

©Copyright 2025

Han Kyou Choi

Bacterial Fitness Optimization Through Noise Robustness and Oscillatory Regulation

Han Kyou Choi

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Paul A. Wiggins, Chair

Andrew H. Laszlo

Beth A. Traxler

Program Authorized to Offer Degree:

Physics

University of Washington

Abstract

Bacterial Fitness Optimization Through Noise Robustness and Oscillatory Regulation

Han Kyou Choi

Chair of the Supervisory Committee:

Paul A. Wiggins

Physics and Bioengineering

Bacterial cells thrive in stochastic environments despite the inherent constraints of microscopic scale and biochemical noise. This thesis investigates how bacteria optimize fitness through different regulatory strategies against various sources of stochasticity. We first establish a regulatory principle, the Robustness-Load Trade-Off model, which provides three key predictions: (1) the majority of essential proteins are expressed in vast excess, (2) essential genes are transcribed above one message per cell cycle, and (3) protein expression is tuned by load balancing transcription and translation. Experimental validation in *Acinetobacter baylyi* confirms these predictions: single-cell and genome-wide analyses demonstrate that 70% of essential proteins are expressed in vast overabundance and that overabundance is higher for low-expression proteins. Finally, we explore an unforeseen phenomenon of homeostatic regulation: temporal oscillations of metabolites. We observe metabolic oscillations in deoxynucleotide triphosphate (dNTP) synthesis via replication fork velocity measurements. By developing a periodically driven temperature-based synchronization technique, we hope to show that oscillations emerge across bacterial species. Collectively, this work uncovers how bacteria are robust to intrinsic noise and dynamics fluctuations of complex biochemical networks.

Table of Contents

	Page
List of Figures	v
List of Tables	vii
Chapter 1: Introduction	1
1.0.1 Noise robustness and metabolic load determine the principles of central dogma regulation	2
1.0.2 Protein overabundance is driven by growth robustness on a threshold-like fitness landscape	3
1.0.3 Measuring metabolic oscillations	4
1.1 Bibliography	6
Chapter 2: Noise robustness and metabolic load determine the principles of central dogma regulation	7
2.1 Introduction	8
2.2 Results	9
2.2.1 Defining RLTO model	9
2.2.2 RLTO predicts protein overabundance	11
2.2.3 Understanding the rationale for overabundance	13
2.2.4 Overabundance is observed in a range of experiments	13
2.2.5 RLTO predicts a one-message transcription threshold	15
2.2.6 A lower threshold is observed for message number	15
2.2.7 Message number distribution is conserved	17
2.2.8 Translation efficiency is predicted to increase with transcription.	18
2.2.9 Load balancing is observed in eukaryotic cells.	20
2.2.10 RLTO model predicts observed noise in yeast.	22
2.2.11 Reducing noise is the rationale for load balancing.	22

2.3	Discussion	23
2.3.1	What are the biological implications of noise?	23
2.3.2	Implications for nonessential genes.	25
2.3.3	Implications of overabundance for inhibitors.	25
2.3.4	The principles that govern central dogma regulation.	26
2.4	Materials and Methods	26
2.4.1	RLTO model.	26
2.5	Supplimentary Information	27
2.5.1	Methods: Detailed description of the noise model	27
2.5.2	Methods: The derivation of the RLTO growth rate	31
2.5.3	Discussion: The fitness landscape of a trade-off.	40
2.5.4	Methods: Central dogma optimization	41
2.5.5	Discussion: Understanding the rationale for overabundance	46
2.5.6	Discussion: RLTO predicts larger overabundance in bacteria.	47
2.5.7	Discussion: RLTO predicts proteins are buffered to depletion.	47
2.5.8	Results: Detailed development of load balancing	48
2.5.9	Results: Translation efficiency is predicted to increase with transcription.	49
2.5.10	Results: RLTO predicts that message number responds to message cost	50
2.5.11	Results: RLTO predicts the yeast global regulatory response	50
2.5.12	Methods: Prediction of the proteome fraction	52
2.5.13	Results: Load balancing is observed in eukaryotic cells	54
2.5.14	Discussion: Relation between load balancing and previous results	54
2.5.15	Methods: Analysis of translational limits & gene-specific load analysis	57
2.5.16	Results: Increased protein-specific cost reduces the optimal translation efficiency.	58
2.5.17	Discussion: Translation limits in <i>E. coli</i>	59
2.5.18	Methods: Estimate of the message cost and metabolic load	60
2.5.19	Model robustness & exploring alternatives to RLTO	64
2.5.20	Results: Overabundance is a robust prediction	65
2.6	Quantitation of central dogma parameters for one-message-rule	67
2.6.1	Methods: Selection of central dogma parameter estimates	67
2.6.2	Methods: Quantitative estimates of central dogma parameters	69
2.6.3	Results: Histograms of central dogma transcriptional statistics	72

2.6.4	Discussion: <i>E. coli</i> essential genes below the one-message-rule threshold	72
2.7	Analysis of gene-expression noise	72
2.7.1	Results: RLTO model predicts non-canonical noise scaling	74
2.7.2	Methods: Analysis of gene expression noise	74
2.7.3	Results: Non-canonical noise scaling is observed in yeast.	78
2.7.4	Methods: Parameter-free prediction of noise from protein-message relation.	79
2.7.5	Results: Parameter-free prediction of noise-abundance in yeast	84
2.7.6	Discussion: Implications of noise	84
2.8	Data Tables	85
2.8.1	Datasets for one-message-rule analysis.	85
2.8.2	Datasets for load-balancing analysis.	85
2.8.3	Datasets for noise analysis.	86
2.8.4	Datasets for one-message-rule exceptions.	86
2.9	Bibliography	86
Chapter 3:	Protein overabundance is driven by growth robustness on a threshold-like fitness landscape	93
3.1	Introduction	94
3.2	Results	95
3.2.1	Natural competence facilitates knockout-depletion	95
3.2.2	Target proteins are depleted by dilution	98
3.2.3	Replication persists during DnaN depletion	98
3.2.4	Many essential knockouts undergo transient growth	99
3.2.5	The fitness landscape is threshold-like	100
3.2.6	Protein overabundance	100
3.2.7	The RLTO model predicts protein overabundance	101
3.2.8	TFNseq determines overabundances genome-wide	101
3.2.9	Many essential proteins have vast overabundance	103
3.2.10	Low-expression genes are highly overabundant	104
3.3	Discussion	105
3.3.1	The shape of the fitness landscape	105
3.3.2	The rationale for a threshold abundance	106

3.3.3	The rationale for overabundance	106
3.3.4	Biological implications	107
3.4	Conclusion	107
3.5	Supplementary material	108
3.5.1	Supplementary Discussion	108
3.5.2	<i>Acinetobacter baylyi</i> strains, manipulation, and culturing	110
3.5.3	Growth models for knockout-depletion experiments	113
3.5.4	Imaging-based knockout-depletion experiments	116
3.5.5	Statistical analysis of TFNseq trajectories	134
3.5.6	RNA-Seq analysis of transcription	140
3.5.7	Robustness Load Trade-Off (RLTO) Model	144
3.5.8	Methods: Statistical procedures	151
3.5.9	Distributions and conventions	153
3.5.10	Description of supplementary data	154
3.6	Bibliography	156
Chapter 4:	Measuring metabolic oscillations	162
4.1	Introduction	163
4.2	Testing the dNTP-oscillation model by characterizing fork-velocity oscillations. 164	
4.2.1	Metabolic pathway of dNTP	164
4.2.2	Motivation	164
4.2.3	Preliminary work	165
4.2.4	Approach	165
4.3	Cell-cycle dependent measurements through cell synchronization	166
4.3.1	Motivation	166
4.3.2	Approach	167
4.3.3	Preliminary work	168
4.3.4	Future work	168
4.4	Bibliography	170

List of Figures

Figure Number	Page
2.1 The RLTO Model.	10
2.2 The RLTO model predicts overabundance is optimal.	12
2.3 A lower threshold for transcription: The one message rule.	16
2.4 How are transcription and translation balanced?	19
2.5 RLTO predicts the magnitude of noise in yeast.	21
2.6 Central dogma regulatory principles: overabundance, one-message rule, and load balancing.	24
2.7 The protein abundance is approximately gamma distributed.	30
2.8 Four perspectives on the fitness landscape.	38
2.9 Four perspectives on the dependence of optimal overabundance on relative load Λ	42
2.10 Four perspectives on load balancing.	43
2.11 Optimal expression levels are buffered.	45
2.12 The message cost affects transcription genome wide.	51
2.13 Increased protein cost decreases optimal translation efficiency.	57
2.14 Exploring the mathematical mechanism of overabundance.	63
2.15 The one message rule. Panel A: One-message-rule for essential genes.	70
2.16 Yeast noise fit against canonical noise model, with a noise floor.	75
2.17 Fit to rescale fluorescence intensity to protein number.	80
2.18 Load balancing predicts the scaling of noise.	83
3.1 Knockout-depletion experiments in ADP1.	96
3.2 The fitness landscape.	97
3.3 A proteome-wide analysis of protein overabundance.	102
3.4 How rate-limited kinetics shapes the fitness landscape.	105
3.5 Panel A: Mutant abundances for trajectory models.	114
3.6 Knockout-depletion experiment: IS element (Non-essential).	128

3.7	Knockout-depletion experiment: $\Delta dnaA$.	130
3.8	Knockout-depletion experiment: $\Delta dnaN$.	131
3.9	Knockout-depletion experiment: $\Delta ftsN$.	133
3.10	Knockout-depletion experiment: $\Delta murA$.	135
3.11	Message number distributions for essential and non-essential genes in <i>E. coli</i> and <i>A. baylyi</i> .	136
3.12	Transcriptome comparisons.	137
3.13	Panel A: Threshold and message number.	138
3.14	Proteome-wide analysis of proliferation dynamics.	139
3.15	Modeling the asymmetric fitness landscape and stochastic gene expression for RLTO framework.	149
4.1	Preliminary results for attempt to synchronize temperature sensitive mutants(<i>dnaC2</i>) through temperature oscillation.	169

List of Tables

Table Number		Page
2.1	RLTO model parameters by figure.	40
2.2	Central dogma parameters for three model organisms with detailed references.	66
2.3	Below-threshold essential genes identified in <i>E. coli</i>.	73
3.1	Measured overabundance for sequencing- versus imaging-based approaches.	99
3.2	Summary of strains used in this study.	111
3.3	Detailed results from fitting imaging-based knockout-depletion experiments.	126
3.4	Gene ontology classifications and terms.	140

Acknowledgments

As I wrap up my PhD program with this thesis, I would like first to thank my mentor, **Paul A. Wiggins**. Your knowledge of biology, physics, and statistics never ceases to amaze me. Time after time, I have asked you countless questions, and you have always been patient, explaining and answering them in detail. You have taught me to become a better scientist (hopefully). Without you, I would not have found biophysics as fascinating as it is.

I want to thank my parents, **Byung Sun Choi** and **In Soon Kim**. You have always nurtured my curiosity and supported me both emotionally and financially throughout my PhD journey. Your unconditional love and sacrifice have been the foundation of my progress and success.

To my forever girlfriend and now wife, **Amy Chiu**, you have been by my side during the most demanding phase of my career. You have pushed me beyond my comfort zone, inspiring me to strive for more. I am especially grateful for your help in editing and reviewing my thesis, emails, applications, and so much more. Without your support and writing skills, I could not have completed this PhD.

To my dearest former colleagues, **Dean Huang**, thank you for helping me through my journey in PhD. You have provided countless advice that helped me throughout my research. I enjoyed your brain teasers during our time in the lab. Although our whimsical conversations often evolved into heated debates, our endless discussions about hypothetical scenarios provided much-needed breaks from research.

I give my thanks to the rest of my former colleagues: **Dani Koch**, it was a pleasure to talk about interesting topics without having to be judged (ahem, unlike Dean). **Teresa Lo**, it was nice to work with you, and I appreciate all the feedback during presentations. **Kevin**

Cutler, I appreciate your timely response to my non-ending questions regarding omnipose: it has been a lifesaver.

In the last year of my PhD research, I was fortunate to have several undergraduate students assist me. **Ivy LeGassick** was the first to experience my mentorship and provided much-needed help, especially when I was starting my oscillations project. **Joshua Beard**, my most physicist-like student, it was a pleasure to have you—the combination of your experience in electronics, physics, and mathematics was especially valuable and inspiring as a fellow physicist. **Ulises Espino Loza**, it was inspiring to see how much progress you made in such a short time. **Federico Vargas**, I appreciate your thoughtful input in developing and improving our experiment; your contributions have been tremendously helpful. **Isaac Briefer**, my latest Physics student, it was wonderful to have you on board, especially with how quickly you learned and contributed. I hope you continue to pursue this field with enthusiasm.

Equally important in my journey, my gratitude extends to my college friends group, fondly known as the **Hot Pot Gang**, for all the countless nights of bottomless hot pot. To the members: **Gina So**, thank you for always ensuring we had more than enough protein; **Josh Lee**, thank you for pushing my spice tolerance in honor of my Korean heritage; **Angela Zhu**, our “hot pot sleeper agent,” thank you for the ~ 6 lb (~ 2.7 kg) nappa cabbage; and **Johan Velo**, thank you for the questionable relationship advice that, surprisingly, led to my current married life. **Alex Tsai**, thank you for the unforgettable memories and for bringing our group closer together.

