond, and translocate). Thus, at steady-state, flow J_{mk} is proportional (up to a constant encoding the number of ribosomes in the system) to π_{mk} / μ_{mk}, where we use d_{mk} throughout as our observed proxy for π_{mk}. Figure 1 shows the relationship between the variables.

We use the counts ***d*** to estimate the quantities $\{\mu_{mk}\}$ and $\{J_m\}$ in a novel probabilistic regression accounting for flow conservation and assuming steady-state and no ribosome fall-off. Briefly, we optimize over two terms:

$$\max_{\mu_m^c, \mu^c} log \prod_{m,k} {\mu_m^c}^{(d_{mk}/J_m)} \exp(-\mu_m^c) - \left[\sum_{m,c} w_m^c (\log \mu_m^c - \log \mu^c)^2 \right]$$

The first term is a standard likelihood term for the data, using a model encoding flow conservation. Since a single ribosome profiling dataset does not contain enough data to robustly infer a separate μ_{mk} for each *(m,k)*, we use the same dwell time μ_m^c for every occurrence of the same codon *c* within message *m*, making μ_m^c an expected dwell time for codon *c* on message *m*. The second term additionally softly constrains μ_m^c to be similar to a global codon dwell μ^c, based on the intuition that the same codon behaves similarly throughout the cell. To optimize the objective, we (1) estimate the dwell times μ_m^c and μ^c with flow J_m fixed and (2) set flow J_m to be the average of the flows J_{mk} (namely, the dwell-corrected footprint counts d_{mk} / μ_{mk}) across each message: $J_m = \Sigma_{k \text{ in } m} (d_{mk} / \mu_{mk}) / L_m$ (see **Materials and Methods** for details).

We ran our model on a ribosome profiling dataset gathered for *Saccharomyces cerevisiae* in rich medium, using a flash-freezing technique as described before (Ingolia et al, 2012) (see **Materials and Methods**). To verify the validity of our estimated parameters, we compared our protein synthesis rate J_m to two external measures of protein abundance – mass-spectrometry-based levels from de Godoy (2008) and GFP-based levels from Newman et al (2006) – and obtained strong correlations (Pearson r = 0.787 and 0.682, respectively, p = 0). These improve on the protein abundance estimates from Ingolia et al (2009), computed as the simple average of (uncorrected) footprint counts per message (Figure E1). While correlation with these standard estimates of protein abundance is reassuring, these methods have general limitations such as ascertainment bias for less abundant proteins as well as technical limitations such as the impact of fusion tags on protein levels. In addition, ribosome profiling measures translation and protein synthesis, but steady-state protein abundance is also affected by rates of protein degradation.

While the protein synthesis flux is perhaps the most obvious interesting quantity that can be extracted from profiling data, we can also derive other quantities of interest from our learned model parameters. We compute translation efficiency TE_m of a given mRNA molecule m by dividing protein synthesis rate J_m by mRNA transcript levels M_m, derived from mRNA fragment data collected separately in the ribosome footprinting experiment. We can identify codon-dependent effects on translation from differences in μ^c. By looking at footprint count deviation from expected dwell time at each (m,k), we can also examine differences among codons on the same message. In the following sections, using the parameters estimated under our robust probabilistic framework, we perform a comprehensive analysis of the biological factors influencing local and global dynamics of translation.

Codon translation is not affected by tRNA abundance or body sequence

A number of studies in *Escherichia coli* initially identified codon usage and the availability of tRNA as the dominant force for codon translation rate (Varenne et al, 1984; Sorensen et al, 1989). Later studies found no correlation between measured rates and tRNA abundance or codon frequency (Bonekamp et al, 1989; Curran & Yarus, 1989; Sorensen & Pedersen, 1991). However, all of these studies measured translation speed indirectly, on individual and potentially idiosyncratic reporter systems. We explore these competing hypotheses in the physiological conditions of our yeast data set. If tRNA abundance were rate-limiting for elongation, we would expect a positive correlation between codon translation rate and tRNA abundance. However, as shown in Figure 2, the correlation is insignificant (Spearman r = 0.135, p = 0.413 for Cy5 and r = 0.125, p = 0.448 for Cy3 from microarray tRNA measurements (Dittmar et al, 2004)). A similar result (r = 0.220, p = 0.088) is also obtained when comparing to tAI, a measure of codon bias based on tRNA gene copy number relative to the overall collection of isoacceptor tRNAs (dos Reis et al, 2004). If we restrict the analysis to the slowest synonymous codon (in terms of tAI), to the fastest, or to the average per amino acid, the correlation to tAI does not improve: r = - 0.11 (p = 0.64), r = -0.26 (p = 0.27), and r = -0.36 (p = 0.12), respectively. Finally, the same insignificant correlation exists in the raw footprint data (r = 0.109, p = 0.401; baseline method for rate described in **Materials and Methods**) and was also observed in another analysis of the yeast data set from Ingolia et al (2009), in which codon dwell time was estimated as the ratio of observed codon frequencies in the footprint data relative to expected codon frequencies in the mRNA fragment data (Qian et al, 2013).

Our analysis of elongation rates on endogenous mRNAs in the context of the co-adapted cellular tRNA pool addresses the effects of codon usage in natural physiology, but may be confounded by this co-adaptation and cannot directly test the causal links between various correlated mRNA features. To measure the effect of tRNA abundance on codon translation rate directly, we created three mutant yeast species to test whether (1) tRNA over-expression speeds up translation, (2) the tRNA body itself causes the tRNA-dependent rate effect observed in other studies, or (3) depletion of tRNA slows down ribosomes. In our first mutant, AGG-OE, the tRNA recognizing AGG (namely, tRNA$^{\mathrm{Arg(CCU)}}$) was over-expressed on a high-copy plasmid; in mutant AGG-QC, the body sequence of the tRNA recognizing AGG was swapped with the body of a more preferred tRNA (as measured by tAI); and in mutant ACA-K, 3 out of 4 copies of the tRNA recognizing ACA were deleted from the genome. The AGG mutants had a URA marker and were compared against a wild-type sample with a URA plasmid (see **Materials and Methods**). For ACA-K, we checked that the abundance of the tRNA for ACA (namely, tRNA$^{\mathrm{Thr(UGU)}}$) did decrease to about 30% of wild-type (Table E1). In the AGG-OE mutant, we measured the amount of total and aminoacylated tRNA for tRNA$^{\mathrm{Arg(CCU)}}$ (see **Materials and Methods**) and verified that the tRNA was over-expressed by 13.8-fold (+/- 0.4), based on an analysis of two independently derived RNA samples, and remained charged at a level similar to wild-type (87%) (Figure E2). For the AGG-QC mutant, we similarly verified that the amount of charged tRNA$^{\mathrm{Arg(CCU)}}$ was similar to wild-type (Figure E2). We generated ribosome profiling data and ran our model on these mutants to test whether AGG codons are translated faster in AGG-OE and AGG-QC and whether ACA codons are translated slower in ACA-K. We observe no significant change in the elongation rates of the affected codon in any of the three mutants compared to wild-type (Figure 3, E3); the overall correlation between ACA-K and wild-type is not as tight as for other mutants, but this is due to changes affecting all codons, not only ACA. We verified the result by inspecting the footprint counts at the perturbed codon relative to adjacent counts in the mutants compared to wild-type and saw no unusual increase or decrease (Figure E4). One prevailing hypothesis (Welch et al, 2009) is that the amount of charged as opposed to total tRNA is the true predictor of codon elongation; our measurements of aminoacylated tRNA suggest that these levels were manipulated as expected and that this is not a confounding factor in the mutant samples. Hence, our results suggest that several-fold changes in tRNA abundance do not affect ribosome dwell time.

Translation efficiency is mildly affected by tRNA knockdown but not by overexpression

One of the major goals of codon optimization in biotechnology is an increase in protein yield. Studies done on transgenes expressed at a large fraction of cellular mRNA abundance report increased protein abundance when the mRNA was optimized for codon bias (Gustafsson et al, 2004; Lavner & Kotlar, 2005; Burgess-Brown et al, 2008), suggesting that codon usage contributes to efficiency (Supek & Smuc, 2010; Tuller et al, 2010b). However, other studies observed that optimizing codon adaptation of a reporter does not significantly improve TE or protein yield (Wu et al, 2004; Kudla et al, 2009; Welch et al. 2009; Hense et al, 2010; Letzring et al, 2010; Shah et al, 2013). Our experiments likewise provide support for the view that the TE of endogenous mRNAs is unchanged by effective codon optimization achieved by changes in the tRNA pool (Figure 4). We find that increasing tRNA abundance or replacing the tRNA body sequence by one with higher tAI does not improve efficiency: most genes remain unchanged in TE between the wild-type and mutant samples (Pearson r = 0.96 for AGG-OE and r = 0.95 for AGG-QC). Further, the top 200 genes that do deviate most in TE relative to the wild-type sample have mutant TE that is both lower (*reduced TE genes*) and higher (*increased TE genes*) compared to wild-type, with bias towards reduced TE genes (134 reduced vs 66 increased for AGG-OE and 141 vs 59 for AGG-QC). In AGG-

OE, we observe no correlation between the fraction of AGG codons per message and the change between mutant and wild-type TE (Spearman r = -0.0057, p = 0.6775); we would expect a positive correlation if increasing tRNA abundance increased TE. Further, despite the many-fold overexpression of tRNA, the correlation between TE and fraction of codon per message for AGG is not higher than the correlation for any of the other codons (Figure 4). AGG-QC behaves similarly, such that manipulating the tRNA to be "faster" does not lead to a scenario where AGG outperforms other codons in affecting translation efficiency. Finally, these observations also hold if we look at protein synthesis rates instead of TE (Figure E5).

While improving codon optimization by changes in tRNA structure or abundance does not seem to causally affect TE, we do see evidence for a modest impact from tRNA depletion (Figure 4). Mutant and wild-type TEs are generally correlated in the ACA-K mutant (Pearson r = 0.95). Although there are more reduced TE genes than increased TE genes (107 versus 93), this difference is not significant via a permutation test (see **Materials and Methods**). However, we find a negative correlation, the lowest of all codons, between the fraction of ACA codons per message and the change in TE between mutant and wild-type (Spearman r = -0.066, $p < 10^{-5}$), as we would expect if decreasing tRNA abundance decreases TE through a direct effect on its cognate codon. One explanation is that tRNA reduction could compromise TE if the demand is higher than the supply – the number of ACA occurrences in the genome is about the average number of occurrences over all codons, but we reduced its levels below those of any other tRNA. However, if protein synthesis and thus TE are controlled by initiation, this implies some feedback from slowed elongation on initiation, whereby affected ACA codons might stack ribosomes. In particular, reduced TE genes compared to increased TE genes have slower-than-expected codons, including ACA, closer to the 5' end and stronger pausing in the first 100 codons (Figure E6; significant under Kolmogorov-Smirnov test; see next section for definition of slower-than-expected codons as "outliers"). These confounding factors might contribute to the decrease in TE for ACA-heavy genes. Alternatively, ribosome stacking at ACA codons could induce fall-off and reduced processivity that manifests as decreased TE.

To situate our results in the context of many previous studies on codon bias and tRNA abundance, we note that our observation focuses on endogenous messages with physiological or near-physiological tRNA levels. When the tRNA pool is limited compared to the number of free ribosomes, as in strong overexpression of transgenes, simulations indeed show that large demand for tRNAs can be rate-limiting (Chu et al, 2011; Chu & von der Haar, 2012; Shah et al, 2013). Experiments showing rate-limiting effects of tRNA abundance likely operated in this non-physiological regime. In addition, manipulation of codon usage rather than the tRNA abundance can perturb mRNA structure and other non-coding sequence features; our experiment is less susceptible to those issues.

Factors for elongation efficiency

The notably modest effect of dramatic changes to the tRNA pool motivates the question: what signals do affect elongation efficiency and translation efficiency? We first take advantage of the ribosome profiling data to understand elongation efficiency – the time for a ribosome to finish translating a transcript once initiated – by studying rate-limiting elongation signals via inspection of outliers in the footprint counts. Based on the observed footprint counts and our model parameters for expected codon dwell time, we define *slow outliers* and *fast outliers* at each position *k* along a message *m* as positions where ribosomes are stalled more or less than expected, respectively. We denote their deviation from expected dwell time as *outlier strength* Δ_{mk} (see **Materials and Methods**). We considered a broad array of potential correlates of Δ_{mk}, based on literature hypothesizing their association with variation in codon translation rate or pausing, classified into eight categories (Table E2): position on message, structure in downstream windows, protein folding, wobble basepairs, reuse of tRNAs from nearby codons, downstream RNA binding protein motifs, nascent peptide effects, and global features. Table E3 shows these correlations, which include significant features in the position, structure, wobble, and nascent peptide categories. We discuss these below and in Note E1.

The strongest correlation to outlier strength for slow outliers is proximity to the 5' end, with larger pauses occurring closer to the beginning of a message, even relative to gene length or even when aligned by stop codon as opposed to start codon (position from 5' correlates to Δ_{mk} with Spearman $r = -0.038$; position from 5' per length with $r = -0.136$; and position from 3' end with $r = 0.118$, $p \approx 0$ for all). Similar observations of increased ribosome occupancy at the 5' end have produced various hypotheses for the causal basis. In the "ramp" model (Tuller et al, 2010a), the presence of more slow codons (low tAI) at the beginning of a message is thought to separate ribosomes early to avoid the wasteful expenditure of resources on stacked, idling ribosomes. However, we observe a correlation between position from 5' end and slow outlier strength even when conditioning on the codon (Figure 5), and thereby controlling for differences in codon usage at different positions within the gene, suggesting that there is an initial low translation speed, regardless of codon usage, which gradually increases as translation proceeds. Additionally, our model helps account for length, position, and abundance biases when calculating outliers in a particular message in two ways: first, we include message-specific codon dwell times, and, second, we exclude the first 100 codons from each gene during model learning (see **Materials and Methods**) to avoid inflating or otherwise biasing the expected rates μ_m^c and μ^c. Our analysis indicates that pausing occurs at the 5' end, even after accounting for major factors such as codon bias and gene length.

Other explanatory signals have been suggested for pausing in ribosome profiling datasets (Stadler & Fire, 2011; Li et al, 2012; Charneski & Hurst, 2013). Our analysis shows a (mild) correlation between pausing and computationally-predicted downstream mRNA secondary structure (Spearman r = 0.021, p ≈ 0 with structure measured by the density of stems). This correlation is reproduced when considering experimentally derived *in vivo* structure data from high-throughput DMS probing of unpaired A and C bases (Rouskin et al, 2013) (r = -0.017). It is also maintained when we restrict our analysis to slow outliers in the first 100 codons (r = 0.016 for density of stems, though only r = -0.007 for *in vivo* energy, potentially due to genes with short UTRs and the decreased reliability of DMS structure probing data at ~20nt or less from the 5' end), and so the effect is not necessarily caused by structure elsewhere on the strand. Single molecule experiments with bacterial ribosomes (Chen et al, 2013) found that some hairpin and pseudoknot constructs at varying distances downstream of the active codon can slow down the ribosome; structural energy could therefore potentially contribute to the excess ribosome density at the 5’ end. We also see a positive correlation on that same order of magnitude between slow outliers and the number of proline codons in the two sites upstream of the active codon ($r = 0.078$, $p \approx 0$), as observed in other organisms (Ingolia et al, 2011; Woolstenhulme et al, 2013). Two correlations that we observed are not expected on the basis of previous studies. A study showing pausing specifically at CGA (Letzring et al, 2010) suggests slower elongation on wobble base pairs, whereas we observe the opposite correlation; this discrepancy might arise because the wobble effect is limited to a few specific codons, or to repeated wobble codons, or because of an incomplete characterization of codon / anticodon pairings which limits our assignment of wobble decoding. The correlation to charge observed by Charneski & Hurst (2013) holds only when considering the number of Arg and Lys residues in a window upstream of the active codon, although this result was later attributed to technical artifacts relating to the strand orientation (Charneski & Hurst, 2014).

Factors correlating with translation efficiency

While elongation efficiency measures time required to synthesize a new protein, translation efficiency measures the throughput of protein synthesis. Besides codon adaptation, which we find to play little or no causal role in improving efficiency, other significant correlates to TE include structural features and the sequence motif around the start codon (Figure E7).

Structure is reduced near the translation start site in many organisms (Gu et al, 2010; Zhou & Wilke, 2011) and, in combination with specific structural motifs downstream, can promote or halt initiation (Kozak, 1990; Kochetov et al, 2007; Robins-Pianka et al, 2010). We performed a sliding window analysis (see **Materials and Methods** and Figure 6) to correlate TE with RNA secondary structure in 40nt windows along the gene, for both experimental *in vitro* and *in vivo* structural energy (Rouskin et al, 2013). The window near the start codon is most significant, as reported previously for computational and *in vitro* structure measurements (Kudla et al, 2009; Kertesz et al, 2010; Tuller et al, 2010b; Keller et al, 2012); the positive correlation indicates that increased TE corresponds to loose structure in this region. Indeed, this is also the window with highest energy, corresponding to the lowest structure, as averaged over all genes (first red line in Figure 6). Interestingly, the correlation to TE for *in vivo* structure is less pronounced and the window is shifted 3 codons downstream. We call this Window A.

