

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Setting and Maintaining Precisely Controlled Initiation of Protein Synthesis in Escherichia Coli

Permalink

<https://escholarship.org/uc/item/4n73x3dk>

Author

Burkhardt, David Hutcheson

Publication Date

2014

Peer reviewed|Thesis/dissertation

Setting and maintaining precisely controlled initiation of protein
synthesis in *Escherichia coli*

by

David Hutcheson Burkhardt

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

I've been fortunate to carry out my graduate work in company of a talented group of scientists who have been generous with their time and insights. It is my pleasure to thank them here for their support.

First, thanks to my adviser, Carol Gross. I first walked into Carol's office as a physics student with a dim understanding of how to approach biological problems. Carol worked hard to understand my abilities and interests, and saw that working with the genomics datasets generated by ribosome profiling would be the right fit. I will miss the long sessions in Carol's office, whittling experimental ideas to the core of what needed to be done and envisioning big-picture approaches to problems. Carol Gross' role as my mentor has extended far beyond her scientific advising: throughout my time in the lab, she has shown sincere commitment to understanding my personal aspirations, and I am thankful for this support.

Thanks also to the Gross Lab members who provided a great scientific environment. Virgil Rhodius, my rotation adviser, first showed me how to tackle scientific problems in the lab. Monica Guo cohabited the lab with me for my entire grad school journey, and I drew on her knowledge and intuition at many turns. Bentley Lim sat directly behind me for much of my graduate career and was a constant source of advice on matters large and small. Bentley set up a Friday lunch / journal club with Monica and me early on that filled out much of my knowledge about techniques and approaches. Thanks to Katya Orlova for building the Δcsp strains. The high-throughput genomics group, Nassos Typas, Bob Nichols, Anthony Shiver, B'young-Mo Koo, Andrew Gray, Jason Peters, and John Hawkins, have been a constant source of advice and inspiration. More recently, it has been a pleasure to help

guide Natalie Baggett and Brian Sharon in developing their projects. Thanks also to all of my other baymates, Rachna Chaba, Takashi Yura, Kenneth Tong, Matylda Zietek, Candy Lu, and Angelo Cabal for great company, and thanks to Sonya Diaz, and Danh Le for keeping our lab well stocked and operational.

Thanks to Jonathan Weissman, who has been a second advisor to me. Our many productive sessions hammering out the details of our paper in his office were an incredible education in formulating and presenting a story. Watching Jonathan approach problems in his group meetings has been fascinating, and his ability to provide vision, drive, and a path to execution to any project is manifest in the depth and breadth of the work that has come out of his group. Thanks also to the many members of the Weissman Lab who helped me along the way. Thanks to Eugene Oh, who taught me ribosome profiling, and was full of insights into planning and carrying out complicated experiments. I was able to study the relationship between RNA structure and translation because of years of work that Silvi Rouskin put into developing DMS-seq. Silvi was a fantastic collaborator to work with, and watching Silvi's tenacity in developing DMS-seq into an assay that delivers very high quality data was inspiring. Gene-Wei Li was really a third mentor to me, providing both big-picture vision and fine-grained guidance on most everything that I did. It was fascinating to follow the growth of two stories that Gene pulled from our complex of our ribosome profiling data, and I look forward to see what path his new lab will take.

Thanks to Hana El-Samad, the third member of my thesis committee, and the El-Samad lab. While working on an initial project, I got a lot of support from Hana and her lab; it was in El-Samad lab meetings that I first developed my taste for how to approach quantitative biology problems.

Thanks also to the other members of my qualifying exam committee, Orion Weiner and Hao Li, and to my rotation advisors, Chao Tang, Wenzhe Ma, Wendell Lim, Andrew Horwitz, Chris Voigt, Jeff Tabor, and Howard Salis. Ideas taken up in all of these labs guided my eventual work.

Thanks to David Agard and Matt Jacobson, leaders of the UCSF Biophysics program, whose tireless efforts created a fantastic intellectual environment. And, of course, thanks to all of my brilliant classmates, in particular Brittany Belin, Jaline Gerardin, Jenny Hsiao, Monica Hsu, Charlie Kehoe, Noah Ollikainen, Jack Peterson, Gabe Rocklin, and Chris Wen, with whom I shared many dinners, board games, and camping trips that made the whole grad school experience much more enjoyable. I look forward to seeing the great things all of you will achieve in science and beyond.

Thanks also to the non-scientists who helped along the way. Adam, Rick and Colin took part in a trip that set me off on this journey, and were always there when needed. Travis and Andrew shared a couple of great adventures at critical junctures. The rotating roster of 1069 Carolinaers, Rich, Ricardo, Jon, Mike, Bennett, Seth, Daniel, Daniel, and Dan, helped keep our Potrero perch a great rent-controlled home.

Thanks to my family, who have been a constant source of support. My brother Mark led the Burkhardt western expedition, and was a constant sounding board as I progressed through grad school, having faced most of the same challenges I dealt with just a couple of years earlier. My parents have faithfully supported all of my endeavors, and I am grateful to them for this support. Thanks also to the Lee/Daniel and Duhn families for constant encouragement and for keeping me well fed.

And, thanks to Alex, for more than I could put into words.

Abstract

Protein synthesis is the most energetically expensive process in prokaryotes. Understanding how protein synthesis is regulated is critical both for decoding natural systems and for engineering synthetic protein synthesis. Protein synthesis in prokaryotes occurs on mRNAs organized into operons consisting of discrete open reading frames (ORFs) that are differentially translated by as much as 100- fold. We have applied ribosome profiling, which enables the quantitative determination of the rates of protein synthesis genome-wide in *E. coli*, to understand the rules that guide these differential rates of protein synthesis. We then combined ribosome profiling with DMS-seq, which monitors mRNA structure genome-wide, to monitor the relationship between mRNA structure and translation on endogenous messages, enabling us to understand the mRNA features that instruct translation efficiencies.

We find precisely tuned synthesis rates for a wide variety of proteins —members of multi-protein complexes are made in proportion to their stoichiometry, and components of functional modules are produced differentially according to their hierarchical role. Additionally, several principles of design optimization emerge from the absolute copy number measurements. These include how the distribution of levels of different transcription factors is optimized to enable rapid responses and how a metabolic pathway (methionine biosynthesis) balances the cost of enzyme production with the requirement for its activity.

Structural probing of mRNAs reveals 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.

Together, this work indicates protein synthesis is precisely controlled in prokaryotes, and this precise control requires mRNA structures designed to reflect synthesis rates. This lays the framework for both future efforts to computationally determine complex stoichiometry, and for computational design of protein synthesis rates.

Table of contents

Preface	Acknowledgements	iii
	Abstract	vii
	Table of Contents	ix
	List of Figures	x
Chapter One	Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources	1
Chapter Two	Operon mRNAs are organized into ORF-centric structures that specify translation efficiency	75

List of Figures

Chapter One

Figure 1	Absolute Quantification of Protein Synthesis Rates	25
Figure 2	Proportional Synthesis of Multi-Protein Complexes	27
Figure 3	Proportional Synthesis for Complexes in Yeast	30
Figure 4	Hierarchical Expression for Functional Modules	33
Figure 5	Composition of the <i>E. coli</i> Proteome	36
Figure 6	Abundance of Transcription Factors (TFs)	38
Figure 7	Quantitative Analysis of the Methionine Biosynthesis Pathway	40
Figure S1	Adjustment to ribosome density based on sequence- and position-specific variation in translation elongation rates.	42
Figure S2	Comparison of published quantitative proteomics measurements and individually measured protein copy number.	45
Figure S3	Proportional synthesis for other multi-protein complexes	47
Figure S4	Proportional synthesis at 10°C, mRNA levels, and gene order	49
Figure S5	Predicted strength of ribosome binding sites and observed translation efficiency.	51
Figure S6	Effect of MetE level on growth rate.	53

List of Figures, continued

Chapter Two

Figure 1	DMS-seq effectively probes RNA structures in <i>E. coli</i>	98
Figure 2	mRNA structure is organized around open reading frames	100
Figure 3	Cold induces a defect in translation instigated by an increase in mRNA structure	104
Figure 4	RNase R and Csps facilitate cold recovery	106
Figure 5	Csp expression is controlled by an auto-regulatory feedback loop	108
Figure 6	Model of operon structural organization and surveillance	110
Figure S1	DMS-seq effectively probes RNA structures in <i>E. coli</i>	112
Figure S2	mRNA structure is organized around open reading frames	115
Figure S3	Structure and translation efficiency remain correlated at 10°C	119
Figure S4	Csp deletion increases mRNA structure and reduces translation efficiencies	120
Figure S5	CspB UTR structure is modulated by cold shock	121
Figure S6	CspB UTR structure is modulated during cold recovery	123

Chapter One:
Quantifying absolute protein synthesis
rates reveals principles underlying
allocation of cellular resources

Gene-Wei Li^{1-3*}, David Burkhardt^{2,4}, Carol Gross^{2,4,5}, and Jonathan S. Weissman^{1-3*}

¹Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute,

²California Institute of Quantitative Biosciences,

³Center for RNA Systems Biology,

⁴Department of Microbiology and Immunology,

⁵Department of Cell and Tissue Biology,

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

The text of this chapter is a reprint of the material that appears in *Cell*, April 2001, volume 2, number 4, pages 247-51 by Gene-Wei Li, David Burkhardt, Carol A. Gross, and Jonathan S. Weissman.

Introduction

Protein biosynthesis is by far the largest consumer of energy during cellular proliferation; translation by ribosomes is estimated to account for ~50% of the energy consumption of a rapidly growing bacterial cell, and ~30% of that for a differentiating mammalian cell (Buttgereit and Brand, 1995; Russell and Cook, 1995). The tremendous cost associated with protein synthesis makes it a key step for regulating diverse cellular functions. Therefore, determining how a cell allocates its synthesis capacity for each protein provides foundational information for systems biology.

A fundamental question is whether it is necessary for the cell to exert tight control over the synthesis of individual protein components. For example, the levels of stoichiometric components of protein complexes could be established by differential degradation of excess subunits (Blikstad et al., 1983; Lehnert and Lodish, 1988), rather than by precise synthesis. Moreover, precise control of steady-state protein abundance may not be critical for the performance of cellular circuits. The architectures of several signaling and metabolic pathways have been shown to be robust against variation in protein levels through post-translational feedback (Alon et al., 1999; Barkai and Shilo, 2007; Batchelor and Goulian, 2003; Hart et al., 2011; Shinar et al., 2007; von Dassow et al., 2000). It remains to be explored whether these post-translational mechanisms are the dominant strategy for maintaining proper functions, or are simply fail-safe mechanisms added on to fine-tuned protein synthesis. More generally, defining such design principles is key to both understanding and manipulating quantitative behavior of a cell.

Efforts to monitor protein synthesis rates at the global level have mainly relied on pulsed metabolic labeling followed by two-dimensional gel electrophoresis, or more recently by mass spectrometry (Dennis, 1974; Lemaux et al., 1978; Schwanhausser et al., 2009). While relative changes in synthesis rates for the same protein are attainable (Selbach et al., 2008), absolute rates are more difficult to evaluate. Additionally, the precision of pulsed metabolic labeling is limited by requirement for nutrient shifts, which affect instantaneous rates of protein synthesis. Alternative methods for expression profiling by determining global mRNA levels (e.g. by high density microarrays or RNA-seq) do not report the extensive regulation present at the level of translation. These constraints point to a need for a label-free method with unbiased and deep coverage of cellular proteins.

Ribosome profiling—deep-sequencing of ribosome protected mRNA fragments—directly captures protein synthesis in natural settings (Ingolia et al., 2009). It is a general tool for monitoring expression as well as enabling identification of novel translational events (Brandman et al., 2012; Brar et al., 2012; Ingolia et al., 2011; Li et al., 2012; Oh et al., 2011; Stern-Ginossar et al., 2012). Here, we exploited the ability of ribosome profiling to provide quantitative measurements of absolute protein synthesis rates, covering >96% of cellular proteins synthesized in a single experiment. For stable proteins in bacteria, we then estimated and verified absolute protein copy numbers.

This analysis revealed precise tuning of protein synthesis rates at the level of translation, including a broadly used “proportional synthesis” strategy in which components of multi-protein complexes are synthesized with ratios that quantitatively reflect their subunit stoichiometry. Optimized translation rates are also prevalent among members of functional modules—differential expression pertinent to their functional hierarchy, i.e. when the activity of one

member is controlled by the other, was widely observed in our dataset. The protein copy numbers inferred from synthesis rates also revealed rules that govern the abundance of transcription factors, and allowed quantitative characterization for the methionine biosynthesis pathway, for which we identified a bottleneck enzyme whose expression level is optimized for maximal growth rate. More broadly, our approach and datasets provide a foundation for quantitative understanding of both cellular physiology and precise biological engineering.

Results

Genome-wide measurement of absolute protein synthesis rates and protein copy numbers

The ribosome profiling approach involves freezing of cellular translation followed by digestion of all mRNA regions that are not protected by the ribosome (Ingolia et al., 2012; Ingolia et al., 2009). Each ribosome-protected mRNA fragment is then identified by massively parallel next-generation sequencing (Ingolia et al., 2012; Ingolia et al., 2009). Because each ribosome is producing one protein molecule, the rate of protein synthesis is proportional to the ribosome density of a given gene as measured by the footprint density (number of footprint per unit length of the gene), provided that all ribosomes complete a full length protein and have similar average rates of elongation across genes. Both criteria are broadly met in our dataset. During exponential growth in *E. coli*, there is little drop-off in ribosome density for the vast majority of genes (Li et al., 2012; Oh et al., 2011) (Fig 1A). The few genes that display large drop-off could represent novel events of translational regulation (Fig. S1A). We have previously demonstrated that rare codons are generally translated at similar speed as abundant codons, indicating that differences in codon usage between transcripts do not cause differences in the average rates of elongation (Ingolia et al., 2011; Li et al., 2012). Moreover, sequence dependent

pausing of ribosomes (Li et al., 2012) does not appear to broadly distort the average density of ribosomes along a message, as similar ribosome densities are observed in the first and second halves of each gene. Most genes differ by <30% (standard deviation of the mean, Fig 1A). Additionally, correcting for sequence- and position-specific variation in elongation rates has only modest effect on average ribosome density (Fig. S1). Together, these results indicate that local variations in translation speed do not strongly impact synthesis rates measurements based on average ribosome density.

To broadly evaluate the rates of protein synthesis, we performed ribosome profiling in *E. coli* grown in different growth conditions with high sequencing depth (90 million fragments per sample) using a modified protocol that enables more complete capture of footprints (Methods). Within each dataset, synthesis rates were calculated as the average ribosome density in the gene body, with correction factors for elevated ribosome density at internal Shine-Dalgarno sequences and towards the beginning of open reading frames (Methods). The corrections were small (Fig. S1D), but were nonetheless important for the quantitative analysis described below. We determined the absolute rates of synthesis (in units of molecules produced per generation) by normalizing the average ribosome density for each protein in the proteome by the total amount of proteins synthesized during the cell doubling time (Methods). For growth in a rich defined medium (Neidhardt et al., 1974), we evaluated 3,041 genes which account for >96% of total proteins synthesized. A similar number of genes were evaluated for glucose-supplemented minimal media. All of these genes have >128 ribosome footprint fragments sequenced, with an error of less than 1.3-fold across biological replicates. The lowest expression rate among these genes correspond to ~10 molecules per generation. The complete list of protein synthesis rates can be obtained at <http://ecoliwiki.net/tools/proteome/> (Table S1).

We validated our results by comparing our data against published measures of specific protein copy numbers for *E. coli*. Because the overwhelming majority of proteins are long-lived compared to the cell cycle during exponential growth (Larrabee et al., 1980), the absolute copy number of a protein can be estimated as the synthesis rate times generation time (21.5 min in rich defined media, see Methods). We compiled a list of 62 proteins that have been quantified individually in 21 independent laboratories (Table S2). Although each measurement is associated with its own uncertainty, we argue that collectively they represent the current standard for quantification. Our results agreed well with these published copy numbers with a Pearson correlation coefficient $R^2 = 0.96$ (Fig. 1B). Deviations from the identity line in Fig. 1B likely reflect biological phenomenon. For example, the strongest outlier is σ^{32} , the heat shock transcription factor that is known to be actively degraded (Grossman et al., 1987). Our measures based on synthesis rates thus provide an upper bound for the protein levels for the small subset of proteins that are rapidly degraded. Differences in growth conditions and strain backgrounds contribute to other small differences between literature values and our results (see Methods). Existing efforts to globally quantify protein abundance in *E. coli* using mass spectrometry or fluorescent reporter show less concordance and dynamic range (Fig. S2). In conclusion, our genome-wide synthesis rate measurements and the resulting estimate of protein abundance are supported by classic biochemical measurements across 5 orders of magnitude of protein abundance.

Proportional synthesis of multi-protein complexes

We next used our measurements to evaluate the extent to which fine-tuned synthesis rates are a general feature of cellular physiology, focusing initially on members of stable multiprotein complexes with known stoichiometry. The subunits of these complexes require balanced steady

state levels, as excess components are often prone to misfolding or aggregation (Tyedmers et al., 2010). Although quality control mechanisms for removing uncomplexed proteins exist (Shemorry et al., 2013), it was unclear whether the stoichiometry balance is generally established first at the synthesis level.

We first examined the F_0F_1 ATP synthase complex, which consists of 8 subunits, each with different stoichiometry, expressed from a single polycistronic transcript (the "ATP operon"). Despite sharing the same message, the ribosome density of each open reading frame is clearly distinct (Fig. 2A), and qualitatively agrees with the differential synthesis rates previously reported (Brusilow et al., 1982; Quax et al., 2013). Remarkably, the synthesis rates quantitatively reflect the stoichiometry of the complex; the ATP operon has evolved to synthesize the appropriate ratio of subunit proteins, ranging from 1- to 10-fold.

Rather than the ATP operon being a specialized case, we found that tuning of synthesis rates to the subunit stoichiometry, or “proportional synthesis”, is a broadly used strategy for protein complexes. We systematically assembled a list of stable multi-protein complexes with well-characterized stoichiometry in *E. coli* (Table S3). Of the 64 complexes (comprising 212 different proteins) that are expressed in our growth conditions, 59 (92%) adhere to proportional synthesis. The majority (55%) are synthesized at levels that are indistinguishable from the stoichiometry (smaller than the experimental uncertainty of 1.3-fold difference). The ratio of synthesis rates exceeds the ratio of stoichiometry by a factor of two in only five complexes (Fig. S3D), and these small number of exceptions could suggest dominant control at the level of degradation or the existence of dynamic sub-complexes, as in the case of the outer membrane protein assembly complex (BAM) (Rigel et al., 2013).

Proportional synthesis applies to both cytosolic and membrane proteins. For complexes with more than two components, the agreement between synthesis rates and subunit stoichiometry is plotted in Fig. 2B and Fig. S3. We also observed very similar synthesis rates for complexes with two equimolar subunits (Fig. 2C and Fig. S3A-C). Notably, proportional synthesis is robust against temperature; similar ratios in synthesis rates were observed both at 37°C and at 10°C (Fig. S4A). Furthermore, both abundant and scarce proteins have evolved strict tuning of synthesis rates, as the expression levels of these complexes ranges over four orders of magnitude.

Proportional synthesis in *E. coli* is predominantly achieved through translational, rather than transcriptional control. The majority of multi-protein complexes encode their subunits on a single polycistronic mRNA, with each subunit translated from its own initiation site (47/64 complexes, Fig. 2B-C and Fig. S3A). RNA-seq analysis confirms that the mRNA levels of the genes in these operons are similar, whereas the different translation efficiency (synthesis rate per mRNA) reflects the stoichiometry (Fig. S4BC and Table S4). Moreover, gene order does not explain differential synthesis rates (Fig. 2A and 2C and Fig. S4D), consistent with our previous observation that translation rates among genes in the same operon are only weakly correlated (inset, Fig. 2C) (Oh et al., 2011). Protein synthesis rates are generally determined by the frequency of translation initiation (Andersson and Kurland, 1990). However, our current understanding of what determines translation initiation rates is highly incomplete as existing models for either the strength of ribosome binding site or the Shine-Dalgarno sequence alone do not predict proportional synthesis (Fig. 2C) (Salis et al., 2009). Translational auto-regulation (Nomura et al., 1984), coupling (Baughman and Nomura, 1983) or specific RNA secondary structures (McCarthy and Gualerzi, 1990) are factors that could contribute to precise tuning of

synthesis rates. Our discovery of proportional synthesis in polycistronic messages should help guide efforts to dissect the molecular mechanism of translation initiation quantitatively, as well as aid the precise engineering of synthetic biological networks.

The use of translational control and polycistronic operons to achieve proportional synthesis has important potential advantages. In particular, setting the ratios of subunit expression levels exclusively at the translational level greatly simplifies transcriptional regulation; the cell needs only to control the overall expression of the complex and not the relative amounts within the complex. Additionally, sharing the same polycistronic mRNA reduces stochastic imbalance among components of the complex. Because transcription originates from a single gene locus and is thus inherently noisy (Li and Xie, 2011), the ratio of proteins encoded on different mRNAs would be subject to much higher noise levels (Elowitz et al., 2002; Swain, 2004). The use of polycistronic mRNAs circumvents this issue, but translational tuning becomes necessary to achieve different expression levels.

Evidence for proportional synthesis in budding yeast

We found evidence that the budding yeast *S. cerevisiae* also exhibits tightly controlled synthesis of stably associated protein complexes, as indicated by our analysis of a subset of highly characterized complexes (Fig. 3A-B). Genomic duplication events in *S. cerevisiae* have led to numerous paralogous genes, which in some cases can substitute for each other in multi-protein complexes. Interestingly, we found that proportional synthesis is maintained by tuning the synthesis rates for duplicated genes that encode the same subunit. For example, the two α -tubulin genes together are translated at a similar rate as the single β -tubulin gene (Fig. 3C). Similarly, for the COPII Sec23/24 heterodimer, the production rate of Sec23 matches that of

Sec24 and its two homologs (Sfb2 and Sfb3) combined (Fig. 3C). A notable exception for proportional synthesis is the signal recognition particle, for which four subunits are translated at 1:1:2:2 ratio and the other two subunits are in excess (Fig. 3A). It has also been shown that vertebrates produce uneven amounts of α - versus β -spectrin and immunoglobulin light chains versus heavy chains (Blikstad et al., 1983; Lehnert and Lodish, 1988; Shapiro et al., 1966). Understanding the rationale behind the unequal synthesis in these exceptions could provide insights into their physiological functions.

Yeasts must employ distinct mechanisms to achieve proportional synthesis, as subunits are encoded on different mRNAs in eukaryotes. For example, the dynamics of nuclear localization of transcription factors and their affinity to promoter sites could provide independent control for complex levels and subunit ratios (Cai et al., 2008). Given the fundamentally different molecular mechanisms for prokaryotic and eukaryotic expression, these observations argue that proportional synthesis is a result of convergent evolution that maximizes protein synthesis efficiency while minimizing the adverse effects of having uncomplexed subunits.

The broad use of proportional synthesis has important implications for the effect of aneuploidy. Most genes do not possess feedback mechanisms for controlling their expression levels (Springer et al., 2010). Thus a sudden changes in gene dosage would lead to a large imbalance of subunits (Papp et al., 2003). Because cells normally do not face large imbalances in the synthesis rate of multiprotein complexes, aneuploidy would lead to a strong challenge to the protein folding and chaperone networks, consistent with the findings of Amon and co-workers that general proteotoxic stress is a hallmark of aneuploidy (Oromendia et al., 2012; Torres et al., 2008).

Taken together, our findings argue that the relative expression of members of multiprotein complexes is primarily determined at the synthesis level, and that targeted degradation of excess subunits is a secondary layer of control. Indeed components of multiprotein assemblies whose uncomplexed subunits have been shown to be degraded, including the ribosomal L8 complex and the SecYEG translocon in *E. coli* and Fas1/2 in *S. cerevisiae*, also show proportional synthesis (Akiyama et al., 1996; Petersen, 1990; Schuller et al., 1992).

Hierarchical expression of functional modules

Stable protein complexes are only one of a wide range of functional modules that are organized into operons in bacteria, leading us to ask whether translational control also sets expression of other types of functional modules. Because our genome-wide ribosome profiling dataset covers many different modules in the same functional class, we can use our data to identify common expression patterns strategies that are selected through evolution. Our studies of several different modules identified a 2nd pattern: hierarchical expression, in which components are differentially expressed according to their hierarchical role.