Lastly, but not least, I would like to thank all the members of my supervisory committee: **Beth Traxler**, **Andrew Laszlo**, **Jason Detwiler**, and **Armita Nourmohammad**. I am deeply grateful for your guidance and constructive feedback throughout my PhD.

Chapter 1

Introduction

There are many decisions we need to make in our lives. When and what should I have my next meal? How many rolls of toilet paper do I need to stock up on in preparation for a global catastrophe? When should I write my thesis? There is much to consider when trying to optimize our lifestyles. Throughout my journey studying bacteria, I have been asking the following big question: How do bacteria optimize for fitness in spite of the complex and stochastic network of biochemical reactions?

Bacterial cells are fundamentally constrained by their microscopic size, and the unpredictable nature of their environment. They must constantly adapt to fluctuating nutrients, temperature, and many other stressors. To adapt in such a fluctuating environment, bacterial cells have evolved diverse strategies, such as bet-hedging and regulatory motifs that allow populations to persist in various conditions. Yet, beyond these environmental challenges, bacteria face an even more fundamental issue: the inherent stochasticity of gene expression. Protein expression is subject to intrinsic noise, the random noise as a consequence of two stochastic processes: transcription followed by translation. Despite these challenges, bacteria can be found in nearly every corner of our planet. My research seeks to shed some light on how bacteria are able to optimize fitness in such a complex environment.

For the rest of this chapter, I will provide an overview of chapters two, three, and four: which are my work for the last 6 years. Starting with the Robustness-Load Trade-Off model, I will describe a regulatory principle of micro-organism against intrinsic noise. This gave

us insight into how cells have to optimize against the unavoidable consequences in the microscopic scale: stochasticity in single molecular processes. Secondly, I will describe and discuss the observation of overabundance, a major prediction from our Robustness-Load Trade-Off model. This has been my major work in my PhD program. And lastly, I will briefly describe an alternative regulatory strategy : feedback regulation. This has been my research of last years in my PhD program. I will briefly explore feedback regulation and temporal oscillations in such systems, and discuss methods to test our predictions.

1.0.1 Noise robustness and metabolic load determine the principles of central dogma regulation

At the scale of bacteria and other single-celled organisms, all cellular processes are subject to noise due to stochasticity of single-molecular processes. So, how do bacteria optimize protein expression levels in light of this unavoidable noise? This is the question that got us into studying and analyzing the fitness associated with protein expression and protein noise. To address this, we developed the Robustness-Load Trade-Off (RLTO) model, which provides three novel predictions: (i) optimal protein expression levels for many genes are in vast overabundance, (ii) essential genes are transcribed above a lower limit of one message per cell cycle, (iii) gene expression is regulated by load balancing: a principle in which gene expression is controlled by tuning both transcription and translation rates.

The RLTO model is built on three critical components: essential protein functions have a threshold-like dependency on protein abundance, which is stochastic; translation and transcription generate metabolic load; and cell growth depends on the proper function of all essential proteins. These simple components allow for a tractable mathematical model of optimal protein expression.

The rationale behind the overabundance and one-message rule can be explained with the following analogies; the abundance of socks found in households. Although we need only a one pair of socks per day, we keep extra spare of socks in our homes. This is because not having enough socks (a pair of socks) disrupts our morning routine, while having extra pairs of

socks is not costly. The same can be said for essential proteins; not having enough essential protein leads to loss of function, while making excess protein does not lead to significant metabolic burden for the cell. We can also use the same analogy to explain the one-message rule: although we only need a pair of socks for a portion of our day, we always have at least one pair available. Likewise, cells synthesize at least one essential mRNA molecule during the cell cycle to avoid periods of no expression.

The rationale behind load balancing is rather straightforward. Protein synthesis can be interpenetrated as a two-stage amplification involving transcription and translation. A given protein abundance is dependent on mRNA per cell cycle and translation efficiency. Higher (lower) mRNA per cell cycle results in a lower (higher) noise associated with protein abundance but at the cost of higher (lower) metabolic load. Ultimately optimal protein overabundance is achieved by balancing both amplifications; mRNA number per cell cycle, and translation efficiency. High (low) expression levels are achieved by high (low) levels of transcription followed by high (low) levels of translation per message.

1.0.2 Protein overabundance is driven by growth robustness on a threshold-like fitness landscape

One of the major predictions of the RLTO model is that most essential proteins should be expressed in vast excess. How can we test this prediction? We observe and quantify overabundance of essential proteins using the naturally-competent bacterium *Acinetobacter baylyi* by characterizing the proliferation dynamics of essential-gene knockouts at a single-cell scale (by imaging) as well as at a genome-wide scale (TFNseq and RNA-seq).

In these experiments, essential genes are knocked out and we observe the cells proliferate for multiple generations as target protein levels are diluted as the essential genes are no longer expressed. The crux of this experiment is that essential proteins are expressed in much greater excess than required for function. Due to the overabundance essential proteins, cells assume normal cellular process until the protein product of the knocked out essential gene is diluted below the critical concentration required for function. Going back to our

previous analogy of socks, it is equivalent to taking away the laundry machine and counting how many days one can continue with their sock reserves. By counting the days before all the clean socks run out, we can obtain a quantitative number of overabundance of socks. Similarly, in our experiment, we count how long the cells are able to proliferate before the cells arrest due to insufficient amount of essential proteins.

Through these experiments, we are able to verify and observe major predictions of the RLTO model: many essential proteins exhibit steep fitness thresholds, about 70% of essential proteins are overabundant, and essential genes are transcribed above a lower limit of about one message per cell cycle. Furthermore, we observe that the overabundance of essential proteins are higher for genes that are transcribed in lower numbers, which is key principles predicted by the RLTO model. Our quantitative measurement of overabundance supports the general regulatory principles of RLTO model.

1.0.3 Measuring metabolic oscillations

During my last year in the PhD program, I have switched gears and have been studying an alternative regulatory principle found in bacterial systems: feedback regulation. In order to reach homeostasis, many cellular processes rely on a complex network of regulatory feedback systems. Although many specific cellular feedback systems have now been characterized, important examples exhibit an unexpected phenomenon: temporal oscillations. For example, ATP abundance is thought to be tightly maintained since they are critical to nearly every biological process; however, recent evidence demonstrates fluctuations in metabolite levels [1]. This raises a question: is there a fundamental limit on the precision of cellular homeostasis?

Consider a feedback-based regulatory control system similar to adjusting the shower tap for optimal water temperature. Most of us know roughly the position of the shower tap; however, we often need to adjust the position for the optimal water temperature. If the temperature of the water is too low, we will increase the water temperature; however, we often overshoot the temperature. Over time we will correct the temperature of the water by adjusting the tap, experiencing overshoots and undershoots until we reach the

desired temperature. This phenomenon is more pronounced when we are in a rush or desire a rapid response. Similarly, we anticipate a trade-off between rapidity and precision in feedback-based regulatory systems of metabolites within bacteria.

We will study the dynamics of metabolic oscillations with two different approaches: (i) studying oscillating fork velocity through lag-time analysis of cultures of different RNR expression levels, and (ii) bulk assay measurements of synchronized cells. Approach (i) involves performing similar experiments those done by Dean *et al.* [2] but we will use our plasmid-based IPTG-inducible *nrdAB* mutant strain to express *nrdAB* at non-endogenous levels. Because replication appears to be nucleotide-limited *in vivo*: Increases in the pool levels lead to increased fork velocity, whereas decreases in the pool levels have the opposing effect [3], we hope to use replication fork velocity as an indicator of dNTPs pool levels. By circumventing the endogenous expression of *nrdAB*, we expect to observe a different oscillatory dynamics in replication-fork velocity as a consequence of perturbing the native feedback regulation.

Method (ii) involves developing a novel synchronization technique: synchronization of temperature sensitive alleles through oscillating temperature. Currently, there are two methods for achieving synchronization: *arrest-based* and *baby-cell-enrichment-based* methods. Each method has their its advantages and disadvantages. In the **cell-cycle-arrest-based approaches**, the cell cycle is arrested at initiation and then released. Although this method results in a sample with replication synchronization, other cellular process can be perturbed significantly. On the other hand, **baby-cell-enrichment-based approaches** do not interfere with cell physiology; however, due to the stochasticity in the duration of the B period[4], the method may have poorer synchronization of replication. To achieve high synchronization and with minimal interfere with cell physiology, we propose synchronization through oscillating temperature. By culturing temperature sensitive alleles (*dnaC2* and *dnaC28*) while applying temperature oscillations between permissive and nonpermissive temperatures, we hope to achieve a physiologically synchronous sample, thus allowing us to perform bulk assays of synchronized cells of different cell-cycle stages.

1.1 Bibliography

- [1] W.-H. Lin and C. Jacobs-Wagner, “Connecting single-cell ATP dynamics to overflow metabolism, cell growth, and the cell cycle in *Escherichia coli*,” *Curr Biol*, vol. 32, no. 18, 3911–3924.e4, Sep. 2022. DOI: [10.1016/j.cub.2022.07.035](https://doi.org/10.1016/j.cub.2022.07.035).
- [2] D. Huang, A. E. Johnson, B. S. Sim, T. W. Lo, H. Merrikh, and P. A. Wiggins, “The in vivo measurement of replication fork velocity and pausing by lag-time analysis,” *Nature Communications*, vol. 14, no. 1, 2023. DOI: [10.1038/s41467-023-37456-2](https://doi.org/10.1038/s41467-023-37456-2).
- [3] M. Zhu *et al.*, “Manipulating the bacterial cell cycle and cell size by titrating the expression of ribonucleotide reductase,” *mBio*, vol. 8, no. 6, Nov. 2017. DOI: [10.1128/mBio.01741-17](https://doi.org/10.1128/mBio.01741-17).
- [4] S. M. Mangiameli, B. T. Veit, H. Merrikh, and P. A. Wiggins, “The replisomes remain spatially proximal throughout the cell cycle in bacteria,” *PLoS Genet*, vol. 13, no. 1, e1006582, Jan. 2017. DOI: [10.1371/journal.pgen.1006582](https://doi.org/10.1371/journal.pgen.1006582).

Chapter 2

Noise robustness and metabolic load determine the principles of central dogma regulation

Originally published as: [1] T. W. Lo, H. J. Choi, D. Huang, and P. A. Wiggins, “Noise robustness and metabolic load determine the principles of central dogma regulation,” *Science Advances*, vol. 10, no. 34, eado3095, 2024. DOI: [10.1126/sciadv.ado3095](https://doi.org/10.1126/sciadv.ado3095).

Author contributions: T.W.L., H.J.C., D.H., and P.A.W. conceived the research. T.W.L. and H.J.C. are listed as co-first authors, jointly contributed to the conception and overall design of the study. T.W.L., H.J.C., D.H., and P.A.W. wrote the paper. H.J.C. primary contribution were in developing the RLTO mathematical model, performing the analyses that revealed optimal overabundance as a consequence of the model, and articulating the broader implications of the RLTO framework, particularly in the context of bacterial systems.

Abstract

The processes of gene expression are inherently stochastic, even for essential genes required for growth. How does the cell maximize fitness in light of noise? To answer this question, we build a mathematical model to explore the trade-off between metabolic load and growth robustness. The model provides insights for principles of central dogma regulation: Optimal protein expression levels for many genes are in vast overabundance. Essential genes are transcribed above a lower limit of one message per cell cycle. Gene expression is achieved by load balancing between transcription and translation. We present evidence that each of these regulatory principles is observed. These results reveal that robustness and metabolic

load determine the global regulatory principles that govern gene expression processes, and these principles have broad implications for cellular function.

2.1 *Introduction*

What rationale determines the optimal transcription and translation level of a gene in the cell? Protein expression levels optimize cell fitness [2, 3]: Too low of an expression level of essential proteins slows growth by compromising the function of essential processes [4, 5], whereas the overexpression of proteins slows growth by increasing the metabolic load [6]. This trade-off naïvely predicts that the cell maximizes its fitness by a Goldilocks principle in which cells express just enough protein for function [7]; however, achieving growth robustness is nontrivial, since all processes at the cellular scale are stochastic, including gene expression [8]. This biological noise leads to significant cell-to-cell variation in protein numbers, even for essential proteins that are required for growth [9, 10]. The optimal expression program must therefore ensure robust expression of hundreds of distinct essential gene products. In this chapter, I present and analyze a minimal mathematical model, Robustness-Load Trade-Off model, to explore the consequences of growth robustness on the central dogma regulatory program.

The Robustness-Load Trade-Off(RLTO) model includes three critical components: (i) Protein levels are stochastic and the single-cell growth rate depends upon them, (ii) gene transcription and translation generate a metabolic load, and (iii) cell growth is dependent on a large number of essential genes. These model characteristics result in a highly-asymmetric fitness landscape. The optimization of expression on this asymmetric landscape predicts unique phenomenology that is absent from previous models (*e.g.* [11]). In this chapter, we provide three major predictions of the RLTO model: (a) optimal protein expression levels for many genes are in vast overabundance, (b) essential genes are transcribed above a lower limit of one message per cell cycle, (c) optimal gene expression is achieved by load balancing between transcription and translation. We will also provide evidence to support

these predictions. Overall, the RLTO model predicts global regulatory principles for essential genes.