Our attention was also drawn to the window downstream of the start codon at ~60nt *in vitro* and ~80nt *in vivo* (second red line in Figure 6) with the lowest energy (more structure) compared to neighboring positions. We call this Window B. The most likely role for this energy barrier seems to be a stalling mechanism. Ribosome density is high nearby: at 135nt (approximately two to three ribosome footprints downstream), our model-estimated ribosome density has a notable peak that is reduced when we exclude outliers, which capture positions where sufficient pausing could stack ribosomes (Figure E8). Although properly placed downstream structure can improve the efficiency of initiation by stalling the scanning pre-initiation complex (Robins-Pianka et al, 2010), or might be selected for heavy structure in order to prevent other regions (namely, around the start codon) from being paired, the lack of significant correlation to TE for Window B suggests that ribosome flow control here optimizes other aspects of translation besides throughput.

In addition to low structure at the start codon, initiation may be assisted by recognition of a 12-mer motif around the start codon called the Kozak sequence in eukaryotes (Kozak, 1981), derived in yeast based on a sequence consensus from highly expressed genes by Hamilton et al (1987). As expected, due to a tight correlation between mRNA abundance and TE (Figure E7), similarity to the Kozak motif correlates strongly to TE (Spearman r = -0.21, $p < 10^{-45}$) (measuring similarity by Kullback-Leibler divergence to the position-weight matrix where 0 divergence means a closer match). The 3rd nucleotide preceding AUG is the most significant (Spearman r = -0.16, $p < 10^{-25}$), consistent with experimental measures of initiation efficiency after modifying positions in the Kozak site (Yun et al, 1996; Looman & Kuivenhoven, 1993). Using a linear regression model for predicting TE based on a set of correlates suggested in literature (see **Materials and Methods**), we learn a refined Kozak motif to reflect highly *efficient* genes (Figure 7). Our learned Kozak motif reduces the error of our regression model predictions relative to an equivalent model using the original motif (from 0.83 to 0.75, averaged over 100 test sets selected randomly, compared to a null model error of 0.96) (Table E4). This indicates that our refined motif better corresponds to highly translated genes, likely because it was trained directly on translation efficiency measurements rather than on a proxy such as mRNA abundance.

Finally, we tested the correlation between translation efficiency and other mRNA features often discussed in literature (Figure E7). We find a negative correlation to evolutionary rate that is suggestive of the intuitive fact that more conserved genes are more highly translated. The positive correlation we find with mRNA abundance suggests a model of co-expression where the need for high protein abundance drives high translation of abundant transcripts. Consistent with previous studies (Ingolia et al, 2009), we observe a negative correlation to length, but it is not significant. We also find a positive correlation (although weaker than that for tAI) to the codon translation rates geometrically averaged over the codons within a gene. Lastly, RNA-binding proteins (RBPs) have recently received attention for their roles in post-transcription regulation, and we also see high Spearman correlations between RBP occupancy and TE. When looking at enrichment of 15 proteins, we find the expected correlation to translation efficiency (as suggested by literature) in eight of ten cases. One of the two “unexpected” proteins, scp160, was recently reported to be required for translational efficiency of particular mRNAs in yeast (Hirschmann et al, 2014), even though it correlates negatively to ribosome occupancy in Hogan et al (2008); our analysis encouragingly suggests the former correlation. Note E1 has further discussion.

Discussion

In this paper we present a statistical model to extract codon translation rates and protein synthesis levels from ribosome profiling data. This robust framework allows us to shed new light on causality in regulation of translation and characterize the features associated with efficient elongation and translation. Although codon usage is a strong correlate to TE (Figure E7), our mutant experiments suggest (via the correlation between codon bias and tRNA abundance) that codon usage may not causally influence efficiency. The direct impact of codon usage on efficiency and the basis of the selective force underlying codon bias has remained a topic of controversy for decades. Some authors have proposed that codon optimization serves directly to enhance the translational efficiency of specific genes, perhaps by speeding elongation on their mRNAs. Our work provides direct experimental evidence against this view. Rather, our work is consistent with an alternative model, aligned with previous results for *Escherichia coli* (Kudla et al, 2009), in which codon bias in highly translated genes results from selection to optimize utilization of the translational machinery, whose abundance and production represents a major limitation on cell growth (Andersson & Kurland, 1990; Bulmer, 1991; Kudla et al, 2009); this selection induces a correlation without implying that increasing codon bias optimizes efficiency on individual genes (Welch et al, 2009). In this view, initiation is rate-limiting and thereby determines translational efficiency. When the demand-supply balance for a tRNA is not compromised by extremely high expression of a transgene not adapted to the host organism, we propose that selective forces beyond the TEs of individual messages guide the distribution of codons. The positive correlation between elongation rate and TE suggests that one contributor could be selection for efficient use of ribosomes and translation factors and that this selective force is strongest for high-expression, high-TE genes. Such selection pressure is consistent with studies of overall cell growth and protein synthesis, which indicate that the translational apparatus is rate-limiting for cell growth and that reduction in the amount of ribosome time devoted to producing an abundant protein can speed cell growth (Andersson & Kurland, 1990; Arava et al, 2003; Kudla et al, 2009; Mitarai & Pedersen, 2013). As elongation rate is not the strongest correlate to TE, other mechanisms also deserve further study. For example, there may be selective pressures on the mRNA sequence itself (e.g., to induce certain secondary structures), which in turn create pressure in the cell to ensure a sufficient supply of tRNAs for efficient translation of the highly translated messages. Our results are also consistent with the prevalent view that initiation is typically the rate-limiting step in protein synthesis, which does not provide a clear mechanism for codon usage in the body of a gene to affect its efficiency, and particularly not through increased elongation rates. Instead, tRNA levels are likely forced to match the lack of disfavored codons by selection against the cost of tRNA production or against poor decoding accuracy.

Our model is designed to account for the complexities of ribosome profiling data while keeping parameter estimation tractable. Although average footprint density on a gene is well correlated to protein abundance, outliers can pull the estimate provided by the mean away from the true level, especially when ribosome stacking is common. Thus, properly accounting for differential elongation rates can improve inference of protein synthesis levels from this data. We maintain a simple translation model (for example, we do not explicitly include a rate of ribosome falloff or an analytical treatment of codons being processed in series), but our design choices trade-off for model simplicity, algorithmic stability, and smoothing of noisy data. Using one model parameter for all codon instances in a gene, as opposed to an individual dwell per position, has several advantages: it averages out sequence biases in footprint fragments, makes the optimization algorithm less susceptible to local minima and hence robust to parameter initialization, and allows us to infer parameters even for low abundance genes by offsetting the lack of data with soft prior constraints. We reassuringly find qualitatively similar results when we replace our refined protein synthesis rates with a simple average of the footprints per gene, while obtaining better quantitative estimates compared to existing protein abundance datasets. More physics-based or simulation models (Zhang et al, 1994; Reuveni et al, 2011; Tuller et al, 2011) require knowledge of the kinetic parameters of translation, can necessitate grossly simplifying assumptions such as a single codon translation rate per gene, base certain model quantities on a limited set of features, or directly assume that codon rate is correlated to codon adaptation. In comparison, our method reduces the number of assumptions made by directly modeling the experimental processing and fitting the model parameters to the data under the single concept of flow conservation. On the other hand, methods that aggregate the data directly (Qian et al, 2012; Charneski & Hurst, 2013; Gardin et al, 2014), similar to our baseline method for calculating codon translation rates, do not readily lend themselves to computing other quantities. For example, because we have an underlying model, detection of outlier codon positions follows easily within our framework, whereas other works rely on choosing an adjacent window of appropriate size to compare counts. Similarly, we can easily study other potentially interesting effects, such as codon translation rate variance within genes and among genes. Finally, our method would particularly be useful in situations where ribosomal profiling data is scarce or noisy. By using a probabilistic model, we infer rates of interest from the observed, noisy data without needing to exclude genes with sparse information. With the growing usage of ribosome profiling, a robust framework for studying rates of elongation and synthesis is essential.

Our resulting analyses address the contributions of initiation versus elongation to efficiency (Arava et al, 2003; Lackner & Bahler, 2008; Shah et al, 2013). While efficient usage of ribosomes and elongation factors influence the overall amount of protein produced from the whole genome, initiation may dictate differences between genes (Firczuk et al, 2013). We characterize two initiation signals that could play a role in translation regulation via a two-stage metering-light model: reduced structure around the start codon and favorable sequence context to promote ribosome binding, followed by an increase in structure that could, in turn, serve to reduce misfolding of the emergent polypeptide by allowing sufficient time for recruitment of chaperones to the ribosome exit tunnel (Fredrick & Ibba, 2010). This barrier could reflect the observed universal per-gene effect, independent of codon identity, whereby the strengths of slow outlier positions correlate to 5’ end proximity. Since translation is resource-heavy, requiring tRNAs, mRNAs, and ribosomes, with the latter being especially costly to produce, we intuit that the cell must balance use of these finite resources while at the same time producing functional protein products. Structure around the 5’ end could be one of the key mechanisms through which the cell regulates translation so as to avoid wasting resources.

The region of slow elongation at the 5’ end certainly merits further exploration. In contrast to the slow-codon ramp proposed in Tuller et al (2010a), our model shows that while there may be an abundance of low tAI codons near the 5’ end, these codons do not cause slow elongation (Figure E9). We find (mild) correlations between pausing and downstream structure, between tAI and downstream structure over the first 50 codons of all genes (Spearman r = 0.007, p = 0.03 *in vitro* and r = 0.009, p = 0.003 *in vivo*), but not between codon usage and codon translation rate. A study performed over diverse bacteria, controlling for GC content, proposes that structure drives codon usage early at the 5’ end (Bentele et al, 2013); in yeast, there may be similar selection whereby structure-related constraints induce a low-tAI ramp.

The impact of secondary structure on translation is complex. In addition to a role in initiation, high structure regions could also act by influencing elongation (Chen et al, 2013). Outliers in the high-variance ribosome profiling data can differ from expected dwell times by a factor of 40, and are distributed throughout the message (Figure E10). One explanation is the presence of downstream structural features that create an energy barrier to elongation; these correlate (more weakly) to outlier strength when ignoring the first 100 codons (whole gene versus truncated gene has r = -0.017 versus r = -0.019 for downstream *in vivo* energy and r = 0.021 versus r = 0.009 for density of stems), precluding the possibility that high ribosome density (based on the 5’ end as a proxy) drives the effect. In addition, mRNA-binding factors can interact with structure (Dethoff et al, 2012), but whether structure performs any common genome-wide functions is not yet established. One possibility is that secondary structure slows the ribosome during elongation to promote correct folding of the nascent protein during its vectorial synthesis by the ribosome.

The significant but mild correlation to structure suggests that other factors also play important roles in pausing. Experiments suggest that the wobble base in the CGA codon causes significant pausing (Letzring et al, 2010; Stadler & Fire, 2011), clusters of slowly translated codons could stall ribosomes more than the sum of their individual decoding times (Zhang et al, 2009), and effects from the nascent peptide could stall elongation, for instance at prolines (Ingolia et al, 2011; Woolstenhulme et al, 2013). It is likely that a compendium of biological features interact to dictate elongation rate. Although our genome-wide outlier analysis shows promising correlations between pausing and features, the small magnitude of the correlation could be improved by looking at more restrictive or genetically meaningful sets of positions. The growing interest in ribosome profiling poses exciting directions for further investigation of the interactions between these features and the changes that may occur in different conditions. With this additional data and measurements from single-molecule experiments (Wen et al, 2008; Uemura et al, 2010), our model could be extended to include finer-grained parameters for codon translation rates, partitioned in various ways, in order to better understand how rate changes over a transcript. Further analysis is also needed into how structure and the sequence around the initiation site work together or against each other. For example, heavy structure can promote initiation in spite of weak initiation context, but the ways that they interact are still unknown.

In this paper, we present a method that provides a rigorous perspective for analyzing the increasing number of ribosome profiling data sets, and thereby addressing these important questions. We illustrate the use of the method in the context of one of these data sets to create a high-level view of the mechanisms involved in initiation and elongation, to study the factors affecting initiation as the rate-limiting step for translation, and to support a model in which the direction of causality goes from translation efficiency to codon usage rather than the opposite.

Materials and Methods

Ribosome profiling datasets

All experiments were done on yeast strain 288C. Cells were collected for ribosome profiling by filtering ~250ml culture of OD= 0.6 and immediately flash freezing on liquid nitrogen. For all ribosome-profiling experiments, footprints were obtained as described before (Ingolia et al, 2012). Three out of four copies of Threonine tRNA (tT(UGU)G2, tT(UGU)H, tT(UGU)P), recognizing the ACA codon were knocked out using the standard technique of homologous recombination from a plasmid PCR product. The resulting strain was marked with nourseothricin, kanamycin, and hygromycin B resistance respectively. Successfully transformed yeast were identified by check PCR. tRNA arginine (tR(CCU)J) recognizing the AGG codon was overexpressed by cloning into a URA marked 2micron plasmid (pRS426) and transforming wild-type yeast using –URA selection. For the tRNA body swap, tRNA sequence from tR(UCU)B was mutated in the anticodon to CCU using QuikChange site-directed mutagenesis kit (Stratagene) in order for the tRNA product from tR(UCU)B to recognize the AGG codon. The mutated tRNA was then cloned in the 2micron plasmid pRS426 and transformed into 288C.

Ribosome-protected fragments were aligned against *Saccharomyces cerevisiae* assembly R63 from the Saccharomyces Genome Database (SGD, http://www.yeastgenome.org) and we kept uniquely mapped reads with no more than 2 mismatches and lengths between 28 and 31. To identify the active codon for ribosome-protected fragments, we let 0 be the first nucleotide of the read and if the read begins on the first/last/middle nucleotide of a codon, the active codon starts at nucleotide 15/16/17, respectively. An mRNA fragment was mapped to a gene if it begins less than 16nt upstream of the start codon and more than 16nt upstream of the stop codon. Genes were ignored if they did not have an AUG start codon, had internal stop codons, had less than 50% of positions on the coding sequence with at least one mapped mRNA count, or if all the footprint counts were 0 over the gene length used in the translation model (see below), leaving around 5000 genes in each sample. When comparing mutants to wild-type samples, we used the intersection of the valid genes in each sample. The AGG mutants were compared against the wild-type sample with a URA plasmid.

Analysis of tRNA charging and relative RNA levels

For analysis of charging levels of tRNAs, duplicate samples of each strain were grown under conditions used for ribosome profiling, followed by harvesting of ~4 OD-ml of cells. Then, bulk RNA was prepared from each pellet under acidic conditions (pH 4.5) using glass beads, and RNA was resolved on a 6.5% acrylamide gel at pH 5 for 15 h at

4°C, transferred to Hybond N+ membrane, and hybridized with appropriate 5'-labeled oligonucleotide probes, as described (Alexandrov et al, 2006). Charging levels were visualized on a Typhoon PhosphorImager (GE Healthcare) and quantified using Imagequant, and relative levels of tRNA$^{Arg(CCU)}$ were measured by normalization to levels of tRNA$^{Leu(CAA)}$ in the corresponding lane.

Feature calculations

Gene copy numbers for tRNA were obtained from the tRNAscan-SE database (Lowe et al, 1997). To measure codon usage bias, we use tAI, which ranges from 0 to 1 for more preferred codons, calculated as in dos Reis et al (2004) with refined weights described in Tuller et al (2010a).

Experimentally derived structure data from DMS probing (Rouskin et al, 2013) was normalized in windows of size 150nt by the minimum count in the top 5% of A and C nucleotides, and the top 5% of counts were set to 1. Windows with less than ten A and C nucleotides in the top 5%, windows with a zero normalization constant, genes without data, and genes without a characterized UTR (Nagalakshmi et al, 2008) were ignored in analyses. In the sliding window energy analysis, energy windows were normalized per gene by the mean over windows on each gene. In the energy profile, normalized windows were then averaged across positions without missing data, aligned by start codon. In the energy-TE correlation profile, we applied a conservative Bonferroni correction by multiplying the p-values by the number of windows (30 upstream of the start codon and 250 downstream, since this span covered the maximum number of genes). To calculate the location of the dip in the energy profile, we identified global minimums within spans of 90nt and took the first minimum.

The correlation between tAI and downstream energy is for tAI over windows of 3 codons in the first 50 codons of all genes and the associated average of the 40nt energy windows 15nt downstream from each nucleotide in the tAI window. Energy windows are calculated as above using DMS *in vitro* and *in vivo* energy.