Bacterial toxin-antitoxin modules (TA) are widely utilized two-gene systems that control cellular survival (Yamaguchi et al., 2011). The role of antitoxin is to bind to and inhibit its cognate toxin. *E. coli* contains at least 12 type II TA systems, each consisting of a toxin protein and an antitoxin protein in a bicistronic operon (Yamaguchi et al., 2011). For every well-characterized type II TA system, we found that the antitoxin is synthesized at a much higher rate than the toxin (Fig. 4A), which would allow *E. coli* to produce sufficient amount of antitoxin to avoid triggering cell death or growth arrest during unstressed growth. The hierarchical

expression between antitoxin and toxin is irrespective of their relative order in the operon (Fig. 4A). Because most toxins target global translation, the translational control observed for hierarchical expression of TA modules may provide insight into how the system switches to a toxin-dominated state via translational feedback—a central question in antibiotic persistence (Gerdes and Maisonneuve, 2012).

s/anti-s modules are conceptually similar to TA modules. Both are usually encoded in the same operon, and anti-s inhibits the transcriptional activity of the s by direct binding. Interestingly, anti-s's, like antitoxins, are produced in excess compared to s's (Fig. 4B). In both cases, the uncomplexed antagonists (antitoxins and anti-s's) are also subject to regulated degradation (Ades et al., 1999; Yamaguchi et al., 2011). Thus the hierarchical expression would not be evident by measuring protein levels, even though cells ensure an excess of inhibitor during synthesis.

Translationally controlled hierarchical expression appears to be common for a diverse range of functional modules. ATP-binding cassette (ABC) transporters, are comprised of core transmembrane proteins and corresponding substrate-binding periplasmic proteins. Whereas the core membrane complex components follow the proportional synthesis principle elucidated above (Fig. 2B-C), we found that the periplasmic binding proteins are always in large excess (Fig. 4D), suggesting that substrate binding is slower than transport across the membrane. Two-component signaling systems, consisting of a histidine kinase (HK) and its substrate, a response regulator (RR), also exhibit hierarchical translation. For each of the 26 two-component systems in *E. coli*, the substrate is synthesized at a much higher level than the kinase (Fig. 4C). Using mathematical modeling and experimental validation, it has been demonstrated that large excess of RR relative to HK promotes robustness against variations in RR and HK levels (Batchelor and

Goulian, 2003; Shinar et al., 2007). Here we show that this strategy is universally employed for all two-component systems.

Taken together, these results show that hierarchical expression within operons is a key design principle for many diverse functional modules. As illustrated in the four examples above, the same hierarchy of expression levels is repetitively used for the same type of module, pointing to a common quantitative property that is critical for the execution of each task. The examples here are certainly an incomplete list; more quantitative design principles could be uncovered by identifying commonalities among similar systems in such genome-wide datasets.

Bacterial proteome composition

Because the large majority of proteins are stable in *E. coli* (Larrabee et al., 1980), our protein synthesis rate data provides a comprehensive view of proteome composition, allowing us to probe how cells allocate resources (Fig. 5). By far the largest fraction of the protein synthesis capacity is dedicated to making the machinery needed for further translation (41% for growth in rich media and 21% in minimal media), whereas transcription-related proteins account for only 5%. This disparity illustrates the importance of understanding the translational control systems that allow cells to allocate their translational capacity. The ability to monitor the partitioning of protein synthesis capacity under different conditions will provide a critical tool for quantitative characterization of cellular physiology.

The expression level of every protein in the cell is subject to two opposing constraints: the requirement of its function and the cost associated with having an excess that consumes limited resources, such as protein synthesis capacity, quality control machineries, and space (Dekel and Alon, 2005). Our dataset opens up the possibility of broadly investigating how these

competing constraints govern protein expression levels. We select two specific cellular functions (transcription factors and methionine biosynthesis) for further study.

Copy numbers of transcription factors reveal their mode of action

The bacterial chromosome is densely covered with transcription factors (TFs) that bind DNA both specifically and non-specifically (Li et al., 2009). The crowded space on DNA imposes constraints on the abundance of TFs, as overcrowding by non-specifically associated DNA-binding proteins could drastically reduce the overall binding kinetics (Hammar et al., 2012; Li et al., 2009). Thus, although higher concentrations of any given TF would allow it to find its cognate DNA sites more rapidly (von Hippel, 2007), too many TFs in total would mask binding sites. Based on our protein abundance estimates, we found that the average distance between DNA-binding proteins is only ~36 basepairs on the *E. coli* chromosome (assuming most DNA-binding proteins are associated with DNA nonspecifically and randomly distributed throughout the genome, see Extended Experimental Procedures), which is close to the theoretically optimal density for rapid binding (Li et al., 2009). How cells allocate the limited space on DNA to maximize rapid regulation by each TF remained obscure.

Our data indicates that the ~200 well-characterized TFs in *E. coli* show a wide variation in level—more than 60% of the TFs are found to have an upper bound of fewer than 100 monomers per genome equivalent (Fig. 6A-B). A low copy number for a TF implies a slow association rate to DNA, which could lead to slow transcriptional responses (Winter et al., 1981). For example, single-molecule imaging *in vivo* previously revealed that it takes six minutes for one Lac repressor to find a single binding site in a cell (Elf et al., 2007). Compared to the cell doubling time, which can be as short as 20 minutes, the binding kinetics for a low copy number

TF would make it difficult to achieve timely regulation. This can be circumvented with the use of TFs that are always bound to their target but whose ability to recruit RNA polymerase depends on the presence of ligands, as the kinetics of regulation would be determined by diffusion of the small ligand rather than by diffusion of the bulky and far less abundant protein. We therefore hypothesize that the low copy number TFs have evolved to bind to DNA independent of their activity.

To test this hypothesis, we mined the literature for the biochemical properties of 102 TFs in *E. coli* (Table S5). We found that abundant TFs bind to DNA only in response to ligands (Fig. 6C). By contrast, the large majority of low abundance TFs bind to the target sites independent of the corresponding ligands (Fig. 6C). Therefore, cells optimize the limited space on DNA and the need for rapid regulation by requiring that TFs with low abundance always bind to their target sites. This mode of DNA binding for low copy number TFs also supports the model that TFs have evolved to occupy their target sites in native environments (Savageau, 1977; Shinar et al., 2006). This class of TFs can be exploited to build transcriptional circuits with fast response time without incurring extra synthesis cost and nonspecific interactions. A potential downside, however, is increased gene expression noise due to stochastic TF dissociation.

Precise control of enzyme production required for methionine biosynthesis

The expression of metabolic enzymes similarly faces two constraints: the requirement for function and the cost of synthesis. Metabolic control analysis suggests that enzymes are generally made in excess amounts, such that small changes in the level for each enzyme have moderate effects on the output (Fell, 1997). On the other hand, the pools of bacterial enzymes in related metabolic pathways are strictly dependent of growth rates (You et al., 2013), arguing for

precise control of expression based on cellular need. Thus, the principal determinant of expression remained obscure. Here, we show that our quantification of the proteome composition makes it possible to globally analyze the relationship between the levels of metabolic enzymes and their actual reaction fluxes.

We focused on the well-characterized L-methionine biosynthetic pathway for *E. coli* grown in media devoid of methionine (Met). We first calculated the cellular demand for this pathway ($31,000 \text{ s}^{-1}$ Met per cell), i.e. the rate of Met consumption by protein synthesis, by summing up the absolute rates of protein synthesis we determined for each protein times the number of methionine residues in that protein. The other major pathway that consumes Met, which is the synthesis of S-adenosyl-L-methionine, was estimated to contribute to a small fraction of the overall flux (Feist et al., 2007) (see also Methods). We then compared the rate of Met consumption with the maximum velocity (V_{\max}) for its biosynthetic pathway. For each reaction in the pathway, we calculated V_{\max} by multiplying the enzyme abundance we determined by its published turnover number (k_{cat}) (Schomburg et al., 2002). The maximum velocity varies by more than one order of magnitude among the reactions in Met biosynthesis, suggesting that most reactions do not operate at saturating substrate concentration. The last step that is catalyzed by MetE has among the smallest V_{\max} (Fig. 7A), suggesting that it may be a bottleneck in this pathway. Remarkably, we found that the maximal Met production rate allowed by MetE ($V_{\max} = 34,000 \text{ s}^{-1}$ per cell) matches the Met consumption rate. Therefore, under these growth conditions, MetE catalyzed conversion of L-homocysteine to L-methionine is a bottleneck step that operates at maximal velocity with saturating substrate concentration.

Given that methionine biosynthesis by MetE is limiting the overall rate of protein synthesis, why do cells not simply make more MetE protein? MetE is a large and slow enzyme,

whose production consumes ~8% of the total protein synthesis capacity in media devoid of methionine. We investigated whether the cost of increasing MetE production further would outweigh its benefit. To do so, we constructed a simple analytical model for the effect of MetE expression on growth rate (Fig. 7B, Methods). The model considers the cost and benefit of MetE synthesis independently, and allows us to evaluate the level of synthesis where the tradeoff between cost and benefit is optimized. The benefit of producing MetE arises from our observation that it is a bottleneck for the methionine supply for protein synthesis. Hence, devoting more protein synthesis capacity to MetE increases growth rate linearly (Methods). The cost of producing excess proteins, independent of their function, comes from competition for ribosomes—an effect that has been widely studied for *E. coli* (Dekel and Alon, 2005; Dong et al., 1995; Scott et al., 2010). To evaluate this cost, we used the well validated numerical relationship described by Scott and Hwa (Scott et al., 2010).

These two constraints predict that the fastest growth rate, a 28 min doubling time, is achieved at an optimal MetE level of 7% of protein synthesis capacity (Fig. 7B). Remarkably, these predictions were in close agreement with the actual values observed for cells lacking methionine: 27 min doubling time and 8% of protein synthesis capacity devoted to MetE. We verified experimentally that both decrease and increase in MetE production lead to slower growth (Fig. S5). Therefore, the expression of the key enzyme MetE is accurately tuned to allow the highest possible growth rate. Furthermore, the cost of expressing MetE is the main determinant for the slower growth rate when Met is limiting.

Our quantitative analysis of the Met pathway revealed a bottleneck step and its relationship to fitness. The same approach should be applicable for a broad range of cellular and engineered metabolic pathways, for which the control points are still largely unknown. In

addition, the global analysis of maximum reaction velocity (V_{\max}) can be used in concert with flux balance analysis (Price et al., 2004; Schuetz et al., 2012) to identify possible routes of metabolic flux at a given condition. More broadly, the global quantification of absolute enzyme concentration provides a transformative tool for studying cellular metabolism.

Discussion

We illustrate here the capacity to measure absolute synthesis rates for cellular proteins and its utility for deciphering the logic behind the design principles of biological networks. We identify the rules underlying the observed synthesis rates for many distinct classes of proteins. These include proportional synthesis for multi-protein complexes and hierarchical expression for common functional modules, both of which are made possible by finely tuned rates of translation initiation. We anticipate that there are many more principles embedded in this and similar datasets which will both elucidate the regime in which biochemical reactions operate, and provide a foundation for rational design of synthetic biological systems.

Our genome-wide dataset on protein synthesis rates also allows in-depth analysis of how cells optimize the use of limited resources. Specifically, these data revealed strategies for allocating limited space on DNA and limited protein synthesis capacity—transcription factors can be kept at low abundances without kinetic penalties by pre-binding to target sites, and the synthesis rate of a key enzyme that limits metabolic flux in the methionine biosynthetic pathway is optimized to achieve a maximal growth rate. Limited resources of various kinds pose constant challenges to all cells. Our approach reveals how the translational capacity of a cell is allocated in the face of these challenges, greatly expanding our ability to perform systems level analyses that were previously limited to selected proteins and pathways.

While our studies illustrate the role of precisely tuned protein synthesis rates in bacteria, our knowledge of how this translational control is achieved remains highly limited.

Understanding the control of translation initiation is both of fundamental importance and a prerequisite for quantitative design in synthetic biology. Yet our current approaches for predicting translation rates, based on strength of Shine-Dalgarno site and computed RNA structure (Salis et al., 2009), fail to accurately account for the observed differences in translation initiation rates (Fig. S6). Empirical measures of mRNA structures as they exist in the cell, in combination with our measures of translation efficiency (Table S4), could be a key tool in addressing this deficiency.

Although we focus on bacterial cells in this work, our approach to globally measure absolute protein synthesis rates has broader applicability. Any species that is amenable to ribosome profiling and has an annotated genome can be subject to this line of investigation; the growing list currently includes both gram-negative and gram-positive bacteria, budding yeast, nematodes, fruit fly, zebra fish, and mammals. For eukaryotes and multi-cellular organisms, our approach will likely reveal a distinct set of principles and constraints for optimizing the allocation of biosynthetic capacities. Furthermore, the breakdown of these principles under stress conditions, such as aneuploidy and temperature and chemical shock, will provide critical insight into the modes of failure and their rescue mechanisms.

References

Ades, S.E., Connolly, L.E., Alba, B.M., and Gross, C.A. (1999). The *Escherichia coli* sigma(E)-dependent extracytoplasmic stress response is controlled by the regulated proteolysis of an anti-sigma factor. *Genes Dev* 13, 2449-2461.

Akiyama, Y., Kihara, A., Tokuda, H., and Ito, K. (1996). FtsH (HflB) is an ATP-dependent protease selectively acting on SecY and some other membrane proteins. *J Biol Chem* 271, 31196-31201.

Alon, U., Surette, M.G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature* 397, 168-171.

Andersson, S.G., and Kurland, C.G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev* 54, 198-210.

Barkai, N., and Shilo, B.Z. (2007). Variability and robustness in biomolecular systems. *Mol Cell* 28, 755-760.

Batchelor, E., and Goulian, M. (2003). Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system. *Proc Natl Acad Sci U S A* 100, 691-696.

Baughman, G., and Nomura, M. (1983). Localization of the target site for translational regulation of the L11 operon and direct evidence for translational coupling in *Escherichia coli*. *Cell* 34, 979-988.

Blikstad, I., Nelson, W.J., Moon, R.T., and Lazarides, E. (1983). Synthesis and assembly of spectrin during avian erythropoiesis: stoichiometric assembly but unequal synthesis of alpha and beta spectrin. *Cell* 32, 1081-1091.

Brandman, O., Stewart-Ornstein, J., Wong, D., Larson, A., Williams, C.C., Li, G.W., Zhou, S., King, D., Shen, P.S., Weibezahn, J., et al. (2012). A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* 151, 1042-1054.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335, 552-557.

Brusilow, W.S., Klionsky, D.J., and Simoni, R.D. (1982). Differential polypeptide synthesis of the proton-translocating ATPase of *Escherichia coli*. *J Bacteriol* 151, 1363-1371.

Buttgereit, F., and Brand, M.D. (1995). A hierarchy of ATP-consuming processes in mammalian cells. *Biochem J* 312 (Pt 1), 163-167.

Cai, L., Dalal, C.K., and Elowitz, M.B. (2008). Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* 455, 485-490.

Dekel, E., and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436, 588-592.

- Dennis, P.P. (1974). In vivo stability, maturation and relative differential synthesis rates of individual ribosomal proteins in *Escherichia coli* B/r. *J Mol Biol* 88, 25-41.
- Dong, H., Nilsson, L., and Kurland, C.G. (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *J Bacteriol* 177, 1497-1504.
- Elf, J., Li, G.W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* 316, 1191-1194.
- Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science* 297, 1183-1186.
- Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., and Palsson, B.O. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3, 121.
- Fell, D. (1997). *Understanding the control of metabolism* (London, Portland Press).
- Gerdes, K., and Maisonneuve, E. (2012). Bacterial persistence and toxin-antitoxin loci. *Annu Rev Microbiol* 66, 103-123.
- Grossman, A.D., Straus, D.B., Walter, W.A., and Gross, C.A. (1987). Sigma 32 synthesis can regulate the synthesis of heat shock proteins in *Escherichia coli*. *Genes Dev* 1, 179-184.
- Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E.G., Berg, O.G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. *Science* 336, 1595-1598.
- Hart, Y., Madar, D., Yuan, J., Bren, A., Mayo, A.E., Rabinowitz, J.D., and Alon, U. (2011). Robust control of nitrogen assimilation by a bifunctional enzyme in *E. coli*. *Mol Cell* 41, 117-127.
- Hu, J.C., Sherlock, G., Siegele, D.A., Aleksander, S.A., Ball, C.A., Demeter, J., Gouni, S., Holland, T.A., Karp, P.D., Lewis, J.E., et al. (2014). PortEco: a resource for exploring bacterial biology through high-throughput data and analysis tools. *Nucleic Acids Res* 42, D677-684.
- Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 7, 1534-1550.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., 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.

- Larrabee, K.L., Phillips, J.O., Williams, G.J., and Larrabee, A.R. (1980). The relative rates of protein synthesis and degradation in a growing culture of *Escherichia coli*. *J Biol Chem* 255, 4125-4130.
- Lehnert, M.E., and Lodish, H.F. (1988). Unequal synthesis and differential degradation of alpha and beta spectrin during murine erythroid differentiation. *J Cell Biol* 107, 413-426.
- Lemaux, P.G., Herendeen, S.L., Bloch, P.L., and Neidhardt, F.C. (1978). Transient rates of synthesis of individual polypeptides in *E. coli* following temperature shifts. *Cell* 13, 427-434.
- Li, G.W., Berg, O.G., and Elf, J. (2009). Effects of macromolecular crowding and DNA looping on gene regulation kinetics. *Nat Phys* 5, 294-297.
- Li, G.W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 538-541.
- Li, G.W., and Xie, X.S. (2011). Central dogma at the single-molecule level in living cells. *Nature* 475, 308-315.
- McCarthy, J.E., and Gualerzi, C. (1990). Translational control of prokaryotic gene expression. *Trends Genet* 6, 78-85.
- Neidhardt, F.C., Bloch, P.L., and Smith, D.F. (1974). Culture medium for enterobacteria. *J Bacteriol* 119, 736-747.
- Nomura, M., Gourse, R., and Baughman, G. (1984). Regulation of the synthesis of ribosomes and ribosomal components. *Annu Rev Biochem* 53, 75-117.
- 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.
- Oromendia, A.B., Dodgson, S.E., and Amon, A. (2012). Aneuploidy causes proteotoxic stress in yeast. *Genes Dev* 26, 2696-2708.
- Papp, B., Pal, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194-197.
- Petersen, C. (1990). *Escherichia coli* ribosomal protein L10 is rapidly degraded when synthesized in excess of ribosomal protein L7/L12. *J Bacteriol* 172, 431-436.
- Price, N.D., Reed, J.L., and Palsson, B.O. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2, 886-897.
- Quax, T.E., Wolf, Y.I., Koehorst, J.J., Wurtzel, O., van der Oost, R., Ran, W., Blombach, F., Makarova, K.S., Brouns, S.J., Forster, A.C., et al. (2013). Differential translation tunes uneven production of operon-encoded proteins. *Cell Rep* 4, 938-944.

- Rigel, N.W., Ricci, D.P., and Silhavy, T.J. (2013). Conformation-specific labeling of BamA and suppressor analysis suggest a cyclic mechanism for beta-barrel assembly in *Escherichia coli*. *Proc Natl Acad Sci U S A* 110, 5151-5156.
- 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. *Nat Biotechnol* 27, 946-950.
- Savageau, M.A. (1977). Design of molecular control mechanisms and the demand for gene expression. *Proc Natl Acad Sci U S A* 74, 5647-5651.
- Schomburg, I., Chang, A., and Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* 30, 47-49.
- Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science* 336, 601-604.
- Schuller, H.J., Fortsch, B., Rautenstrauss, B., Wolf, D.H., and Schweizer, E. (1992). Differential proteolytic sensitivity of yeast fatty acid synthetase subunits alpha and beta contributing to a balanced ratio of both fatty acid synthetase components. *Eur J Biochem* 203, 607-614.
- Schwanhausser, B., Gossen, M., Dittmar, G., and Selbach, M. (2009). Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* 9, 205-209.
- Scott, M., Gunderson, C.W., Mateescu, E.M., Zhang, Z., and Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. *Science* 330, 1099-1102.
- Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature* 455, 58-63.
- Shapiro, A.L., Scharff, M.D., Maizel, J.V., and Uhr, J.W. (1966). Synthesis of excess light chains of gamma globulin by rabbit lymph node cells. *Nature* 211, 243-245.
- Shemorry, A., Hwang, C.S., and Varshavsky, A. (2013). Control of protein quality and stoichiometries by N-terminal acetylation and the N-end rule pathway. *Mol Cell* 50, 540-551.
- Shinar, G., Dekel, E., Tlusty, T., and Alon, U. (2006). Rules for biological regulation based on error minimization. *Proc Natl Acad Sci U S A* 103, 3999-4004.
- Shinar, G., Milo, R., Martinez, M.R., and Alon, U. (2007). Input output robustness in simple bacterial signaling systems. *Proc Natl Acad Sci U S A* 104, 19931-19935.
- Springer, M., Weissman, J.S., and Kirschner, M.W. (2010). A general lack of compensation for gene dosage in yeast. *Mol Syst Biol* 6, 368.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T., Hein, M.Y., Huang, S.X., Ma, M., Shen, B., Qian, S.B., Hengel, H., et al. (2012). Decoding human cytomegalovirus. *Science* 338, 1088-1093.

Swain, P.S. (2004). Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J Mol Biol* 344, 965-976.

Torres, E.M., Williams, B.R., and Amon, A. (2008). Aneuploidy: cells losing their balance. *Genetics* 179, 737-746.

Tyedmers, J., Mogk, A., and Bukau, B. (2010). Cellular strategies for controlling protein aggregation. *Nat Rev Mol Cell Biol* 11, 777-788.

von Dassow, G., Meir, E., Munro, E.M., and Odell, G.M. (2000). The segment polarity network is a robust developmental module. *Nature* 406, 188-192.

von Hippel, P.H. (2007). From "simple" DNA-protein interactions to the macromolecular machines of gene expression. *Annu Rev Biophys Biomol Struct* 36, 79-105.

Winter, R.B., Berg, O.G., and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli* lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry-Us* 20, 6961-6977.

Yamaguchi, Y., Park, J.H., and Inouye, M. (2011). Toxin-antitoxin systems in bacteria and archaea. *Annu Rev Genet* 45, 61-79.

You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C.W., Wang, Y.P., Lenz, P., Yan, D., and Hwa, T. (2013). Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature*.

Figure 1. Absolute Quantification of Protein Synthesis Rates

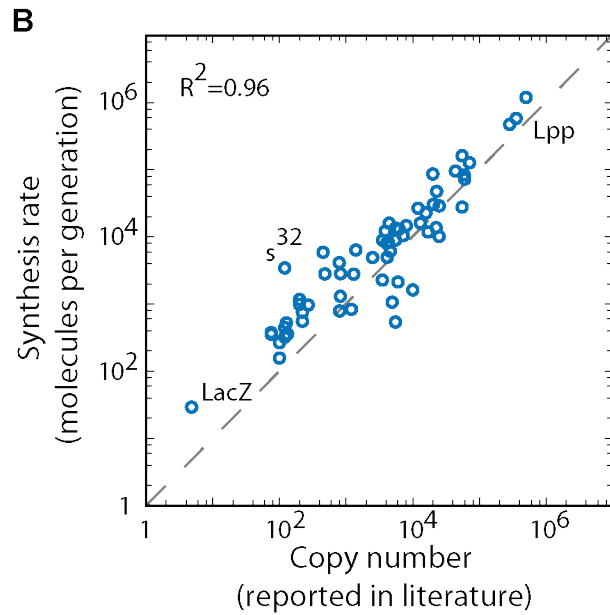
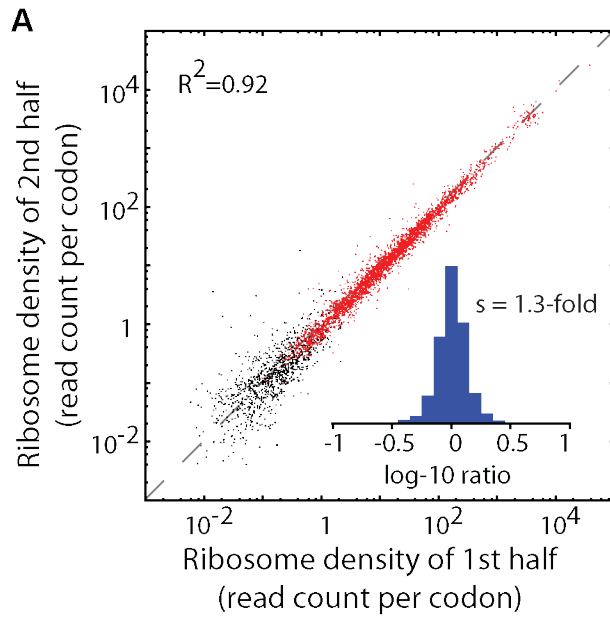


Figure 1. Absolute Quantification of Protein Synthesis Rates

(A) Effect of translational pausing on average ribosome density. Average ribosome density is plotted for the first and second half of each gene. The Pearson correlation for genes with at least 64 reads aligned to both halves (red) is $R^2 = 0.92$. The inset shows the distribution of the fold-difference between the second and the first halves ($N = 2,870$, $SD = 1.3$ fold).

(B) Agreement between published protein copy numbers and absolute synthesis rates. The copy numbers of 62 proteins which have been individually quantified in the literature are plotted against the absolute protein synthesis rates (Pearson correlation $R^2 = 0.96$).