2.2 Results

2.2.1 Defining RLTO model

The protein number N_p expressed from gene i is the product of two sequential stochastic processes: transcription and translation [12], leading to cell-to-cell variation in protein number, which we will refer to as *noise*. In our analysis, we will model gene expression using the canonical steady-state noise model [13]. (See Fig. 2.1A.) In this model, the numbers of proteins N_p for gene i are predicted to be gamma-distributed [14], in close agreement with observation [10]. The distribution is described by two gene-specific statistical parameters: *message number* (μ_m), defined as the mean number of messages transcribed per cell cycle for gene i , and the *translation efficiency* (ε), the mean number of proteins translated from each message transcribed for gene i . The mean protein abundance is their product: $\mu_p = \mu_m \varepsilon$. These parameters can be expressed in terms of ratios of the rates of the underlying gene expression processes, as described in Supplementary Material Sec. a..

How should the effect of essential protein expression on growth rate be modeled in the context of the RLTO model? Much recent work has focused on cellular resource allocation to functional sectors (*e.g.* [15]). In this approach, an optimization is performed by the *coordinated* modulation of the abundance of all proteins in a particular sector, leading to a trade-off between functional capacities of the cell. However, in the RLTO model, the optimization is fundamentally different: We consider the *uncoordinated* modulation of the abundance of protein species i due to noise. For these incoherent changes, we generically expect proteins to exhibit rate-limited kinetics: Increases in the protein number N_p above a threshold level n_p has minimal effect on the rate since other chemical species (proteins, metabolites, *etc.*) are rate limiting [16]. However, if the protein number N_p falls below the threshold n_p , then protein species i becomes rate limiting and leads to a significant slowdown

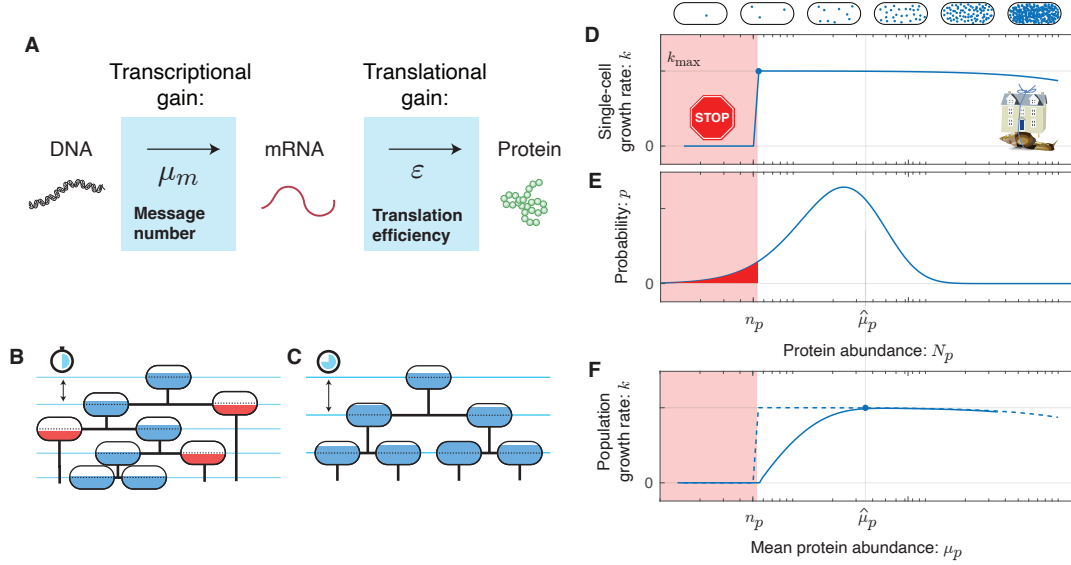


Figure 2.1: **The RLTO Model. Panel A: Gene expression is stochastic.** The central dogma describes a two-stage stochastic process where genes are first transcribed and then translated. The transcription process transcribes an average of μ_m messages per cell cycle. The translation process translates an average of ϵ proteins per message. **Panel B & C:** A schematic cell lineage tree is shown during exponential growth. For a specific protein i , the cell fill represents the protein number N_p relative to its threshold number n_p required for cell growth. **Panel B:** Reducing the mean expression level reduces doubling time; however, stochasticity in expression results in below-threshold cells (red fill) which grow slowly. **Panel C:** Increasing protein expression increases the doubling time; however, all cells are above threshold (blue fill). **Panel D: The fitness landscape as a function of protein number.** Growth arrests for protein number N_p smaller than the threshold level n_p (red) due to the failure of essential processes. High expression levels are penalized due to the metabolic cost of protein expression. This trade-off leads to a highly asymmetric fitness landscape: The relative metabolic cost of overabundance is small relative to the cost of growth arrest due to the large size of the total metabolic load N_0 . **Panel E: The gene expression process is stochastic.** There is significant cell-to-cell variation in protein abundance (N_p) around the mean level (μ_p). Even for mean expression levels significantly above the threshold level n_p , some cells fall below threshold (red). The distribution in protein number is modeled using a gamma distribution [10], parameterized by message number μ_m and translation efficiency ϵ . **Panel F: The robustness load trade-off determines the optimal expression level.** The population growth rate depends on the distribution of the protein number. The asymmetry of the fitness landscape drives the optimal expression level far above the threshold level due to the high fitness cost of low protein abundance.

of the growth rate. In the RLTO model, we coarse-grain the details of this growth slowdown as growth arrest. (See Fig. 2.1.) There is already some precedent for the use of this type of threshold (*e.g.* [17]), but we will demonstrate that the detailed form of the fitness landscape is not important. (See Supplementary Material Sec. 2.5.19.) Although sufficiently detailed knowledge of the relevant molecular and cellular biology could be used to predict the protein thresholds n_p , we will treat these as gene-specific unknown parameters.

As shown in Materials and Methods of this chapter, the relative cellular fitness with respect to the expression of gene i can be computed by combining the fitness losses associated with robustness (Eq. 2.6) and metabolic load (Eq. 2.8):

$$\frac{\Delta k}{k_0} = -(\Lambda + \frac{\varepsilon}{N_0})\mu_m - \frac{1}{\ln 2}\gamma(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}), \quad (2.1)$$

where the first term represents fitness loss due to metabolic load of transcription and translation while the second term represents loss due to the arrest of essential processes. γ is the regularized incomplete gamma function and the central distribution function (CDF) of the gamma distribution. (See Supplementary Material Sec. a. and 2.5.2.) In summary, the model depends only on a single global parameter: the relative metabolic load Λ and three gene-specific parameters: the threshold number n_p , the message number μ_m and the relative translation efficiency ε/N_0 . We propose that the cell is regulated to maximize the growth rate with respect to transcription (message number) and translation (translation efficiency). The fitness landscape predicted by the RLTO model for representative parameters is shown in Fig. 2.2A.

2.2.2 RLTO predicts protein overabundance

The optimal regulatory program (μ_m and ε values) can be predicted analytically. They depend on only a single global parameter, the relative load Λ , and the gene-specific threshold number n_p . Since the threshold number is not directly observable experimentally, we will instead predict the optimal overabundance o , defined as the ratio of the mean protein number

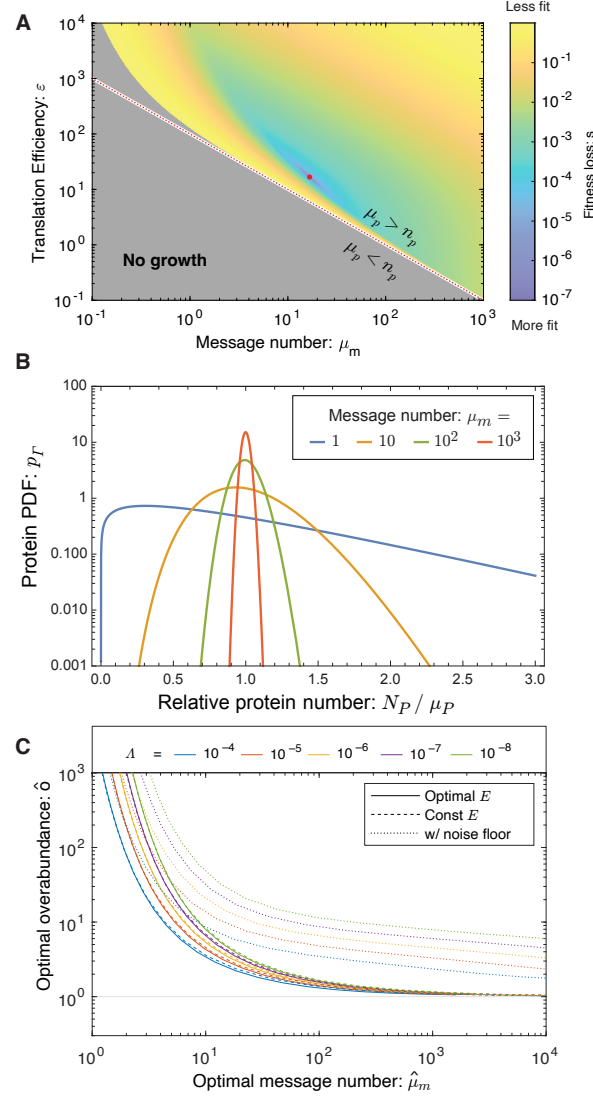


Figure 2.2: **The RLTO model predicts overabundance is optimal.** **Panel A: Fitness landscape determines optimal message number and translation efficiency.** The fitness loss ($s \equiv \ln k_{\max}/k$) is shown as a function of message number (μ_m) and translation efficiency (ε). The red dotted curve represents programs where the mean protein number is equal to the threshold ($\mu_p = n_p$) and the red dot represents the optimal regulatory program $(\hat{\mu}_m, \hat{\varepsilon})$. **Panel B: Gene-expression noise.** Due to the stochasticity of gene expression processes at equilibrium, the protein number N_p is gamma-distributed [10]. For high-expression genes, expression has low noise and the protein number is tightly distributed around its mean; however, for low-expression genes, expression is noisy and the distribution is extremely wide. **Panel C: Overabundance is optimal for all genes.** For high-expression genes, low overabundance is optimal ($\mu_p \approx n_p$); however, for low-expression genes, vast overabundance is optimal ($\mu_p \gg n_p$). From a quantitative perspective, overabundance depends on the relative load Λ ; however, the qualitative dependence is invariant to over an orders-of-magnitude range of values.

to the threshold number:

$$o \equiv \mu_p/n_p. \quad (2.2)$$

As shown in Fig. 2.2C, the RLTO model generically predicts that the optimal protein fraction is overabundant ($o > 1$); however, the overabundance is not uniform for all proteins. For highly-transcribed genes ($\mu_m \gg 1$) like ribosomal genes, the overabundance is predicted to be quite small ($o \approx 1$); however, for message numbers approaching unity, the overabundance is predicted to be extremely high ($o \gg 1$). At a quantitative level, the relation between optimal overabundance and message number depends on the relative load (Λ), but its phenomenology is qualitatively unchanged over orders of magnitude variation in Λ .

2.2.3 Understanding the rationale for overabundance

To explore both the robustness of the protein overabundance prediction and to understand its mathematical rationale, we explored a collection of more complex models numerically. (Supplementary Material Sec. 2.5.19.) The key mathematical feature that drives overabundance is not the assumption of growth arrest, but rather the strong asymmetry of the fitness landscape: the high cost of protein *underabundance* and the low cost of protein *overabundance*. (See Fig. 2.1EF.) The population growth rate (Panel 2.1F) can be understood qualitatively as the convolution of the single-cell growth rate (Panel 2.1D) with the probability density function (PDF) of the protein abundance (Panel 2.1E). In the RLTO model, this asymmetry is parameterized by the relative load (Λ), defined as the relative metabolic cost of transcribing an additional message. Since we estimate that $\Lambda < 10^{-5}$, this cost is very low relative to the total metabolic cost of the cell, therefore we expect this asymmetry, and the prediction of the RLTO model, to be robust.

2.2.4 Overabundance is observed in a range of experiments

The RLTO Model predicts that all essential proteins are overabundant. In general, the RLTO model predicts that protein numbers have very significant robustness (*i.e.* buffering)

to protein depletion. Although this result is potentially surprising, it is in fact consistent with many studies. For instance, Belliveau *et al.* have recently analyzed the abundance of a wide range of metabolic and other essential biological processes, and conclude that protein abundance appears to be in significant excess of what is required for function [7]. Likewise, CRISPRi approaches have facilitated the characterization of essential protein depletion. The qualitative results from these experiments are consistent with overabundance: Large-magnitude protein depletion is typically required to generate strong phenotypes [4, 18, 19]. In particular, Peters *et al.* engineered a complete collection of CRISPRi essential-gene depletion constructs in *Bacillus subtilis*. Importantly, when *dcas9* is constitutively expressed, these constructs deplete essential proteins about three-fold below their endogenous expression levels [4]; however, roughly 80% grew without measurable fitness loss in log-phase growth despite the depletion. When grouped by functional category, only ribosomal proteins were found to have statistically significant reductions in fitness [4]. As shown in Fig. 2.2C, the RLTO model predicts that all but the highest expression proteins are expected to show minimal fitness reductions in response to a three-fold depletion of essential enzymes. The optimality of protein overabundance explains the paradox of protein expression levels being simultaneously optimal [2] and in excess of what is required for function [4, 5, 7, 19].