Translation model

As discussed in the main text, we optimize our objective over the parameters μ_m^c and μ^c and solve for J_m. Since individual footprint counts can be noisy and sparse, we smooth the data in three ways. First, we use a single μ_m^c for every copy of codon *c* on message *m*. These dwells softly agree with the global μ^c in a weighted geometric average with weight w_m^c, the number of codons *c* on gene *m* normalized by the number of codons *c* over all genes. Hence, genes with more copies of codon *c* get a larger vote in the average estimating μ^c. Second, we add a pseudo-count of 1 to all footprint counts and use the logarithm of normalized counts in the Poisson term (similar to a more robust geometric average as opposed to an arithmetic average that is easily skewed by outliers), first scaling the flow-normalized counts by a single factor over all *(m,k)* so that the lowest one is 1. Third, during model training, we ignore the first 100 codons (or the first 25% for genes shorter than 100 codons) since this region may have unusual flow conservation properties. If it doesn't, excluding these codons should not affect the learned rates. The second term in the objective function is multiplied by a constant C = 100 so as to not be greatly outweighed by the data term. Altogether, we solve the following optimization problem (where *k'* is restricted and *d'* are scaled as described above):

$$\max_{\mu_m^c,\mu^c} log \prod_{m,k'} {\mu_m^c}^{log(d'_{mk}/J_m)} \exp(-\mu_m^c) - C\left[\sum_{m,c} w_m^c(\log\mu_m^c - \log\mu^c)^2\right]$$

We verified that the constant C did not affect our results by running the main analyses again – correlations for codon bias measures, protein abundance, and outliers – on several other values (1, 10, 1000, 10000, 100000). We note no significant change (Table E5), except for some outlier correlations for 100000: pos-from-end is now significant; multi-down is not; is-in-domain is significant, suggesting slow outliers lie outside of protein domains; dist-prev-codon and dist-prev-trna are significant, suggesting that slow outliers are associated with nearby codons of the same type or using the same tRNA; pair-Pro-down is now significant, suggesting that slow outliers are not associated with pair prolines downstream. Similar to taking the limit of the constant to infinity, we also considered a model with only μ^c parameters and no μ_m^c (and hence no second term in the objective function) (Table E5). Again, no extreme change exists in the correlation between codon translation rate and codon bias measures. Perhaps because we have removed a layer of parameters, we do see a slight decrease in correlation to protein abundance and some changes to outlier correlations: pos-from-end is now significant; hairpins-down, multi-down, stems-down15, stemsGC-down15 are no longer significant but still show a similar correlation strength; is-in-domain is significant, suggesting again that slow outliers lie outside of protein domains.

The optimization algorithm is as follows: J_m is fixed to $D_m = \Sigma_{k\ in\ m}\ d_{mk}\ /\ L_m$ and μ_m^c and μ^c are initialized to dwells from the baseline method (see below), shifted in log space so that the mean is log(6.8), plus a small random number. The value 6.8 is the mean over all *(m,k)* of the flow-normalized counts normalized and smoothed as described above for the wild-type sample. The appropriate mean value was replaced for each of the mutant samples. The parameters are estimated via coordinate descent by iterating through codons c and learning the associated μ_m^c and μ^c. Optimization per c used an L-BFGS method (Byrd et al, 1995; Matlab wrapper, http://www.cs.toronto.edu/~liam/software.shtml) with the following stopping criteria: max number of iterations 5000; gradient tolerance 1e-5; function tolerance 1e3. Coordinate descent was stopped when the difference in weights was less than 5e-5 or the difference in function value was less than 1e-5. Codons not appearing in a particular gene m did not have an associated μ_m^c and we also excluded the stop codons. We then compute $J_m = \Sigma_{k\ in\ m}\ (d_{mk} / \mu_{mk}) / L_m = \Sigma_{k\ in\ m}\ (d_{mk} / \mu_m^{c=codon(m.k)}) / L_m$. The optimization is not sensitive to initialization (Figure E11).

Although less robust, we also optimized a model with a separate dwell time μ_{mk} for every *(m,k)* with the following initialization of weights: $\mu_{mk} = \mathrm{d}_{mk} / D_m$, with 0 counts replaced by the mean of all non-zero counts, shifted in log space so that the mean is log(6.8); μ^c are dwells from the baseline method (see below) shifted in log space so that the mean is log(6.8); all weights perturbed by a small random value. The value 6.8 was chosen as above. L-BFGS settings were as above. Coordinate descent was stopped when the difference in weights was less than 1e-2 or the difference in function value was less than 1e-1. The overall codon dwell times μ^c were well correlated to those in the original model (Pearson $r = 0.99$, $p < 10^{-68}$), but analyses based on dwell times per *(m,k)* could be impacted, since these parameters are more sensitive to initialization. So we verified all qualitative observations presented still hold. The correlation between codon translation rate and codon bias measures is insignificant (r = 0.171, p = 0.295 for Cy5; r = 0.158, p = 0.336 for Cy3; r = 0.249, p = 0.053 for tAI). Protein abundance estimates correlate similarly to external measures (r = 0.787 for de Godoy (2008) data and r = 0.670 for Newman et al (2006) data, p = 0 for both). In the outlier analysis, all correlations still hold except position from 3' end is now also significant, for the structure features only the density of stems 9nt downstream is significant but the others are on the same order of magnitude, and the protein domain feature is significant for bases inside a domain. Correlations between TE and gene-level features are similar except length is now barely significant, experimental *in vitro* energy for the mRNA sequence is barely not significant, and Npl3 is significant (in the expected direction). The energy-TE correlation profile is the same except the window at 18nt for *in vivo* energy is barely not significant. The ribosome density graph has the same peak at 135nt and decreases when outliers are removed. The refined Kozak motif has the same dominant bases except position 6 in Figure 7 has A swapped with C and the non-dominant T at positions 2 and 3 are swapped with A. Finally, the error when replacing the learned Kozak motif with the original similarly drops from 0.77 to 0.68.

Baseline method for codon translation rate

To get dwell time per codon c from the raw data, we average over counts (m,k) for which $codon(m,k) = c$, normalized by the average per gene ($D_m = \Sigma_{k\ in\ m}\ d_{mk} / L_m$). Rate is the reciprocal of dwell time. As above, we first add a pseudo-count of one to each d_{mk} and ignore the first 100 codons (or the first 25% for genes shorter than 100).

Analysis of translation efficiency in mutants

To test if the difference in the number of reduced TE genes versus increased TE genes (107 versus 93) in ACA-K is significant, we permuted the mutant TE values 1000 times and calculated the number of reduced TE versus increased TE genes for each permutation. There were 0 cases where the difference was less than the original difference, indicating the original difference is not statistically significant.

Model for translation efficiency

We used a regression model to predict TE of an mRNA message based on various features:

$$\min_{w} \sum_{m} (TE_m - w^T f_m)^2 + \lambda_1 \sum_{p} |w_p| + \lambda_1 \sum_{p} w_p^2$$

The first term fits an optimal set of weights ***w*** to the TE of a set of genes *{m}* using a linear combination of the set of features $\boldsymbol{f_m}$. The last two terms enforce sparsity (so that features that do not explain the data well receive a weight of 0) and shrinkage (so that weights are kept at a small scale). Under a standard machine learning framework, we divide the genes in our yeast dataset into a test set (size 400 genes) and a training set (the remaining genes). The hyperparameters λ_1 and λ_2 are learned via cross-validation: we further divide the training set into fifths, and evaluate the error for a grid of hyperparameter values on each fifth of the training set. The weights ***w*** are then learned on the whole training set with the best hyperparameters (with lowest cross-validation error). Test set error is the squared norm difference between predicted and actual TE, averaged over all genes in the test set. For reference, we create a null model where the weights are learned from TEs randomly permuted among the genes. The final weights are the average over all training/test combinations. The features used are minimal in order to maximize the number of genes that have these characterized: tAI of gene; computationally predicted energy of 5' UTR, 3' UTR, mRNA, and window around the start codon with highest correlation to TE; length of coding sequence; mRNA abundance; identity of bases overlapping the Kozac site (genes without a characterized UTR (Nagalakshmi et al, 2008) were excluded).

To compute the weights for the refined Kozak site, we include a feature f^k in $\boldsymbol{f}$ defined as $f^k = 1/(1 + \exp(x{*}g))$. The vector g has 36 indicators, 4 per each of the 9 positions in the Kozak site (excludes the start codon). The vector x has the corresponding weights for each indicator, is included in the shrinkage term, and is learned iteratively with $\boldsymbol{w}$. The refined Kozak motif in Figure 7 is the average of the 100 values of x learned separately for each training set. To create a position-weight matrix from these weights, we shift the weights for each position so that the most negative value (if any) is 0 and normalize by the sum of the four weights at that position. The sequence logo was generated by seqLogo (Bembom O, seqLogo: Sequence logos for DNA sequence alignments, R package v1.28.0, http://bioconductor.org/packages/release/bioc/html/seqLogo.html).

To test whether the refined motif provides better TE predictions than the original Kozak motif, within each of 100 training sets, we fix f^k for each sequence with x set to the original motif (scaling the weights so that the sum at each position matches the sum of the learned motif) and learn the remaining weights as before. We then compute accuracy on the corresponding test set.

Outlier model

The strength of an outlier Δ_{mk} at position *(m,k)* is defined as the difference between the observed count (d_{mk}) and the expected count (J_{mk} * μ_m^c), divided by s_{mk}, a standard deviation representing the variance in that count due to the abundance of the gene and the codon it corresponds to. For s_{mk}, we divide the genes into 30 quantiles by abundance and compute the standard deviation of the counts in each bin per codon. Thirty was chosen as the maximum number that still gave at least three counts in each bin per codon and no zero-valued s_{mk}. This normalization helps distinguish true biological outliers from outliers arising due to differential mRNA sampling and abundance depths across genes. Counts are as in the optimization setup (d_{mk} have a pseudo-count of 1 and J_{mk} are scaled by a single factor). A slow outlier is an *(m,k)* with $\Delta_{mk} > T$ for some threshold T. Non-outliers are *(m,k)* with $-1 < \Delta_{mk} < 1$, excluding slow outliers.

Since there is a small uncertainty in the position of the active codon within ribosome-protected fragments of certain lengths, what we might see as a fast outlier (a position *(m,k)* where $\Delta_{mk} < -T$ and, for example, a wrongly-labeled count of 0) could actually have a fragment that was falsely associated with an adjacent slow position. The opposite is much less likely; an observed slow outlier has many more counts than expected, making it unlikely that so many fragments were wrongly attributed and belong instead to an adjacent fast outlier. For that reason, we compare slow outliers only to non-outliers.

When correlating features to outlier strength (Table E3), we call features significant only if they pass a stringent set of conditions: Pearson and Spearman correlations must have the same sign for all slow outlier thresholds (T = 0, 0.5, 1, 1.5, 2, 2.5) and be significant; the correlation when binned by codons must have at least 30 significant codons; the sign of the correlation must match the direction suggested by the comparison of means for slow versus non-outliers. When referring to significant features in Table E3, we cite the correlation for T = 0 since all thresholds are significant. For a more stringent set of outliers, we use T = 1 in analyses requiring a fixed T (Figure E6, E8, E10).

Accession numbers

Pending on GEO.

Acknowledgements

We thank members of the Koller and Weissman labs, especially Gene-Wei Li, for useful discussions. This research was supported by NSF Graduate Research Fellowship (to C.P.), NIH Broad Institute Grant #5410353-5500000575 (to D.K.), Howard Hughes Medical Institute (to J.S.W), and NIH Grant GM052347 (to E.M.P).

Author Contributions

D.K. and J.S.W. conceived of the study. C.P. and D.K. created the computational methods. S.R., N.T.I., and J.S.W. designed and conducted the profiling experiments. L.H. and E.M.P. designed and conducted the aminoacylation experiments. C.P. analyzed the data. C.P. and D.K. wrote the manuscript. All authors reviewed and commented on the manuscript.

References

Alexandrov A, Chernyakov I, Gu W, Hiley SL, Hughes TR, Grayhack EJ, Phizicky EM (2006) Rapid tRNA decay can result from lack of nonessential modifications. *Mol Cell* 21: 87-96

Andersson SG & Kurland CG (1990) Codon Preferences in Free-Living Microorganisms. *Microbiol Rev* 54(2): 198–210

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-Wide Analysis of mRNA Translation Profiles in Saccharomyces Cerevisiae. *Proc Natl Acad Sci* 100(7): 3889–94

Arava Y, Boad FE, Brown PO, Herschlag D (2005) Dissecting Eukaryotic Translation and Its Control by Ribosome Density Mapping. *Nucleic Acids Res* 33(8): 2421–32

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N (2013) Efficient Translation Initiation Dictates Codon Usage at Gene Start. *Mol Syst Biol* 9: 675

Bonekamp F, Dalboge H, Christensen T, Jensen KF (1989) Translation Rates of Individual Codons Are Not Correlated with tRNA Abundances or with Frequencies of Utilization in Escherichia Coli. *J Bacteriol* 171(11): 5812–16

Bulmer M (1991) The Selection-Mutation-Drift Theory of Synonymous Codon Usage.

Genetics 129(3): 897–907

Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O (2008) Codon Optimization Can Improve Expression of Human Genes in Escherichia Coli: A Multi-Gene Study. *Protein Expr Purif* 59(1): 94–102

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J Sci Comput* 16(5): 1190–1208

Charneski CA & Hurst LD (2013) Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biol* 11(3): e1001508

Charneski CA & Hurst LD (2014) Positive Charge Loading at Protein Termini Is Due to Membrane Protein Topology, Not a Translational Ramp. *Mol Biol Evol* 31(1): 70–84

Chen CA, Zhang H, Broitman SL, Reiche M, Farrell I, Cooperman BS, Goldman YE (2013) Dynamics of Translation by Single Ribosomes through mRNA Secondary Structures. *Nat Struct Mol Biol* 20(5): 582–88

Chu D, Barnes DJ, von der Haar T (2011) The Role of tRNA and Ribosome Competition in Coupling the Expression of Different mRNAs in Saccharomyces Cerevisiae. *Nucleic Acids Res* 39(15): 6705–14

Chu D & von der Haar T (2012) The Architecture of Eukaryotic Translation. *Nucleic Acids Res* 40(20): 10098–106

Curran J F & Yarus M (1989) Rates of Aminoacyl-tRNA Selection at 29 Sense Codons in Vivo. *J Mol Biol* 209(1): 65–77

Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM (2012) Functional Complexity and Regulation through RNA Dynamics. *Nature* 482(7385): 322–30

Ding Y, Shah P, Plotkin JB (2012) Weak 5'-mRNA Secondary Structures in Short Eukaryotic Genes. *Genome Biol Evol* 4(10): 1046–53

Dittmar KA, Mobley EM, Radek AJ, Pan T (2004) Exploring the Regulation of tRNA Distribution on the Genomic Scale. *J Mol Biol* 337(1): 31–47

Firczuk H, Kannambath S, Pahle J, Claydon A, Beynon R, Duncan J, Westerhoff H, Mendes P, McCarthy JEG (2013) An in Vivo Control Map for the Eukaryotic mRNA Translation Machinery. *Mol Syst Biol* 9: 635

Fredrick K & Ibba M (2010) How the Sequence of a Gene Can Tune Its Translation. *Cell* 141(2): 227–29

De Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M (2008) Comprehensive Mass-Spectrometry-Based Proteome Quantification of Haploid versus Diploid Yeast. *Nature* 455(7217): 1251–54

Dos Reis M, Savva R, Wernisch L (2004) Solving the Riddle of Codon Usage Preferences: A Test for Translational Selection. *Nucleic Acids Res* 32(17): 5036–44

Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B (2014) Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*: e03735

Gu W, Zhou T, Wilke CO (2010) A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. *PLoS Comp Biol* 6(2): e1000664

Gustafsson C, Govindarajan S, Minshull J (2004) Codon Bias and Heterologous Protein Expression. *Trends Biotechnol* 22(7): 346–53

Hamilton R, Watanabe CK, de Boer HA (1987) Compilation and Comparison of the Sequence Context around the AUG Startcodons in Saccharomyces Cerevisiae mRNAs. *Nucleic Acids Res* 15(8): 3581–93

Hense W, Anderson N, Hutter S, Stephan W, Parsch J, Carlini DB (2010) Experimentally Increased Codon Bias in the Drosophila Adh Gene Leads to an Increase in Larval, but Not Adult, Alcohol Dehydrogenase Activity. *Genetics* 184(2): 547–55

Hershberg R & Petrov DA (2008) Selection on Codon Bias. *Annu Rev Genetics* 42: 287–99

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 324(5924): 218–23

Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* 147(4): 789–802

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The Ribosome Profiling Strategy for Monitoring Translation in Vivo by Deep Sequencing of Ribosome-Protected mRNA Fragments. *Nat Protoc* 7(8): 1534–50

Keller, Thomas E, S David Mis, Kevin E Jia, and Claus O Wilke (2012) Reduced mRNA Secondary-Structure Stability near the Start Codon Indicates Functional Genes in Prokaryotes. *Genome Biol Evol* 4(2): 80–88

Kertesz MW, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467(7311): 103–7

Kochetov AV, Palyanov A, Titov II, Grigorovich D, Sarai A, Kolchanv NA (2007) AUG_hairpin: Prediction of a Downstream Secondary Structure Influencing the Recognition of a Translation Start Site. *BMC Bioinform* 8: 318

Kozak M (1981) Possible Role of Flanking Nucleotides in Recognition of the AUG Initiator Codon by Eukaryotic Ribosomes. *Nucleic Acids Res* 9(20): 5233–52