Figure 2. Proportional Synthesis of Multi-Protein Complexes

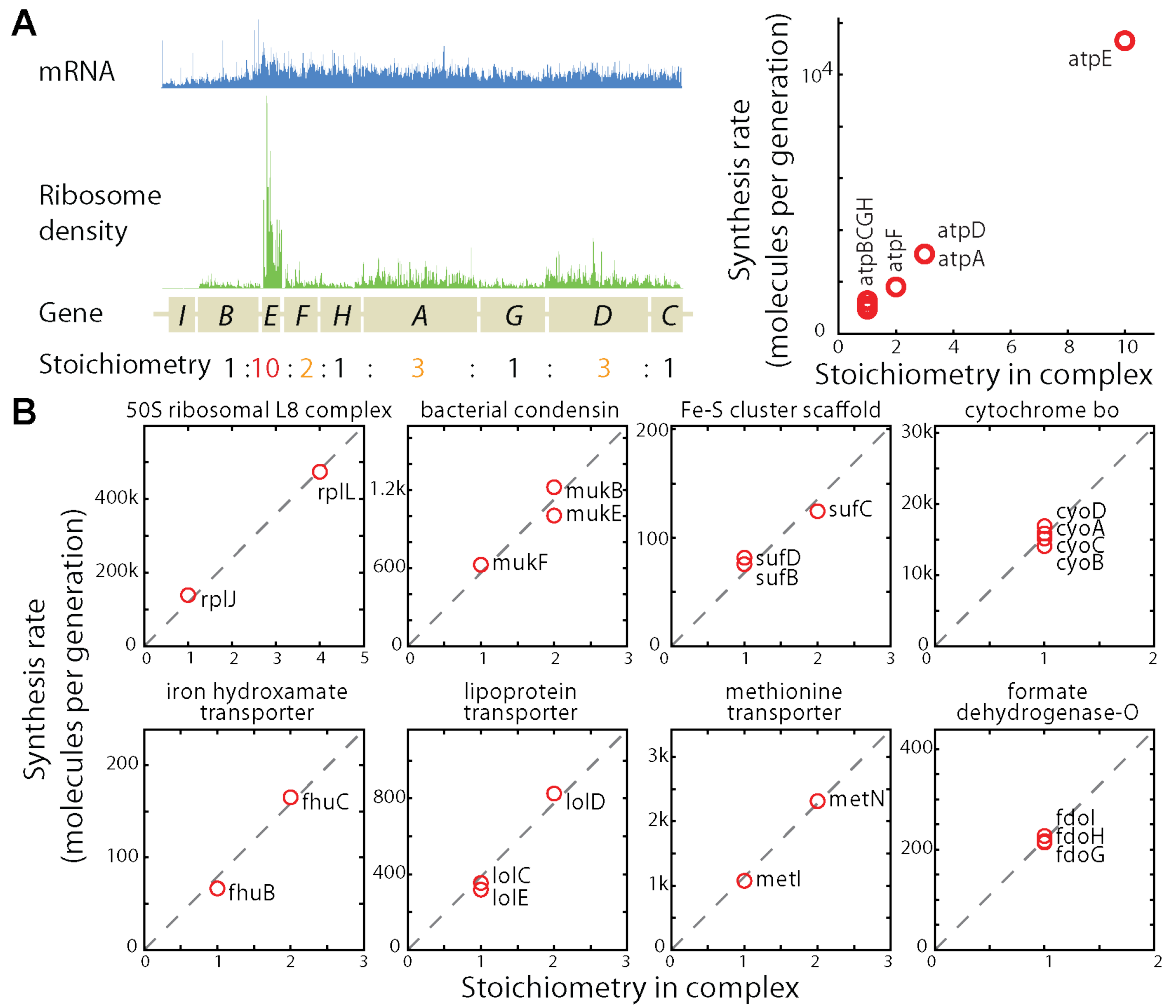


Figure 2. Proportional Synthesis of Multi-Protein Complexes (continued)

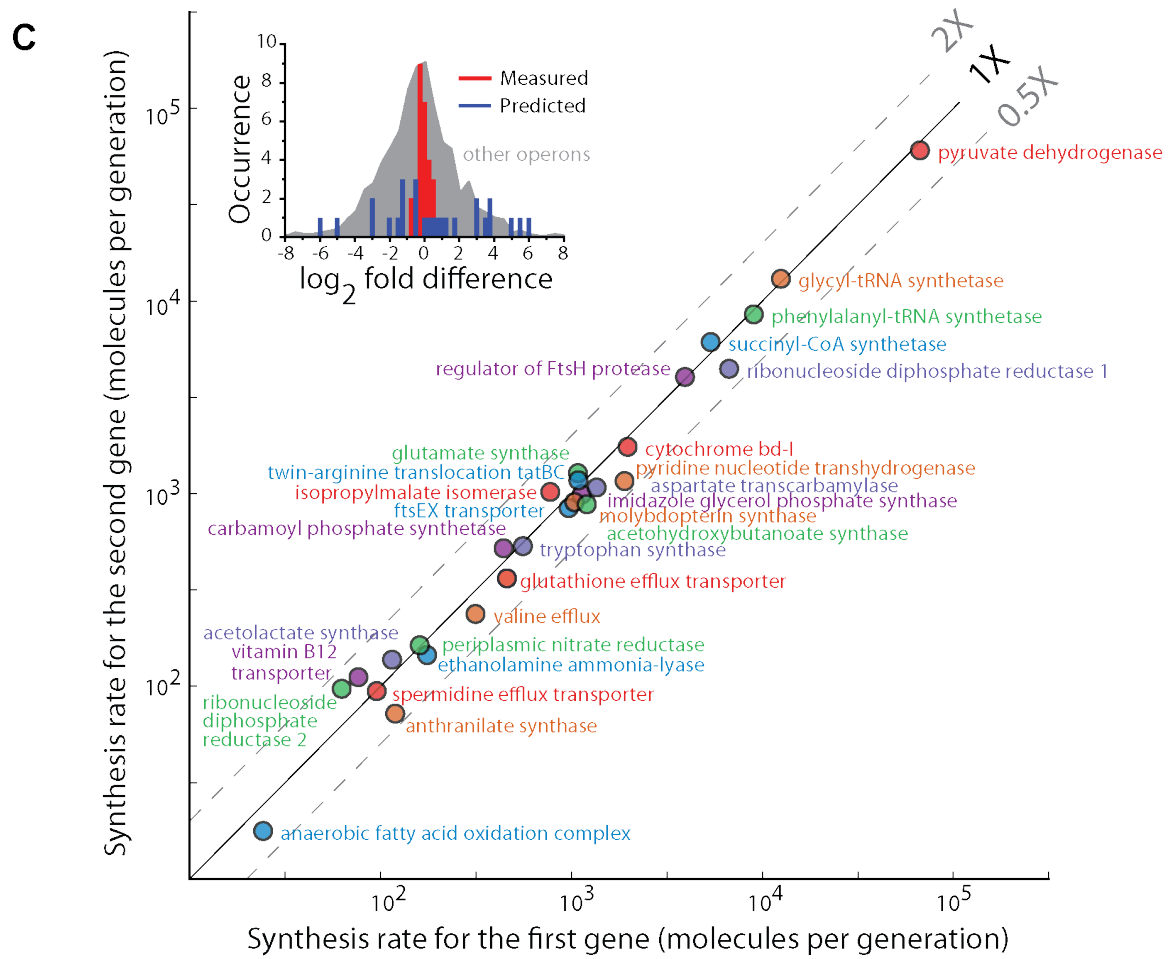


Figure 2. Proportional Synthesis of Multi-Protein Complexes

(A) Translation rates reflecting subunit stoichiometry for the ATP operon. Eight subunits of the F_0F_1 ATP synthase are expressed from a polycistronic mRNA, whose level as measured by RNA-seq is shown in blue. Each subunit is associated with different levels of ribosome density (green), and the average density is proportional to the subunit stoichiometry (right).

(B) Proportional synthesis for a diverse range of complexes. Synthesis rates are plotted as a function of the subunit stoichiometry for multi-protein complexes whose subunits are encoded in the same operon. Complexes with different subunit stoichiometry or more than two subunits are included here (also see panel (C)). The dashed line indicates the best-fit that crosses the origin.

(C) Proportional synthesis for complexes with two equimolar subunits. Each complex is plotted for the synthesis rates of the two subunits, with the earlier (later) gene in the operon on the horizontal (vertical) axis. 28 equimolar and co-transcribed complexes, covering 4 orders of magnitude in expression level, are plotted here. Inset shows the histogram of fold-difference between the synthesis rates of the two subunits. Our experimental results are shown in red, and the predicted values based on a thermodynamic model considering the sequence surrounding translation initiation sites are shown in blue (Salis et al., 2009). The distribution of the differences in translation rates for all other operons is shown in gray. Panels B and C show complexes whose subunits are encoded on a single polycistronic operon.

Figure 3. Proportional Synthesis for Complexes in Yeast

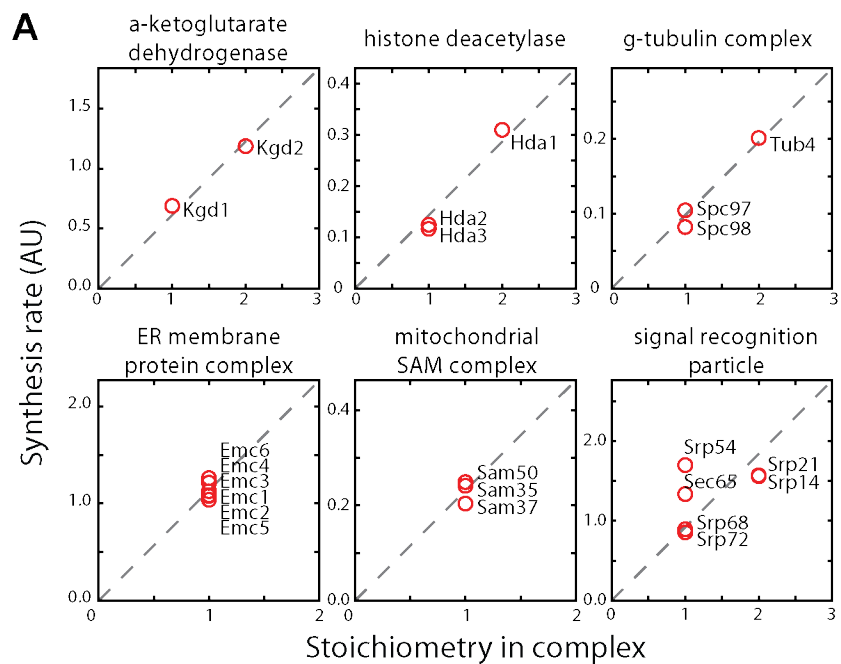


Figure 3. Proportional Synthesis for Complexes in Yeast (continued)

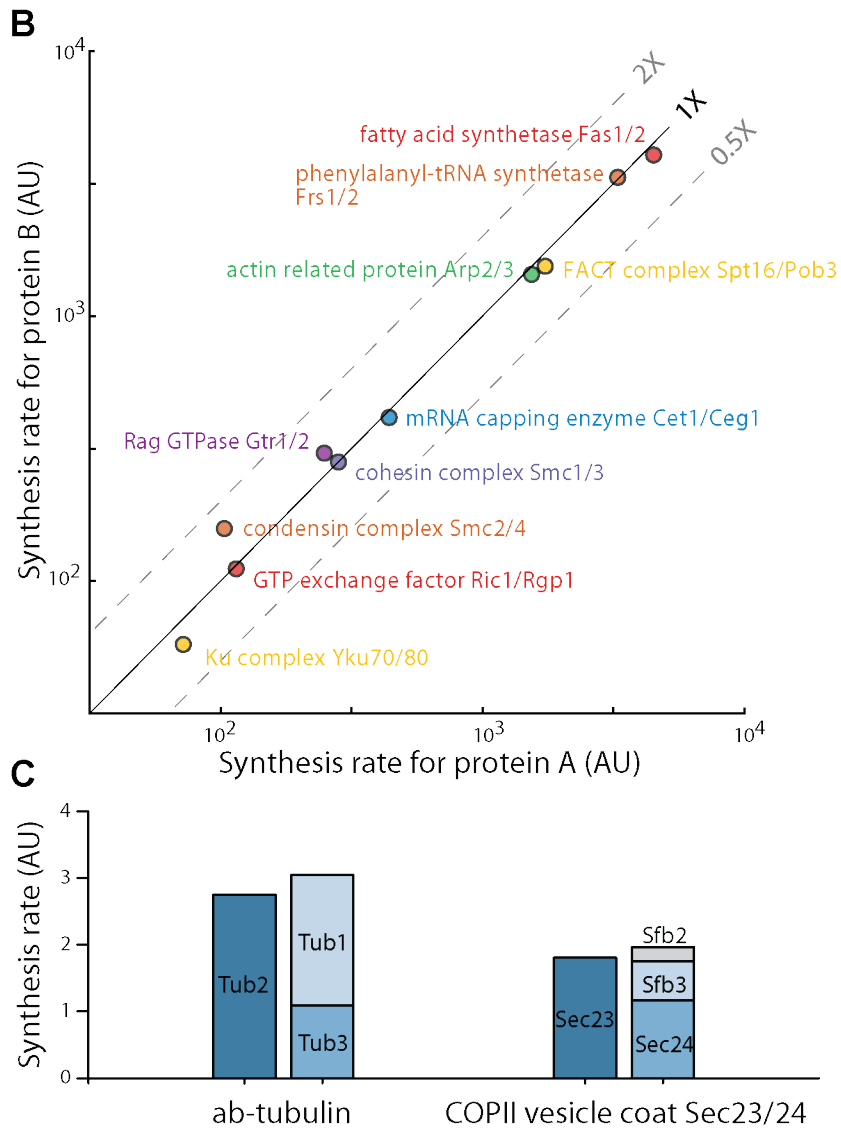


Figure 3. Proportional Synthesis for Complexes in Yeast (continued)

(A) Proportional synthesis for multi-protein complexes in *S. cerevisiae*. Synthesis rates are plotted as a function of the subunit stoichiometry for complexes with more than two subunits. For the signal recognition particle, four subunits (Srp14/21/68/72) are synthesized according to their stoichiometry, and the other two are exceptions.

(B) Proportional synthesis for heterodimeric complexes in *S. cerevisiae*. Each complex is plotted for the synthesis rate of the two subunits.

(C) Proportional synthesis for complexes with paralogous subunits. For each complex, the subunits that can substitute each other are plotted in the same column.

Figure 4. Hierarchical Expression for Functional Modules

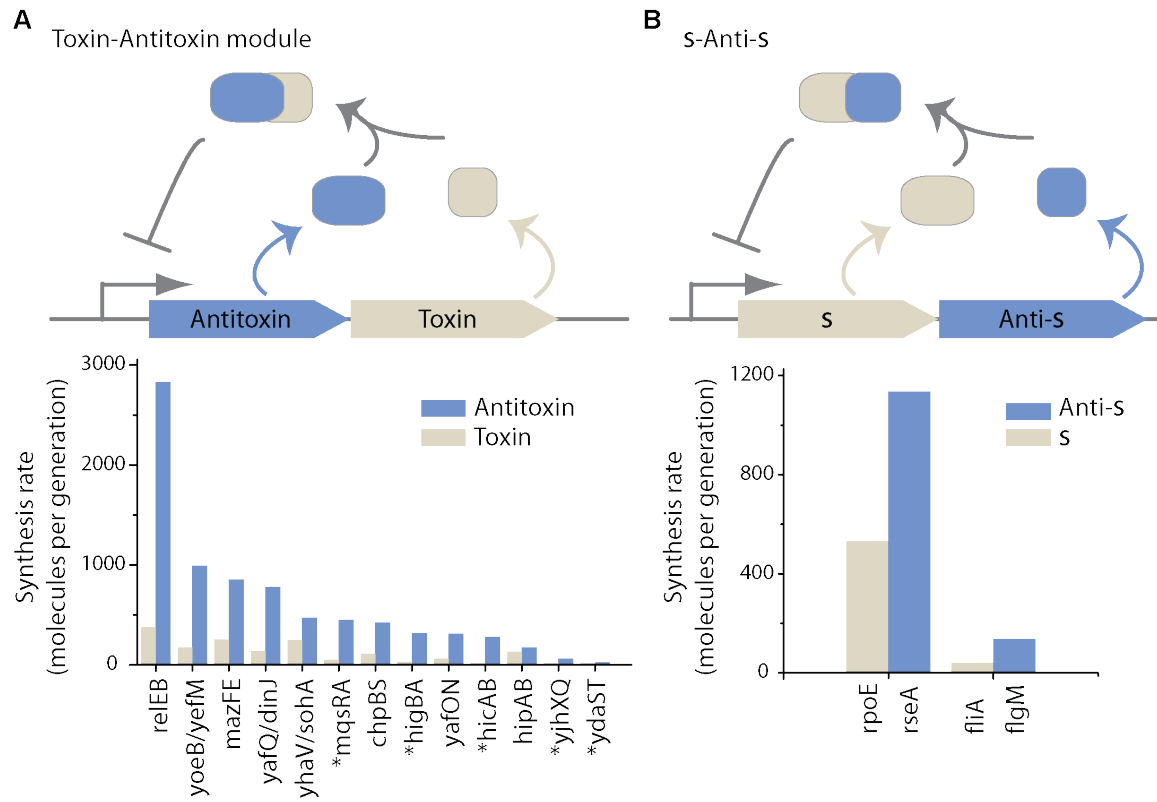


Figure 4. Hierarchical Expression for Functional Modules (continued)

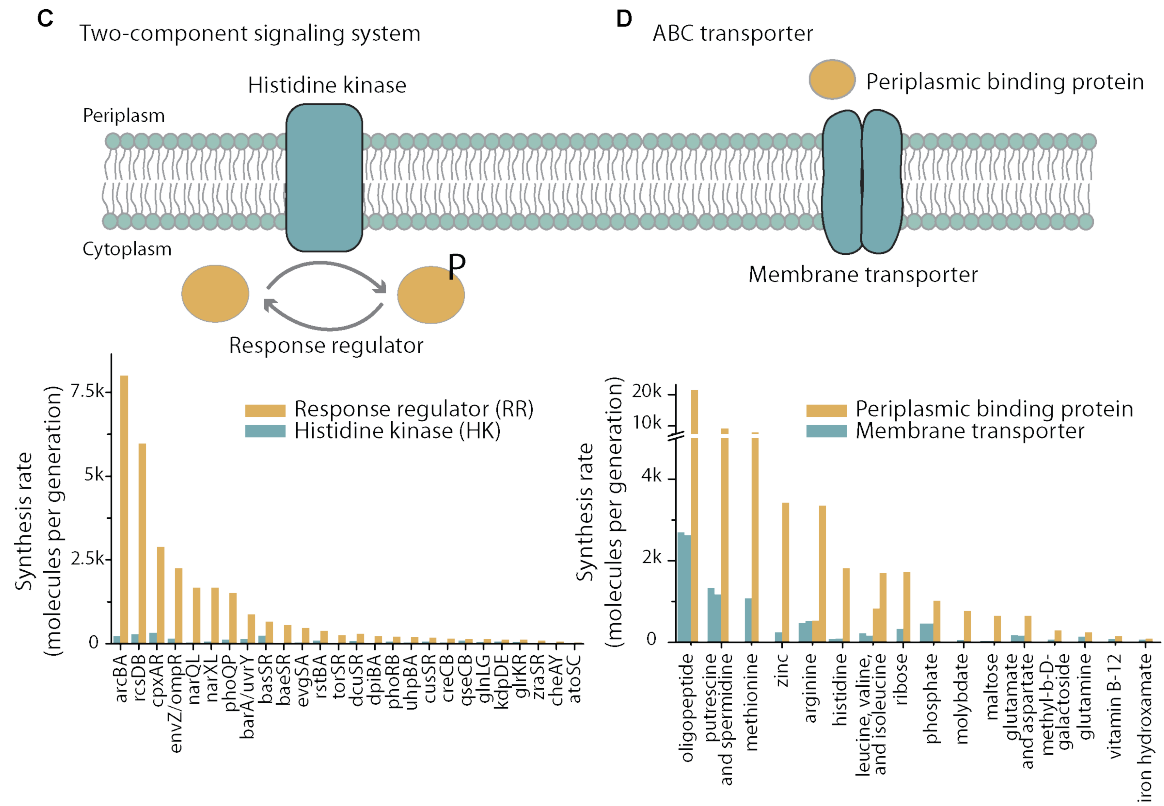


Figure 4. Hierarchical Expression for Functional Modules

(A) Synthesis rates for toxin-antitoxin (TA) modules. *E. coli* contains 12 type II TA systems that are each expressed from a polycistronic mRNA. (The order of genes differs among systems.) The anti-toxin protein binds to and inhibits the toxin protein, while repressing its own transcription. The synthesis rates for each system are plotted (bottom). Modules with the toxin gene preceding the antitoxin gene in the operon is marked by asterisk.

(B) Synthesis rates for sigma-anti-sigma factors modules. The anti-sigma factor binds to and inhibits the sigma factor, preventing transcription from the promoter driven by the corresponding sigma factor. The synthesis rates for each systems are plotted (bottom).

(C) Synthesis rates for two-component signaling systems. Bacterial two-component signaling system consists of a membrane-bound histidine kinase and the cognate response regulator. The synthesis rates for 26 two-component systems in *E. coli* are plotted (bottom).

(D) Synthesis rates for ATP-binding cassette (ABC) transporters. An ABC transporter consists of a core membrane transporter, an ATP-binding domain, and the corresponding periplasmic binding proteins. The synthesis rates for each transporter are plotted (bottom).

Figure 5. Composition of the *E. coli* Proteome

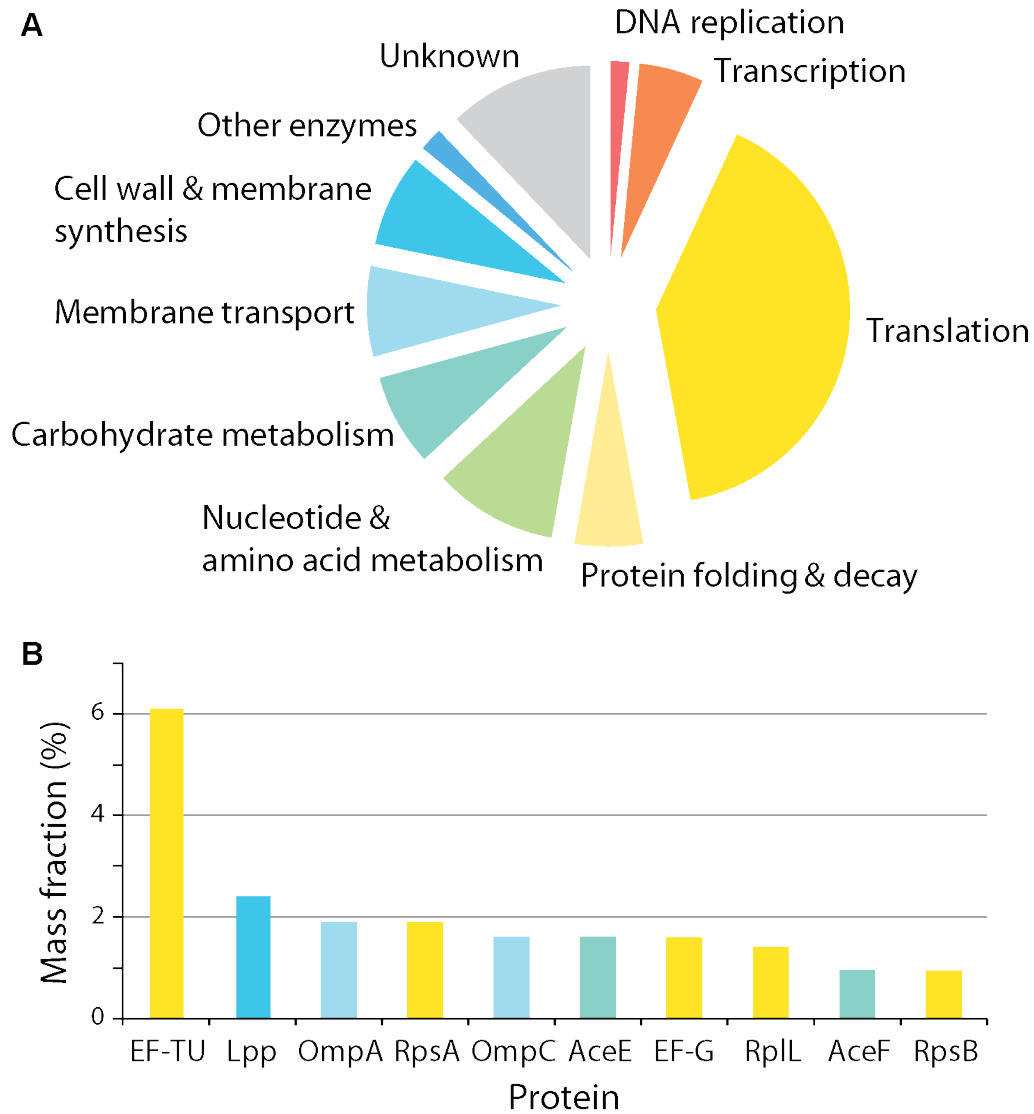


Figure 5. Composition of the *E. coli* Proteome

(A) Break down of the proteome by functions. The mass-fraction of the proteome that is devoted to specific biological functions is plotted as a pie chart. The copy numbers were estimated for *E. coli* grown in rich defined medium (Methods).