Although this qualitative picture of essential protein overabundance is clear, there has yet to be a quantitative and detailed measurement of protein overabundance, and in particular, an analysis of the relationship between protein overabundance and message number. However, by studying the essential-gene knockout proliferation dynamics at single-cell scale (by imaging) and genome-wide scale (by TFNseq [5]), we are able to observe and quantify the overabundance of all essential proteins in the naturally-competent bacterium in *Acinetobacter baylyi*. In our experiment we find that protein function has threshold-like dependence on protein abundance. As predicted by the RLTO model, we observe that (i) roughly 70% of essential proteins are overabundant, (ii) overabundance increases as the expression level decreases and (iii) the lowest abundance proteins are in vast excess ($>10\times$)

of what is required for growth in the typical cell. The details of the experiment and the results will be discussed in Chapter 3.

2.2.5 RLTO predicts a one-message transcription threshold

The RLTO model predicts protein overabundance, but is there a clear transcriptional signature? To analyze this question, we define the message threshold $n_m \equiv \mu_m/o$. (This parameterization is convenient since it is independent of the translation efficiency.) We can then analyze the relation between optimal message number and threshold message number, as shown in Fig. 2.3A. The model predicts that even for genes that have extremely small threshold message numbers (*e.g.* $n_m = 10^{-2}$), the optimal message number stays above one message transcribed per cell cycle. Qualitatively, expressing messages below this level is simply too noisy even for proteins needed at the lowest expression levels. (See the blue curve in Fig. 2.2B corresponding to the protein number distribution of $\mu_m = 1$.) The model therefore predicts a lower floor on transcription for essential genes of one message per cell cycle.

2.2.6 A lower threshold is observed for message number

To identify a putative transcriptional floor, we first analyzed the transcriptome in *Escherichia coli*. We hypothesize that cells must express essential genes above the one-message threshold for robust growth. The distinction between essential and nonessential genes is critical in this context, since nonessential genes can be inducibly expressed. For instance, in *E. coli*, the *lac* operon is repressed in the absence of lactose and therefore need not satisfy the one-message threshold. The transcriptional threshold is only hypothesized to apply to genes whose products are required to maintain cell fitness under the measured conditions.

We generated histograms for *E. coli* growing rapidly on rich media for these two classes of genes. (See Supplementary Material Sec. 2.6.) The message numbers for *nonessential* genes are widely distributed, with a significant fraction of genes falling below the one-message threshold; however, only one *essential* gene is expressed below the one-message threshold

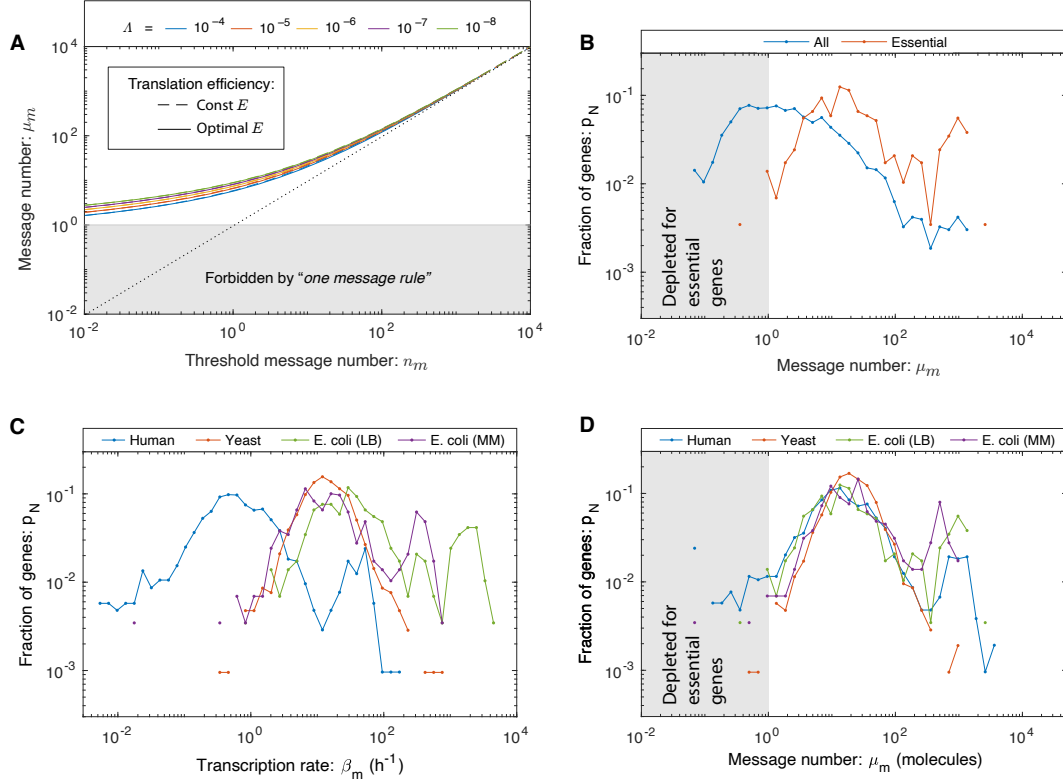


Figure 2.3: **A lower threshold for transcription: The one message rule.** Panel **A**: RLTO predicts the one-message rule. For high-expression genes, overabundance is low and the message number μ_m is predicted to be comparable to the threshold level n_m (dotted line); however, for low-expression genes there is a lower threshold ($\mu_m \geq 1$) below which expression is too noisy for robust growth. The threshold is weakly dependent on relative load Λ . Panel **B**: **A one-message threshold is observed in *E. coli* for essential genes.** A histogram shows the distribution of gene message numbers for all genes (blue) versus essential genes (orange). As predicted by the RLTO model, virtually all essential genes are expressed above the one-message-per-cell-cycle threshold. Panel **C**: **The distribution of transcription rates for essential genes.** No alignment is observed between the distributions of transcription rates in three evolutionarily-divergent organisms. For instance, the per gene transcription rate is significantly lower in human cells relative to *E. coli*. Panel **D**: **The distribution of message numbers for essential genes in three evolutionarily-divergent organisms.** The alignment of distributions of message number per gene between human, yeast, and *E. coli* (under two distinct growth conditions) reveals a nontrivial commonality between central dogma regulatory programs. We propose that the rationale for this alignment is the one-message rule that predicts that all essential genes must be expressed above one message per cell cycle. Both yeast and *E. coli* come very close to satisfying this proposed threshold; however, a greater proportion of genes in human break the one-message threshold. We speculate that this is due in part to the *ad hoc* nature of the essential-gene classification in the context of complex multicellular organisms.

(0.3% of essential genes). (See Fig. 2.3B.) The threshold is not sharp, but rather a smooth depletion relative to a median of 18 messages per cell cycle. This observation is consistent with the predictions of the RLTO model.

To further test this prediction, we then analyzed *E. coli* transcription under slow-growth conditions. Since these cells are less transcriptionally active, we hypothesized that this analysis would constitute a more stringent test of the one-message rule. (See Supplementary Material Sec. 2.6.) To our surprise, although the transcription rate is indeed reduced in slow growth, the essential gene message numbers still satisfy the one-message rule (with a two gene exception, 0.7%), again consistent with the predictions of the RLTO model. (See Fig. 2.3D.)

Next, we analyzed eukaryotic transcriptomes in *Saccharomyces cerevisiae* (yeast) and *Homo sapiens* (human). (See Supplementary Material Sec. 2.6.) For yeast, there is a well-defined notion of essential genes [20]. As predicted, yeast essential genes obey the one-message threshold (with two exceptions, 0.2%). (See Fig. 2.3D.) The interpretation is less clear-cut in human cells: An essential gene classification has been generated in the context of proliferation in cell culture [21]. In order to try to capture a generic picture, we average the human transcriptome of cell types. We find that the vast majority of essential genes obey the one-message rule; however, there are significantly more genes that break the rule (81 genes, 8%) than in the other organisms.

2.2.7 Message number distribution is conserved

To what extent is this human data consistent with the RLTO model? For human cells, our test of the one-message rule is too simplistic in two respects: (i) We ignore the significant transcriptional differences associated with distinct cell types and (ii) the essential gene classification itself is defined by the ability of mutants or knockdowns to proliferate in cell culture; in marked contrast to the *in vivo* context where cell proliferation is tightly regulated [21]. Due to these subtleties, we decided to take a complementary approach: We considered the distribution of three different transcriptional statistics for each gene: transcription rate,

cellular message number, defined as the average number of messages instantaneously, and message number (μ_m), defined as the number of messages transcribed in a cell cycle. (See Supplementary Material Sec. 2.6.) The RLTO model predicts a one-message threshold with respect to message number, but not the other two statistics. We therefore predict that the message number distributions in each organism (*E. coli*, yeast, and human) should align for low expression genes with respect to message number, but not for the other two transcriptional statistics. Consistent with the predictions of the RLTO model, there is a striking alignment of message number for essential genes between all three model organisms and growth conditions for message number. (See Fig. 2.3D.) This alignment is non-trivial: It is not observed with respect to other transcriptional statistics (Fig. 2.3C and Supplementary Material Fig. 2.3).

What is the significance of the similarity in the distributions of message number between organisms? Another strategy for satisfying the one message rule would be for transcription to be increased. For instance, mammalian cells have about 1000 times the number of proteins relative to bacterial cells. (See Supplemental Tab. 2.2.) One might therefore naively predict that the message number should be increased 1000-fold as well. This is not observed. In fact, the message number distributions of all the model organisms analyzed about the one message threshold. The proximity to the threshold suggests that organisms do as little transcription as possible while satisfying the one message rule. This appears to be a conserved transcriptional regulatory strategy from *E. coli* to human.

2.2.8 Translation efficiency is predicted to increase with transcription.

What does the RLTO model predict about how the cell should balance the gene expression process between transcription and translation? Minimizing transcription (at fixed protein abundance) reduces the metabolic load; however, it decreases robustness. Growth rate maximization balances these two costs. Quantitatively, the maximization of the growth rate (Eq. 2.1) with respect to the translation efficiency can be performed analytically, predicting the optimal translation efficiency, shown in Fig. 2.4A. We provide an exact expression in

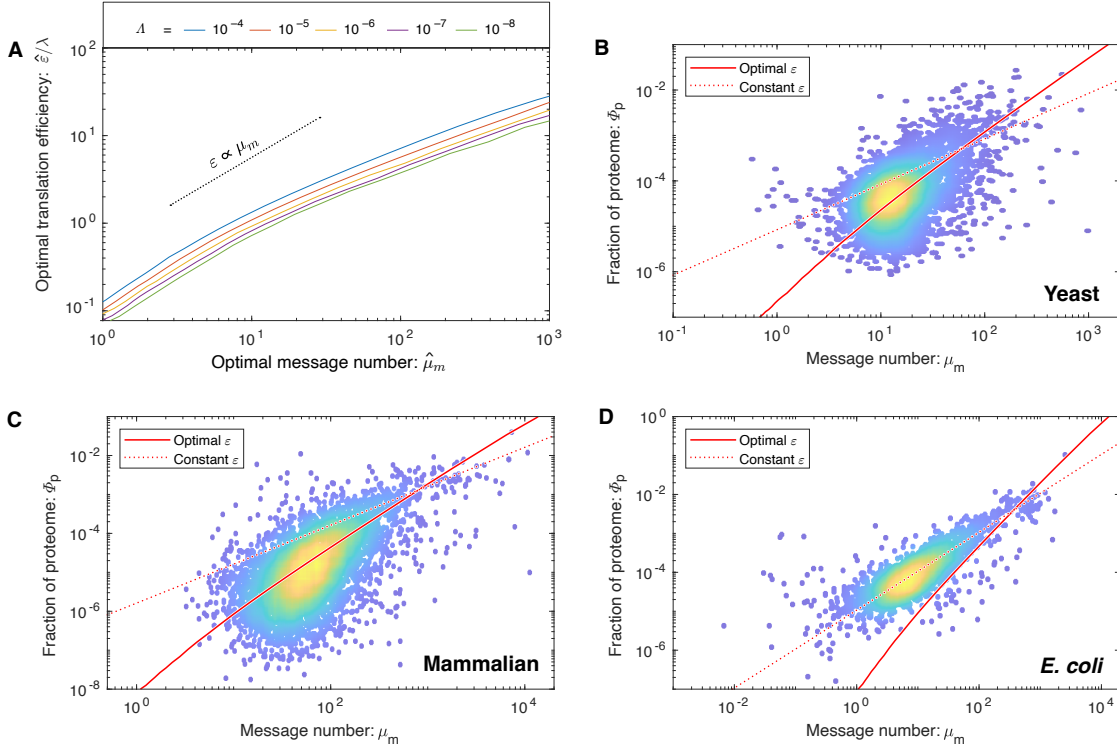


Figure 2.4: **How are transcription and translation balanced? Panel A: The RLTO model predicts load balancing.** The ratio of the optimal translation efficiency ($\hat{\varepsilon}$) to the message cost (λ) is roughly independent of the relative load (Λ). The translation efficiency ε is predicted to be roughly proportional message number μ_m . **Panel B: RLTO predicts the protein-message-abundance relation in yeast.** The observed proteome fraction is compared to two models: the RLTO optimal model (solid red line) and constant-translation-efficiency model (dotted red line). Both models make parameter-free predictions. The RLTO optimum predicts the global trend. (Data from Ref. [22].) **Panel C: Mammalian proteome fraction.** The RLTO prediction (solid) is superior to the constant-translation-efficiency prediction (dashed). **Panel D: *E. coli* proteome fraction.** In contrast, the constant-translation-efficiency prediction (dashed) is superior to RLTO prediction (solid).

the Supplementary Material Sec. b.; however, an approximate expression for the translation efficiency is more clearly interpretable:

$$\hat{\varepsilon} \approx 0.1\lambda\hat{\mu}_m. \quad (2.3)$$

The optimal translation efficiency has two important qualitative features for central dogma regulation. The first prediction is that as the message cost (λ) rises, the optimal translation efficiency ($\hat{\varepsilon}$) increases in proportion while the message number decreases. We present evidence for this prediction in the Supplementary Material Sec. 2.5.11.