Kozak M (1990) Downstream Secondary Structure Facilitates Recognition of Initiator Codons by Eukaryotic Ribosomes. *Proc Natl Acad Sci* 87(21): 8301–5

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-Sequence Determinants of Gene Expression in Escherichia Coli. *Science* 324(5924): 255–58

Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, Preiss T, Bahler J (2007) A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol Cell* 26(1): 145–55

Lackner DH & Bahler J (2008) Translational control of gene expression from transcripts to transcriptomes. *Int Rev Cell Mol Biol* 271: 199–251

Lareau LF, Hite DH, Hogan GJ, Patrick OB (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife*:

e01257

Lavner Y & Kotlar D (2005) Codon Bias as a Factor in Regulating Expression via Translation Rate in the Human Genome. *Gene* 345(1): 127–38

Letzring DP, Dean KM, Grayhack EJ (2010) Control of Translation Efficiency in Yeast by Codon-Anticodon Interactions. *RNA* 16(12): 2516–28

Li G-W, Oh E, Weissman JS (2012) The Anti-Shine-Dalgarno Sequence Drives Translational Pausing and Codon Choice in Bacteria. *Nature* 484(7395): 538–41

Looman AC & Kuivenhoven JA (1993) Influence of the Three Nucleotides Upstream of the Initiation Codon on Expression of the Escherichia Coli lacZ Gene in Saccharomyces Cerevisiae. *Nucleic Acids Res* 21(18): 4268–71

Lowe TM & Eddy SR (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* 25(5): 955–64

Man O & Pilpel Y (2007) Differential Translation Efficiency of Orthologous Genes Is Involved in Phenotypic Divergence of Yeast Species. *Nat Genet* 39(3): 415–21

Mitarai N & Pedersen S (2013) Control of Ribosome Traffic by Position-Dependent Choice of Synonymous Codons. *Phys Biol* 10(5): 056011–7

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* 320(5881): 1344–49

Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS (2006) Single-Cell Proteomic Analysis of S. Cerevisiae Reveals the Architecture of Biological Noise. *Nature* 441(7095): 840–46

Plotkin JB & Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 12(1): 32–42

Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J (2012) Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. *PLoS Genet* 8(3): e1002603

Reuveni S, Meilijson I, Kupiec M, Ruppin E, Tuller T (2011) Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model. *PLoS Comp Biol* 7(9): e1002127

Robbins-Pianka A, Rice MD, Weir MP (2010) The mRNA Landscape at Yeast Translation Initiation Sites. *Bioinform* 26(21): 2651–55

Rouskin S, Zubrady M, Washietl S, Kellis M, Weissman JW (2013) Genome-Wide Probing of RNA Structure Reveals Active Unfolding of mRNA Structures in Vivo.

Nature 505(7485): 701–5

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB (2013) Rate-Limiting Steps in Yeast Protein Translation. *Cell* 153(7): 1589–1601

Sorensen MA, Kurland CG, Pedersen S (1989) Codon Usage Determines Translation Rate in Escherichia Coli. *J Mol Biol* 207(2): 365–77

Sorensen MA & Pedersen S (1991) Absolute in vivo translation rates of individual codons in Escherichia coli: The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol* 222(2): 265–80

Stadler M & Fire A (2011) Wobble Base-Pairing Slows in Vivo Translation Elongation in Metazoans. *RNA* 17(12): 2063–73

Supek F & Smuc T (2010) On Relevance of Codon Usage to Expression of Synthetic and Natural Genes in Escherichia Coli. *Genetics* 185(3): 1129–34

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y (2010a) An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* 141(2): 344–54

Tuller T, Waldman YY, Kupiec M, Ruppin E (2010b) Translation Efficiency Is

Determined by Both Codon Bias and Folding Energy. *Proc Natl Acad Sci* 107(8): 3645–50

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M (2011) Composite Effects of Gene Determinants on the Translation Speed and Density of Ribosomes. *Genome Biol* 12(11): R110

Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD (2010) Real-Time tRNA Transit on Single Translating Ribosomes at Codon Resolution. *Nature* 464(7291): 1012–17

Varenne S, Buc J, Lloubes R, Lazdunski C (1984) Translation Is a Non-Uniform Process. Effect of tRNA Availability on the Rate of Elongation of Nascent Polypeptide Chains. *J Mol Biol* 180(3): 549–76

Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson (2009) Design Parameters to Control Synthetic Gene Expression in Escherichia Coli. *PloS One* 4(9): e7002

Wen J-D, Lancaster L, Hodges C, Zeri A-C, Yoshimura SH, Noller HF, Bustamante C, Tinoco Jr I (2008) Following Translation by Single Ribosomes One Codon at a Time. *Nature* 452(7187): 598–603

Woolstenhulme CJ, Parajuli S, Healey DW, Valverde DP, Petersen EN, Starosta AL, Guydosh NR, Johnson WE, Wilson DN, Buskirk AR (2013) Nascent Peptides That Block Protein Synthesis in Bacteria. *Proc Natl Acad Sci* 110(10): E878–87

Wu X, Jornvall H, Berndt KD, Oppermann U (2004) Codon Optimization Reveals Critical Factors for High Level Expression of Two Rare Codon Genes in Escherichia Coli: RNA Stability and Secondary Structure but Not tRNA Abundance. *Biochem Biophys Res Commun* 313(1): 89–96

Yun DF, Laz TM, Clements JM, Sherman F (1996) mRNA Sequences Influencing Translation and the Selection of AUG Initiator Codons in the Yeast Saccharomyces Cerevisiae. *Mol Microbiol* 19(6): 1225–39

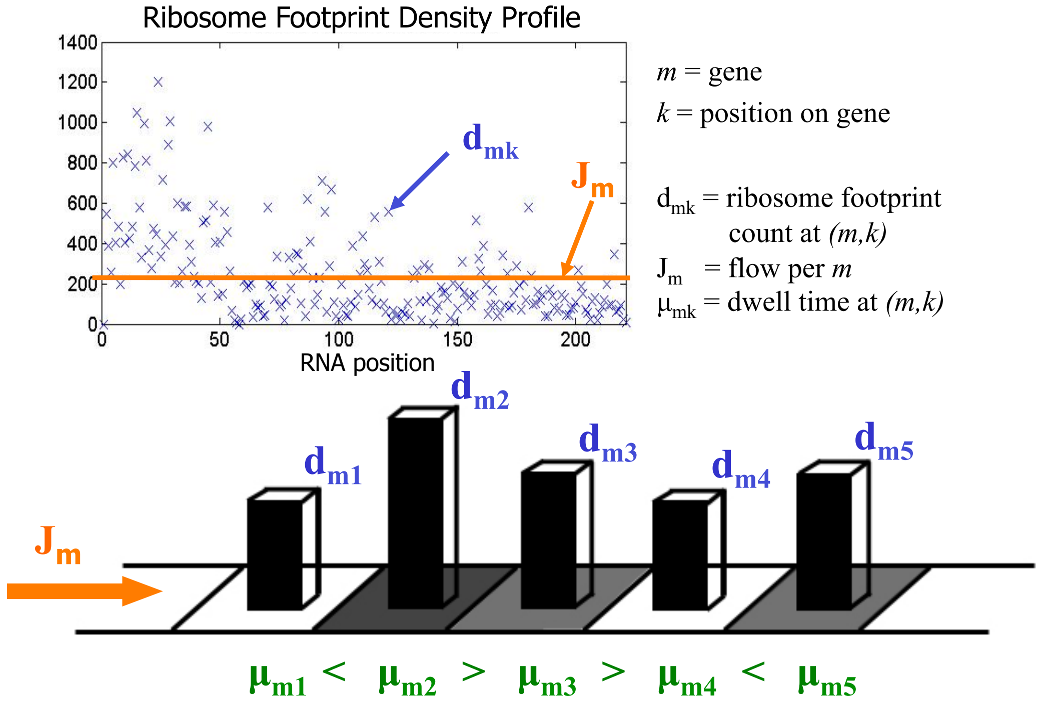
Zhang G, Hubalewska M, Ignatova Z (2009) Transient Ribosomal Attenuation Coordinates Protein Synthesis and Co-Translational Folding. *Nat Struct Mol Biol* 16(3): 274–80

Zhang S, Goldman E, Zubay G (1994) Clustering of Low Usage Codons and Ribosome Movement. *J Theor Biol* 170: 339-354

Zhou T & Wilke CO (2011) Reduced Stability of mRNA Secondary Structure near the Translation-Initiation Site in dsDNA Viruses. *BMC Evol Biol* 11: 59

Zur H & Tuller T (2012) Strong Association between mRNA Folding Strength and Protein Abundance in S. Cerevisiae. *EMBO Rep* 13(3): 272–77

Figure1



m = gene

k = position on gene

d_{mk} = ribosome footprint count at *(m,k)*

J_m = flow per *m*

μ_{mk} = dwell time at *(m,k)*

count at position = flow * dwell at position

$$d_{mk} = J_m \mu_{mk}$$

Figure2

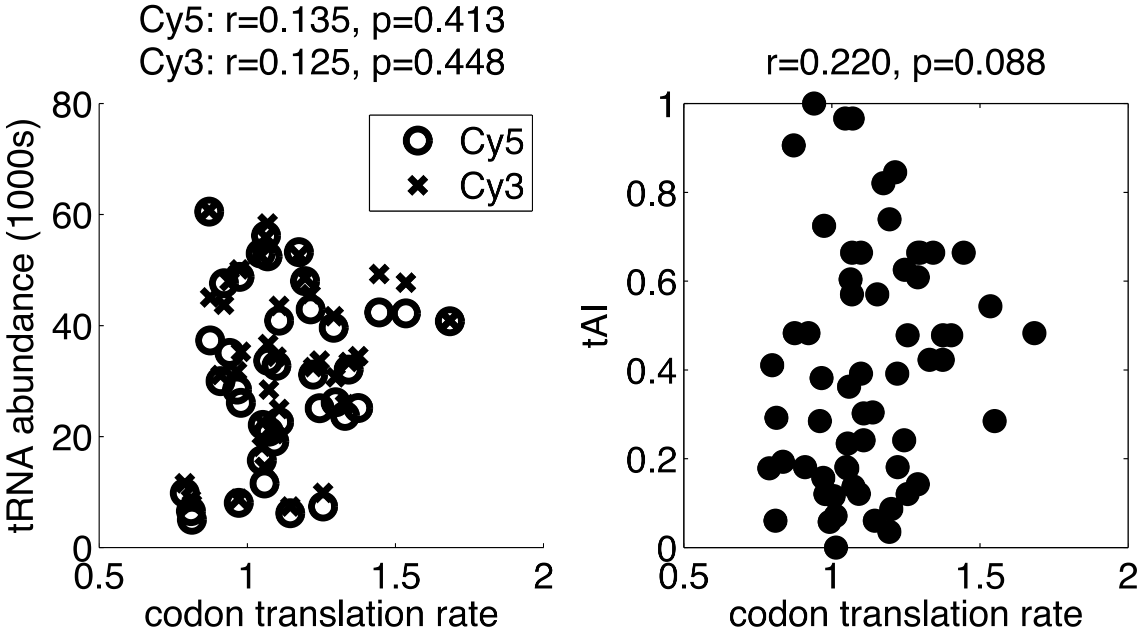

Cy5: r=0.135, p=0.413
Cy3: r=0.125, p=0.448
tRNA abundance (1000s)
80
60
40
20
0
0.5
1
1.5
2
codon translation rate
Cy5
Cy3
r=0.220, p=0.088
tAI
1
0.8
0.6
0.4
0.2
0
0.5
1
1.5
2
codon translation rate

Figure3

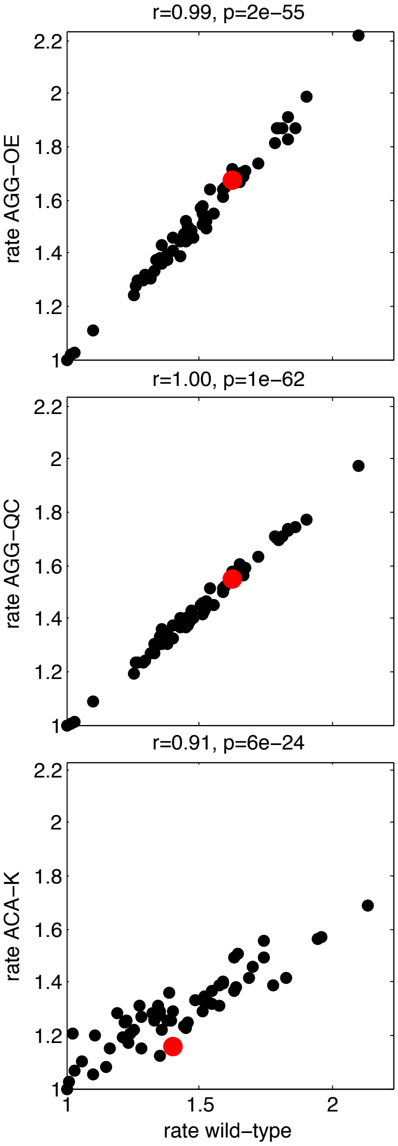
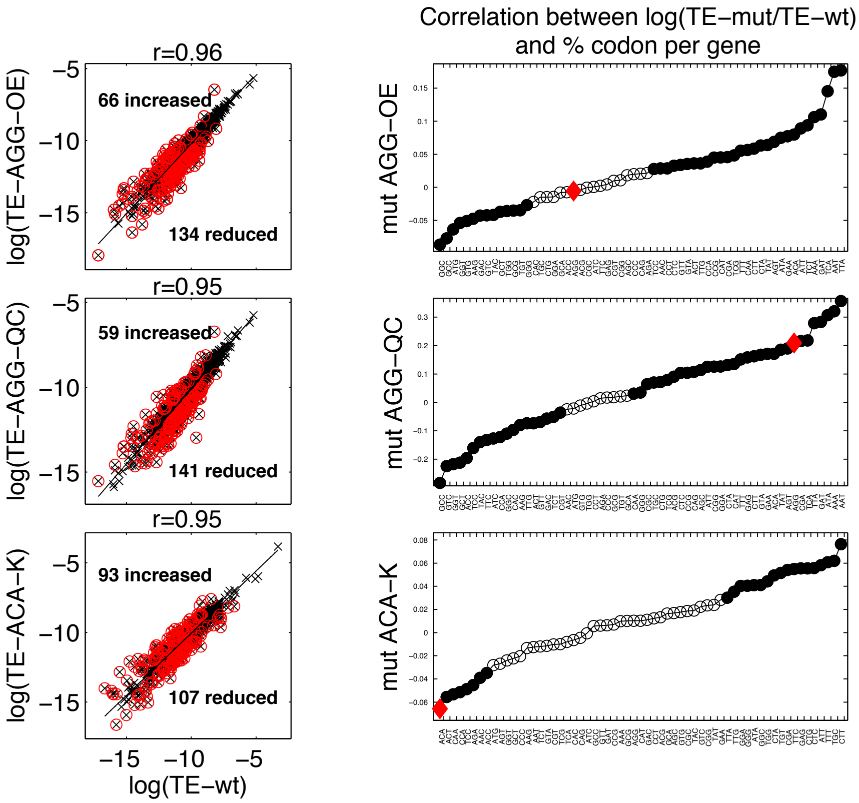

r=0.99, p=2e−55
rate AGG−OE
2.2
2
1.8
1.6
1.4
1.2
1
r=1.00, p=1e−62
rate AGG−QC
2.2
2
1.8
1.6
1.4
1.2
1
r=0.91, p=6e−24
rate ACA−K
2.2
2
1.8
1.6
1.4
1.2
1
1
1.5
2
rate wild−type

Figure 4



Correlation between log(TE−mut/TE−wt)
and % codon per gene
r=0.96
log(TE−AGG−OE)
66 increased
134 reduced
−5
−10
−15
r=0.95
log(TE−AGG−QC)
59 increased
141 reduced
−5
−10
−15
r=0.95
log(TE−ACA−K)
93 increased
107 reduced
−5
−10
−15
−15
−10
−5
log(TE−wt)
mut AGG−OE
0.15
0.1
0.05
0
−0.05
GGC GCC ATG GGT GTG AAG GAC GTC TAC GCT TGG GCG TGT GGG CAC TGC CTG GGA GCA ACC AGG ACG CGC ATC TTC GAG CGT CGG AGC CCC CAG AGA TCC AAC CCT CTC GTT GTA ACT TTG CCA CCG CAT CGA TCG TTT CAA CTT CTA TAT AGT ATA GAA ACA ATT TCT AAA GAT TCA AAT TTA
mut AGG−QC
0.3
0.2
0.1
0
−0.1
−0.2
GCC GTC GGT GCT ACC TCC TAC TTC ATC CCA GGC CAC AAG TTG ACT GTT GAC TCT CGT AAC ATG GTG TGG CCT AGA CCC GCG TGT GCA CAA GGG CGC TGC CTG TCG ACG CTC CCG CAG AGC ATT CGG GGA CTA CAT TTT GAG CTT GTA GAA ACA TAT AGT AGG CGA TCA TTA GAT ATA AAA AAT
mut ACA−K
0.08
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
ACA ACT CAA CCA TCC AGA AAC ACC ATG AGT GGT GCT CCC AAG AAT TCT GTA CGT TCG TCA CAC CAG ATC GCC GTT GAT CCG AAA GCG AGG CAT GAC CCT ACG GCA AGC GTG CGC TAC GTC CGG TAT GAA TTA TTG GGA GGG ATA GGC TGG CTA TGT CGA TTC GAG CTG CTC ATT TTT TGC CTT