(B) Ten proteins with the largest mass-fraction in the proteome. The color used for each protein corresponds to the biological function indicated in A.

Figure 6. Abundance of Transcription Factors (TFs)

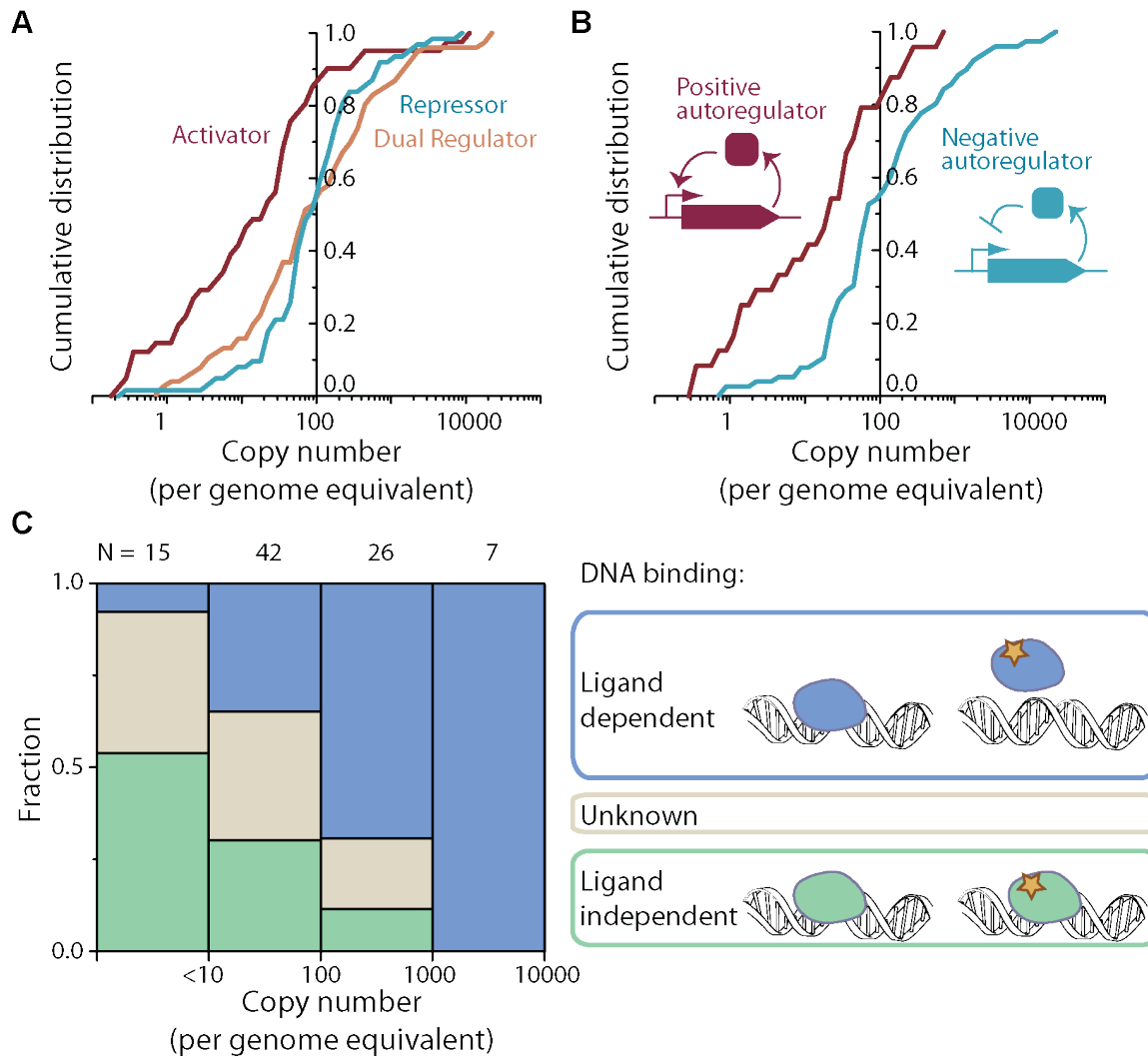


Figure 6. Abundance of Transcription Factors (TFs)

(A) Cumulative distribution of abundance for transcriptional activators, repressors, and dual regulators. The cumulative distribution for each class of TF is plotted as a function of the copy number per genome equivalent.

(B) Cumulative distribution of abundance for autoregulators. The cumulative distributions for positive- and negative-autoregulators are plotted as a function of the copy number per genome equivalent.

(C) Ligand dependence of target binding. Among TFs whose abundance fall into a given range, the fraction that binds to the target site in a ligand-dependent way is shown in blue, and the fraction that binds to the target site independent of ligands is shown in green. The number of transcription factors analyzed is indicated above each bin.

Figure 7. Quantitative Analysis of the Methionine Biosynthesis Pathway

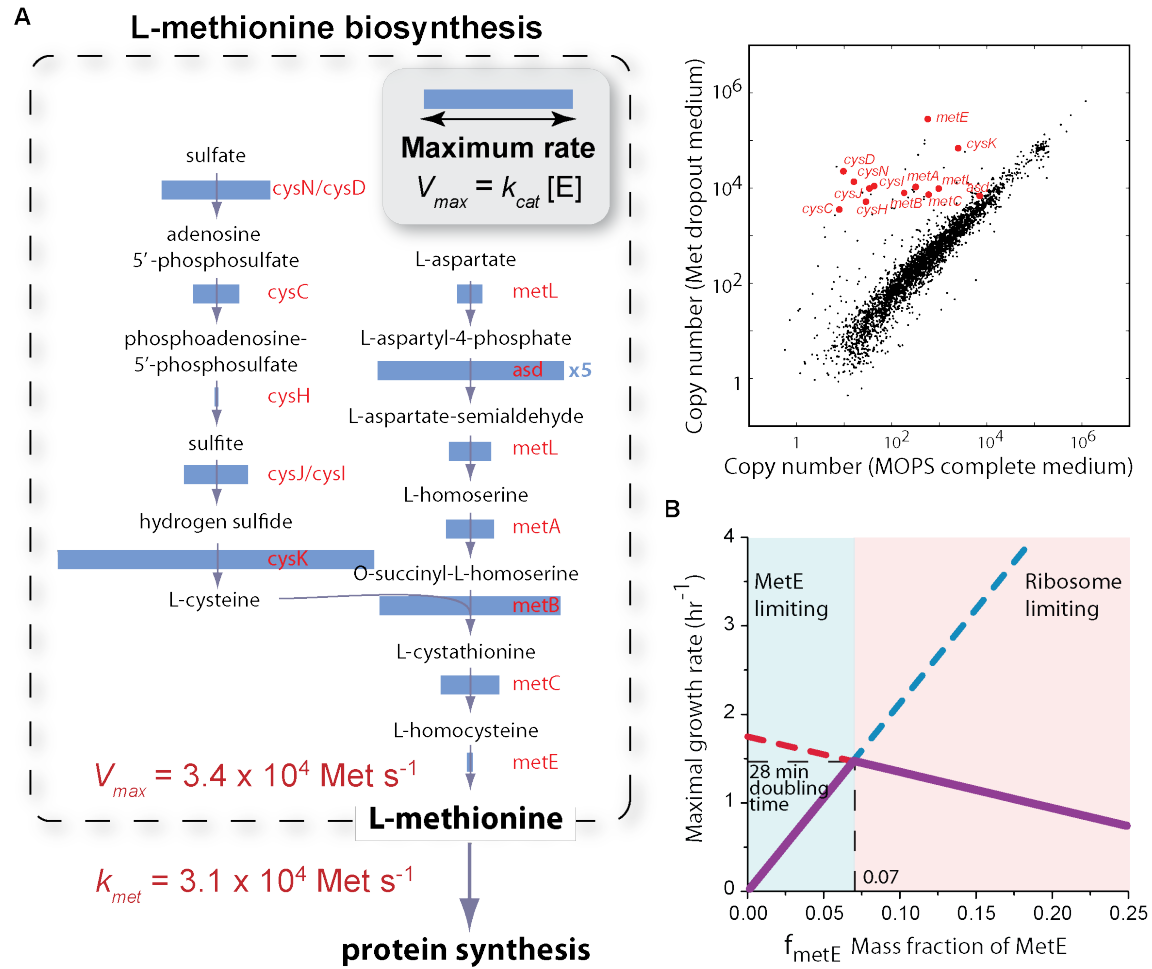


Figure 7. Quantitative Analysis of the Methionine Biosynthesis Pathway

(A) Maximal reaction rates for the intermediate steps. For each step of the pathway, the maximal reaction rate (V_{max}), inferred from the enzyme abundance *in vivo* and the turnover number measured *in vitro*, is shown as the width of the blue bar, unless no *in vitro* data were available. The last step that is catalyzed by the enzyme MetE has $V_{max} = 34,000$ Met/s/cell, whereas the flux of methionine into protein synthesis is 31,000 Met/s/cell. The scatter plot on the right shows up-regulation of these enzymes in media without methionine.

(B) Model predicting the optimal MetE level. In a model that considers the cost and benefit of MetE expression, the maximal growth rate is plotted as a function of the mass fraction of MetE in the proteome. The cost due to competition with new ribosome synthesis is shown in red, and the benefit from increased methionine flux is shown in blue. The maximal growth rate is highest (28 min) when the mass fraction of MetE is $\sim 7\%$. This prediction agrees with experimental results.

Figure S1. Adjustment to ribosome density based on sequence- and position-specific variation in translation elongation rates.

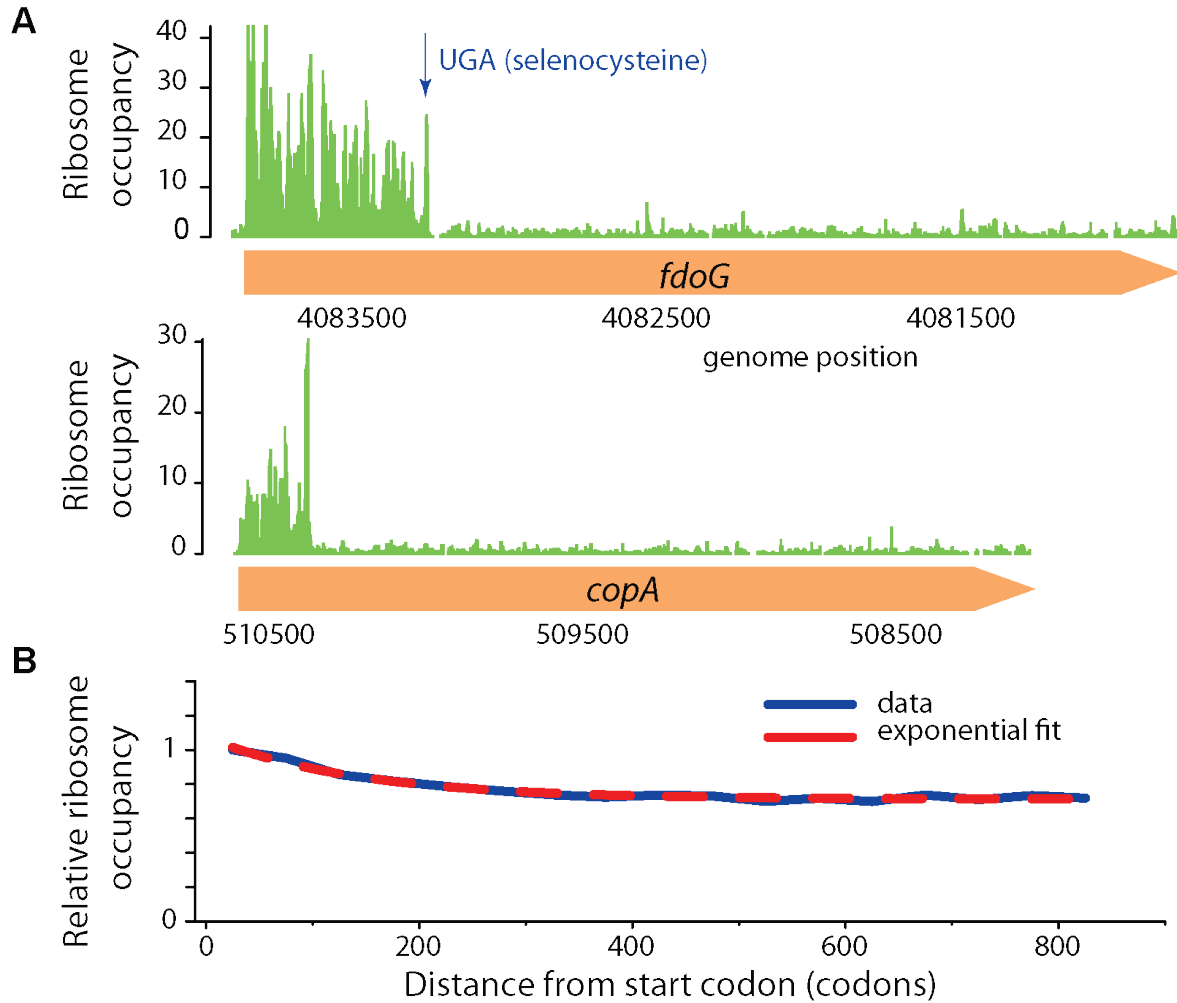


Figure S1. Adjustment to ribosome density based on sequence- and position-specific variation in translation elongation rates. (continued)

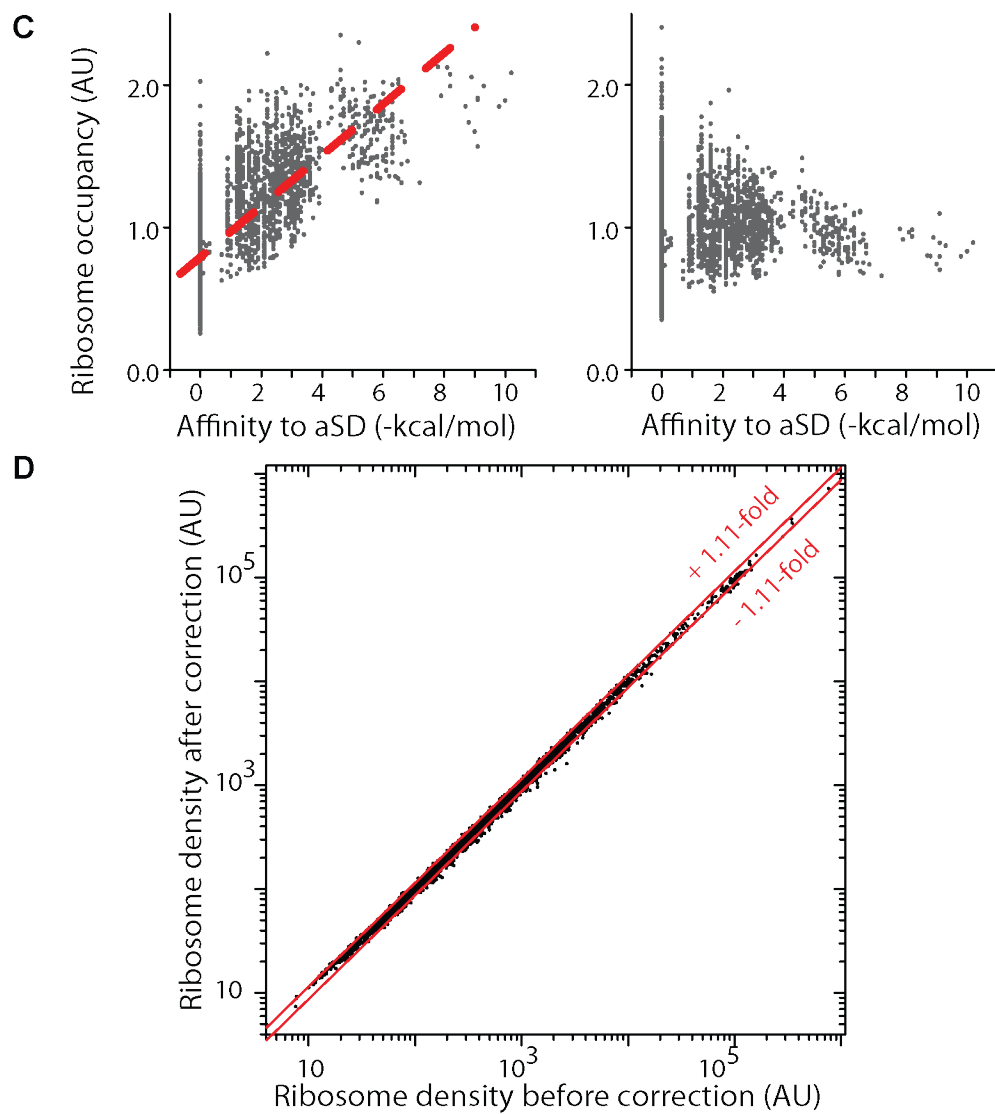


Figure S1. Adjustment to ribosome density based on sequence- and position-specific variation in translation elongation rates.

(A) Atypical genes with large drop-off in ribosome occupancy. For selenoproteins, e.g. FdoG, we observed reduced ribosome occupancy after the selenocysteine codon. Only the region after the selenocysteine codon was used to calculate the average ribosome density. Abrupt decrease in ribosome occupancy on a few other genes, such as *copA*, could indicate novel translational events.

(B) Correction for elevated ribosome occupancy towards the beginning of open reading frames. The slight increase in occupancy (blue) was modeled as an exponential function (red). The fitting parameters were used to adjust the position-dependent ribosome occupancy.

(C) Correction for translational pausing induced by internal Shine-Dalgarno-sequences. The average ribosome occupancy downstream from a hexanucleotide sequence is plotted against its affinity to the anti-Shine-Dalgarno sequence. The observed relationship (left) was fitted with a linear function (red). The fitting parameters were used to adjust the sequence-dependent ribosome occupancy, so that the result is independent of the strength of Shine-Dalgarno sequences (right).

(D) Effects of the corrections for local variation in translation elongation rates. For each gene, the average ribosome density before and after corrections is plotted. The standard deviation for the differences is 1.11 fold.

Figure S2. Comparison of published quantitative proteomics measurements and individually measured protein copy number.

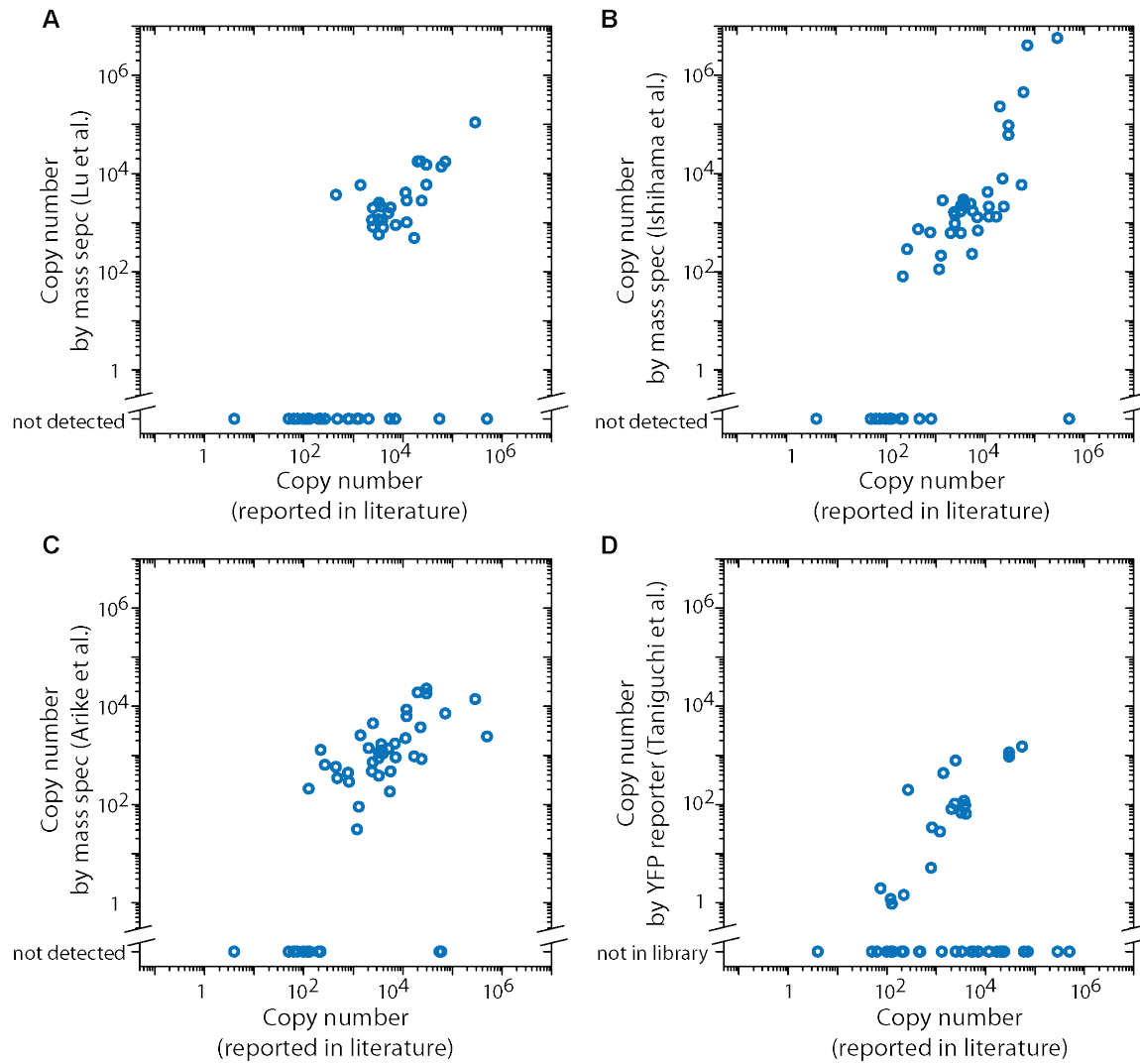


Figure S2. Comparison of published quantitative proteomics measurements and individually measured protein copy number.

(A) Proteomics data using absolute protein expression (APEX) profiling based on mass spectrometry (Lu et al., 2007).

(B) Proteomics data using exponentially modified protein abundance index (emPAI) based on mass spectrometry (Ishihama et al., 2008).

(C) Proteomics data using intensity-based absolute quantification (iBAQ) based on mass spectrometry (Ariike et al., 2012). We note that the data in (A-C) were obtained using label-free quantification. Current development in other absolute quantification methods using isotopic labeling and synthetic peptides as standards could provide improvements in accuracy and coverage (Hanke et al., 2008; Picotti et al., 2009).

(D) Proteomics data using a YFP-fusion library (Taniguchi et al., 2010). The library was constructed for ~25% of the genome. The measurements were performed at a lower growth rate (150 minutes per doubling) compared to other reports, which gave rise to lower protein abundance in general.

Figure S3. Proportional synthesis for other multi-protein complexes

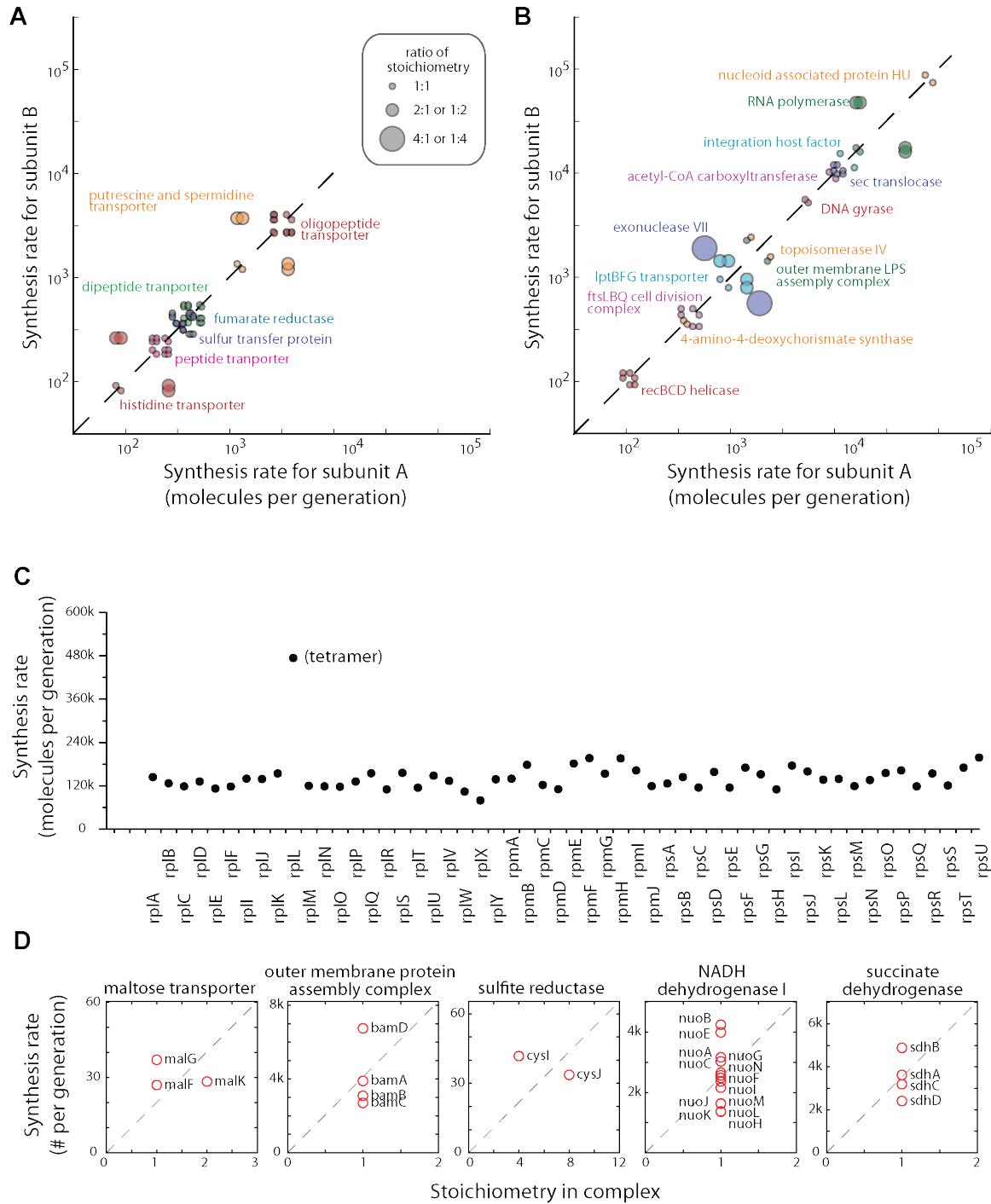


Figure S3. Proportional synthesis for other multi-protein complexes.