The second prediction is that the optimal translation efficiency is also approximately proportional to message number ($\hat{\varepsilon} \propto \mu_m$). Therefore, the RLTO model predicts that low expression levels should be achieved with low levels of transcription and translation, whereas high-expression genes are achieved with high levels of both. We call this relation between optimal transcription and translation the *load balancing principle*. The most direct test of load balancing is measuring the protein-message abundance relation. Due to load balancing, the RLTO model predicts protein number (and proteome fraction) to scale like:

$$\hat{\mu}_p \propto \hat{\mu}_m^2, \quad (2.4)$$

whereas a constant-translation-efficiency model has linear scaling ($\mu_p \propto \mu_m$). Computing proteome fraction, rather than protein number, results in a parameter-free prediction. (See Supplemental Sec. 2.5.12.)

2.2.9 Load balancing is observed in eukaryotic cells.

To test the RLTO predictions, we compare observed proteome measurements in three evolutionarily divergent species, *E. coli* [23], yeast [22] and mammalian cells [24], to two models: the RLTO and the constant-translation-efficiency models. The results of the parameter-free predictions are shown in Fig. 2.4BCD for each organism. The RLTO model clearly captures the global trend in the proteome-fraction message-number relation

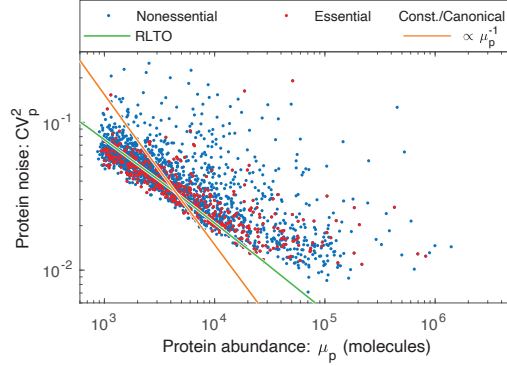


Figure 2.5: **RLTO predicts the magnitude of noise in yeast.** The observed gene expression noise in yeast is shown for essential and nonessential genes. Two protein-message abundance models are compared to the data: The RLTO model (green) versus the constant-translation-efficiency (canonical model, orange). The RLTO model predicts both the magnitude of the noise, as well as its scaling with protein abundance. The reduced slope of the RLTO model is the consequence of load balancing, which reduces the noise for the noisiest, low-expression genes. (Data from Ref. [9].)

in eukaryotic cells and a direct fit to a power law with an unknown exponent is consistent with Eq. 2.4 (Supplementary Material Sec. c.).

In *E. coli*, the constant-translation-efficiency model better describes the data. Why does this organism appear not to load balance? In the supplementary material, we demonstrate that the observed translation efficiency is consistent with the RLTO model, augmented by a ribosome-per-message limit. Hausser *et al.* have proposed just such a limit, based on the ribosome footprint on mRNA molecules [11]. (See Supplementary Material Sec. 2.5.17.) Although this augmented model is consistent with central dogma regulation in *E. coli*, it is not a complete rationale. This proposed translation-rate limit could be circumvented by increasing the lifetime of *E. coli* messages, which would increase the translation efficiency. A more in-depth analysis specific to *E. coli* is needed to understand why the observed message lifetime is so short.

2.2.10 RLTO model predicts observed noise in yeast.

Although the protein fraction measurements support the RLTO predictions for the translation efficiency in eukaryotic cells, these measurements do not provide a compelling rationale for why load balancing maximizes the growth rate. To understand its rationale, we explore its implications for noise.

In a typical biological context, $\mu_m \ll \varepsilon$ and as a result, noise production is dominated by the transcription step of the gene expression process [13, 14]. (A table of central dogma parameters for each model organism appears in the Supplementary Material Tab. 2.2.) Quantitatively, the canonical steady-state noise model predicts that the noise should be inversely related to the message number [13, 14]:

$$CV_p^2 = \frac{\ln 2}{\mu_m}, \quad (2.5)$$

however, it is the relation between mean protein abundance μ_p and noise (CV_p^2) which is typically reported [9, 10]. Based on the scaling of the optimal translation efficiency with the message number in eukaryotic cells (Eq. 2.3), we find the protein number to scale with message number (Eq. 2.4), which predicts that noise should scale with protein abundance $CV_p^2 \propto \mu_p^{-1/2}$ in yeast (see Supplementary Material Sec. 2.7.1); however, due to the observed absence of translation-efficiency scaling in bacteria, the noise should scale as $CV_p^2 \propto \mu_p^{-1}$ in bacteria, as observed [10]. Does the yeast noise show the predicted scaling? The parameter-free RLTO noise prediction closely matches the observed noise in both magnitude and scaling, as shown in Fig. 2.5.

2.2.11 Reducing noise is the rationale for load balancing.

This noise analysis also provides a conceptual insight into the rationale for load balancing. The load balanced (RLTO-green) and constant-translation-efficiency (orange) predictions for the noise are shown in Fig. 2.5. Load balancing results in decreased noise for low-expression, noisy genes over what is achieved with constant translation efficiency. This decreased noise is predicted to increase growth robustness. In principle, the noise could be reduced further

by tipping the balance even more towards transcription; however, the RLTO model predicts that this approach is too metabolically costly, and the optimal strategy is that observed for noise scaling in yeast.

2.3 Discussion

2.3.1 What are the biological implications of noise?

Many important proposals have been made, including bet-hedging strategies, the necessity of feedback in gene regulatory networks, *etc.* [8]. Our model suggests that overcoming cell-to-cell variation may fundamentally reshape the metabolic budget: Typically, proteins constitute 50-60% of the dry mass of the cell [6] and therefore overabundance could increase the overall protein budget by a significant factor. Why does the cell tolerate this significant increase in metabolic load above what would be predicted by a resource allocation analysis (*e.g.* [25])? This strategy dramatically reduces the consequence of stochastic expression of proteins on the rate of single-cell proliferation.

A second source of stochasticity, environmental fluctuations, has been proposed as a rationale for overabundance [26], especially in the context of metabolic genes [27]. In short, cells express protein to hedge against starvation [26] or changes in the carbon source, *etc.* [27]. How does this hypothesis compare to our growth robustness hypothesis? There are some similarities between these environmental-fluctuation models and the RLTO model: In both models, it is a fluctuations-based mechanism that drives overabundance; however, there are important distinctions between the model predictions. In the environmental fluctuation model, there is a trade-off between log-phase fitness and the rapidity of adaptation [26]; whereas in the RLTO model, overabundance corresponds to the log-phase optimum. Organisms experiencing prolonged periods of balanced growth would therefore be expected to reduce overabundance. Furthermore, the environmental fluctuation model most naturally explains overabundance for proteins related to metabolic processes, whereas the RLTO model

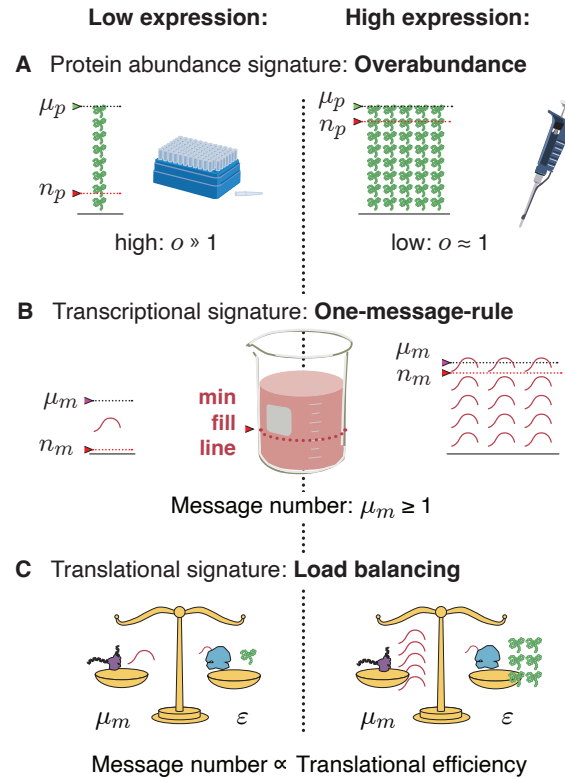


Figure 2.6: **Central dogma regulatory principles.** **Panel A: Overabundance.** Low-expression essential genes are expressed with high overabundance; whereas, high-expression essential genes are expressed with low overabundance. Lab supply analogy: Low-cost items that are used stochastically (*e.g.* pipette tips) are purchased in great excess, while the higher cost items that are less stochastic (*e.g.* pipette) are purchased as needed. **Panel B: One-message rule.** Robust expression of essential genes requires them to be transcribed above a threshold of one message per cell cycle. **Panel C: Load balancing.** In eukaryotic cells, optimal fitness is achieved by balancing transcription and translation: The optimal message number is proportional to the optimal translation efficiency. High (low) expression levels are achieved by high (low) levels of transcription followed by high (low) levels of translation per message.

predicts overabundance generically, dependent only on message number, which appears to be much more consistent with experiments exploring essential-protein depletion [5].

2.3.2 Implications for nonessential genes.

In our analysis, we have focused on essential genes in order to motivate the growth-threshold in the RLTO model. To what extent do nonessential genes share the same optimization? In support of the proposal that RLTO optima describe nonessential genes is the success of the model in predicting the translation efficiency for all genes, not just essential genes. (See Fig. 2.4 and Fig. 2.5.) Furthermore, the definition of a gene as *essential* depends on context: For instance, in the context of *E. coli* growth on lactose, the gene *lacZ* is essential, although it is nonessential on other carbon sources [28]. Under growth conditions where the *lacZ* gene is essential, we predict that LacZ should be overabundant, consistent with observation [27]. Finally, our modeling suggested that RLTO model phenomenology is the results of asymmetry of the cost of under versus overabundance. For nonessential genes whose activity significantly increases fitness, we still expect fitness asymmetry due to the low relative metabolic cost of increased expression. We therefore expect all gene products, most especially those with low expression, to be overabundant, under conditions where their activity increases fitness.

2.3.3 Implications of overabundance for inhibitors.

The generic nature of overabundance, especially for low-expression proteins, has important potential implications for the targeting of these proteins with small-molecule inhibitors (*e.g.* drugs). For the highest expression proteins, like the constituents of the ribosome, relatively small decreases in the active fraction (*e.g.* a three-fold reduction) are expected to lead to growth arrest [4]. This may help explain why inhibitors targeting translation make such effective antimicrobial drugs. However, we predict that the lowest expression proteins require a much higher fraction of the protein to be inactivated, with the lowest-expression proteins

expected to need more than a 100-fold depletion. This predicted robustness makes these proteins much less attractive drug targets [29].

2.3.4 The principles that govern central dogma regulation.

We propose that robustness to noise fundamentally shapes the central dogma regulatory program for all genes and predicts a number of key regulatory principles. (See Fig. 2.6.) For high-expression genes, load balancing implies that gene expression consists of both high-amplification translation and transcription. The resulting expression level has low overabundance relative to the threshold required for function. In contrast, for essential low-expression genes, a three-fold strategy is implemented: (i) overabundance raises the mean protein levels far above the threshold required for function, (ii) load balancing, and (iii) the one-message rule ensures that message number is sufficiently large to lower the noise of inherently-noisy, low-expression genes. We anticipate that these regulatory principles, in particular protein overabundance, will have important implications, not only for our understanding of central dogma regulation specifically, but for understanding the rationale for protein expression level and function in many biological processes.

2.4 *Materials and Methods*

2.4.1 RLTO model.

The effect of stochastic cell arrest can be implemented analytically as follows: The probability of growth is the probability that all essential proteins are above threshold, P_+ . The population growth rate k is [30]:

$$\frac{k}{k_0} = 1 + \frac{1}{\ln 2} \ln P_+, \quad (2.6)$$

for a population of cells subject to stochastic arrest with probability $1 - P_+$ per cell cycle where k_0 is the growth rate of the non-arrested cells. For each gene i , the canonical steady-state noise model predicts the protein number CDF in terms of message number μ_m and translation

efficiency ε [13]. Assuming the below-threshold probability is small, the probability that the cell is below threshold for gene i is:

$$\ln P_{+,i} = -\gamma\left(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}\right), \quad (2.7)$$

where γ is the regularized incomplete gamma function and the CDF of the gamma distribution. (See Supplementary Material Sec. 2.5.2.)