Figure 5

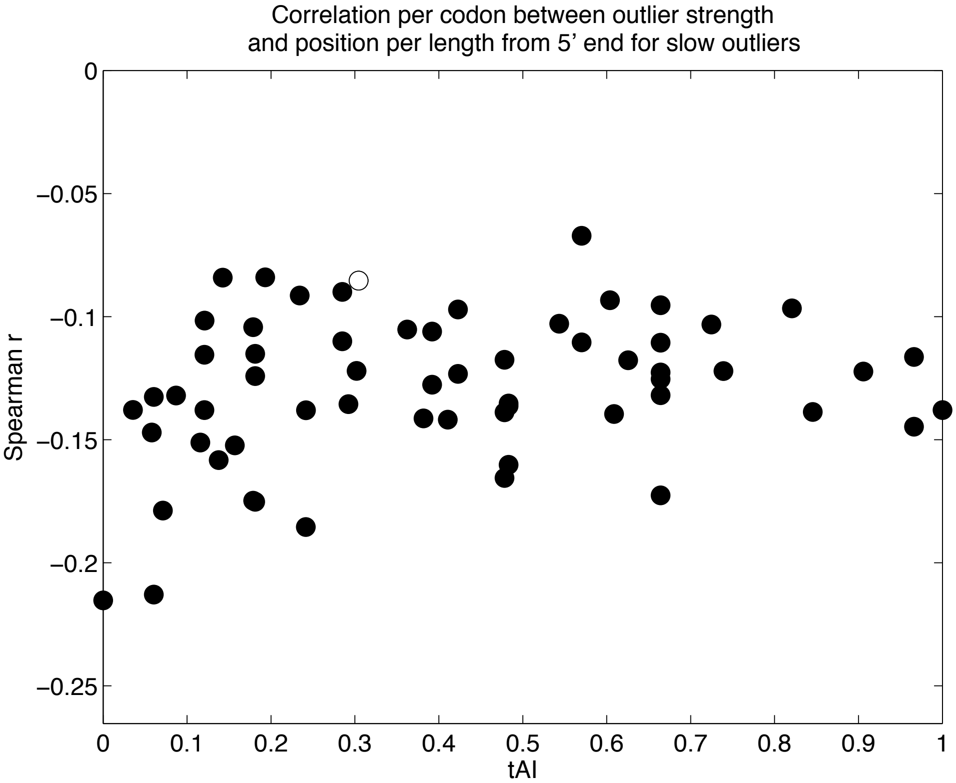

Correlation per codon between outlier strength
and position per length from 5' end for slow outliers
0
−0.05
−0.1
−0.15
−0.2
−0.25
Spearman r
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
tAI

Figure 6

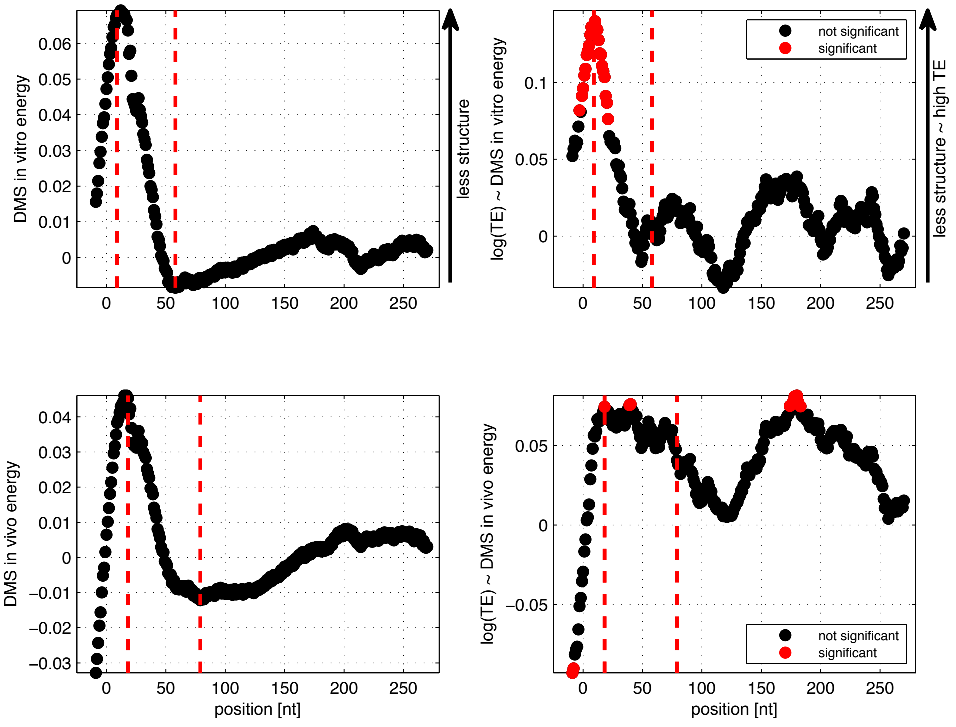

DMS in vitro energy
0.06
0.05
0.04
0.03
0.02
0.01
0
less structure
log(TE) ~ DMS in vitro energy
0.1
0.05
0
● not significant
● significant
less structure ~ high TE
0
50
100
150
200
250
DMS in vivo energy
0.04
0.03
0.02
0.01
0
−0.01
−0.02
−0.03
position [nt]
log(TE) ~ DMS in vivo energy
0.05
0
−0.05
● not significant
● significant
position [nt]

Figure 7

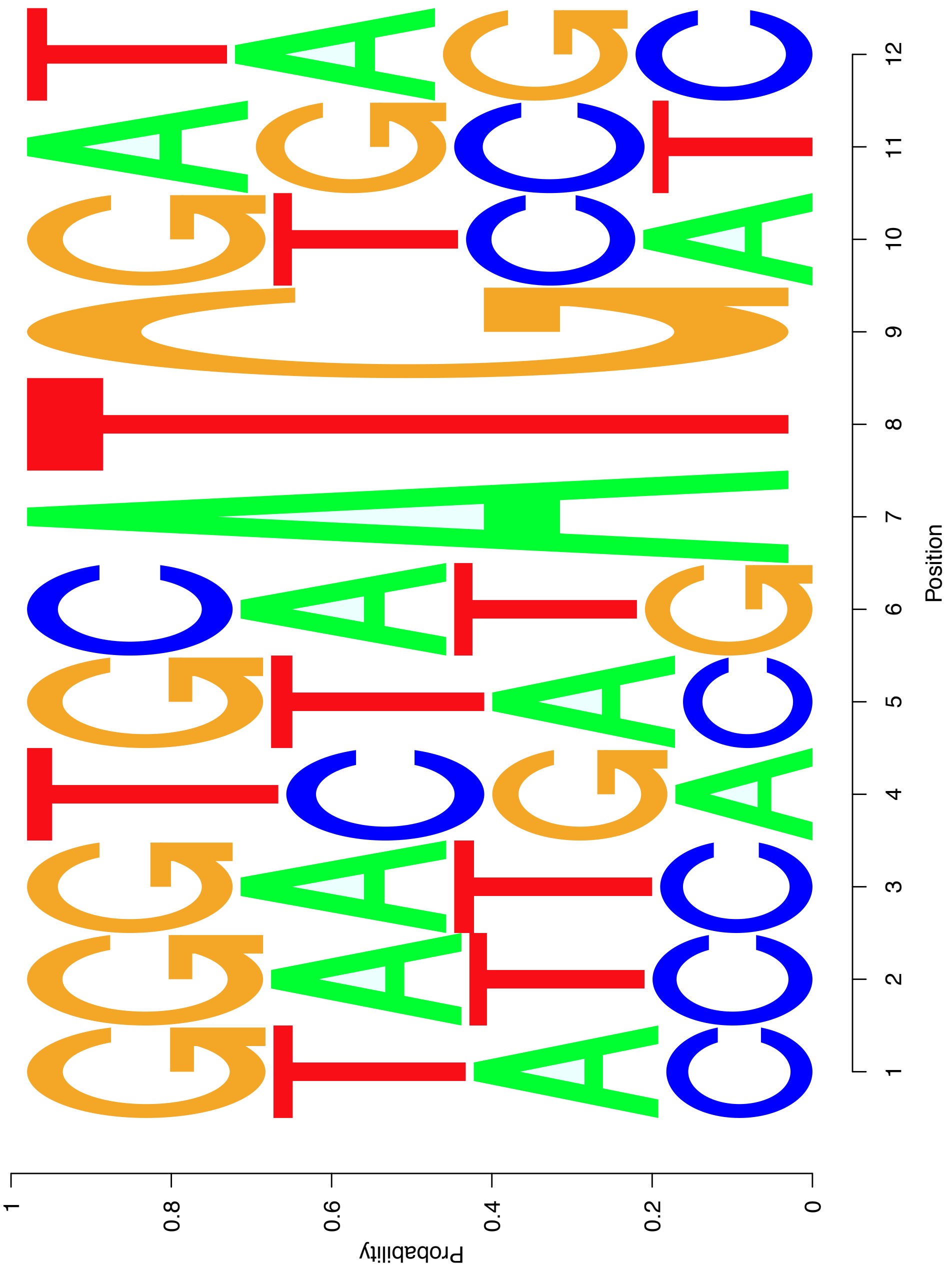


Table E1

Counts of tRNA in RPM (number of reads per million) in ACA-K and wild-type. The threonine tRNA recognizing the ACA codon (highlighted) is reduced to 1/3 of the wild-type level.

tRNA gene name (anticodon)	RPM (ACA-K)	RPM (wt)	RPM (ACA-K) / RPM (wt)
tK(UUU)D	78	80	0.98
tY(GUA)F1	19	16	1.19
tM(CAU)C	11	11	1.00
tD(GUC)B	218	251	0.87
tE(UUC)B	582	428	1.36
tN(GUU)C	225	166	1.36
tS(UGA)P	122	148	0.82
tP(AGG)N	24	27	0.89
tC(GCA)B	82	58	1.41
tQ(UUG)B	103	106	0.97
tW(CCA)G1	35	44	0.80
tG(UCC)O	143	96	1.49
tT(UGU)G1	25	75	0.33
tR(UCU)E	138	172	0.80
tA(AGC)D	72	42	1.71
tT(CGU)K	9	9	1.00
tV(AAC)E1	129	82	1.57
tQ(CUG)M	166	138	1.20
tA(UGC)Q	3	3	1.00
tL(UAA)J	72	81	0.89
tI(AAU)B	98	46	2.13
tH(GUG)E1	328	266	1.23
tT(AGU)B	152	141	1.08
tF(GAA)B	124	115	1.08
tK(CUU)C	1328	1914	0.69

Table E2
Eight categories of potential correlates to outlier strength.

Category	Features
Position	Distance from 5' end (pos) Distance from 5' end per length (pos-per-len) Distance from 3' end (pos-from-end)
Structure in 25nt window 15nt downstream of active (unless indicated, features are derived from computationally-predicted structure)	Minimum free energy (energy-down) Experimentally-derived in vitro energy (vitroDMS-energy-down) (Rouskin et al, 2013) Experimentally-derived in vivo energy (vivoDMS-energy-down) (Rouskin et al, 2013) Experimentally-derived in vitro inverse-energy (PARS-invenergy-down) (Kertesz et al, 2010) Number of hairpins (hairpins-down) Number of internal loops (internal-down) Number of multi-loops (multi-down) Number of stems (stems-down15) Number of GC pairs in stems (stemsGC-down15) Number of stems 12nt downstream (stems-down12) Number of stems 9nt downstream (stems-down9)
Protein folding	Active site is inside a protein domain (is-in-domain) End of protein domain is 30 codons upstream of active (is-end-domain-up-30)
Wobble bases at P-site	Is wobble base (is-wobble)
Reuse of tRNAs	Distance from same codon upstream (dist-prev-codon) Distance from codon with iso-accepting tRNA upstream (dist-prev-trna) Is same codon in 10-codon window upstream (is-prev-codon-close) Is codon with iso-accepting tRNA in 10-codon window upstream (is-prev-trna-close)
KL divergence to RNA binding motifs (Brown et al. 2009) in 3-codon window 5 codons downstream of active	KL divergence over motifs and positions combined via mean (rbp-mean) KL divergence over motifs and positions combined via min (rbp-min)
Nascent peptide	Charge of active codon (charge) Mean charge in 10-codon window ending upstream of active (cluster-charge-up-1) Fraction of Arg or Lys in 10-codon window ending upstream of active (cluster-ArgLys-up-1) Fraction of Pro in the P and E sites (pair-Pro-up) Fraction of Pro in two codons downstream of active (pair-Pro-down)
Global	Length (len) Abundance (abund)

Table E3. Correlation between outlier strength and features. Significant ones (see Materials and **Methods**) are highlighted. The first two rows per feature

Feature	Slow Outlier Threshold (T)	Correlation r-value	Correlation p-value	Mean of Slow Outliers	Std of Slow Outliers	Mean of Non-Outliers	Std of Non-Outliers
Position							
pos	0	-0.038	0	359.626	377.546	421.248	411.324
	2.5	-0.082	0	306.774	371.353	395.611	396.166
	0	-0.044	0	359.626	377.546	421.248	411.324
	2.5	-0.112	0	306.774	371.353	395.611	396.166
pos-per-len	0	-0.136	0	0.488	0.289	0.517	0.287
	2.5	-0.119	0	0.377	0.294	0.513	0.286
	0	-0.127	0	0.488	0.289	0.517	0.287
	2.5	-0.119	0	0.377	0.294	0.513	0.286
pos-from-end	0	0.118	0	385.626	396.552	391.507	394.579
	2.5	0.064	0	511.86	482.595	377.104	385.906
	0	0.136	0	385.626	396.552	391.507	394.579
	2.5	0.055	0	511.86	482.595	377.104	385.906
Structural Features							
energy-down	0	0	0.659	-2.66	1.752	-2.621	1.728
	2.5	-0.016	0	-2.668	1.759	-2.641	1.739
	0	0.006	0	-2.66	1.752	-2.621	1.728
	2.5	-0.013	0.001	-2.668	1.759	-2.641	1.739
vitroDMS-energy-down	0	-0.003	0.009	0.489	0.156	0.49	0.156
	2.5	-0.006	0.13	0.486	0.156	0.489	0.156
	0	0.004	0	0.489	0.156	0.49	0.156
	2.5	-0.007	0.088	0.486	0.156	0.489	0.156
vivoDMS-energy-down	0	-0.017	0	0.512	0.176	0.51	0.176
	2.5	-0.008	0.043	0.502	0.175	0.512	0.176
	0	-0.012	0	0.512	0.176	0.51	0.176
	2.5	-0.015	0	0.502	0.175	0.512	0.176
PARS-invenergy-down	0	-0.015	0	0.327	0.555	0.318	0.543
	2.5	0.036	0	0.32	0.538	0.326	0.554
	0	-0.029	0	0.327	0.555	0.318	0.543
	2.5	0.019	0	0.32	0.538	0.326	0.554
hairpins-down	0	0.021	0	5.898	5.004	5.796	4.981
	2.5	0.02	0	6.232	5.055	5.817	4.985
	0	0.018	0	5.898	5.004	5.796	4.981
	2.5	0.015	0	6.232	5.055	5.817	4.985
internal-down	0	0.014	0	1.128	1.331	1.108	1.324
	2.5	0.014	0	1.187	1.344	1.113	1.326
	0	0.013	0	1.128	1.331	1.108	1.324
	2.5	0.011	0.004	1.187	1.344	1.113	1.326
multi-down	0	0.022	0	0.181	0.433	0.175	0.425
	2.5	0.022	0	0.209	0.463	0.176	0.426
	0	0.019	0	0.181	0.433	0.175	0.425
	2.5	0.014	0	0.209	0.463	0.176	0.426
stems-down15	0	0.021	0	5.898	5.004	5.796	4.981
	2.5	0.02	0	6.232	5.055	5.817	4.985
	0	0.018	0	5.898	5.004	5.796	4.981
	2.5	0.015	0	6.232	5.055	5.817	4.985
stemsGC-down15	0	0.019	0	2.321	2.286	2.276	2.267
	2.5	0.016	0	2.456	2.316	2.286	2.273
	0	0.018	0	2.321	2.286	2.276	2.267
	2.5	0.014	0.001	2.456	2.316	2.286	2.273
stems-down12	0	0.023	0	5.915	5.008	5.791	4.977
	2.5	0.02	0	6.262	5.054	5.821	4.986
	0	0.02	0	5.915	5.008	5.791	4.977
	2.5	0.017	0	6.262	5.054	5.821	4.986
stems-down9	0	0.024	0	5.933	5.013	5.786	4.972
	2.5	0.02	0	6.303	5.07	5.824	4.984
	0	0.021	0	5.933	5.013	5.786	4.972
	2.5	0.017	0	6.303	5.07	5.824	4.984
Protein Folding							
is-in-domain	0	-0.012	0	0.721	0.449	0.728	0.445
	2.5	-0.055	0	0.702	0.458	0.725	0.447
	0	-0.003	0.001	0.721	0.449	0.728	0.445
	2.5	-0.04	0	0.702	0.458	0.725	0.447
is-end-domain-up-30	0	-0.002	0.035	0.004	0.063	0.004	0.064
	2.5	-0.002	0.484	0.003	0.058	0.004	0.063
	0	-0.001	0.186	0.004	0.063	0.004	0.064
	2.5	-0.001	0.872	0.003	0.058	0.004	0.063
Wobble Codons							
is-wobble	0	-0.032	0	0.435	0.496	0.466	0.499
	2.5	-0.017	0	0.395	0.489	0.456	0.498
	0	-0.036	0	0.435	0.496	0.466	0.499
	2.5	-0.01	0.005	0.395	0.489	0.456	0.498
Reuse of tRNAs							
dist-prev-codon	0	-0.047	0	46.003	62.322	44.91	59.684
	2.5	-0.025	0	39.967	53.054	46.051	61.81
	0	-0.052	0	46.003	62.322	44.91	59.684
	2.5	-0.019	0	39.967	53.054	46.051	61.81
dist-prev-trna	0	-0.036	0	37.066	49.947	36.529	48.562
	2.5	-0.022	0	33.118	43.671	37.167	49.809
	0	-0.034	0	37.066	49.947	36.529	48.562
	2.5	-0.017	0	33.118	43.671	37.167	49.809
is-prev-codon-close	0	0.021	0	0.255	0.436	0.259	0.438
	2.5	0.025	0	0.276	0.447	0.255	0.436
	0	0.024	0	0.255	0.436	0.259	0.438
	2.5	0.016	0	0.276	0.447	0.255	0.436
is-prev-trna-close	0	0.015	0	0.295	0.456	0.299	0.458
	2.5	0.025	0	0.311	0.463	0.296	0.456
	0	0.016	0	0.295	0.456	0.299	0.458
	2.5	0.016	0	0.311	0.463	0.296	0.456
Downstream Motifs							
rbp-mean	0	0.001	0.254	11.624	9.696	11.695	9.694