(A) Proportional synthesis for complexes whose members are encoded in the same operon.

Complexes not included in Fig. 2B are shown here. The synthesis rate for each pair of subunits in the complex is plotted, with the identity of the complex indicated by the color code. The size of the symbol reflects the ratio of stoichiometry between the pair. Each pair is plotted twice with different order.

(B) Proportional synthesis for complexes whose members are encoded in more than one operon.

The size of symbols is the same as in (A). Inset shows synthesis rates for ribosomal proteins. For some of the ribosomal protein with equal stoichiometry, proportional synthesis may be achieved by a combination of translational coupling and auto-regulation.

(C) Proportional synthesis for ribosomal proteins. All proteins, except RplL (L7/L12), have the stoichiometry of one per ribosome.

(D) Exceptions to proportional synthesis. Five complexes do not follow proportional synthesis out of 64 complexes. The synthesis rates relative to the stoichiometry are plotted here. Subunits of the maltose transporter and the BAM complex are translated from different mRNA, whereas the other three complexes are translated from the same polycistronic mRNA.

Figure S4. Proportional synthesis at 10°C, mRNA levels, and gene order

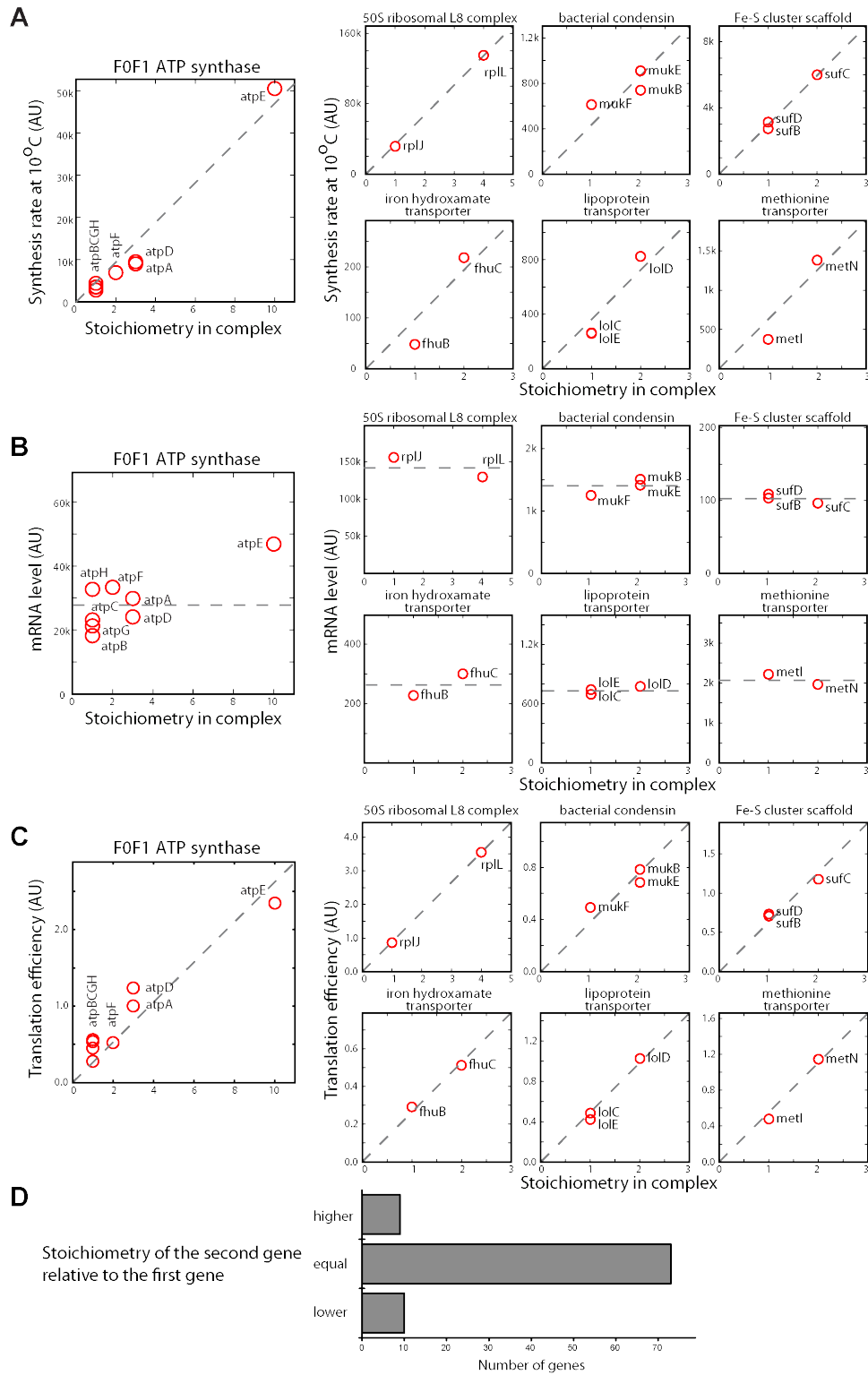


Figure S4. Proportional synthesis at 10°C, mRNA levels, and gene order

(A) Proportional synthesis at 10°C. Synthesis rates relative to stoichiometry are plotted for complexes expressed from the same operon. Experiment was performed at 50 hours after shifting the culture to 10°C. The dashed line indicates the best-fit that crosses the origin.

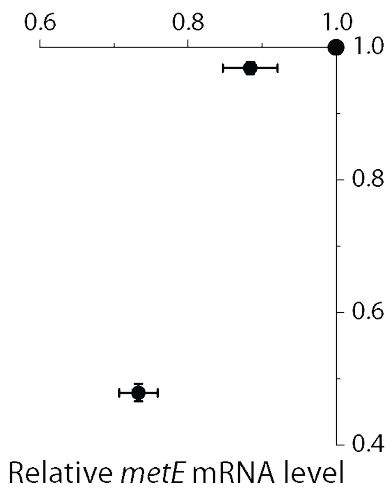
(B) mRNA levels for multi-protein complexes. The average transcript level for each gene is plotted against the stoichiometry in the complex. The dashed lines indicate the average transcript levels. Complexes with subunits expressed from the same polycistronic operon with uneven stoichiometry are shown here. Small variation in mRNA level could be due to alternative transcription start site or differential degradation. Overall, the mRNA levels are similar across different subunit and are not proportional to the stoichiometry.

(C) Translation efficiency for multi-protein complexes. The rate of protein synthesis per mRNA, as measured by protein synthesis rates (from ribosome profiling) divided by mRNA levels (from mRNA-seq), is plotted against the stoichiometry in the complex. The dashed line indicates the best-fit that crosses the origin. In combination with (B), the difference in protein production is mainly determined at the translational level.

(D) Gene order and ratio of translation rates. For the complexes analyzed in this work, the relative stoichiometry between the gene products and their order in the operon is shown in the histogram. The preceding gene product has similar likelihood to have higher and lower stoichiometry relative to the following gene product.

Figure S5. Predicted strength of ribosome binding sites and observed translation efficiency.

A MetE knockdown



B MetE overexpression

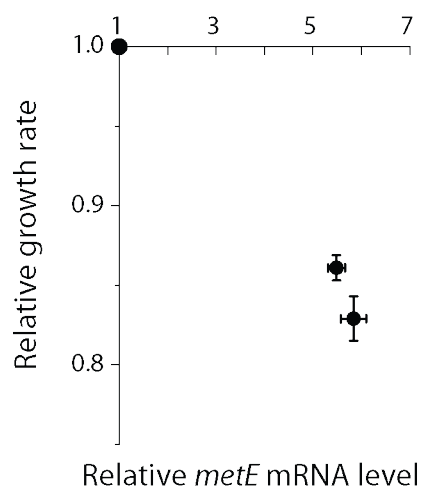


Figure S5. Predicted strength of ribosome binding sites and observed translation efficiency.

Prediction using the sequence near the translational start site was based on the model established by Salis *et al.* (Salis et al., 2009). Translation efficiency was estimated from the ribosome footprint density relative to the mRNA level. The small degree of correlation is mostly explained by the predicted secondary structure of mRNA, and not by the strength of Shine-Dalgarno sequences.

Figure S6. Effect of MetE level on growth rate.

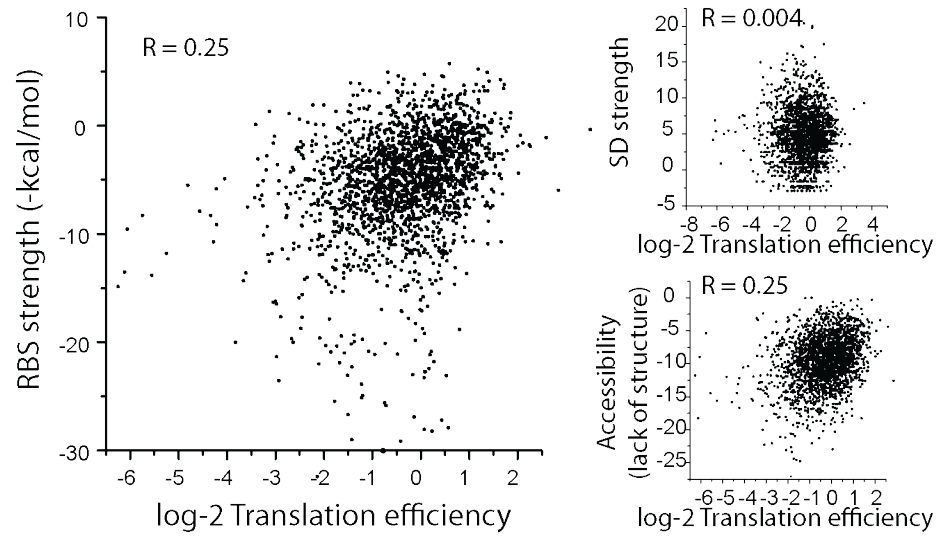


Figure S6. Effect of MetE level on growth rate.

(A) Effect of MetE knockdown. The growth rate relative to control is plotted as a function of MetE level. The transcription of *metE* is reduced using CRISPRi knockdown. Both growth rates and levels of *metE* mRNA are relative to control experiment with sgRNA targeting RFP instead of *metE*. (x-error bars: standard deviation of the mean, N=3, y-error bars: standard error)

(B) Effect of MetE overexpression. The growth rate relative to control is plotted as a function of MetE level. The transcription of ectopic *metE* is induced using a tetracycline-inducible promoter. Both growth rates and levels of *metE* mRNA are relative to control experiments with empty vector. (x-error bars: standard deviation of the mean, N=3, y-error bars: standard error)

Extended Experimental Procedures

Strain and growth conditions

E. coli K-12 strain MG1655 was used for this study. All cultures were based on MOPS media with 0.2% glucose (Teknova), with either full supplement (Neidhardt et al., 1974), full supplement without L-methionine, or no supplement. . An overnight liquid culture was diluted 400-fold into 200 ml fresh media. The culture was kept in a 2.8-liter flask at 37°C with aeration (180 rpm) until OD₆₀₀ reached 0.3. For the experiment at 10°C, the culture in M9 complete media with 0.2% glucose and amino acids except L-methionine was transferred from 37°C to 10°C with continuous shaking when OD₆₀₀ reached 0.12. Cells were harvested 50 hours later when OD₆₀₀ reached 0.43. The doubling time at 37°C is 21.5±0.4 minutes in fully supplemented MOPS media, 26.5±1.1 minutes in the methionine dropout medium, and 56.3±0.5 minutes in minimal medium. The results presented in this work are based on MOPS complete media unless otherwise mentioned.

Ribosome profiling

Bacterial ribosome profiling was performed as described in detail previously (Li et al., 2012; Oh et al., 2011) with the following modifications. 200 ml of cell culture was rapidly filtered at 37°C by passing through a nitrocellulose filter with 220 nm pore size (GE MicronSep). Cell pellets were rapidly collected using a pre-warmed metal table crumber, flash frozen in liquid nitrogen, and combined with 650 µl of frozen droplets of lysis buffer (10 mM MgCl₂, 100 mM NH₄Cl, 20 mM Tris pH 8.0, 0.1% NP-40, 0.4% Triton X-100, 100 U/ml DNase I, 0.5 U/µl Suprase-In, 1 mM chloramphenicol). Cells and lysis buffer were pulverized in 10 ml canisters (Retsch) pre-chilled in liquid nitrogen using Qiagen TissueLyser II (5 cycles of 3 minutes at 15

Hz). Pulverized lysate was thawed on ice and clarified by centrifugation at 20,000 ref for 10 minutes at 4°C. 5 mM CaCl₂ was added to the clarified lysate containing 0.5 mg of RNA which was then digested with 750 U of micrococcal nuclease (Roche) at 25°C for 1 hr. The reaction was quenched by adding EGTA to 6 mM and moved on ice.

The monosome fraction, following nuclease digestion to create footprints, was collected using sucrose gradient and hot-phenol extraction. Ribosome-protected mRNA fragments were isolated by size excision on a denaturing polyacrylamide gel (15%, TBE-Urea, Invitrogen). Fragments with size ranging from 15 to 45 nucleotides were excised from the gel. The 3' end of footprints was dephosphorylated using 20 units of T4 polynucleotide kinase (New England Biolabs) at 37°C for one hour. Five picomoles of footprints were ligated to 1 µg of 5' adenylated and 3'-end blocked DNA oligo (/5rApp/CTGTAGGCACCATCAAT/3ddc, Integrated DNA Technologies) using truncated T4 RNA ligase 2 K277Q at 37°C for 2.5 hours. The ligated product was purified by size excision on a 10% TBE-Urea polyacrylamide gel (Invitrogen). cDNA was generated by reverse transcription using Superscript III (Invitrogen) at 50°C for 30 minutes with primer o225-Link1 (5phos/GATCGTCGGACTGTAGAACTCTGAACCTGTTCGGTGGTCGCCGTATCATT/iSp18 /CACTCA/iSp18/CAAGCAGAAGACGGCATAACGAATTGATGGTGCCTACAG), and isolated by size excision on a 10% TBE-Urea polyacrylamide gel (Invitrogen).

Single-stranded cDNA was circularized using CircLigase (Epicentre) at 60°C for 2 hours. Ribosomal RNA fragments were removed using biotin-linked DNA oligos (5Biosg/TCATCTCCGGGGGTAGAGCACTGTTTCG,5Biosg/GGCTAAACCATGCACCGAA GCTGCGGCAG,5Biosg/AAGGCTGAGGCGTGATGACGAGGCACT,5Biosg/CGGTGCTGA AGCAACAAATGCCCTGCTT) and MyOne Streptavidin C1 Dynabeads (Invitrogen). After

being purified using isopropanol precipitation, the remaining cDNA was amplified using Phusion DNA polymerase (Finnzymes) with o231 primer (CAAGCAGAAGACGGCATAACGA) and indexing primers

(AATGATACGGCGACCACCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNCGACAGGTTTCAGAGTTC). After 8-10 rounds of PCR amplification, the product was selected by size excision on a 8% TB polyacrylamide gel (Invitrogen).

Sequencing was performed on an Illumina HiSeq 2000. Bowtie v. 0.12.0 was used for sequence alignment to the reference genomes NC_000913.fna obtained from NCBI Reference Sequence Bank. The footprint reads with size between 20 to 42 nucleotides in length were mapped to the genome using the center-weighted approach; for each footprint read, the center residues that are at least 10 nucleotides away from either ends were given the same score, which is weighted by the length of the fragment. The dataset is deposited in the Gene Expression Omnibus (GEO) under accession number GSE53767.

Ribosome profiling data for *S. cerevisiae* S288C was obtained from Brandman *et al.* (Brandman et al., 2012). For paralogous genes in Figure 3C, we used regions with at least one nucleotide difference within a 28-nucleotide window to calculate the synthesis rates.

mRNA sequencing

Pulverized cell lysate described in the previous section was collected and RNA was extracted using hot phenol extraction. Ribosomal RNA was removed by subtractive hybridization using MICROBExpress (Ambion). Small RNAs were removed using MEGAclear purification kit (Ambion). The remaining mRNA was fragmented using alkaline hydrolysis (50 μ l of 44 mM NaHCO₃, 6 mM Na₂CO₃, 1 mM EDTA) at 95°C for 23 minutes. RNA was

immediately precipitated. Fragment size between 20-40 nucleotides were selected on a denaturing polyacrylamide gel (15%, TBE-UREA, Invitrogen). These fragments were then ligated and converted to DNA as described in the previous section.

To estimate the abundance of mRNA levels, we calculated the number of sequencing reads mapped to a gene, divided by the length of the gene to yield the number of reads corresponding to the message per thousand bases of message per million sequencing reads (RPKM). Translation efficiency was calculated by dividing the protein synthesis rate by the estimate for mRNA level. Table S5 lists mRNA level and translation efficiency in arbitrary units.

Average ribosome density and correction for translation elongation rate

Average ribosome density was calculated for reads mapped to the gene excluding the first and last five codons to remove effects of translation initiation and termination. A number of genes with unconventional translational events were treated differently, as described below. (1) For genes with translational frameshift (*prfB* and *dnaX*), only the density after the frameshift event was used. The *dnaX* gene uses ~50% frameshift to code for two stable proteins—the larger tau subunit and the smaller gamma subunit. The density for the tau subunit was subtracted from the density before the frameshift to give the density for the gamma subunit. (2) For selenoproteins (*FdhF*, *FdoG*, *FdnG*), we observe higher ribosome density before the codon for selenocysteine, suggesting that most ribosomes terminates at the selenocysteine codon, which is also a stop codon recognized by release factor 2. Only the density after the selenocysteine codon was used in our calculation. (3) For proteins translated without a stop codon, such as the alternative ribosome rescue factor (*ArfA*), only the density before the end of the transcript was considered. (4) For proteins translated with a known ribosome stalling site (*SecM* and *TnaC*), the

ribosome density around the stalling site was not included in the average. (5) For proteins with nearly identical coding sequences, such as TufA and TufB, GadA and GadB, YnaE and YdfK, LdrA and LdrC, YbfD and YhhI, TfaR and TfaQ, RzoD and RzoR, PinR and PinQ, we considered the pair as the same protein and calculated the average ribosome density together.

To test whether the overall measure of ribosome density for the entire gene averages out local variation in elongation rates and sequencing biases, we compared the density for the first half and second half of the gene. Genes with at least 20 codons and 64 reads in either halves are included in the analysis. The result is shown in Fig. 1A. The lack of bias towards the first half confirms the finding that there is little drop-off in ribosome density for most genes (Li et al., 2012; Oh et al., 2011). Genes with special translational events, as discussed above, were excluded from this analysis.

These small effects can be further corrected using the knowledge of the sequence features that drive variations in elongation rates. We first considered the elevated density observed for the first 50-100 codons (Oh et al., 2011). For each gene, we calculated the local ribosome density as a function of length, which was measured as the average ribosome density in a window of 50 codons relative to that in the first 50 codons. The median values of this function among all genes, except those with unconventional translation events and those with less than 128 reads mapped, were fitted with an exponential decay function with an offset. The fitting result was independent of the window size. The resulting function was used to adjust the ribosome density for all genes, similar to the method used by Ingolia et al. (Ingolia et al., 2009), to remove the elevated density at the beginning of open reading frames.

We next correct for the elevated density due to ribosome pausing. We have previously reported that interactions between mRNA and the 16S rRNA cause ribosome pausing at internal Shine-Dalgarno (SD) sequences (Li et al., 2012). We have also found that the affinity between a hexanucleotide sequence and the ribosomal anti-Shine-Dalgarno (aSD) site strongly predicts the duration of ribosome pausing (Li et al., 2012). Here we used this information to correct for the effect of ribosome pausing. We fitted a linear function for the average ribosome occupancy downstream from a hexanucleotide sequence with respect to its affinity to the aSD site. This function is then used to adjust the ribosome occupancy at each position in each gene; at each position, the measured occupancy was divided by the expected pause duration based on the strongest hexanucleotide sequence at 6-11 bases upstream. The adjusted ribosome occupancy is no longer correlated with the SD-aSD interaction.

Finally, we remove the residual variations that are not accounted for using 90% Winsorization (Tukey, 1962). Namely, the top and bottom 5% of the ribosome occupancy for each gene were removed from the calculation for average ribosome density. The results before and after all three corrections listed here are shown in Fig. S1. These corrections together only have moderate effects, as the difference between uncorrected and corrected density is typically below 18% (standard deviation of the mean).

Total protein measurement and the conversion to absolute synthesis rates

Ribosome profiling and the analyses above allow direct comparison of the relative synthesis rates among proteins. To obtain the absolute synthesis rate, we normalized the results by the total weight of proteins synthesized per cell cycle. Because the majority of the proteins in

E. coli are long-lived compared to the generation time during rapid growth, the total weight of proteins synthesized per cell cycle can be approximated as the total weight of proteins per cell.

To measure the total weight of proteins per cell, we grew cells in the same way as those used for ribosome profiling experiments. When OD₆₀₀ reached the same level, we counted the number of cells per unit volume by serial dilution and plating on LB-agar plates. At the same time, 1 ml of culture harvested using centrifugation for 30 seconds. After removing 950 μ l supernatant, cells were resuspended and added to 950 μ l ice-cold PBS with 0.017% deoxycholate and on ice for 5 minutes. We then added 113 μ l of 100% trichloroacetic acid and incubate on ice overnight. Protein precipitation was collected by centrifugation at 20,000 rcf for 15 minutes at 4°C. The amount of proteins was quantified using the Lowry method with Peterson's modification (Sigma-Aldrich). To establish a standard curve, serial dilution of bovine serum albumin was made in the same culture media, and precipitated in the same way. The total weight of protein per cell (P) is calculated as the amount of protein per OD₆₀₀ per ml of culture divided by the number of cells per OD₆₀₀ per ml of culture. We measured 680 fg of protein per cell for culture in MOPS complete media, 450 fg per cell in MOPS complete media without methionine, and 238 fg per cell in MOPS glucose minimal media. These results on protein content and the corresponding estimate for the number of ribosomes (Table S1) at various growth rates are consistent with the estimate by Bremer and Dennis (REF).

To obtain the absolute synthesis rates, we first used the corrected ribosome density of each protein relative to that of all proteins to estimate its mass fraction in the proteome:

$$\phi_i = \frac{MW_i RD_i}{\sum_j MW_j RD_j},$$

where ϕ_i , MW_i , RD_i , are the mass fraction, molecular weight, and ribosome density of protein i ,

respectively. The synthesis rate for protein i (k_i) is given by

$$k_i = \frac{\phi_i P}{MW_i} = \frac{RD_i P}{\sum_j MW_j RD_j},$$

where k_i has the unit of molecules per generation. One generation is 21.5 ± 0.4 minutes in fully supplemented MOPS media. For stable proteins, k_i is also the copy number. The results are listed in Table S1.

Literature mining for published protein copy numbers

We validated our results by comparing with a list of protein copy numbers that have been individually characterized using classic assays. To compile such a list, we combined a community-based approach and computer-aided search. We consulted bacteriologists for relevant publications to their knowledge. We also utilized search engines such as PubMed and Google with relevant keywords. To obtain an unbiased list, every publication we could identify that measured specific protein abundance in *E. coli* was included. Results from other high-throughput studies were excluded in order to avoid bias toward any one method.