While protein underabundance slows cell growth by the arrest of essential processes, protein overabundance slows growth by increasing the metabolic load. To implement the metabolic-load contribution to cell fitness, we use a minimal model that realizes the metabolic cost of both transcription and translation that is analogous to those previously used in the context of resource allocation (*e.g.* [15]). The metabolic load of transcription and translation of gene i is:

$$\frac{k}{k_0} = 1 - \frac{\lambda + \varepsilon}{N_0} \mu_m, \quad (2.8)$$

where k_0 is the growth rate in the absence of the metabolic load of gene i , N_0 is the total cellular metabolic load, and λ is the metabolic message cost. (See the Supplementary Material Sec. 2.5.2 for a detailed development of the model.) The λ -term represents the metabolic cost of transcription and the ε -term represents the metabolic cost of translation of gene i . We define the relative load as $\Lambda \equiv \lambda/N_0$ as the ratio of the metabolic load of a single message to the total metabolic cost of the cell. In *E. coli*, we estimate that Λ is roughly 10^{-5} and it is smaller still for eukaryotic cells.

2.5 Supplementary Information

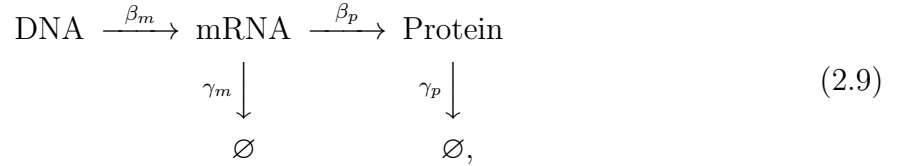
2.5.1 Methods: Detailed description of the noise model

In this section, we provide a detailed development of the RLTO model. First, we describe the stochastic kinetic model for the central dogma, which introduces key quantities for the RLTO model (Sec. a.). Next, we provide a derivation of the growth rate as a function of the model parameters (Sec. 2.5.2) as well as other methods (Secs. b., 2.5.12, 2.5.15, and 2.5.18).

For each of the results discussed in the main paper, we provide more detailed analyses, which include both supplemental results (Secs. 2.5.8, 2.5.9, 2.5.11, 2.5.13, and 2.5.16) that support the story described in the main paper, as well as supplemental discussions (Secs. 2.5.3, 2.5.5, 2.5.6, 2.5.7, 2.5.14, and 2.5.17).

a. Stochastic kinetic model for the central dogma

The canonical steady-state noise model for the central dogma describes multiple steps in the gene expression process [10, 13, 14]: Transcription generates mRNA messages. These messages are then translated to synthesize the protein gene products [12]. Both mRNA and protein are subject to degradation and dilution [31]. At the single cell level, each of these processes are stochastic. We will model these processes with the stochastic kinetic scheme [12]:



where β_m is the transcription rate (s^{-1}), β_p is the translation rate (s^{-1}), γ_m is the message degradation rate (s^{-1}), and γ_p is the protein effective degradation rate (s^{-1}). The message lifetime is $T_m \equiv \gamma_m^{-1}$. For most proteins in the context of rapid growth, dilution is the dominant mechanism of protein depletion and therefore γ_p is approximately the growth rate [10, 32, 33]: $\gamma_p = T^{-1} \ln 2$, where T is the doubling time.

b. Statistical model for protein abundance.

To study the stochastic dynamics of gene expression, we used a stochastic Gillespie simulation [34, 35]. (See Sec. c..) In particular, we were interested in the explicit relation between the kinetic parameters ($\beta_m, \gamma_m, \beta_p, \gamma_p$) and experimental observables.

Consistent with previous reports [13, 14], we find that the distribution of protein number per cell (at cell birth) was described by a gamma distribution: $N_p \sim \Gamma(a, \theta)$, where N_p is the protein number at cell birth and Γ is the gamma distribution which is parameterized by

a scale parameter θ and a shape parameter a . (See Sec. a..) We refer to this distribution as the *canonical steady-state noise model*; The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [14]: The scale parameter:

$$\theta = \varepsilon \ln 2, \quad (2.10)$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \quad (2.11)$$

where β_p is the translation rate and γ_m is the message degradation rate. ε is understood as the mean number of proteins translated from each message transcribed. The shape parameter a can also be expressed in terms of the kinetic parameters [14]:

$$a = \frac{\beta_m}{\gamma_p}; \quad (2.12)$$

however, we will find it more convenient to express the scale parameter in terms of the cell-cycle message number:

$$\mu_m \equiv \beta_m T = a \ln 2, \quad (2.13)$$

which can be interpreted as the mean number of messages transcribed per cell cycle. Forthwith, we will abbreviate this quantity *message number* in the interest of brevity.

c. Gillespie simulation of stochastic kinetic scheme

Protein distributions based on the kinetic scheme defined in Sec. a. were simulated with a Gillespie algorithm, with specific parameter values for *E. coli*. Assuming the lifetime of the cell cycle ($T = 30$ min) [36], mRNA lifetime ($T_m = 2.5$ min) [37], and translation rate ($\beta_p \approx 500$ hr⁻¹), the protein distributions for several mean expression levels were numerically generated for exponential growth with 100,000 stochastic cell divisions, with protein partitioned at division following the binomial distribution.

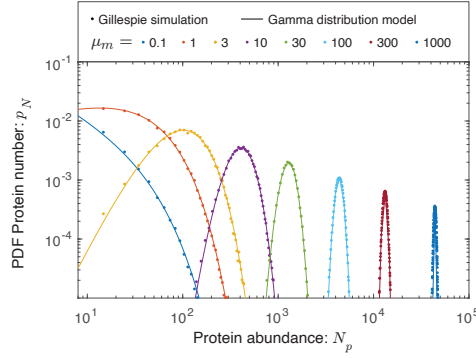


Figure 2.7: **The protein abundance is approximately gamma distributed.** Protein abundance was modeled for eight different transcription rates using a Gillespie simulation, including the stochastic partitioning of the proteins between daughter cells at cell division. The range in abundance matches the observed range of expression levels in the cell. We observed that the simulated protein abundances were well fit by gamma distributions.

The gamma distributions for each mean message number with scale and shape parameters determined by the corresponding translation efficiency and message number ($\theta = \varepsilon \ln 2$, $a = \frac{\mu_m}{\ln 2}$) as used for the Gillespie simulation were also plotted with the protein distributions. We observe an excellent match between these Gillespie simulations and the canonical steady-state statistical noise model (*i.e.* gamma function) as shown in Fig. 2.7.

d. Gamma function and distribution conventions

There are a number of conflicting conventions for the gamma function and distribution arguments. We will use those defined on Wikipedia and the CRC Encyclopedia of Mathematics [38]. The gamma distributed random variable X will be written:

$$X \sim \Gamma(a, \theta), \quad (2.14)$$

where a is the shape parameter and θ is the scale parameter. The PDF of the distribution is:

$$p_{\Gamma}(x|a, \theta) \equiv \frac{x^{a-1}}{\theta^a \Gamma(a)} e^{-x/\theta}, \quad (2.15)$$

where $\Gamma(a)$ is the gamma function. The CDF is therefore:

$$P_{\Gamma}(x|a, \theta) = \int_0^x dx' p_{\Gamma}(x'|a, \theta), \quad (2.16)$$

$$= P_{\Gamma}\left(\frac{x}{\theta}|a, 1\right), \quad (2.17)$$

$$= \int_0^{x/\theta} dx' \frac{x'^{a-1}}{\Gamma(a)} e^{-x'}, \quad (2.18)$$

$$= \gamma(a, x/\theta), \quad (2.19)$$

where γ is the regularized incomplete gamma function.

2.5.2 Methods: The derivation of the RLTO growth rate

a. The metabolic load of protein and the resource allocation model

To model the effect of the metabolic load on cell growth, we will expand on a model used by Hwa and co-workers [15]. For conciseness, we will call the original model the *resource allocation model*.

Consider a cell where the total number of proteins in the cell is N . The synthesis of these proteins requires two sets of processes: (i) the metabolic processes responsible for synthesizing the precursors (*i.e.* amino acids, *etc*) and (ii) the translation process. Proteins involved in the metabolic processes be referred to as the P sector and number N_P . The proteins involved in the translational process will be referred to as the R sector and number N_R . In addition to the P and R sectors, there is a third Q sector with protein number N_Q . The total protein number per cell is therefore:

$$N = N_R + N_P + N_Q. \quad (2.20)$$

The proteome fractions are defined $\Phi_X \equiv N_X/N$ and have the normalization condition:

$$1 = \Phi_R + \Phi_P + \Phi_Q. \quad (2.21)$$

The key assumption in the resource allocation model is that the abundances of the R and P sectors can change in size to accommodate changes in the nutrient quality and translation

load associated with a particular growth condition [15]. In contrast, the Q sector has a fixed proteome fraction, Φ_Q , irrespective of growth conditions. In the resource allocation model, the size of these adjustable R and P sectors are chosen to optimize the growth rate k .

The condition for balanced growth requires that the overall protein output of the translation process match the growth rate:

$$kN = k_R N_{*R}, \quad (2.22)$$

where k_R is the effective translation rate per protein and N_{*R} is the number of productive R sector proteins, which is subset of the total number N_R :

$$N_R = N_{*R} + N_{0R}, \quad (2.23)$$

and N_{0R} represents unproductive R sector protein. We can rewrite Eq. 2.22 in terms of the proteome fraction:

$$k = k_R(\Phi_R - \Phi_{0R}). \quad (2.24)$$

In the expression above, we will assume that the parameters k_R and Φ_{0R} are fixed, but the total fraction Φ_R is chosen to optimize the growth rate k .

For any productive sectors i , we will write analogous equations to Eq. 2.24 linking sector fraction size Φ_i to function:

$$k = k_i(\Phi_i - \Phi_{0i}), \quad (2.25)$$

where, as before, Φ_{0i} represents a fixed-size fraction of unproductive protein.

To determine the unknown optimum growth rate and sector sizes, Eq. 2.25 can be re-written:

$$\Phi_{*i} = \Phi_i - \Phi_{0i} = \frac{k}{k_i}, \quad (2.26)$$

and then summed over all sectors (excluding Q):

$$\Phi_* = k \sum_{i \neq Q} k_i^{-1}, \quad (2.27)$$

where we define the total productive fraction of the proteome:

$$\Phi_* \equiv \sum_{i \neq Q} \Phi_{*i} = 1 - \Phi_0, \quad (2.28)$$

and Φ_0 represents the total fraction of unproductive protein:

$$\Phi_0 = \Phi_Q + \sum_{i \neq Q} \Phi_{0i}, \quad (2.29)$$

including the entire Q sector. Since Φ_Q and the Φ_{0i} are all assumed to be fixed, Eq. 2.27 determines the growth rate. (In *E. coli*, Hwa and coworkers estimate that $\Phi_* \approx 0.55$.)

To understand the meaning of Eq. 2.27, we first define an ideal growth rate as

$$k_{\text{ideal}}^{-1} = \sum_{i \neq Q} k_i^{-1}, \quad (2.30)$$

which would be the growth rate in the absence of unproductive protein; however, due to the presence of the unproductive protein, the growth rate is proportional to the productive fraction:

$$k = k_{\text{ideal}} \Phi_*. \quad (2.31)$$

The optimal protein fractions can be determined using Eq. 2.26:

$$\Phi_i = \frac{k_{\text{ideal}}}{k_i} \Phi_* + \Phi_{0i}. \quad (2.32)$$

How does the growth rate change when the unproductive protein fraction is changed by $\delta\Phi$? The productive fraction is reduced:

$$\Phi_* \rightarrow \Phi'_* = \Phi_* - \delta\Phi. \quad (2.33)$$

The ratio of the new growth rate k' to the original is therefore:

$$\frac{k'}{k} = 1 - \frac{\delta\Phi}{\Phi_*}. \quad (2.34)$$

Note that our generalized resource allocation model is written for arbitrary number of functional sectors $i \neq Q$ and the key determinant of the change in growth rate is the fraction of productive protein Φ_* . This equation will be used to model the fitness cost of the metabolic load.

b. The metabolic load of mRNA

What is the cost of transcription? It is perhaps useful to first consider the estimates of biosynthetic cost of macromolecules in the cell in descending order in *E. coli* [39]:

Macromolecule	Biosynthetic cost (10^9 ATP)
Protein	4.5
Phospholipid	3.2
RNA	1.6
Lipopolysacchride	3.8
DNA	0.35
Peptidoglycan	0.17
Glycogen	0.03

So clearly the cost of RNA is itself not insignificant. Although a significant fraction of the RNA is rRNA rather than mRNA, the mRNA itself in *E. coli* undergoes multiple rounds of transcription due to its short lifetime, increasing its cost to what is required to synthesize the molecules observed in the *E. coli* cell at any time t . Furthermore, transcription is dependent on protein enzymes, which themselves must be synthesized. We therefore conclude from this estimate that the cost of transcription is likely a significant determinant of the metabolic load.