Table E4

Performance of TE regression model (see **Materials and Methods**). Error (should be low) and correlation (should be high) between predicted and actual TE is measured on 100 random test sets of genes not used during model training. Performance drops in a null model learned on randomized TE labels (last column). Performance also drops when using the original Kozak motif (middle column). Error on the training set is included to show that our model generalizes to genes not used in training (it is close to test set error).

	Regression		**Regression (with original Kozak)**		**Null Model**	
	Mean	***Std***	***Mean***	***Std***	***Mean***	***Std***
Error	0.7458	0.059	0.8339	0.0573	0.9602	0.0614
Error (Train)	0.7404	0.0073	0.8392	0.0071	0.9558	0.0076
Spearman r	0.6722	0.0299	0.5265	0.0328	-0.0282	0.0397
Spearman p	0	0	0	0	0.5156	0.2685
Pearson r	0.6263	0.0337	0.5173	0.0373	-0.0098	0.0413
Pearson p	0	0	0	0	0.5569	0.28

Table E5
Summary of main results for model variations. The first five columns are models with different constants for the second term in the objective function and the last column is a model without μ_m^c parameters (see **Materials and Methods**). Rows 1-3 represent correlation between our parameters in our model and in the model variation. Rows 4-6 represent correlation between codon translation rates in model variations and codon bias measures. Rows 7-8 represent correlation between protein synthesis rates in model variation and protein abundance measures. Results are similar to the ones reported for the model used throughout the paper (const = 100).

Result		**const = 1**	**const = 10**	**const = 1000**	**const = 10000**	**const = 100000**	**No μ_m^c**
μ^c (const=100)	r	1.000	1.000	1.000	1.000	1.000	1.000
	p	1E-172	9E-176	5E-133	1E-98	1E-90	5E-92
μ_m^c (const=100)	r	1.000	1.000	1.000	0.981	0.832	NA
	p	0	0	0	0	0	NA
J_m (const=100)	r	1.000	1.000	1.000	1.000	0.999	0.995
	p	0	0	0	0	0	0
tAI	r	0.221	0.221	0.221	0.222	0.224	0.216
	p	0.088	0.088	0.087	0.086	0.085	0.095
tRNA abund (Cy5)	r	0.135	0.135	0.131	0.134	0.134	0.128
	p	0.413	0.413	0.424	0.415	0.415	0.436
tRNA abund (Cy3)	r	0.135	0.135	0.131	0.134	0.134	0.128
	p	0.448	0.448	0.459	0.448	0.448	0.471
PA (Newman et al)	r	0.7875	0.7875	0.7876	0.7882	0.7889	0.7838
PA (de Godoy et al)	r	0.6822	0.6822	0.6822	0.6818	0.6787	0.6701

Figure E1

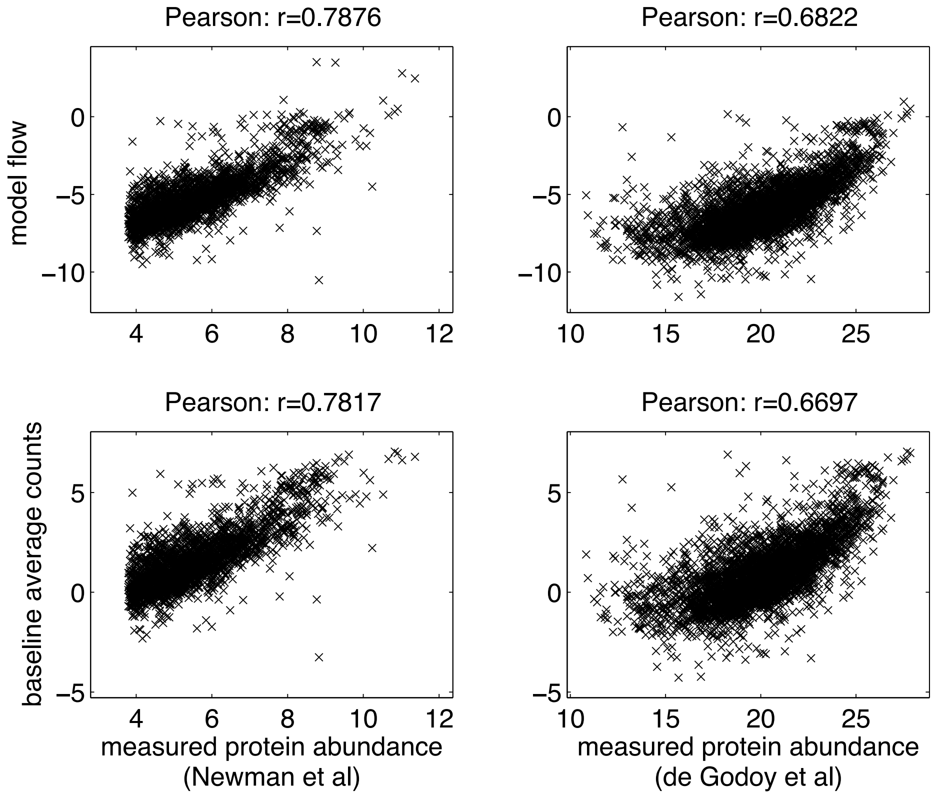

Pearson: r=0.7876
Pearson: r=0.6822
model flow
0
−5
−10
4
6
8
10
12
10
15
20
25
Pearson: r=0.7817
Pearson: r=0.6697
baseline average counts
5
0
−5
measured protein abundance
(Newman et al)
measured protein abundance
(de Godoy et al)

Figure E2

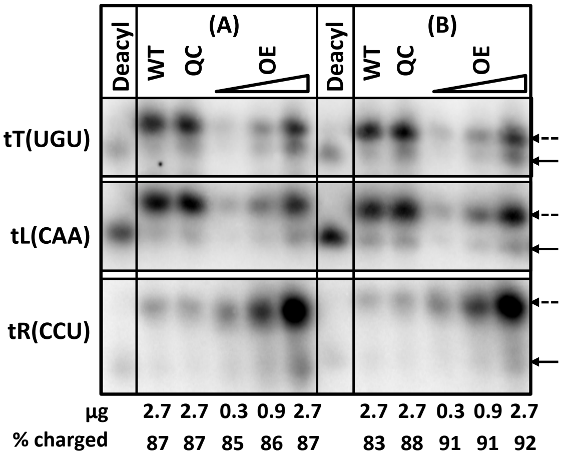

Deacyl
(A)
WT
QC
OE
Deacyl
(B)
WT
QC
OE
tT(UGU)
tL(CAA)
tR(CCU)
µg 2.7 2.7 0.3 0.9 2.7 2.7 2.7 0.3 0.9 2.7
% charged 87 87 85 86 87 83 88 91 91 92

Figure E3

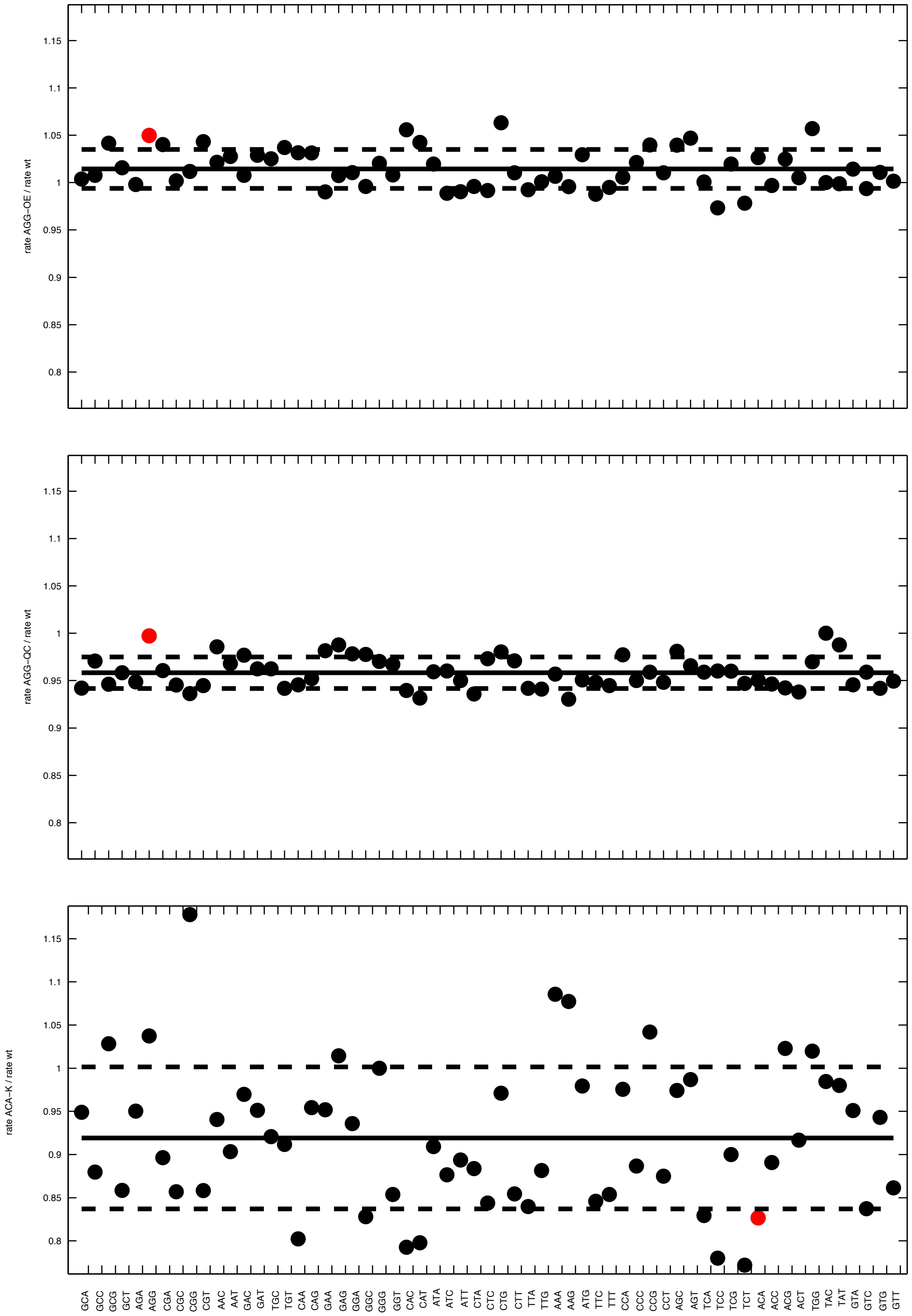

rate AGG–OE / rate wt
rate AGG–OC / rate wt
rate ACA–K / rate wt
1.15
1.1
1.05
1
0.95
0.9
0.85
0.8
GCA GCC GCG GCT AGA AGG CGA CGC CGG CGT AAC AAT GAC GAT TGC TGT CAA CAG GAA GAG GGA GGC GGG GGT CAC CAT ATA ATC ATT CTA CTC CTG CTT TTA TTG AAA AAG ATG TTC TTT CCA CCC CCG CCT AGC AGT TCA TCC TCG TCT ACA ACC ACG ACT TGG TAC TAT GTA GTC GTG GTT

Figure E4

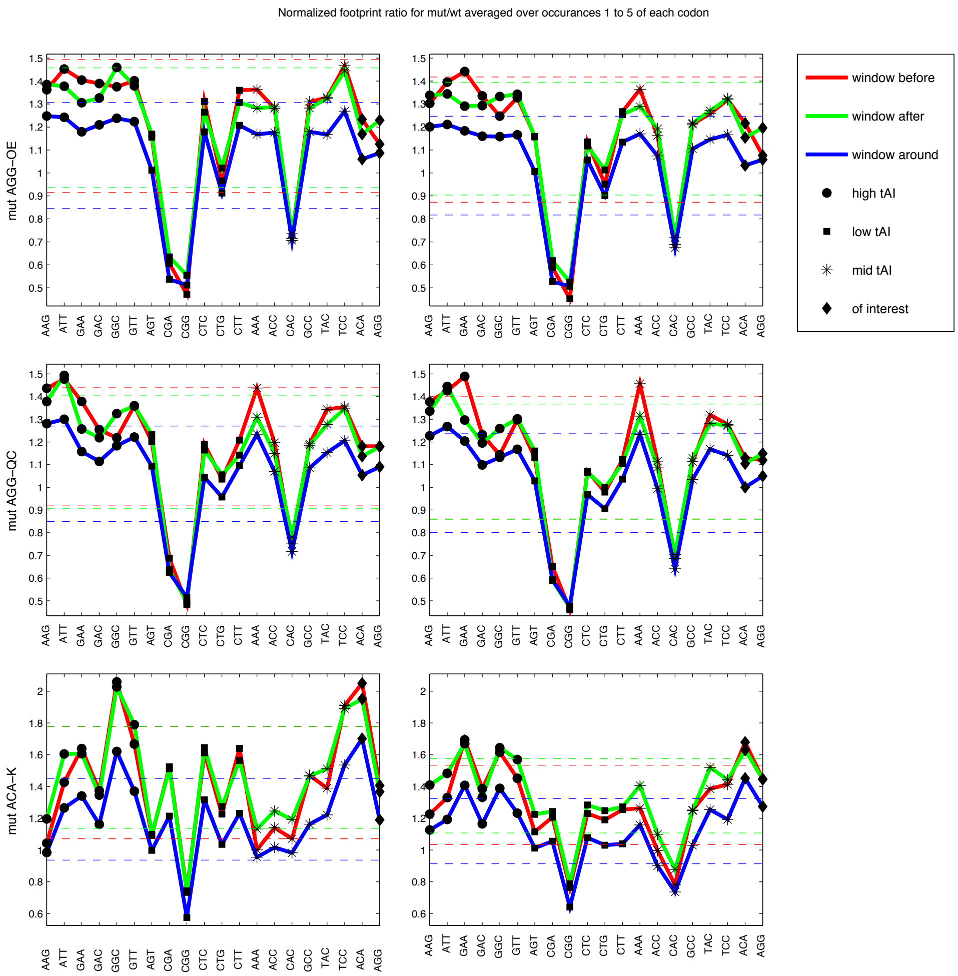

Normalized footprint ratio for mut/wt averaged over occurances 1 to 5 of each codon
mut AGG−OE
mut AGG−QC
mut ACA−K
window before
window after
window around
high tAI
low tAI
mid tAI
of interest
AAG ATT GAA GAC GGC GTT AGT CGA CGG CTC CTG CTT AAA ACC CAC GCC TAC TCC ACA AGG
0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5
0.6 0.8 1 1.2 1.4 1.6 1.8 2