The published quantification used various strain backgrounds, growth media, temperature, and growth phases. If the same protein has been reported for multiple conditions, we chose the values that were measured in the condition closest to ours (MG1655, MOPS complete media with 0.2% glucose, 37°C, and exponential phase growth). Because most of the published quantification was based on cells with slower growth rates and lower cell mass, the protein copy numbers are in general slightly lower than our estimates in rich media. For copy numbers that were reported as molecules per genome equivalent, we multiplied the number by four to approximate our growth condition. The detailed protein information, strain background, media, temperature, growth phase, and PubMed ID for the original publication were listed in Table S2.

Systematic analysis of complex stoichiometry

To our knowledge, there is no curated database for multi-protein complexes and their stoichiometry. We systematically created a list based on the references available in the Ecocyc Database (Keseler et al., 2013). We first obtained a list of proteins that have been annotated as either 'subunit' or 'component'. We then inspected the literature to confirm whether each protein is a stoichiometric component of a larger complex. Because we set out to analyze the synthesis rates for obligate members of stable complexes, several criteria were used for the selection. (1) The subunit has no additional roles outside the complex that have been reported. Several ribosomal proteins do not meet this criterion (even though their synthesis rates are closely matched) because the free proteins function separately in translational repression or ribosome assembly. (2) The complex is formed in the default state, rather than as a response to signals. For example, the DNA repair enzyme *uvrABC* is excluded because the assembly depends on damaged DNA. (3) The stoichiometry of the complex has been well documented. We identified 62 multi-protein complexes that meet these criteria and are expressed in our dataset (more than 128 sequenced reads per protein). The gene names and the corresponding stoichiometry is listed in Table S3. For the budding yeast *S. cerevisiae*, we performed a small-scale analysis on multi-protein complexes using the same criteria.

Among the 64 complexes in *E. coli*, 47 have all the components encoded in the same operon, and their synthesis rates are shown in Fig. 2. The operon structure is based on the experimentally validated annotation in Ecocyc, and confirmed by our RNA-seq data. The rest of the complexes have members expressed from at least two different mRNAs, and their synthesis rates are shown in Fig. S3. The exceptions to proportional synthesis are also shown in Fig. S3C.

Predicted translation rates using RBS calculator

To calculate the predicted translation rates based on the model established by Salis et al. (Salis et al., 2009), we used the RBS Calculator downloaded from github.com/hsalis/Ribosome-Binding-Site-Calculator-v1.0. For each gene, we used the nucleotide sequence from 35 bases upstream to 35 bases downstream from the translation start site as the input, and obtained the predicted strength of ribosome binding site, as well as predicted Shine-Dalgarno sequence and mRNA accessibility. For complexes with two equimolar subunits that are expressed from the same operon, the predicted fold-difference between the translation rates is $e^{-\beta(\Delta G_2 - \Delta G_1)}$, where $\beta = 0.45$ and ΔG_1 and ΔG_2 are the predicted strength of ribosome binding sites for gene 1 and gene 2, respectively.

DNA-binding properties of transcription factors

To test our hypothesis that low abundance transcription factors (TFs) always bind to their cognate sites independent of ligands, we mined the literature to determine the mode of DNA binding of TFs and compared with our results on TF copy numbers. *E. coli* has >400 annotated TFs, most of which are putative with little biochemical characterization. To focus on well-characterized TFs, we based our analysis on TFs that have been shown to regulate their own transcription. For each of these 102 TFs, we first identified the ligand that binds directly to regulate the activity of the TF. We then searched for evidence for whether the binding of ligands alter the ability to bind to the cognate DNA binding sites. TFs that do not have known ligands are not included in Fig. 6. Table S4 lists the ligands, mode of DNA binding, and the PubMed ID for the reference that showed whether DNA-binding is affected by ligands.

To estimate the average spacing between DNA-binding proteins on the chromosome, we divided the length of the *E. coli* chromosome by the total amount of proteins (per genome equivalent) that are annotated with DNA binding activity. Proteins whose main function is not associated with DNA binding, such as alanyl-tRNA synthetase, are not considered in the calculation. The total number of DNA-binding proteins is divided by two because most of them bind to DNA as dimers. A large fraction of these proteins consists of nucleoid-associated proteins and the RNA polymerase (~70%). In this estimation, we assumed that the vast majority of these proteins are nonspecifically associated with DNA *in vivo*. Indeed, the concentrations for nucleoid-associated proteins and the RNA polymerase are higher than the reported nonspecific dissociation constants (Li et al., 2009), indicating that they are bound to the chromosome. This estimation gives ~56 basepairs center-to-center distance between neighboring DNA-binding proteins. The average spacing between adjacent proteins is 36 basepairs if the average footprint size on DNA is assumed to be 20 basepairs (Li et al., 2009).

Quantitative analysis of the methionine biosynthesis pathway

The description of the pathway and the corresponding enzymes were obtained from the Ecocyc Database (Keseler et al., 2013). The methionine synthesis pathway acquires the backbone of the amino acid from aspartate, and sulfur from cysteine. Although both aspartate and cysteine are supplied in our media, we noticed that the enzyme involved in cysteine biosynthesis were also induced. We also noticed that cysteine codons, in addition to the methionine codon, have slightly elevated ribosome occupancy, suggesting an insufficient pool of cysteine in the cell. Therefore, we also included the sulfur assimilation pathway from sulfate in the analysis.

Several reactions in the pathway can be catalyzed by more than one enzyme. For example, for the last step of pathway, *E. coli* has two homocysteine transmethylase: cobalamin-dependent MetH and cobalamin-independent MetE. In our methionine drop-out MOPS media, MetH expression was not induced compared to the level in the complete media, and was 0.6% of the level of MetE. Further, cells with MetE knockdown are unable to grow in media without methionine (see below), suggesting that MetH is not contributing significantly to methionine biosynthesis in this condition. Therefore, we only included MetE and not MetH in the analysis. Similarly for other reactions, we only included enzymes that were up-regulated.

In addition to providing building blocks for protein synthesis, L-methionine is also converted to *S*-adenosyl-L-methionine (SAM). In eukaryotes, the demand for SAM is high due to its involvement in the synthesis of phosphatidylcholine (PC), which is a major component of the cell membrane (Giovanelli et al., 1985; Hirata and Axelrod, 1978). However, this demand is nonexistent in *E. coli* because it does not carry enzymes that synthesize PC (Sohlenkamp et al., 2003), and its lipid composition lacks PC (Oursel et al., 2007). The other pathways that utilize SAM were estimated to account for a small fraction of the methionine synthesis rate (0.4%) (Feist et al., 2007). We therefore consider protein synthesis as the major consumption for L-methionine.

The turnover number (k_{cat}) for each enzyme was obtained from the BRENDA database and the references therein (Schomburg et al., 2002). The only enzyme whose turnover number has not been reported for *E. coli* is sulfate adenylyltransferase (CysND). We instead used the measurement for a different proteobacteria, *Thiobacillus denitrificans*. The maximal reaction velocity (V_{max}) for each step was calculated as the product of k_{cat} and our estimate on the enzyme

copy number. Because we provide an upper bound of the copy number, the maximal reaction velocity is also an upper bound of the actual flux.

The smallest V_{max} was found for two reactions in the pathway, catalyzed by MetE and CysH, respectively. Whereas the V_{max} for MetE matches the methionine consumption rate, the V_{max} by CysH is even smaller, raising the possibilities that either (1) there is an alternative for sulfur assimilation that has not been characterized, or (2) the reported k_{cat} for CysH *in vitro* is lower than that *in vivo*. To distinguish between these possibilities, we first tested whether a *cysH*-null strain (from the KEIO collection) can grow on various sulfur sources. The *cysH*-null strain was unable to grow in media with sulfate as the only sulfur source, but was able to grow when supplemented with 40 mM MOPS (3-morpholinopropane-1-sulfonic acid). Therefore, *E. coli* has an alternative pathway for sulfur assimilation from MOPS, which could explain the extra flux needed in our MOPS-based media.

We then tested which genes are responsible for this uncharacterized pathway, by constructing double-deletion strains with *cysH* and either genes that are up-regulated in the Met drop-out media, or enzymes that are known to utilize other types of sulfonic acids. We found that the ability to grow on MOPS was lost when both *cysH* and *tauD*, which encodes a taurine dioxygenase, were deleted. Therefore, TauD is an essential enzyme for this alternative pathway. However, for wildtype cells in the Met drop-out media, TauD is not up-regulated and is only expressed at a very low level (<50 copies/cell), suggesting that this novel pathway is not active in this condition. Therefore, it is likely that CysH is still the main pathway for cysteine biosynthesis, and that the published k_{cat} for CysH *in vitro* does not reflect the actual turnover number *in vivo*.

Analytical model for optimal level of MetE expression

To understand why cells produce the measured level of MetE when it is limiting methionine and protein synthesis, we modeled the growth rate as a function of the amount of MetE synthesized. The model takes into account the cost and benefit of MetE synthesis separately. The cost function is based on previous observations that synthesis of excess proteins competes with that of other proteins (Scott et al., 2010). As a result of decreased amount of ribosomes that are necessary for auto-catalysis, growth rate decreases. The work by Hwa and co-workers established the relationship between the growth rate and the mass fraction of excess proteins (Scott et al., 2010). Here we used a modified version for the methionine biosynthetic pathway, and express the predicted growth rate (λ) as a function of the mass fraction of MetE in the proteome (ϕ_E):

$$\lambda = \lambda_0 \left(1 - \frac{\phi_{m/c} + \phi_E}{\phi_C}\right),$$

where $\lambda_0 = 1.93 \text{ hr}^{-1}$ is the growth rate in methionine-supplemented media, $\phi_C = 0.48$ is the phenomenological parameter obtained by Scott et al, $\phi_{m/c} = 0.045$ is the mass fraction of all enzymes except MetE in the methionine and cysteine biosynthesis pathways. We chose to fix $\phi_{m/c}$ while varying ϕ_E because these other enzymes appear to be made in excess capacity. This cost function is plotted in red in Fig. 7.

The benefit function is based on our observation that the rate of Met synthesis at maximal MetE activity is equal to the rate of Met consumption by protein synthesis:

$$N_E k_{cat} = f_{met} N_R k_e,$$

where N_E, N_R are the numbers of MetE and translating ribosomes, respectively. k_{cat}, k_e are the turnover number of MetE and translation elongation rate, respectively. f_{met} is the fraction of

translated codons that encodes methionine. We can re-write this equation in terms of the growth rate λ and the mass fraction of MetE ϕ_E . To do so, we first notice that, when the majority of proteins are long-lived compared to the cell doubling time, the mass fraction of MetE is equivalent to the fraction of translating ribosomes making MetE, which is given by

$$\phi_E = \frac{k_{in}^E \frac{l_E}{k_e}}{N_R},$$

where k_{in}^E is the translation initiation rate of MetE and $\frac{l_E}{k_e}$ is the time it takes to synthesize MetE (l_E is the length of MetE polypeptide). The translation initiation rate of MetE is proportional to the amount of MetE in a cell:

$$N_E = \frac{k_{in}^E}{\lambda},$$

Using these relations, the production/consumption equation becomes

$$\lambda = \frac{k_{cat}}{f_{met} l_E} \phi_E.$$

Notice that this relationship is only dependent on well-established parameters: the published turnover number ($k_{cat}=0.12 \text{ s}^{-1}$), the fraction of Met codons (2.7%), and the length of MetE (753 amino acids). This benefit function is plotted in blue in Fig. 7.

Combining the cost and benefit functions, we predict that the maximal growth rate can be achieved when the mass fraction of MetE is

$$\phi_E = \frac{\phi_C - \phi_{m/c}}{1 + \frac{k_{cat} \phi_C}{f_{met} l_E \lambda_0}} = 0.069.$$

And the maximal growth rate is

$$\lambda = \frac{k_{cat}}{f_{met} l_E} \frac{\phi_C - \phi_{m/c}}{1 + \frac{k_{cat} \phi_C}{f_{met} l_E \lambda_0}} = 1.47 \text{ hr}^{-1},$$

which corresponds to a doubling time of 28 min. These predictions match what we observed for cells grown in the Met drop-out media.

MetE repression and overexpression

To test the prediction that the MetE level is optimized in wildtype cells to maximize growth rate in media without methionine, we constructed strains with either MetE repression or overexpression and measured the effect on growth rate. For repression, we used the recently described CRISPRi system, which uses a DNA binding protein, dCas9 to block transcription elongation (Qi et al., 2013). With a short guide RNA (sgRNA), dCas9 represses transcription with high sequence specificity in *E. coli*. We designed a sgRNA with a short guide region (18 nucleotides) targeting +182-198 bases of the metE gene on the nontemplate strand. The sgRNA is expressed from a tetracycline-inducible promoter on plasmid pJW1423, which is derived from pgRNA-bacteria (Addgene #44251) (Qi et al., 2013). To compare the effect on growth rate at various induction levels, we used a control sgRNA targeting RFP, which is absent in the cell, expressed from pgRNA-bacteria (Qi et al., 2013). Tetracycline repressor (TetR) and dCas9 are expressed from pdCas9-bacteria (Addgene #44249) (Qi et al., 2013). The dCas9 and sgRNA plasmids were simultaneously transformed into MG1655 and selected under chloramphenicol and carbenicillin.

Full induction of the sgRNA targeting MetE inhibits cell growth in media lacking methionine, likely due to impaired methionine biosynthesis. This result also indicates that the other methionine synthase, MetH, is not functional in this growth condition. To test whether even modest reduction in MetE level affects growth as our model predicts, we measured growth rates with either uninduced basal expression level, or induced with 10 nM anhydrotetracycline.

Overnight culture in methionine dropout media without anhydrotetracycline was diluted 1:1000 into 20 ml fresh media. The culture was kept at 37°C in waterbath shaker, and OD₆₀₀ was measured until it reaches 0.4. The effect of knockdown is confirmed by qPCR (see below). The results are plotted in Fig. S6.

For overexpression, ectopic MetE was expressed from a tetracycline-inducible promoter (plasmid pJW1424) in MG1655. Because the endogenous metE transcript is regulated by sRNA, which could limit the ability to over-produce MetE proteins (Boysen et al., 2010), we replaced the native 5' UTR with a synthetic sequence. Due to this difference in 5' UTR, the induction of MetE protein levels may not be proportional to the changes in its mRNA levels. We measured growth rates at 37°C with either 432 nM or 4.32 μM anhydrotetracycline. The effect of the overexpression is confirmed by qPCR.

While these results provide strong qualitative agreement with our model (Fig. S6), we caution that quantitative measurements are much more difficult due to the lack of non-perturbative tools for fine-tuning protein expression. Both the knockdown and overexpression systems we used requires accessory proteins, whose expression itself affects growth. Therefore, we are careful to compare the effect of metE level on growth rate should be relative to controls in the presence of these accessory proteins, and not to the wildtype for which our quantitative model is based upon.

Quantitative PCR for metE mRNA levels

RNA was extracted following the RNAsnap protocol (Stead et al., 2012). 1 ml of culture at OD₆₀₀=0.3 was pelleted by centrifugation for 30 seconds. The pellet was added to 100 μl of RNA extraction solution (95% formamide, 1% beta-mercaptoethanol, 500 mM EDTA, 0.026%

SDS), and incubated at 95°C for 7 minutes. After centrifugation at 21 kg for 5 minutes, the supernatant was transferred to new tubes with 150 µl Tris buffer, 25 µl sodium acetate, and 825 µl ethanol. RNA was precipitated by incubating at -80 °C for >1 hour, centrifugation at 4°C for 30 minutes, and the pellet was washed by 250 µl ice-cold 100% ethanol and resuspended in 500 µl Tris buffer pH 7. Remaining debris was removed by centrifugation and RNA was precipitated again and resuspended in 20 µl.

10 µg RNA in 20 µl was treated with 2 µl DNase I (100U/ml, Roche) in supplied buffer at 37°C for 30 min, followed by 75°C for 10 min. RNA was purified using Zymo Clean and Concentrate Columns. cDNA was generated using M-MuLV reverse transcriptase (NEB) with random hexamers. Quantitative PCR was performed in triplicates using DyNAmo HS CYBR Green qPCR kit (Thermo) on Roche LightCycler. Standard curves were generated using concentrated cDNA samples. Primers targeting *cyoA* cDNA was used for normalization. Relative *metE* levels were compared to controls (empty vector for overexpression and sgRNA targeting RFP for repression).

SUPPLEMENTARY REFERENCES

Ariake, L., Valgepea, K., Peil, L., Nahku, R., Adamberg, K., and Vilu, R. (2012). Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. *J Proteomics* 75, 5437-5448.

Boysen, A., Moller-Jensen, J., Kallipolitis, B., Valentin-Hansen, P., and Overgaard, M. (2010). Translational regulation of gene expression by an anaerobically induced small non-coding RNA in *Escherichia coli*. *J Biol Chem* 285, 10690-10702.

Brandman, O., Stewart-Ornstein, J., Wong, D., Larson, A., Williams, C.C., Li, G.W., Zhou, S., King, D., Shen, P.S., Weibezahn, J., *et al.* (2012). A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* 151, 1042-1054.

Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., and Palsson, B.O. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3, 121.

- Giovanelli, J., Mudd, S.H., and Datko, A.H. (1985). Quantitative analysis of pathways of methionine metabolism and their regulation in *lemna*. *Plant Physiol* 78, 555-560.
- Hanke, S., Besir, H., Oesterhelt, D., and Mann, M. (2008). Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *J Proteome Res* 7, 1118-1130.
- Hirata, F., and Axelrod, J. (1978). Enzymatic synthesis and rapid translocation of phosphatidylcholine by two methyltransferases in erythrocyte membranes. *Proc Natl Acad Sci U S A* 75, 2348-2352.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218-223.
- Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J., and Frishman, D. (2008). Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* 9, 102.
- Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., *et al.* (2013). EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* 41, D605-612.
- Li, G.W., Berg, O.G., and Elf, J. (2009). Effects of macromolecular crowding and DNA looping on gene regulation kinetics. *Nat Phys* 5, 294-297.
- Li, G.W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 538-541.
- Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25, 117-124.
- Neidhardt, F.C., Bloch, P.L., and Smith, D.F. (1974). Culture medium for enterobacteria. *J Bacteriol* 119, 736-747.
- 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.
- Oursel, D., Loutelier-Bourhis, C., Orange, N., Chevalier, S., Norris, V., and Lange, C.M. (2007). Lipid composition of membranes of *Escherichia coli* by liquid chromatography/tandem mass spectrometry using negative electrospray ionization. *Rapid Commun Mass Spectrom* 21, 1721-1728.
- Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B., and Aebersold, R. (2009). Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell* 138, 795-806.
- Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173-1183.
- Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27, 946-950.

Schomburg, I., Chang, A., and Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* 30, 47-49.

Scott, M., Gunderson, C.W., Mateescu, E.M., Zhang, Z., and Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. *Science* 330, 1099-1102.

Sohlenkamp, C., Lopez-Lara, I.M., and Geiger, O. (2003). Biosynthesis of phosphatidylcholine in bacteria. *Prog Lipid Res* 42, 115-162.

Stead, M.B., Agrawal, A., Bowden, K.E., Nasir, R., Mohanty, B.K., Meagher, R.B., and Kushner, S.R. (2012). RNAsnap: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. *Nucleic Acids Res* 40, e156.

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.

Tukey, J.W. (1962). Future of Data-Analysis. *Ann Math Stat* 33, 1-&.

Chapter Two:

**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†}

¹Graduate Group in Biophysics,

²Department of Microbiology and Immunology,

³Department of Cell and Tissue Biology,

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

⁵California Institute of Quantitative Biology,

⁶Center for RNA Systems Biology,

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

*These authors contributed equally to this work.

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 ($\rho = 0.48$, Figure 2G, S2B-C), whereas control samples without DMS modification was not correlated ($\rho = .05$, 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 ($\rho = 0.48$, 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 ³⁵S-methionine incorporation, plateauing at 30 min after cold shock when the run-off observed by ribosome profiling is presumably complete. At that point, ³⁵S-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 Δrnr 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 Δrnr strain exhibits a minimal decrease in mRNA level. Moreover, mRNA accumulates to 9.8% of total RNA at 8hr after cold shock in Δrnr 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 ($\Delta cspABEG$) 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, $\Delta cspABEG$ 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 $\Delta cspABEG$ 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 $\Delta cspABCEG$ 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 $\Delta cspABCEG$ strain indicate that the TEs of the best-translated ORFs in WT (\geq 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 ($T_{1/2}$ = 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 Csp 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 $\Delta cspABEG$ strain, which lacks the Csp ORFs but retains their 5'UTRs (Figure 5C-D), but the CspB 5'UTR does change structure in a $\Delta cspBG$ 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 Csp's.

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 uninformative 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 $\sim 2 * 10^6$ /cell, (Xia et al., 2001), which is consistent with Csp-mRNA interaction over the entire length of open reading frames ($\sim 10^7$ 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.

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 growing *Bacillus 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). Acquisition 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.

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

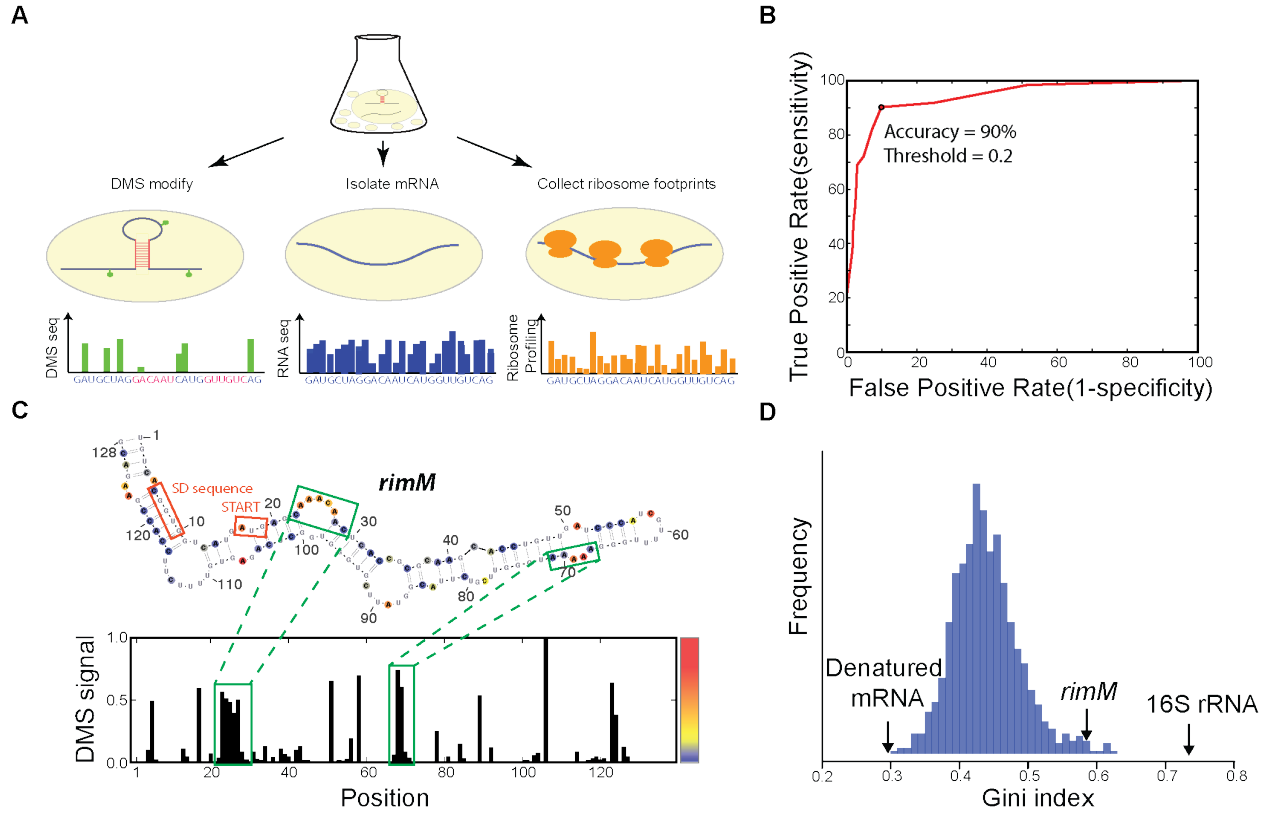


Figure 1: DMS-seq effectively probes RNA structures in *E. coli*

- (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
(continued on next page)**