Experimentally, Kafri and coworkers have measured the fitness cost of transcription and translation independently using the DAmP (Decreased Abundance by mRNA Perturbation) system in *Saccharomyces cerevisiae* [40]. As expected, they report that the metabolic cost of transcription is comparable to translation and that the reduction in growth rate is linear in transcription, in close analogy to Eq. 2.34.

c. Metabolic load in the RLTO model

To produce a minimal model to study the trade-off between robustness and metabolic load, we must consider both the metabolic cost of transcription and translation. We will write that the metabolic load (in protein equivalents) associated with gene i is:

$$\delta N_i = \lambda \mu_{m,i} + \mu_{p,i}, \quad (2.35)$$

where λ is the message cost, the metabolic load associated with an mRNA molecule relative to a single protein molecule of the gene product. $\mu_{m,i}$ is the mean number of messages transcribed per cell cycle (mRNA molecules per cell cycle) for gene i . $\mu_{p,i}$ is the mean number of protein translated per cell cycle for gene i . We will describe the mean protein number in terms of the translation efficiency ε_i , the number of proteins translated per message:

$$\mu_{p,i} = \varepsilon_i \mu_{m,i}. \quad (2.36)$$

How does the cell growth rate change due to the metabolic load associated with the expression of gene i ? The change in the metabolic load is:

$$\delta\Phi_i = \frac{\delta N_i}{N}, \quad (2.37)$$

where N represents the total metabolic load of all components of the cell, in units of protein equivalents. Using Eq. 2.34, the resulting change in growth rate is:

$$\frac{k}{k_0} = 1 - \frac{(\lambda + \varepsilon_i) \mu_{m,i}}{\Phi_* N}, \quad (2.38)$$

where k_0 is the growth rate in the absence of the metabolic load of gene i .

In our analysis, the exact size of the total metabolic load N will not be important. In the interest of simplicity we will therefore adsorb the productive fraction Φ_* into an effective total metabolic load:

$$N_0 \equiv \Phi_* N, \quad (2.39)$$

and write a concise relation between the load from gene i and the growth rate:

$$\frac{k}{k_0} = 1 - \frac{(\lambda + \varepsilon_i) \mu_{m,i}}{N_0}. \quad (2.40)$$

This equation has an intuitive interpretation: growth slows in proportion to the relative added metabolic load. Since Φ_* is order unity, we will ignore the distinction between the N and N_0 quantities hence forth.

Although the global parameters N_0 and λ provide an intuitive representation of the model, the relative growth rate depends on fewer parameters. Let k and k_0 be the growth rates in

the presence and absence of the metabolic load of gene i . The relative growth rate is:

$$\frac{k}{k_0} = 1 - (\Lambda + E_i)\mu_{m,i}, \quad (2.41)$$

where we have introduced two new reduced parameters: the relative load, defined as $\Lambda \equiv \lambda/N_0$, represents the ratio of the metabolic load of a single message to the total load and the relative translation efficiency, defined $E_i \equiv \varepsilon_i/N_0$, which is the ratio of the number of proteins translated per message to the total metabolic load N_0 . (Note that due to the high multiplicity $N_0 \gg (\lambda + \varepsilon_i)\mu_{m,i}$, we can ignore the distinction between N_0 and N'_0 in the denominator.) If we neglect the difference between the total metabolic load and the number of proteins, the proteome fraction for gene i is:

$$\Phi_i = E_i\mu_{m,i}. \quad (2.42)$$

Both reduced parameters, Λ and E_i are extremely small. In *E. coli*, we estimate that both Λ and E_i are roughly 10^{-5} and they are smaller still for eukaryotic cells. (See Sec. 2.5.18.)

d. Growth rate with stochastic arrest

For completeness, we provide a derivation of the growth rate with stochastic cell-cycle arrest that we have previously described [30]. Starting from the exponential mean expression for the population growth rate [30]:

$$k = \frac{\ln 2}{\bar{T}}, \quad (2.43)$$

where k is the population growth rate and

$$\bar{T} \equiv -\frac{1}{k} \ln \mathbb{E}_T \exp(-kT), \quad (2.44)$$

is the exponential mean, where T is the stochastic cell cycle duration and $\mathbb{E}$ is the expectation operator [30, 41]. We take a coarse-grained model which considers changes in growth rate due to fluctuations in protein number to be negligible. Note that Eq. 2.44 is equivalent to the *Euler-Lotka* equation [42, 43].

Let P_+ be the probability of growth. When the cells are growing, the cell cycle duration τ is determined by the metabolic load predictions (Eq. 3.55). The probability mass function is therefore:

$$p_T(t) = \begin{cases} P_+, & t = \tau \\ (1 - P_+), & t \rightarrow \infty \end{cases}. \quad (2.45)$$

Evaluating the expectation in Eq. 2.44 gives:

$$\bar{T} = -k \ln P_+ + \tau. \quad (2.46)$$

Using Eq. 2.43, we can solve for the growth rate k :

$$k = \tau^{-1} \ln(2P_+). \quad (2.47)$$

As expected, the growth rate goes down as the probability of growth P_+ decreases, stopping completely at $P_+ = \frac{1}{2}$. We can then compute the ratio of the growth with (k) and without arrest (k_0):

$$\frac{k}{k_0} = 1 + \frac{1}{\ln 2} \ln P_+. \quad (2.48)$$

where k_0 is computed by evaluating Eq. 3.57 at $P_+ = 1$.

e. RLTO growth rate

In the RLTO model, we will assume the probability of growth is the probability that all essential protein numbers are above threshold. We will further assume that each protein number is independent, and therefore:

$$P_+ = \prod_{i \in \mathcal{E}} \Pr\{N_{p,i} > n_{p,i}\}, \quad (2.49)$$

where $\mathcal{E}$ is the set of essential genes. Clearly, this assumption of independence fails in the context of polycistronic messages. We will discuss the significance of this feature of bacterial cells elsewhere, but we will ignore it in the current context.

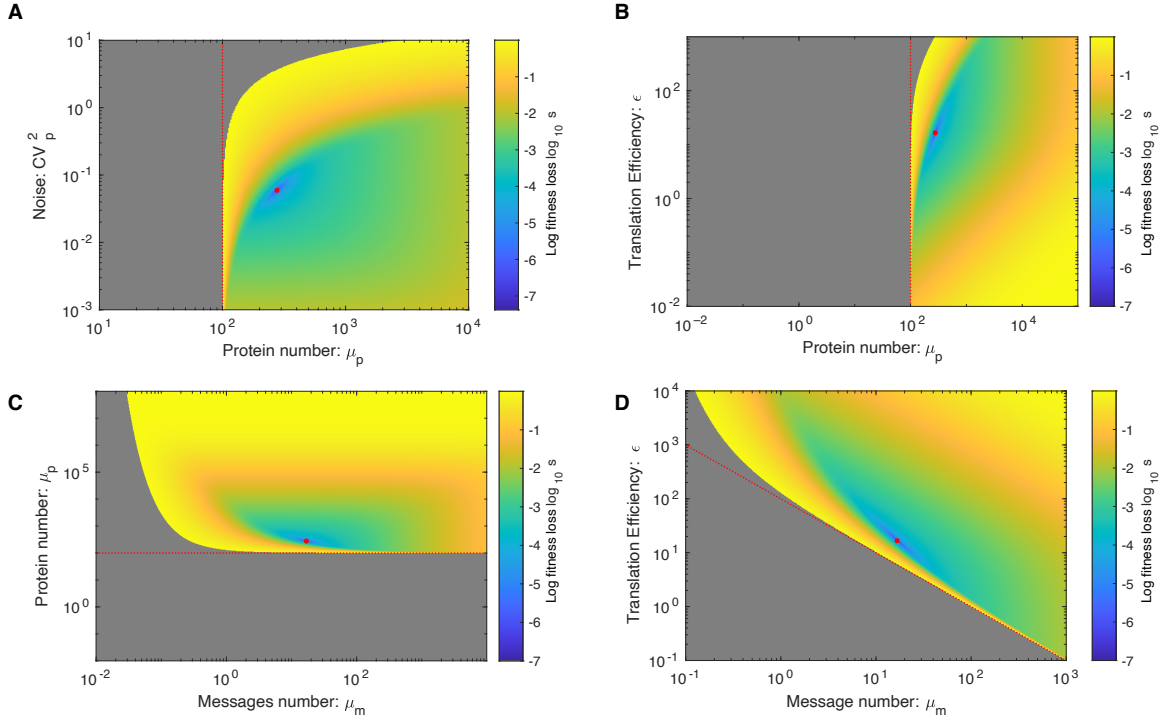


Figure 2.8: **Four perspectives on the fitness landscape.** In each landscape, the red circle represents the fitness optimum. The red dotted line represents the mean protein number equal to the protein threshold $n_p = 10^2$. Here, fitness is quantified by the log growth rate: $s = \ln k_{\max}/k$. **Panel A:** Mean protein number versus noise. **Panel B:** Mean protein number versus translation efficiency. **Panel C:** Mean message versus protein numbers. **Panel D:** Message number versus translation efficiency.

As we will discuss, the probability of arrest of any protein i to be above threshold is extremely small. It is therefore convenient to work in terms of the CDFs which are very close to zero:

$$\ln P_+ = \sum_{i \in \mathcal{E}} \ln(1 - \Pr\{N_{p,i} < n_{p,i}\}), \quad (2.50)$$

$$\approx - \sum_{i \in \mathcal{E}} \Pr\{N_{p,i} < n_{p,i}\}, \quad (2.51)$$

$$= - \sum_{i \in \mathcal{E}} \gamma\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon_i \ln 2}\right), \quad (2.52)$$

where γ is the regularized incomplete gamma function and the CDF of the gamma distribution (see a. and a.).

f. When the ln approximation is avoided...

The approximation discussed in the previous section is extremely well justified at the optimal central dogma parameters; however, there are a set of figures where we cannot use it. In the fitness landscape figures (Fig. 2.8), we compute the fitness not just at the optimal values but far from them. When cell arrest has a large effect on the growth rate, we cannot approximate the natural log with a series expansion, and we must use the full expression in Eq. 2.50.

g. Single-gene equation

By summing the fitness losses from the metabolic load and cell arrest (Eqs. 3.55, 3.58, and 3.60), we can write an expression for the growth rate including contributions from essential gene i :

$$\begin{aligned} \frac{k}{k_0} = & 1 - \frac{1}{N_0}(\lambda + \varepsilon_i)\mu_{m,i} + \dots \\ & - \frac{1}{\ln 2} \gamma\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon_i \ln 2}\right), \end{aligned} \quad (2.53)$$

where the second term on the RHS represents the fitness loss due to the metabolic load and the third term represents the fitness loss due to stochastic cell arrest due to protein i

Figure number and panel:	Relative load: Λ (No units)	Protein cost: λ (molecules ⁻¹)	Protein threshold: n_p (molecules)	Message number: μ_m (molecules/cell cycle)	Translation efficiency: ε (No units)	Noise floor C_0 (No units)
2.2A	10^{-5}	10	10^2	Range	Range	0
2.2C	Range	NA	Range	Local optimum	Local optimum	0
2.4A	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
2.4B	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
2.4C	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
2.4D	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
2.5	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
2.8A	10^{-5}	10	10^2	Range	Range	0
2.8B	10^{-5}	10	10^2	Range	Range	0
2.8C	10^{-5}	10	10^2	Range	Range	0
2.8D	10^{-5}	10	10^2	Range	Range	0
2.9A	Range	10^2	Range	Local optimum	Local optimum	0
2.9B	Range	10^2	Range	Local optimum	Local optimum	0
2.9C	Range	10^2	Range	Local optimum	Local optimum	0
2.9D	Range	10^2	Range	Local optimum	Local optimum	0
2.10A	Range	10^2	Range	Local optimum	Local optimum	0
2.10B	Range	10^2	Range	Local optimum	Local optimum	0
2.10C	Range	10^2	Range	Local optimum	Local optimum	0
2.10D	Range	10^2	Range	Local optimum	Local optimum	0
2.11	10^{-5}	NA	Range	Local optimum	30	0.1
2.12A	Range	NA	Range	Local optimum	Local optimum	0
2.13	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0
2.15A	Range	NA	Range	Local optimum	Local optimum	0
2.18B	Weak dependence (10^{-5})	NA	Range	Local optimum	Local optimum	0

Table 2.1: **RLTO model parameters by figure.** *Range* appears if a range of parameters is used. *NA* appears if the parameter value is irrelevant. *Local optimum* appears if the parameter values in optimized to maximize the growth rate.

falling below threshold. The fitness landscape for different gene expression parameters is shown from four different perspectives in Fig. 2.8. From this point forward, we will drop the subscript i for the sake of brevity unless otherwise noted.

h. Summary of RLTO parameter values for figures

The parameter values for the RLTO model used for each figure in the paper are shown in Tab. 2.1.

2.5.3 Discussion: The fitness landscape of a trade-off.

The fitness landscape predicted by the RLTO model for representative parameters is shown in Fig. 2.8. The figure displays a number of important model phenomena: There is no growth for mean protein number μ_p below the threshold number n_p , and for high noise, μ_p must be in significant excess of n_p . Rapid growth can be achieved by the two mechanisms: (i) high

expression levels (μ_p) are required for high noise amplitude (CV_p^2) or (ii) lower expression levels coupled with lower noise. This trade-off leads to a ridge-like feature of nearly optimal models. The optimal fitness corresponds to a balance between increasing the mean protein number (μ_p) and decreasing noise (CV_p^2). This optimal central dogma program strategy leads to significant overabundance.