Figure E5

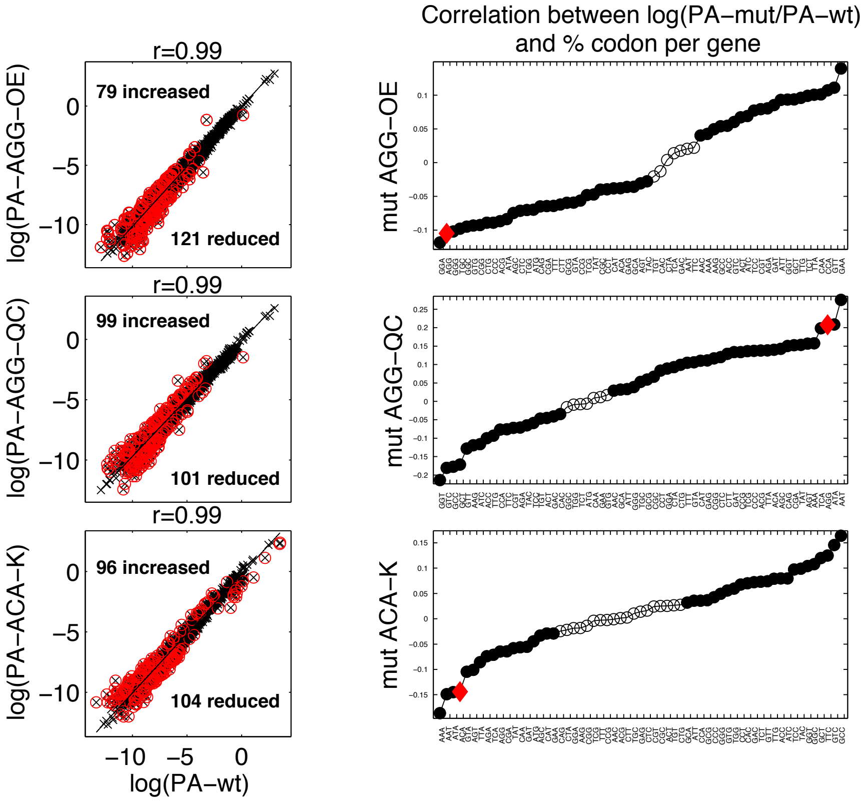
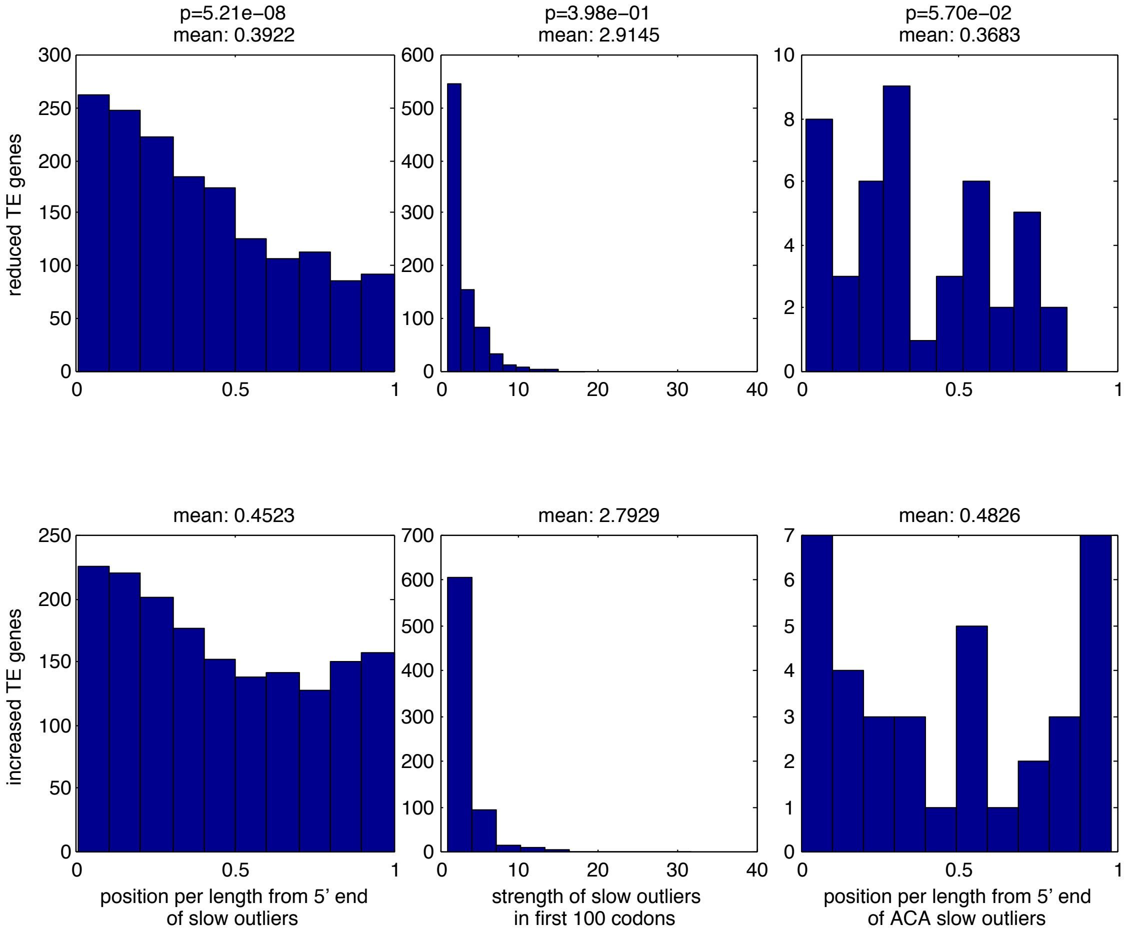

Correlation between log(PA−mut/PA−wt)
and % codon per gene
r=0.99
log(PA−AGG−OE)
79 increased
121 reduced
0
−5
−10
r=0.99
log(PA−AGG−QC)
99 increased
101 reduced
0
−5
−10
r=0.99
log(PA−ACA−K)
96 increased
104 reduced
0
−5
−10
−10
−5
0
log(PA−wt)
mut AGG−OE
0.1
0.05
0
−0.05
−0.1
GGA AGG GGG TGC GGC GTG CGG CTG CCC ACG ATA AGC CTC TGG ATG CAG CGA TTT CTT GCG GTA CCG TCG TAT CGC CCT CAT ACA GAG GCA AGT TAC TGT CAC CTA TCA GAC AAT TTC AAC AAA AAG GCC ACC GTC ACT ATC TCC CGT AGA GAT ATT GGT GCT TTG TCT TTA CAA CCA GTT GAA
mut AGG−QC
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
GGT GTC GCC GCT GTT AAG ATC ACC TTG CCA TTC CGT AGA TAC TCC TGT ACT GAC CAC GGC TGG TCT ATG CAA GAA GTG AAC GCA ATT GGG TGC GCG CGC CCT GGA CTA CTG TTT GTA CAT GAG CGG CTC CTT GAT CCG TCG CCC ACG TTA ACA AGC CAG CGA TAT AGT AAA TCA AGG ATA AAT
mut ACA−K
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
AAA AAT ATA ACA GTA AGT TTA AGA TCA AGG CGA TAT CAA GAT ATG AGC CAT GAA CAG CTA GGA AAG CGG TCG TTT CCG AAC ACG CTT TGC GAG CTC CGT CGC ACT TGT CTG GCA ATT CCA GCG CCC GGG GTG TGG CCT CAC GAC TCT GTT TTG ACC ATC TCC TAC GGT GGC GCT TTC GTC GCC

Figure E6



p=5.21e−08
mean: 0.3922
p=3.98e−01
mean: 2.9145
p=5.70e−02
mean: 0.3683
reduced TE genes
mean: 0.4523
mean: 2.7929
mean: 0.4826
increased TE genes
position per length from 5’ end of slow outliers
strength of slow outliers in first 100 codons
position per length from 5’ end of ACA slow outliers

Figure E7

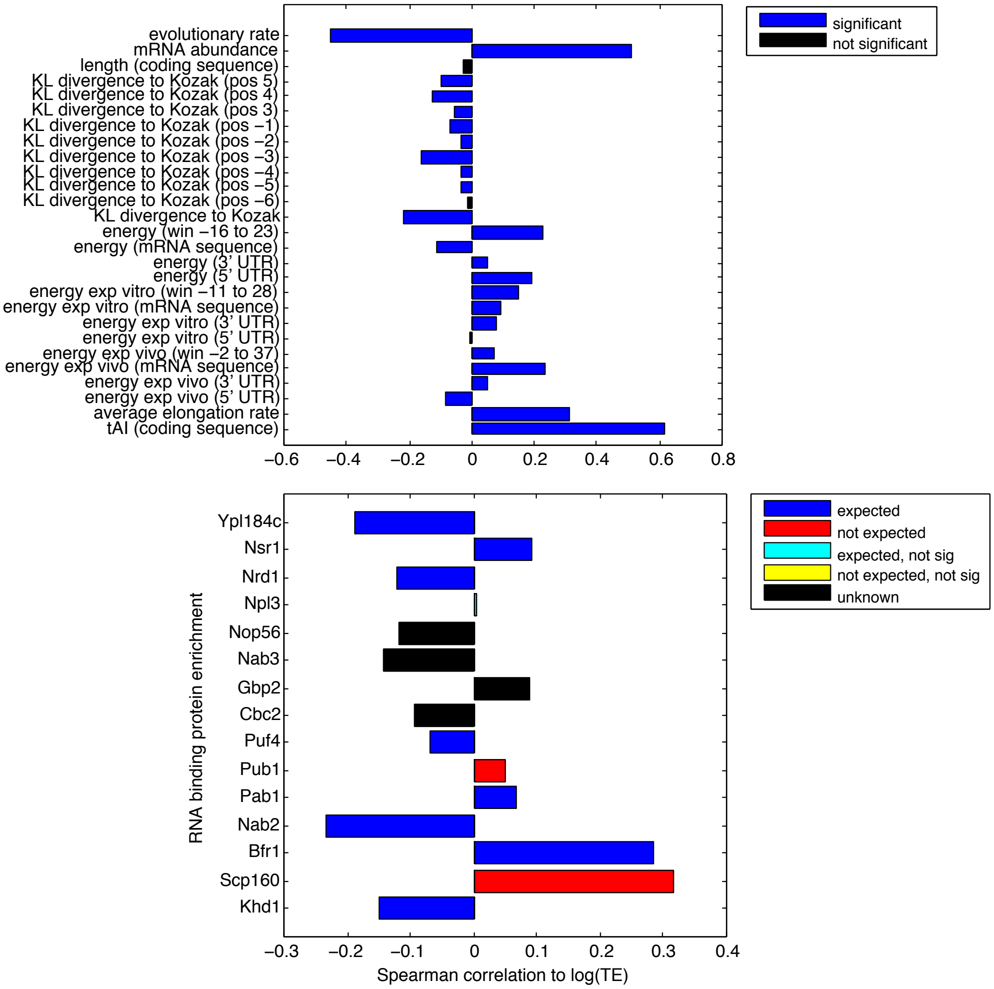

evolutionary rate
mRNA abundance
length (coding sequence)
KL divergence to Kozak (pos 5)
KL divergence to Kozak (pos 4)
KL divergence to Kozak (pos 3)
KL divergence to Kozak (pos −1)
KL divergence to Kozak (pos −2)
KL divergence to Kozak (pos −3)
KL divergence to Kozak (pos −4)
KL divergence to Kozak (pos −5)
KL divergence to Kozak (pos −6)
KL divergence to Kozak
energy (win −16 to 23)
energy (mRNA sequence)
energy (3' UTR)
energy (5' UTR)
energy exp vitro (win −11 to 28)
energy exp vitro (mRNA sequence)
energy exp vitro (3' UTR)
energy exp vitro (5' UTR)
energy exp vivo (win −2 to 37)
energy exp vivo (mRNA sequence)
energy exp vivo (3' UTR)
energy exp vivo (5' UTR)
average elongation rate
tAI (coding sequence)
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
significant
not significant
Ypl184c
Nsr1
Nrd1
Npl3
Nop56
Nab3
Gbp2
Cbc2
Puf4
Pub1
Pab1
Nab2
Bfr1
Scp160
Khd1
RNA binding protein enrichment
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
Spearman correlation to log(TE)
expected
not expected
expected, not sig
not expected, not sig
unknown

Figure E8

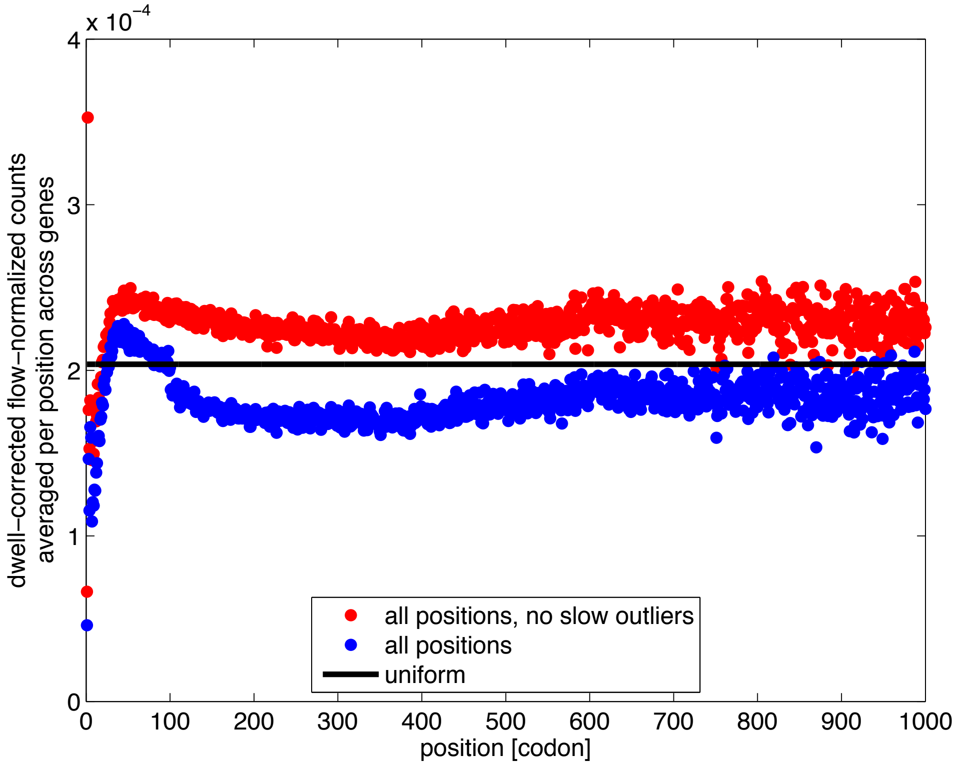

x 10^{-4}
4
3
2
1
0
dwell-corrected flow-normalized counts
averaged per position across genes
0
100
200
300
400
500
600
700
800
900
1000
position [codon]
all positions, no slow outliers
all positions
uniform

Figure E9

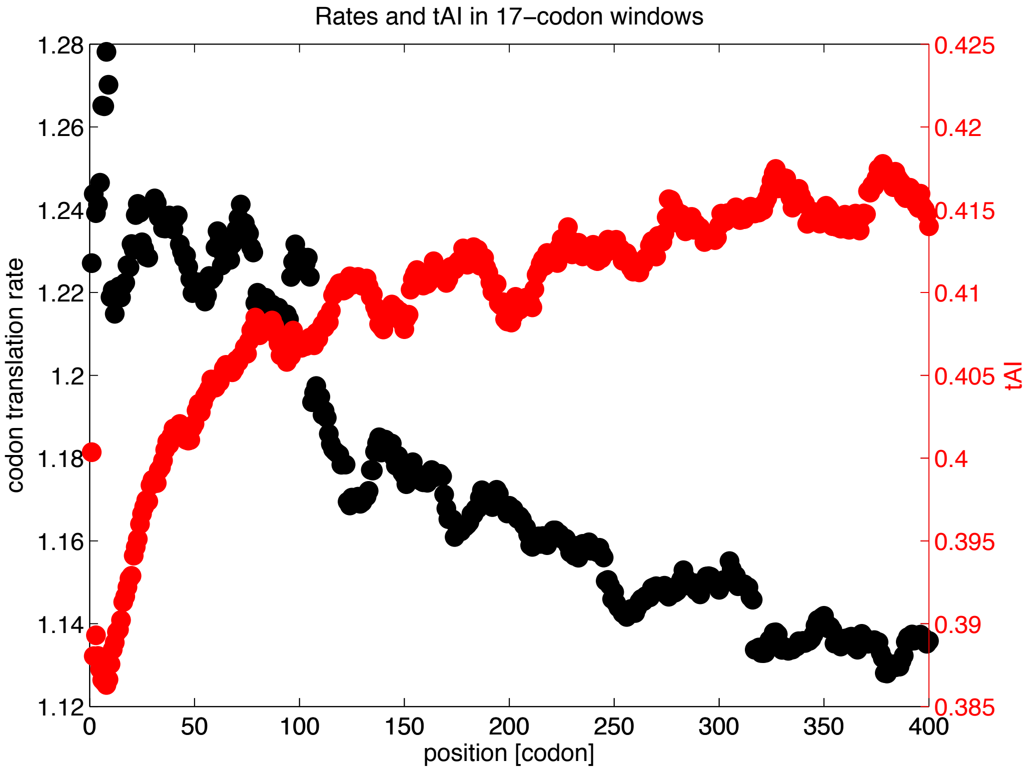

Rates and tAI in 17−codon windows
codon translation rate
1.28
1.26
1.24
1.22
1.2
1.18
1.16
1.14
1.12
0
50
100
150
200
250
300
350
400
position [codon]
tAI
0.425
0.42
0.415
0.41
0.405
0.4
0.395
0.39
0.385