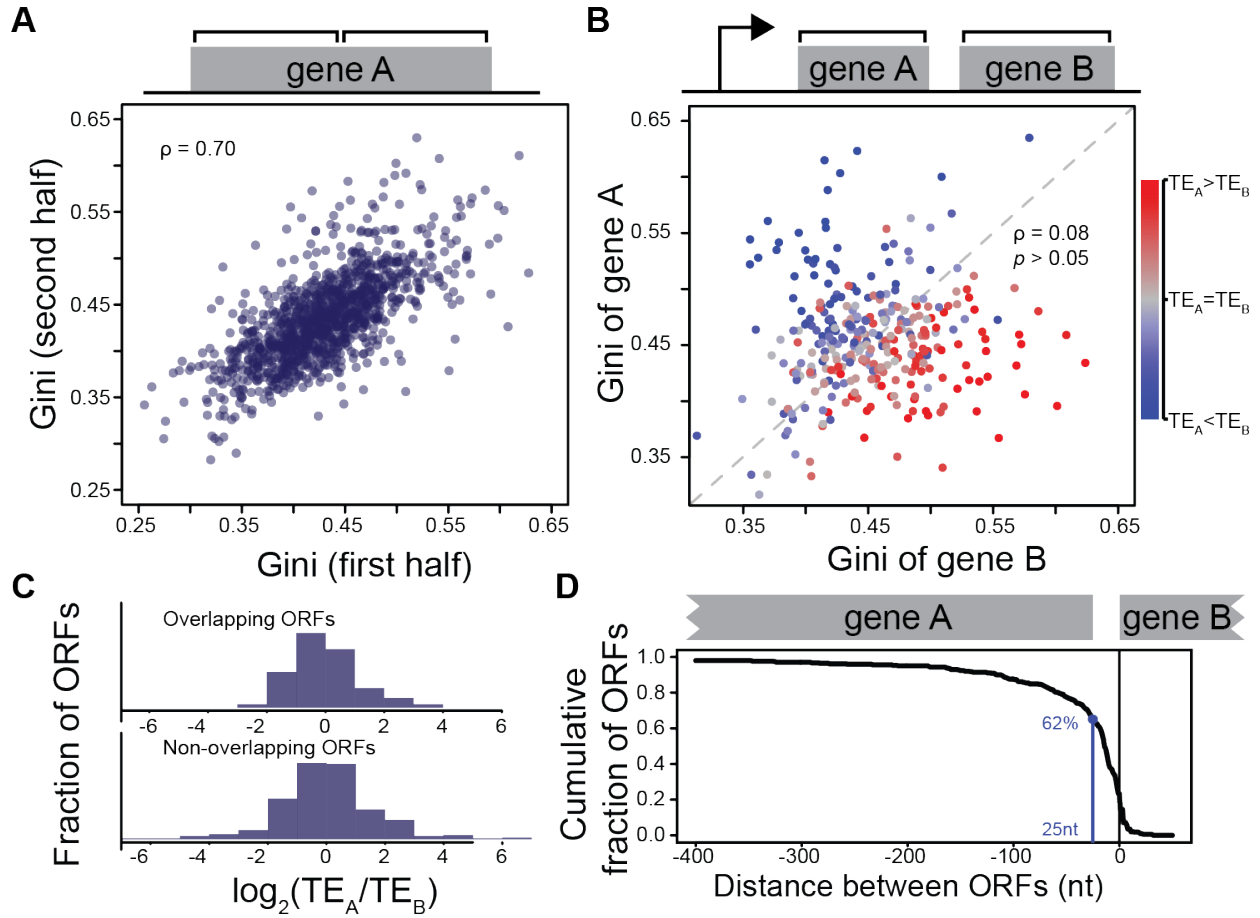


Figure 2: mRNA structure is organized around open reading frames (continued)

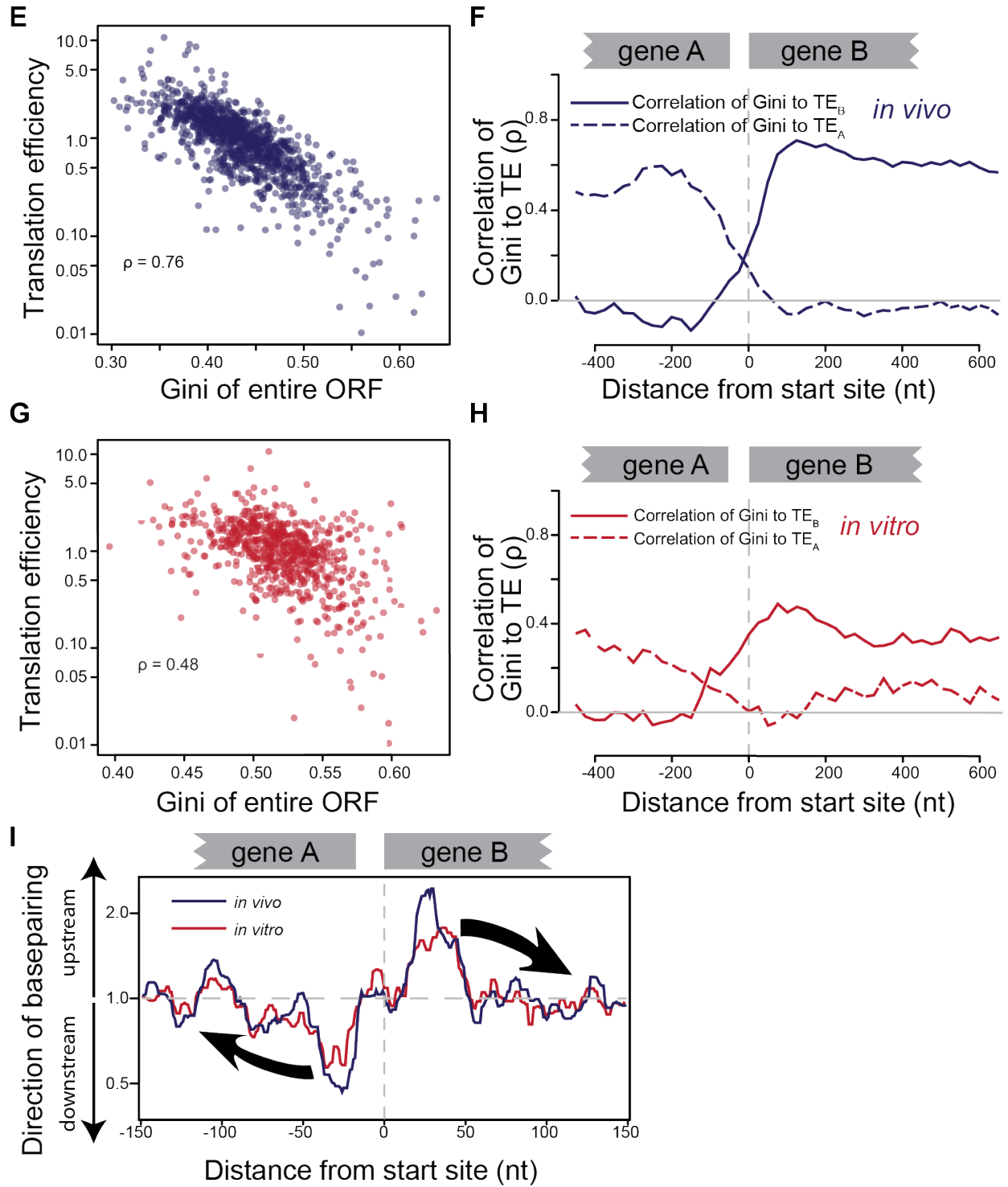


Figure 2: mRNA structure is organized around open reading frames

- (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

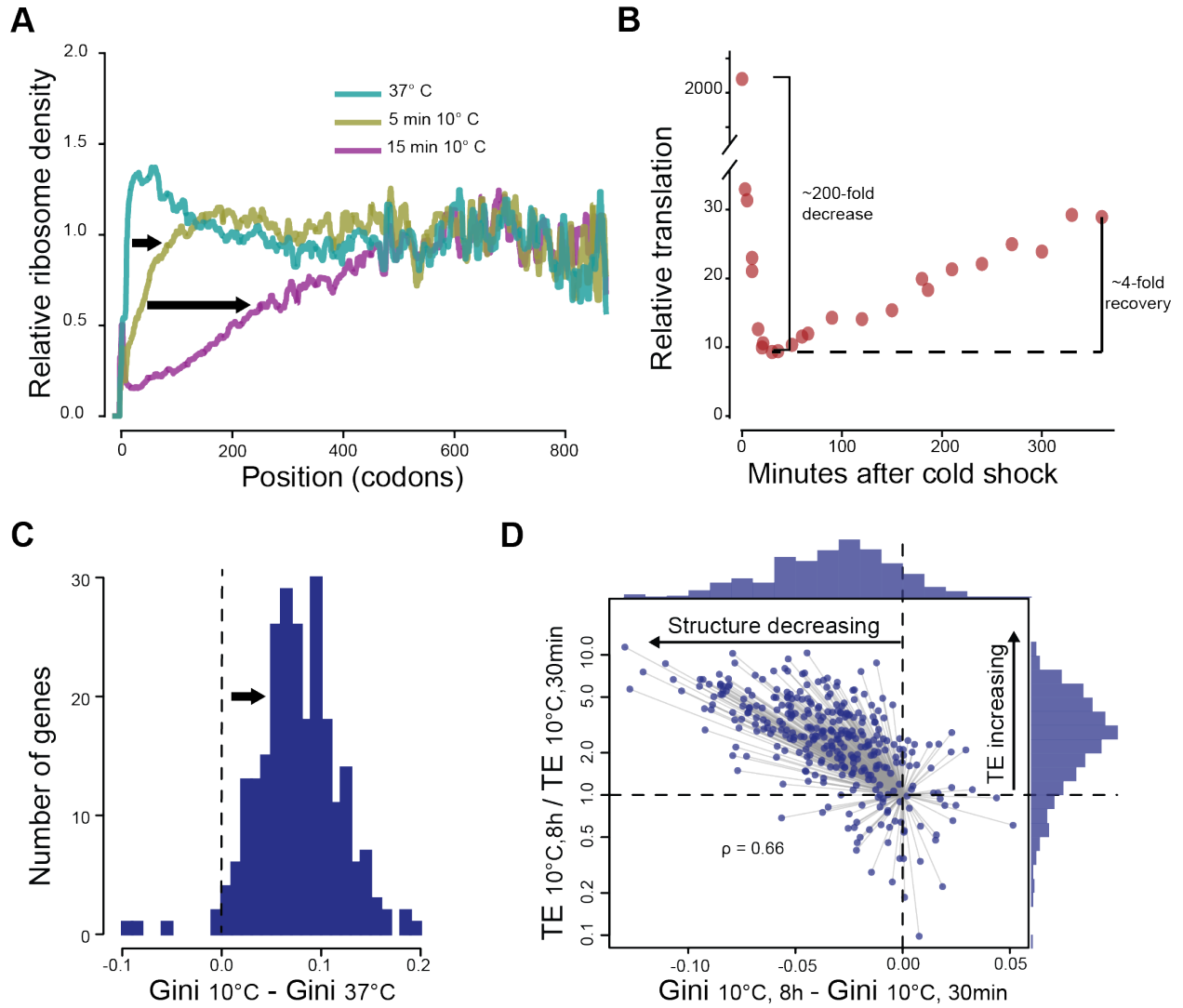


Figure 3: Cold induces a defect in translation instigated by an increase in mRNA structure

- (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 CspA facilitate cold recovery

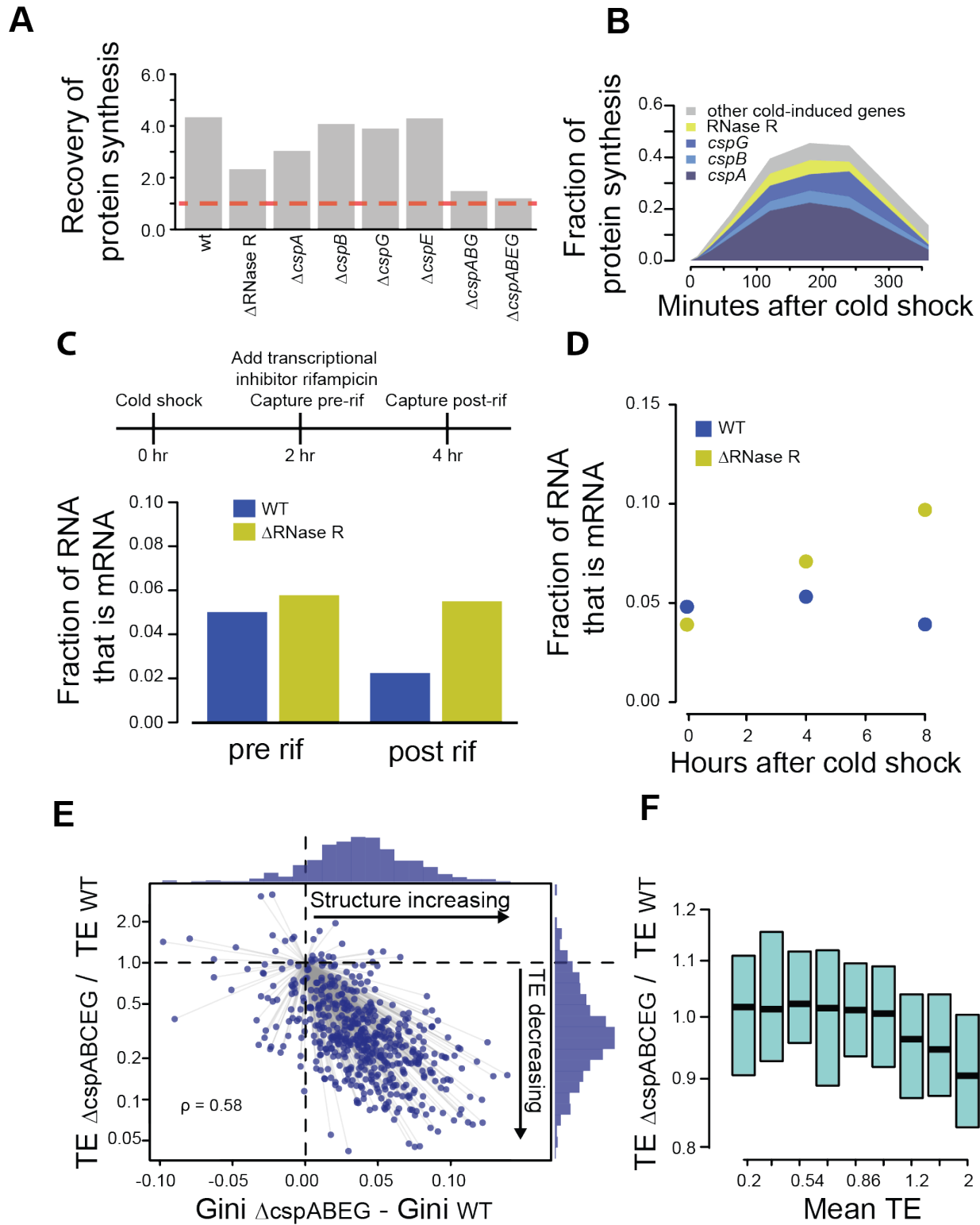


Figure 4: RNase R and Csps facilitate cold recovery

- (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, Δrnr 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 Δrnr 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 Δrnr 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 ($\Delta cspABEG$) 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 $\Delta cspABCEG$ 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 $\Delta cspABCEG$ 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

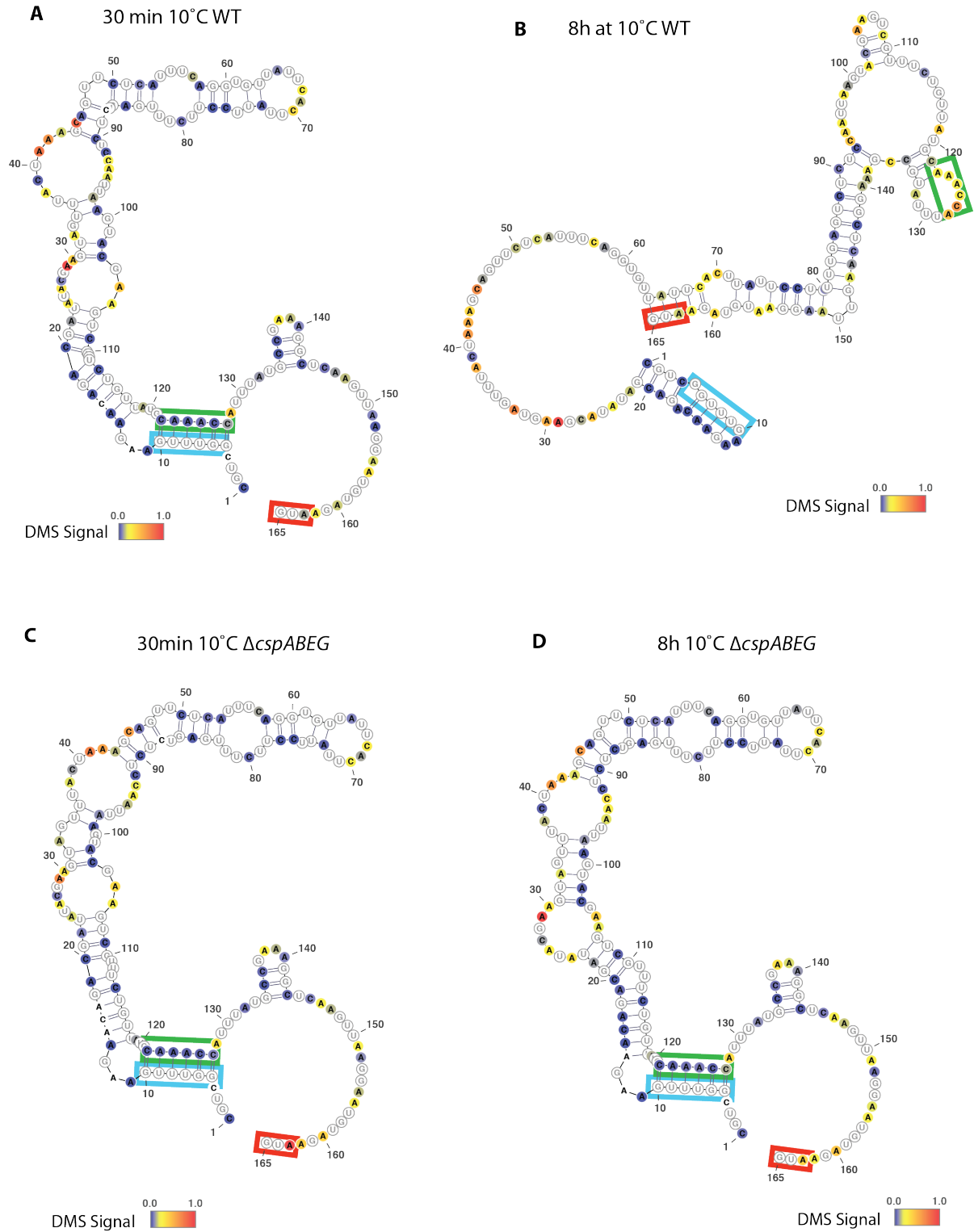


Figure 5: Csp expression is controlled by an auto-regulatory feedback loop

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 $\Delta cspABEG$ (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

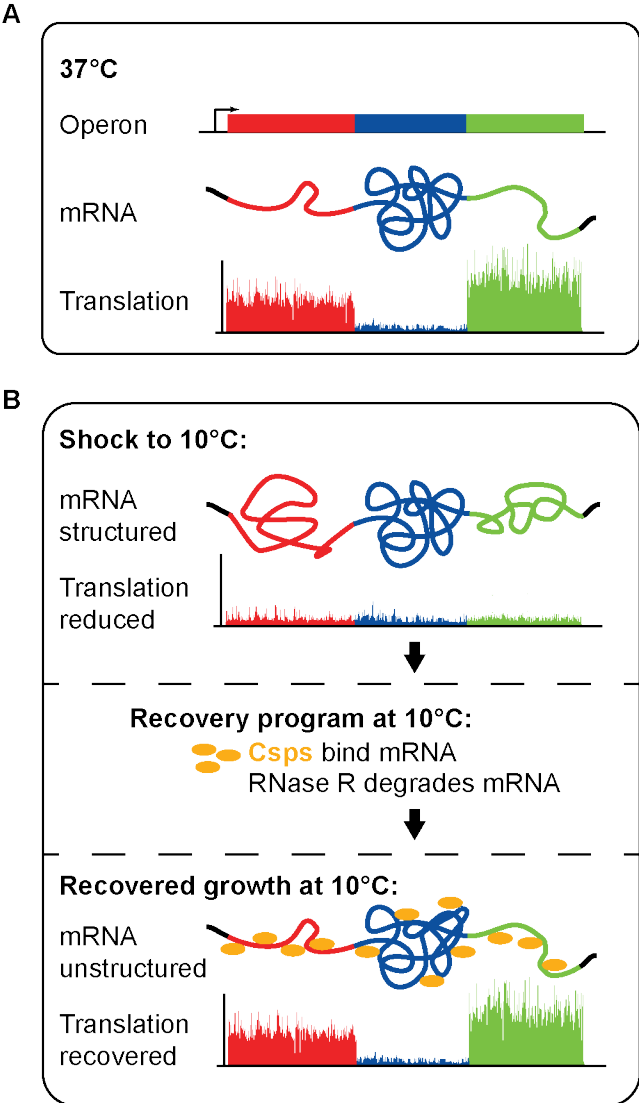


Figure 6: Model of operon structural organization and surveillance

- (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.

Figure S1: DMS-seq effectively probes RNA structures in E. coli

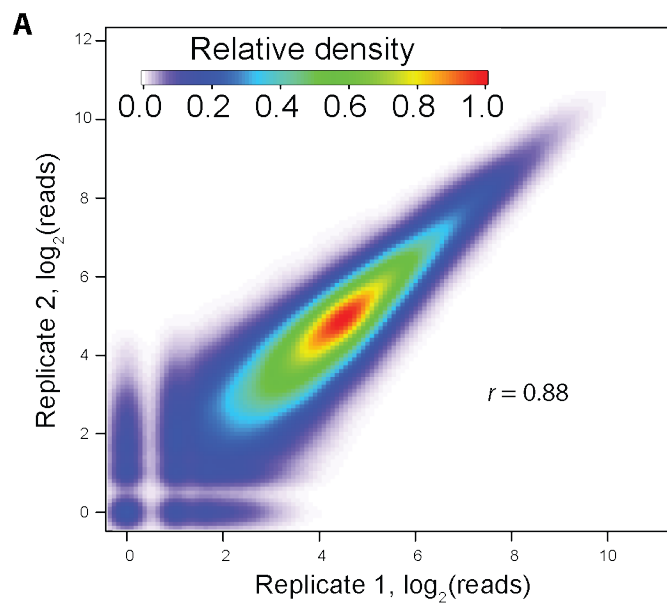


Figure S1: DMS-seq effectively probes RNA structures in E. coli (continued)