Figure E10

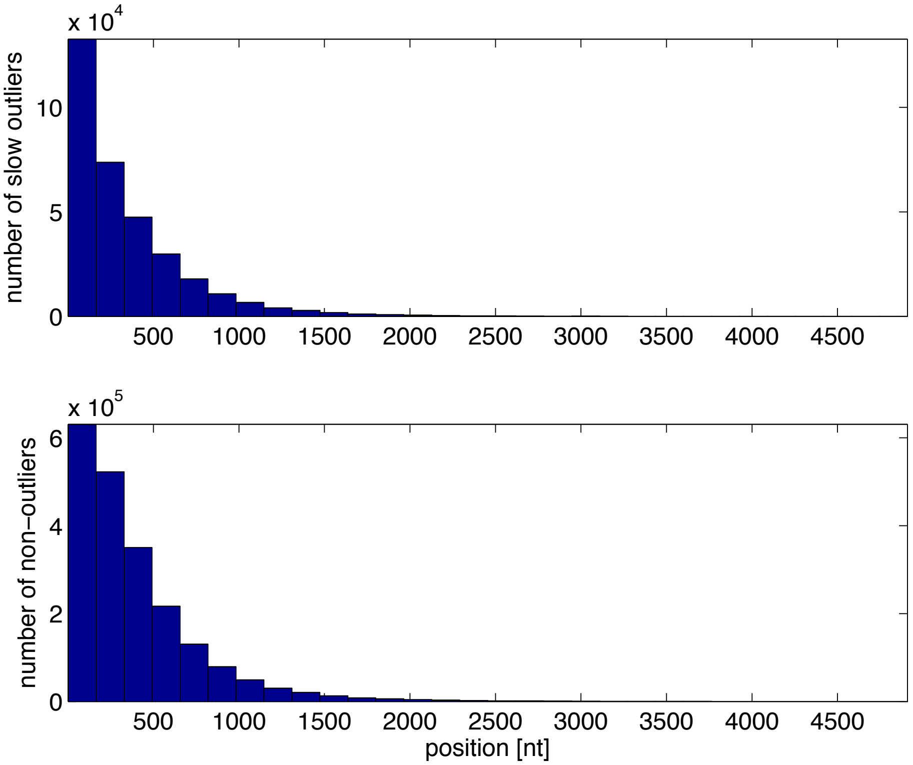
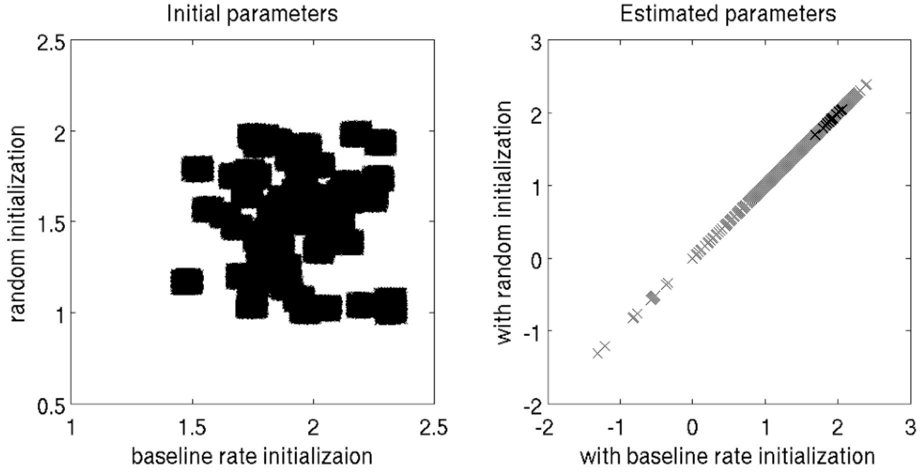

x 10^4
number of slow outliers
10
5
0
500
1000
1500
2000
2500
3000
3500
4000
4500
x 10^5
number of non−outliers
6
4
2
0
500
1000
1500
2000
2500
3000
3500
4000
4500
position [nt]

Figure E11



Initial parameters
2.5
2
1.5
1
0.5
1
1.5
2
2.5
random initialization
baseline rate initializaion
Estimated parameters
3
2
1
0
-1
-2
-2
-1
0
1
2
3
with random initialization
with baseline rate initialization

Figure Legends

Figure 1: Model of protein synthesis.

Ribosomes initiate translation with a protein synthesis rate or flow (*J*) of ribosomes. This is conserved across the strand, so that at each residue *(m,k)* the flow depends on the dwell time of the ribosome (μ) and the ribosome occupancy (proportional to footprint count *d*). Slower positions, for example *(m,2)* compared to *(m,1)*, can inflate the average footprint count per gene and must be accounted for when estimating flow. Dwell times and flow are correlated with local and global *cis*-features.

Figure 2: Correlation between codon translation rates and measures of codon usage bias.

Left: Insignificant Spearman correlation between estimated codon translation rates (scaled up by a factor of 1000) and tRNA abundance from microarray measurements using either fluorophore Cy3 or Cy5 (Dittmar et al, 2004) on 39 codons with measured levels.

Right: The same correlation but to tAI is also not significant.

Figure 3: Comparison between codon translation rates in wild-type and mutants.

Correlation between estimated codon translation rates in wild-type versus mutant for the three mutant samples (the manipulated codon is highlighted in red). Rates are normalized by the minimum one in each sample. Pearson correlations are nearly exact, indicating that the mutant rates are generally unaffected.

Figure 4: Comparison between translation efficiency in wild-type and mutants.

Left: Wild-type TE compared to mutant TE for the three mutant samples. Strong Spearman correlations shown suggest TE is generally unaffected by tRNA manipulation. Right: Spearman correlation, for each codon, between the ratio of mutant TE to wild-type TE and the percent of codon per gene. Significant correlations are shown as filled dots. For AGG mutants, the correlation is not higher for the manipulated codon (highlighted) than for other codons, indicating that optimizing codon usage does not affect TE. For ACA-K, the correlation is negative for the ACA codon, suggesting a mild effect.

Figure 5: All codons show negative correlation between outlier strength and proximity to gene start.

Correlation between slow outlier strength and position per length from 5' end, conditioned by the codon, plotted against codon tAI. For each codon *c*, we calculate the Spearman correlation for outlier strength Δ_{mk} and position per length from 5' end (k / L_m) but restricted to the *(m,k)* that satisfy *codon(m,k) = c*. All codons except one (hollow circle), which has the second lowest abundance in the genome, have a significant negative correlation. This indicates that 5' end outliers are slower even independent of codon bias.

Figure 6: RNA structure energy and its relationship to translation efficiency.

Left: Energy averaged in sliding windows of 40nt (see **Materials and Methods**) across all genes for *in vitro* and *in vivo* measures of energy via DMS probing (Rouskin et al, 2013). The second red line corresponds to the first window with lowest energy (~60nt for in vitro and ~80nt in vivo).

Right: Spearman correlation between the energy windows and TE. The first red line corresponds to the first window with significant correlation (9nt for *in vitro* and 18nt for *in vivo*).

Figure 7: Estimated Kozak motif for efficient genes.

Estimated TE-driven Kozak motif based on a regression model (see **Materials and Methods**). The original Kozak consensus for yeast (Hamilton et al, 1987) is WAMAMAATGTCY.

Expanded View Legends

Figure E1

Correlation between experimental measures of protein abundance – de Godoy et al (2008) and Newman et al (2006) – and estimated flow. As a baseline, we compare against protein abundance calculated as the average footprint count per gene.

Figure E2

Overexpression of tRNA$^{Arg(CCU)}$ does not significantly alter amino acid charging levels. Bulk RNAs from strains as indicated were resolved at pH 5 by PAGE, transferred, and hybridized with oligonucleotide probes specific for tRNA species as indicated, and relative tRNA$^{Arg(CCU)}$ levels and charging levels were evaluated as described in **Materials and Methods**. Solid arrows show deacylated tRNAs; dashed arrows show charged tRNAs; % charged refers to tRNA$^{Arg(CCU)}$.

Figure E3

The ratio between estimated mutant and wild-type rates. The mean (solid black line) and standard deviation (dashed line) are shown. ACA-K has a larger spread, but the manipulated codon (shown in red) is not an outlier in any sample. Codons are grouped and sorted by amino acid.

Figure E4

The ratio of mutant to wild-type footprint count per codon, averaged over the first 5 occurrences of the codon per gene over all genes, presented for the three mutant samples.

Counts are normalized by the average in the 15-codon window before (red line), after (green line), or around (blue line) the codon. We show a subset of the codons: the 5 with lowest tAI (dots), the 5 with highest tAI (squares), and the 6 with middle tAI (stars), in addition to the two codons ACA and AGG (diamonds). In each case, if the manipulated codon of interest induces a change in speed under the common hypothesis (lower for ACA-K and higher for AGG-OE and AGG-QC), we expect a corresponding peak or valley, respectively, in the presented ratio. However, the ratios at ACA and AGG are not significantly higher than 1-standard deviation (dotted line) or than the other representative codons.

Left: Counts are raw footprint counts.

Right: Counts are dwell-corrected footprint counts.

Figure E5

The analysis of Figure 4 repeated on flow instead of TE. As before, wild-type and mutant flows generally agree. Correlations between the ratio of mutant flow to wild-type flow and the percent of codon per gene are not higher for the manipulated codons compared to other codons, despite the dramatic change in tRNA abundance.

Figure E6

Distribution of three features among reduced TE genes and increased TE genes in ACA-K. Distributions are skewed for reduced TE genes (with lower TE in mutant compared to wild-type) toward initiation signals that could confound the TE decrease. Slower-than-expected codons with an excess number of ribosome counts are defined formally as "outliers" (see **Materials and Methods**). Each feature distribution is calculated over all positions in the genes in the specified gene set (either reduced TE genes or increased TE genes) satisfying the specified criteria (a position that is a slow outlier, a position that is a slow outlier in the first 100 codons, or a position that is a slow outlier and an ACA codon). The feature distributions for reduced TE versus increased TE genes are distinct (p-values shown are significant under a Kolmogorov-Smirnov test). Outlier positions are calculated in the ACA-K mutant.

Figure E7

Correlation between log(TE) and gene-level features, including *cis*-features and RNA binding protein enrichment (see **Materials and Methods**). Significant threshold is p = 0.05. (See Note E1 for how expected correlations for the RNA binding proteins were determined.)

Figure E8

Dwell-corrected footprint counts normalized by flow, geometrically averaged per position over all genes aligned by start codon (ignoring 0 footprint counts). Removing slow outliers (red curve) reduces the peak in density at ~45 codons (135 nt).

Figure E9

The tAI in sliding windows of 17-codons averaged across all the genes aligned by start codon (red curve). The same analysis with our estimated codon translation rates (scaled up by 1000) (black curve) show that rates at the 5’ end are not lower compared to the rest of the gene.

Figure E10

Histograms of positions of slow outliers and non-outliers are similar.

Figure E11

Two different initializations of the parameters for the translation model yield estimated parameters that are nearly exact. This demonstrates the model is robust to initialization.

Table E1

Counts of tRNA in RPM (number of reads per million) in ACA-K and wild-type. The threonine tRNA recognizing the ACA codon (highlighted) is reduced to 1/3 of the wild-type level.

Table E2

Eight categories of potential correlates to outlier strength.

Table E3

Correlation between outlier strength and features. Significant ones (see **Materials and Methods**) are highlighted. The first two rows per feature correspond to Pearson correlation; the last two are Spearman. (See Note E1 for more discussion of the correlations.)

Table E4

Performance of TE regression model (see **Materials and Methods**). Error (should be low) and correlation (should be high) between predicted and actual TE is measured on 100 random test sets of genes not used during model training. Performance drops in a null model learned on randomized TE labels (last column). Performance also drops when using the original Kozak motif (middle column). Error on the training set is included to show that our model generalizes to genes not used in training (it is close to test set error).

Table E5

Summary of main results for model variations. The first five columns are models with different constants for the second term in the objective function and the last column is a model without μ_m^c parameters (see **Materials and Methods**). Rows 1-3 represent correlation between our parameters in our model and in the model variation. Rows 4-6 represent correlation between codon translation rates in model variations and codon bias measures. Rows 7-8 represent correlation between protein synthesis rates in model variation and protein abundance measures. Results are similar to the ones reported for the model used throughout the paper (const = 100).

CHAPTER 5

Discussion

DISCUSSION

For a long time, the majority of our knowledge about RNA structure came from X-ray crystallography and nuclear magnetic resonance spectroscopy[1]. Although these methods are very precise and provide great detail of the exact nature of base pairing interactions, both X-day and NMR are limited to mostly short molecules with stable RNA structures that can form a homogeneous population. Yet the majority of cellular RNAs, such as mRNAs and long intervening non-coding RNAs (lincRNAs), are long and likely exist in multiple conformations[2]. Recent applications of chemical and enzymatic probing in a high throughput manner[3,4] have allowed the identification of RNA structure within many long molecules on the order of thousands of bases. Nevertheless, we are still limited to obtaining a structural signature of a population average and are thus ignoring a major aspect of RNA:RNA interactions- their flexibility and sensitivity to the exact cellular environment. Indeed, RNAs can form more than one stable structure, and these distinct conformations often have different biological activities[2,5]. Thus, it is crucial going forward to bring high throughput chemical probing to the single molecule level. Here I will discuss improvements to the DMS-seq assay.

Ideally, it would be best to read out all DMS modifications directly for every molecule in a given population. Interestingly, upon UV crosslinking of RNA protein interactions, the crosslinking sites can be detected in two ways - either as sites that block reverse transcription[6] or as sites that induce errors during RNA to DNA synthesis[7]. Similarly, since DMS reacts with the Watson/Crick positions, thus preventing the correct paring of complementary nucleotides, there can be conditions under which the reverse transcriptase (RT) incorporates a mismatched nucleotide at the DMS modified position.

Indeed it was shown recently that using Mn_2Cl instead of Mg_2Cl causes SuperScript II (a viral-based RT) to be more error prone and read through the DMS modification by introducing mutations. Nevertheless, it is not clear what percent of the time a mutation is made when a DMS modification is present. It is important that most DMS modifications result in a mutation otherwise one would need to extensively modify the RNA and this can lead to RNA unfolding. Another drawback of using Mn_2Cl is that the background error frequency of the RT also increases, which is less than ideal for a high throughput application.

An alternative option is using a non-retroviral RT. Mobile group II introns encode RTs that function in intron mobility by a process that requires reverse transcription of a highly structured, >2kb intron RNA, and may be better suited to read through DMS modifications. Several years ago, two very thermostable (up to 70C) group II intron RTs were purified from T. elongates(Tel4c) and Geobacillus stearothermophilus(GsI-IIC)[8]. In addition, it was shown that sequencing human tRNAs with GsI-IIC RT results in a high frequency of mutation (conversion predominantly to T or G base) at an endogenous, DMS-like methylated A or C. Excitingly, preliminary results show that using GsI-IIC RT in a DMS-seq read-through single molecule strategy results in a better signal to noise ratio than the original Superscript II RT block assay [unpublished]. Moreover, this type of strategy is resistant to ligation, fragmentation and other sequencing library generation biases because it is internally controlled – the mutation signal is calculated by counting the number of mutations at each nucleotide divided by the number of times the correct nucleotide is present in the sequencing library.

Finally, another *in vivo* RNA structure probing chemical was developed recently[9].

This chemical is a derivative of NMIA (N-methylisotoic anhydride) used in the SHAPE (selective 2'OH acylation followed by primer extension) procedure and does not report on the Watson–Crick position directly, but interacts with the 2'OH group of the ribose in a conformation dependent matter[10]. It appears that the size and chemistry of NMIA derivatives make them very sensitive not only to RNA structure but also the presence of RNA binding proteins. Since DMS and NMIA report on overlapping yet distinct properties of RNA it would be very insightful to use both in a single molecule readout approach. Such techniques will allow the investigation of RNA dynamics and how it functions to affect major cellular processes.

References:

1. Cruz, J. A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136,** 604–609 (2009).
2. Dethoff, E. A., Chugh, J., Mustoe, A. M. & Al-Hashimi, H. M. Functional Complexity and Regulation through RNA Dynamics. *Nature* **482,** 322–330 (2012).
3. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (2014).
4. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
5. Montange, R. K. & Batey, R. T. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **37,** 117–133 (2008).
6. Huppertz, I. *et al.* iCLIP: protein-RNA interactions at nucleotide resolution. *Methods San Diego Calif* **65,** 274–287 (2014).

7. Zhang, C. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29,** 607–614 (2011).

8. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA N. Y. N* **19,** 958–970 (2013).

9. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9,** 18–20 (2013).

10. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127,** 4223–4231 (2005).