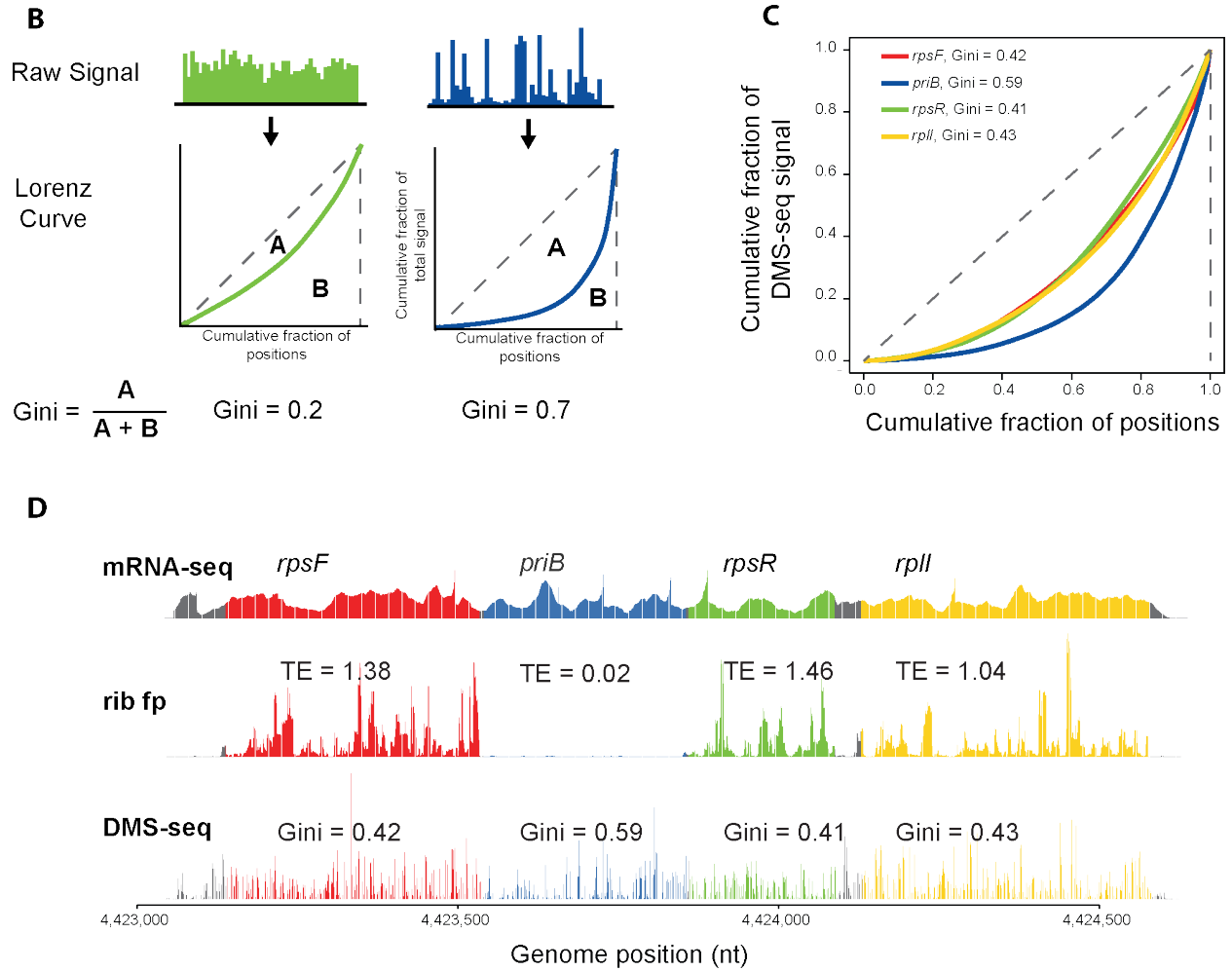


Figure S1: DMS-seq effectively probes RNA structures in *E. coli*

(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

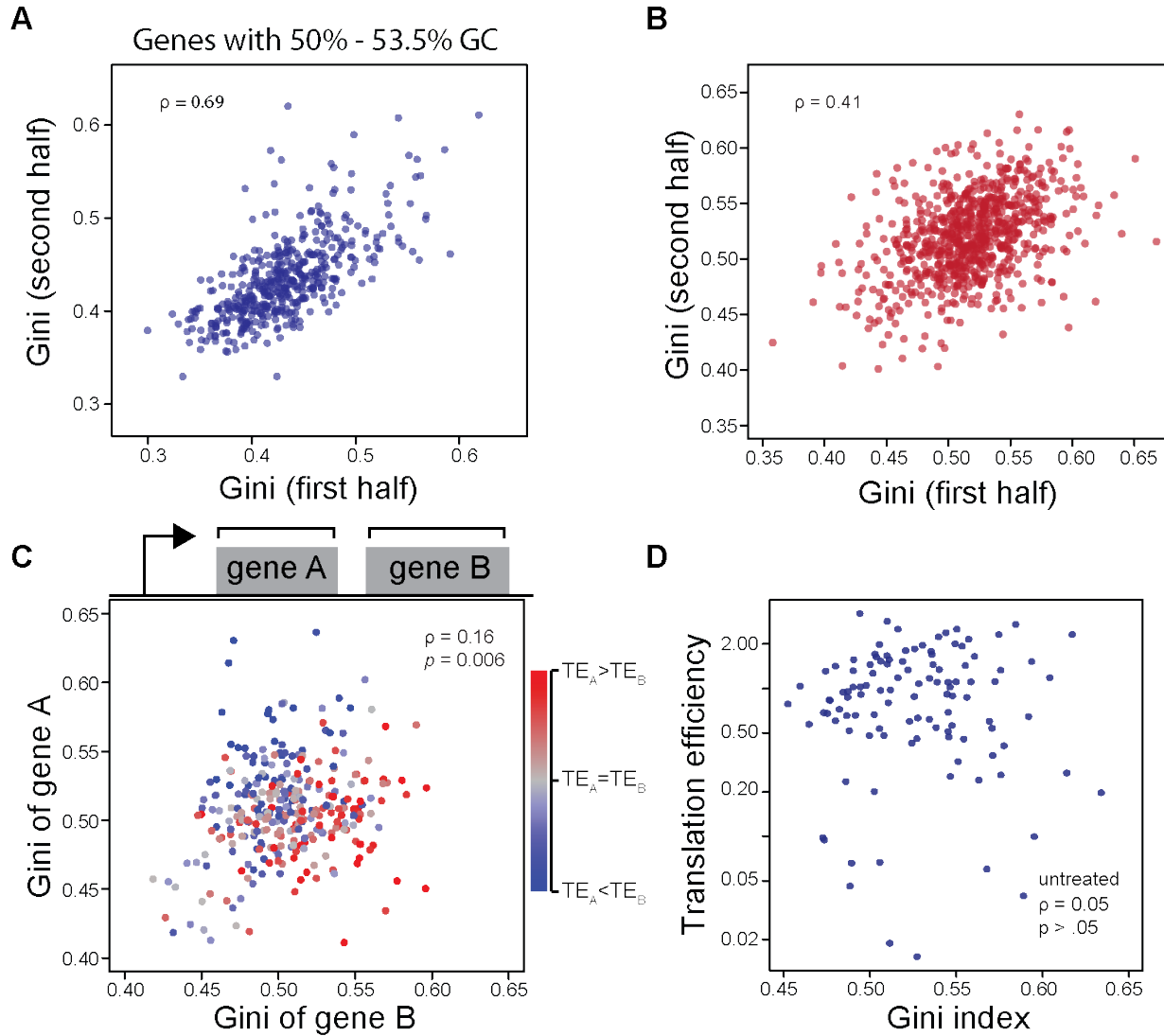


Figure S2: mRNA structure is organized around open reading frames (continued)

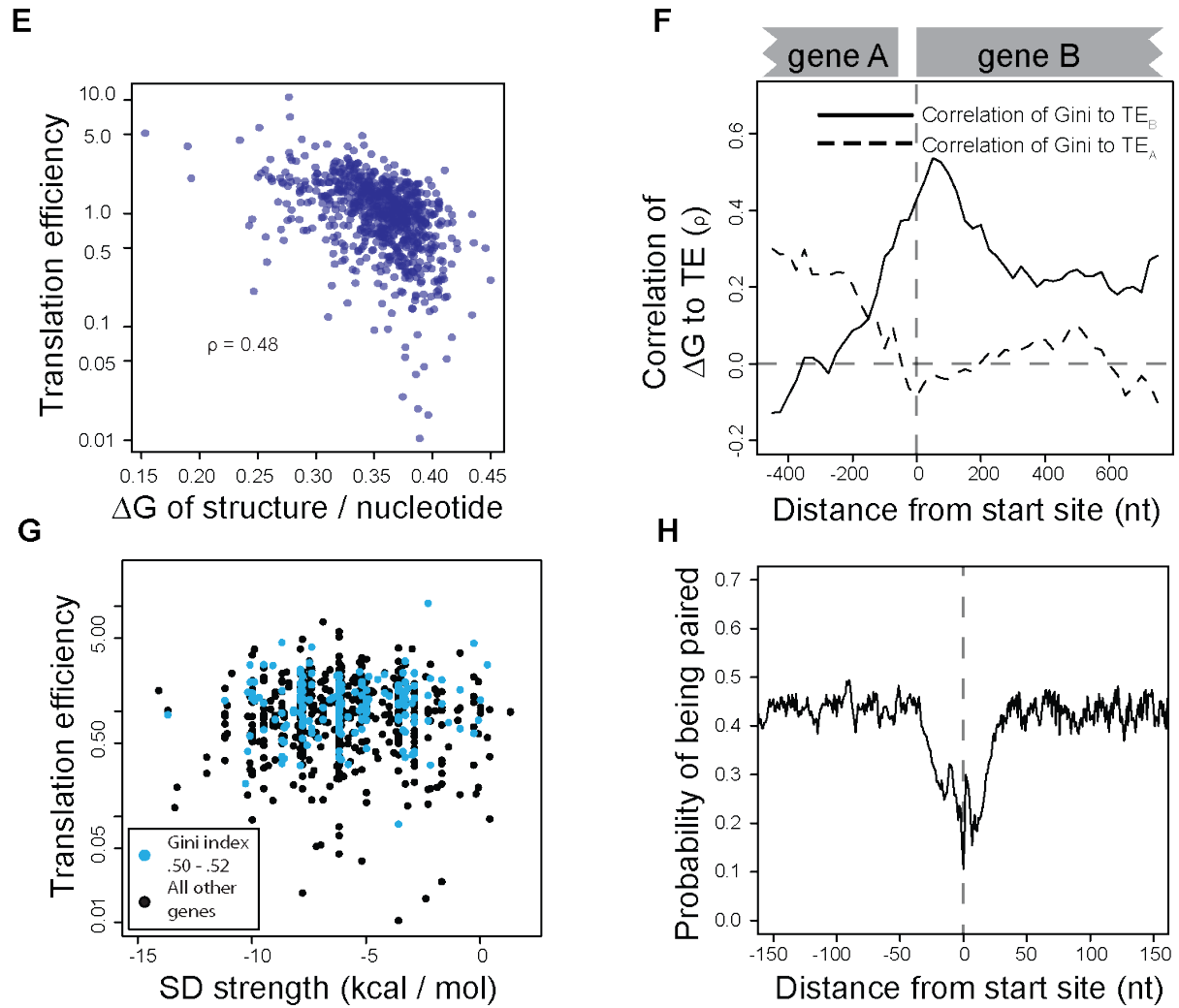


Figure S2: mRNA structure is organized around open reading frames

- (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

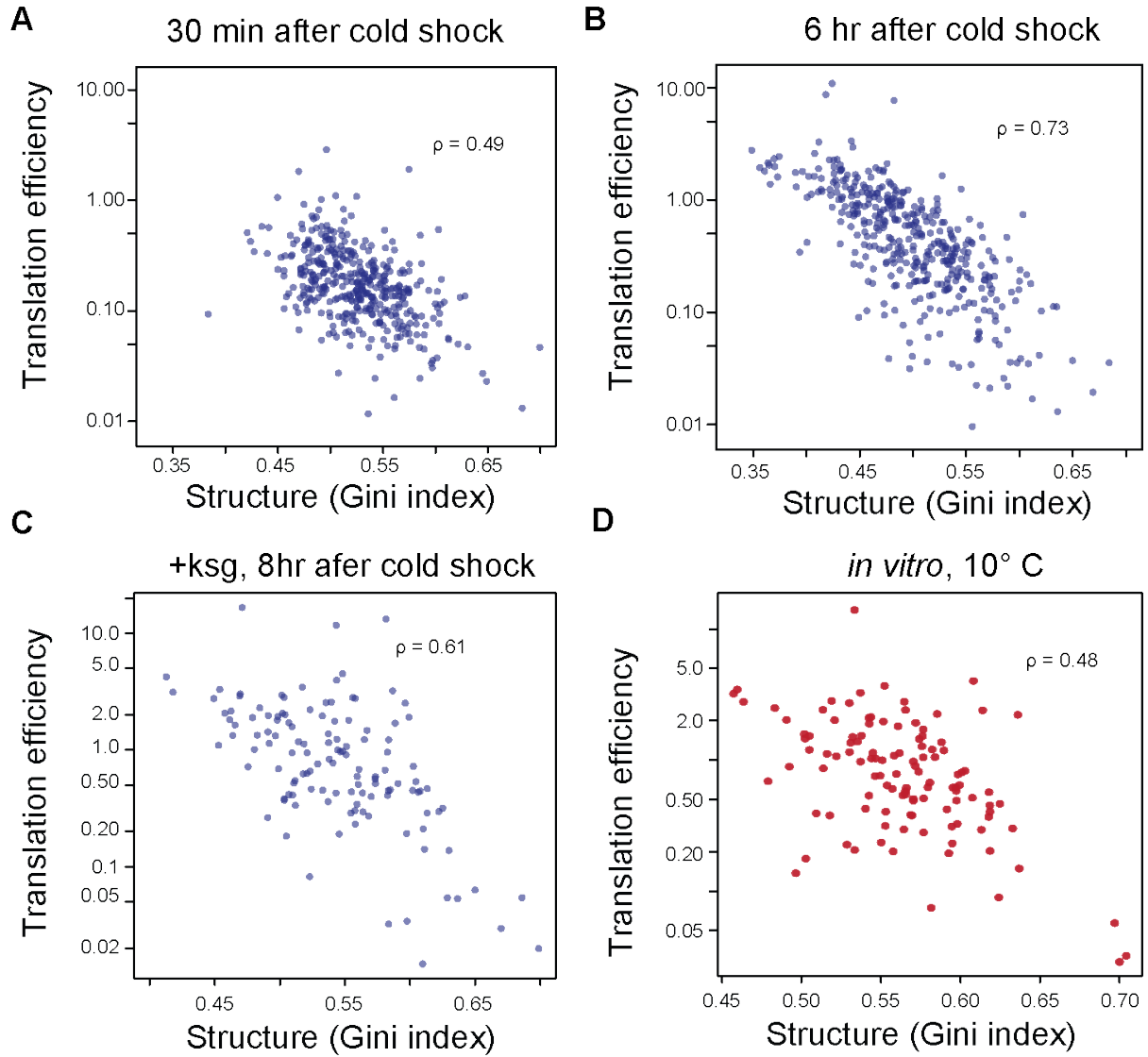


Figure S3: Structure and translation efficiency remain correlated at 10°C

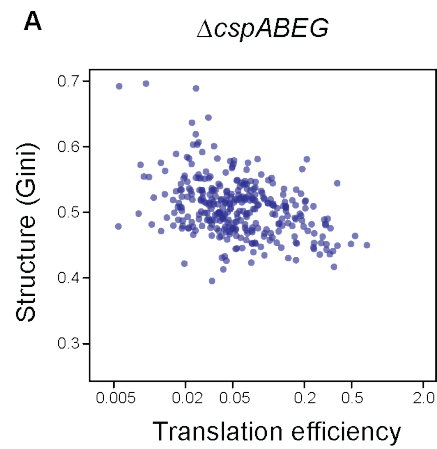
(A) Plot of Gini index of *in vivo* DMS-modified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs, measured 30 min following cold shock to 10°C.

(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

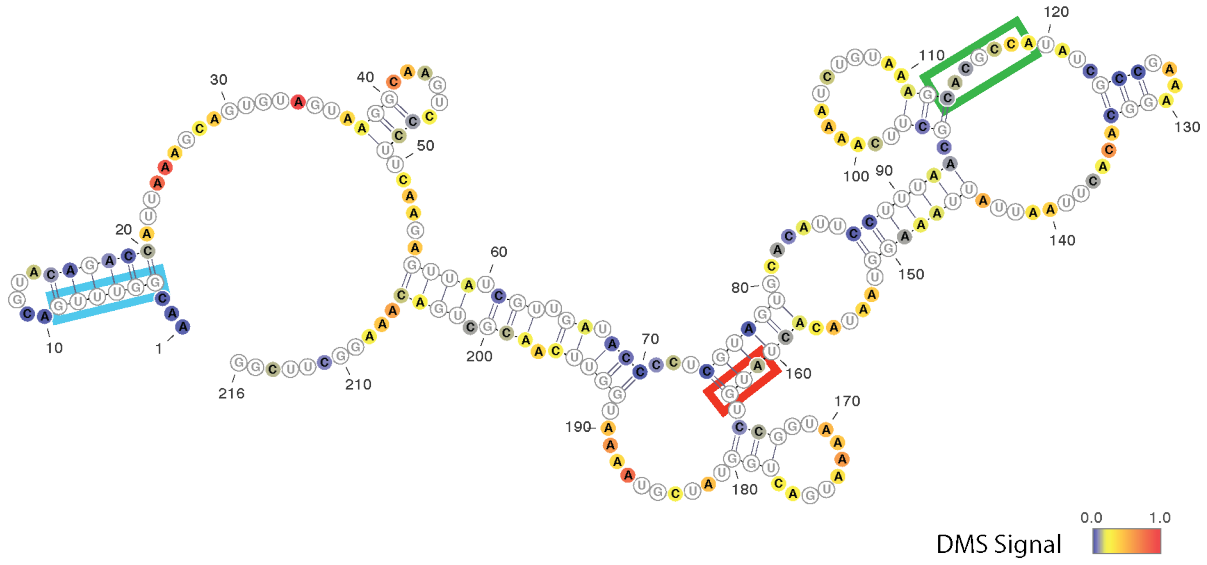


Plot of Gini index of *in vivo* DMS-modified mRNA calculated across the entire ORF body against *in vivo* translation efficiency for well-expressed ORFs, measured in a $\Delta cspABEG$ strain at 6 hr following cold shock.

Fig S5: CspB UTR structure is modulated by cold shock

A

cspA 5'UTR 37°C WT



B

cspA 5'UTR 10°C WT

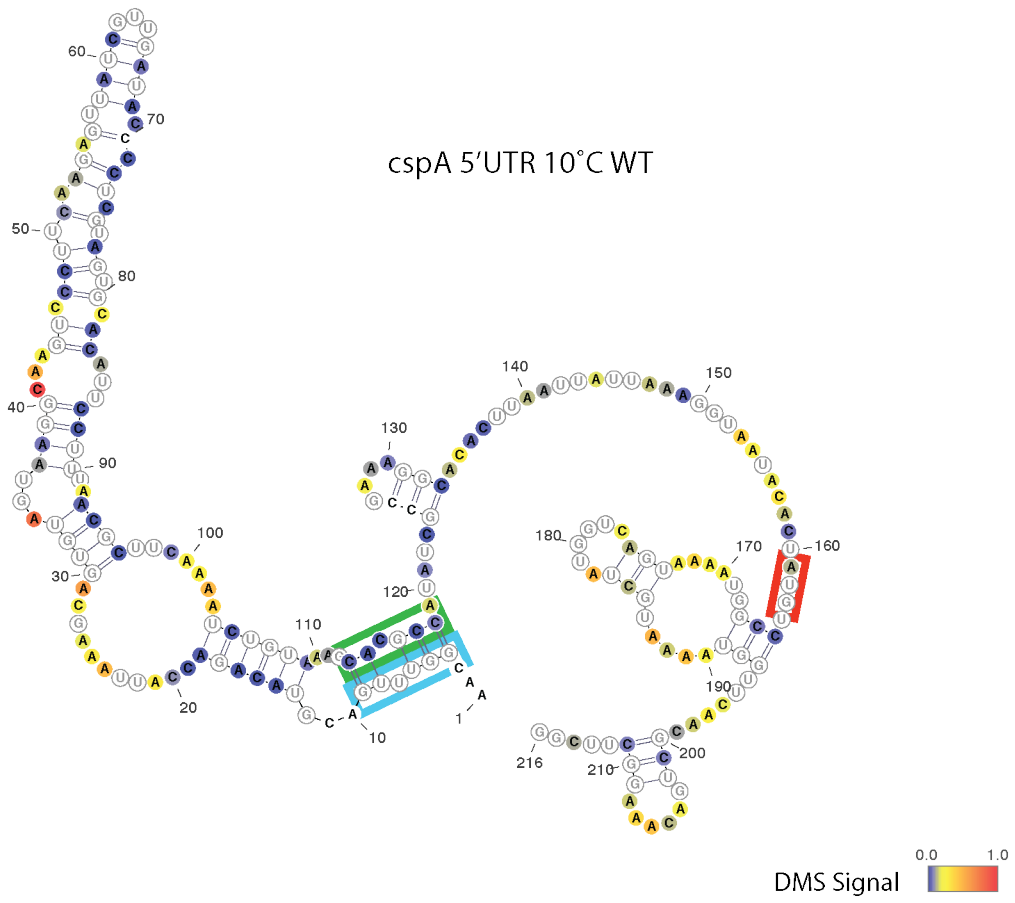
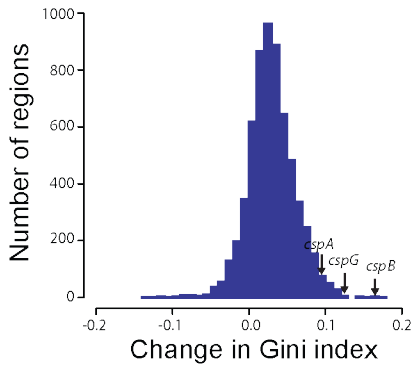


Fig S5: CspB UTR structure is modulated by cold shock

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

A



B

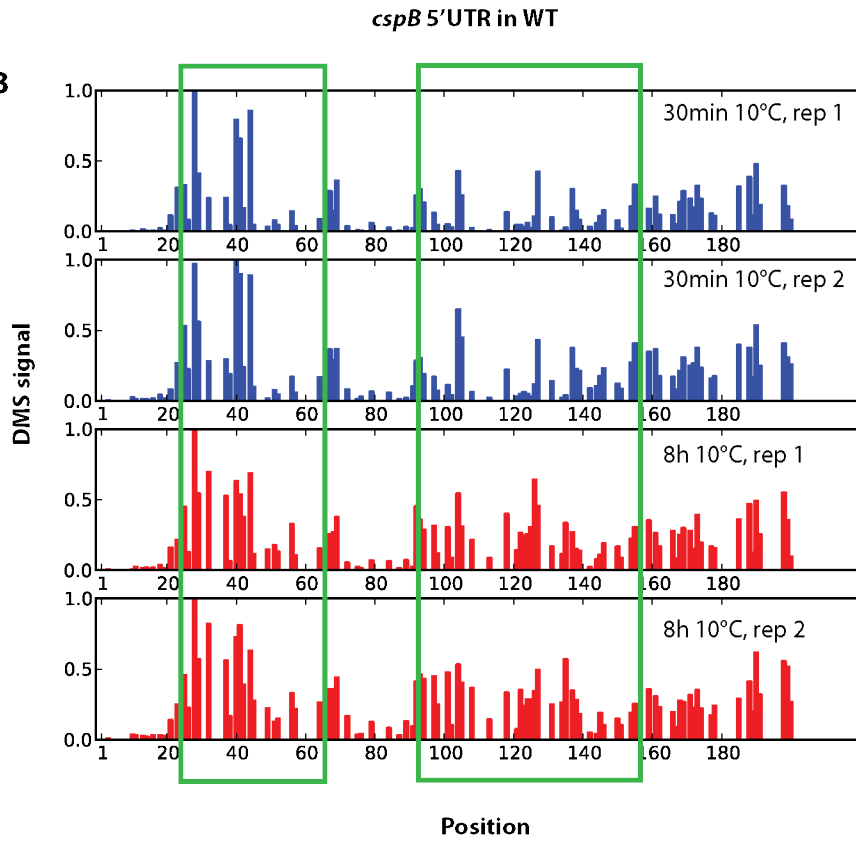


Fig S6: CspB UTR structure is modulated during cold recovery (continued)

C

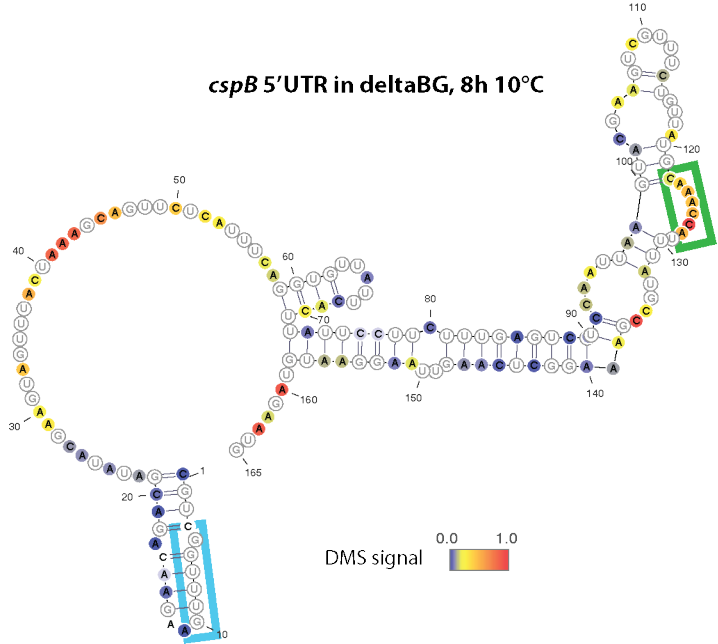


Fig S6: CspB UTR structure is modulated during cold recovery

- (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, the 5' end of the *cspB* message, 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 $\Delta cspBG$, presented as in Figure 5.

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 OD₄₂₀ = .001 in fresh medium and grown until OD₄₂₀ reached 0.4 where samples were collected. For 10°C samples, cultures were grown to OD₄₂₀ = 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 MgCl₂, 100 mM NH₄Cl, 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 profiling experiments, and for total mRNA used for *in vitro* DMS-seq experiments, 4mL of OD₄₂₀ = 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/GATCGTCCGACTGTAGAACTCTGAACCTGTCGGTGGTCGCCGTATCATT/iSp18/CACTCA/iSp18/CAAGCAGAAGACGGCATAACGAATTGATGGTGCCTACAG 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 μ l 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% β -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 μ L 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. 2 μ g of mRNA was denatured at 95 °C for 2 min, cooled on ice and refolded in 90 μ L 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 μ l 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 (Zn²⁺ 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 μ l 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 1 μ C of Perkin Elmer EasyTag 35 S labeled methionine (Product # NEG709A) was mixed with 5 μ L 288 μ mol unlabeled methionine and 24 μ L media. At the time of capture, 900 μ L 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 μ L of 50% trichloroacetic 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 100 μ L of sample through a 25mm APFC glass fiber filter (Millipore APFC02500) pre-wetted with 750 μ L of 5% TCA on a vacuum stand, and washing three times with 750 μ L 5% TCA and three times with 750 μ L 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 modified 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 than 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 Å². 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 that 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 performed 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 experiment. 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 $\ll 1$, 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 F1p-catalyzed excision of the antibiotic-resistance determinant. *Gene* 158, 9–14.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

9-10-2014
Date