2.5.4 Methods: Central dogma optimization

a. Optimization of transcription and translation (eukaryotes)

The relative growth rate is:

$$\frac{\Delta k}{k_0} = -(\Lambda + \frac{\varepsilon}{N_0})\mu_m - \frac{1}{\ln 2}\gamma(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}), \quad (2.54)$$

where γ is the regularized incomplete gamma function, which is the CDF of the gamma distribution and represents the probability of arrest due to gene i . (Note that this equation is identical to Eq. 3.61 but with the gene subscript i implicit.) We set the partial derivative of the growth rate with respect to message number equal to zero:

$$0 = -\frac{\lambda + \hat{\varepsilon}}{N_0} - \frac{1}{(\ln 2)^2}\gamma_{,1}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon} \ln 2}), \quad (2.55)$$

where we use the canonical comma notation to show which argument of γ has been differentiated. Next we differentiate with respect to the translation efficiency to generate a second optimization condition:

$$0 = -\frac{\hat{\mu}_m}{N_0} + \frac{n_p}{\hat{\varepsilon}^2 (\ln 2)^2}\gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon} \ln 2}). \quad (2.56)$$

We will work in the large multiplicity limit where the overall metabolic load is much smaller than the metabolic load associated with any single gene: $N_0 \gg (\lambda + \hat{\varepsilon})\hat{\mu}_m$. Next, we eliminate the threshold n_p in favor of the optimal overabundance:

$$\hat{o} \equiv \frac{\hat{\mu}_p}{n_p} = \frac{\hat{\varepsilon}\hat{\mu}_m}{n_p}, \quad (2.57)$$

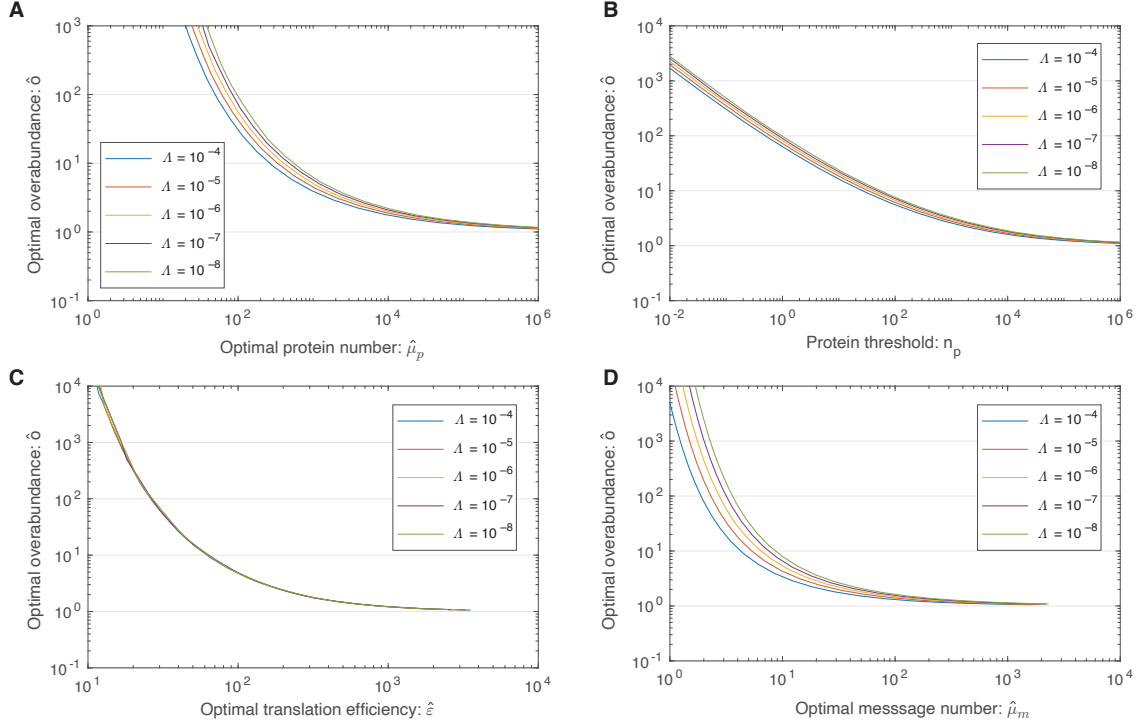


Figure 2.9: **Four perspectives on the dependence of optimal overabundance on relative load Λ .** All these calculations are performed for protein cost $\lambda = 100$ in order to give real numbers in molecules per cell. **Panel A: Overabundance as a function of protein number.** Overabundance decreases as protein number increases. These calculations are λ dependent. **Panel B: Overabundance as a function of the protein threshold.** Overabundance decreases as protein threshold increases. These calculations are λ dependent. **Panel C: Overabundance as a function of translation efficiency.** Overabundance decreases as translation efficiency increases. These calculations are λ dependent. **Panel D: Overabundance as a function of message number.** Overabundance decreases as message number increases. These calculations are λ independent.

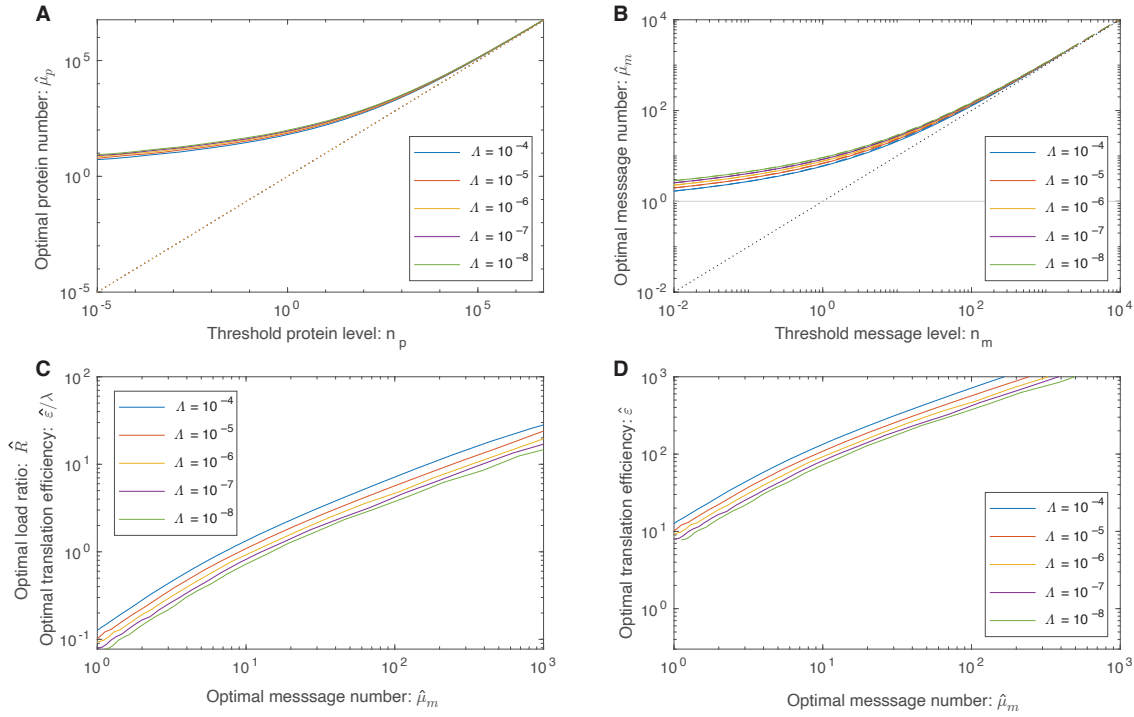


Figure 2.10: **Four perspectives on load balancing.** All these calculations are performed for protein cost $\lambda = 100$ in order to give real numbers in molecules per cell. **Panel A: Protein number versus protein threshold.** At high expression levels, the protein number tracks the protein threshold; however, the one message rule forces the protein number to threshold for low expression levels. These calculations are λ dependent. **Panel B: Message number versus message threshold.** At high expression levels, the message number tracks the message threshold; however, the one message rule forces the message number to a threshold close to $\mu_m = 1$ for low expression levels. These calculations are λ independent. **Panel C & D: Message number versus translation efficiency.** The optimal translation efficiency grows almost linearly with the optimal message number. The scaled translation efficiency ($\hat{\varepsilon}/\lambda$) is independent of λ while the translation efficiency ($\hat{\varepsilon}$) is dependent on λ . The ratio ε/λ has a second interpretation: the load ratio R . R is defined as the metabolic cost of translation over transcription of the gene.

in both Eqs. 2.55 and 2.56. Eq. 2.56 can now be solved for the optimal translation efficiency:

$$\hat{\varepsilon} = \frac{N_0}{\hat{o}(\ln 2)^2} \gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}). \quad (2.58)$$

If we reinterpret γ as the CDF of the gamma distribution, we can rewrite this equation in terms of the gamma distribution PDF:

$$\hat{\varepsilon} = \frac{N_0}{\ln 2} p_{\Gamma}(\mu_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \quad (2.59)$$

which will be the optimization equation for the translation efficiency.

To derive the optimization condition for the message number μ_m , we substitute Eq. 2.58 into Eq. 2.55:

$$\frac{\lambda \ln 2}{N_0} = -\frac{1}{\hat{o} \ln 2} \gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}) - \frac{1}{\ln 2} \gamma_{,1}(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}). \quad (2.60)$$

The two terms on the RHS can now be collected as the single partial derivative of message number μ_m :

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m} \gamma(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}), \quad (2.61)$$

where the relative load is $\Lambda \equiv \lambda/N_0$.

The two optimization equations are summarized below:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m} \gamma(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}), \quad (2.62)$$

$$\frac{\hat{E}}{\Lambda} = \frac{\hat{\varepsilon}}{\lambda} = \frac{1}{\Lambda \ln 2} p_{\Gamma}(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2). \quad (2.63)$$

The optimal overabundance is shown for a range of relative loads in Fig. 2.9. The optimal translation efficiency and scaled translation efficiency are shown for a range of relative loads in Fig. 2.10.

b. Optimization of message number only

Consider the special case of optimizing the message number only at fixed translation efficiency. Eq. 2.55 is the condition; however, in this case it makes sense to adsorb both the

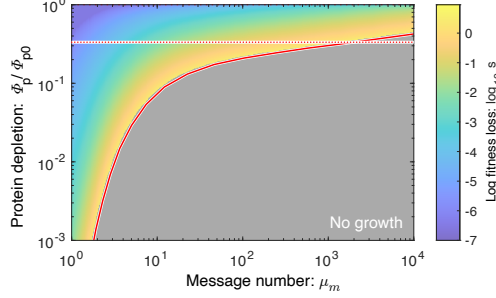


Figure 2.11: **Optimal expression levels are buffered.** The predicted fitness loss as a function of protein depletion level and message number for bacterial cells (including the noise floor). Due to the overabundance phenomenon, all proteins are buffered against depletion, but low-expression genes are particularly robust due to higher overabundance. The solid red line represents $1/o$, and predicts the range of depletion values for which cell growth is predicted. The dotted red line represents a three-fold depletion.

message and protein metabolic load into a single metabolic load. The optimum message number satisfies the equation:

$$\frac{(\lambda+\varepsilon) \ln 2}{N_0} = -[\partial_{\hat{\mu}_m} \gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\phi}}. \quad (2.64)$$

We define a modified relative load:

$$\Lambda' \equiv \frac{(\lambda+\varepsilon)}{N_0}, \quad (2.65)$$

and substitute this into the optimum message number equation:

$$\Lambda' \ln 2 = -[\partial_{\hat{\mu}_m} \gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\phi}}. \quad (2.66)$$

which is clearly closely related to Eq. 2.61.

We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 2.2C and demonstrate that the two make nearly identical predictions.

c. Inclusion of the noise floor

In bacterial cells, the noise is dominated by the noise floor for high expression genes. Including the noise floor, the coefficient of variation squared is ([10] and Sec. 2.7):

$$\text{CV}_p^2 = \tilde{a}(\mu_m)^{-1} = \frac{\ln 2}{\mu_m} + C_0, \quad (2.67)$$

where $C_0 = 0.1$ for bacterial cells [10]. In spite of the addition of noise from the noise floor, the observed distribution of protein number is still well described by the gamma distribution [10]; however, we need to modify the statistical parameters to account for the noise floor. (See the definition of the *statistical noise model* in Sec. a..) The modified gamma parameters are:

$$a = \tilde{a}, \quad (2.68)$$

$$\theta = \varepsilon \frac{\mu_m}{\tilde{a}}, \quad (2.69)$$

chosen such that the noise is determined by Eq. 2.67 but the protein number remains:

$$\mu_p = \varepsilon \mu_m, \quad (2.70)$$

the product of the message number and translation efficiency.

The qualitative effect of the noise floor is to increase the noise, especially for low-copy messages. Above $\mu_m = 7$ messages, the noise is dominated by the noise floor. Increases in transcription above this point have little effect on reducing the noise. As a consequence, the overabundance stays high, even for high copy messages. We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 2.2C and demonstrate that bacterial cells are predicted to have much higher overabundance at high expression levels.

2.5.5 Discussion: Understanding the rationale for overabundance

Essential protein overabundance is the signature prediction of the RLTO model. Its mathematical rationale is the highly-asymmetric fitness landscape. To understand why

we expect this rationale to be generic, consider the form of the optimization condition for message number:

$$\Lambda \ln 2 = -\frac{\partial}{\partial \hat{\mu}_m} \gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\delta \ln 2}\right). \quad (2.71)$$

The growth rate is maximized when the probability of slow-growth (*e.g.* arrest) is roughly equal to the relative load of adding one more message. Since the cell makes roughly 10^5 messages per cell cycle, the relative load is extremely small and therefore the probability of slow growth must be as well. Making this probability very small requires vast overabundance for the inherently-noisy, low-expression proteins.

The reason we expect the RLTO model protein abundance predictions to be robust is that we generically expect the fitness cost of overabundance to be small due to the high multiplicity (*i.e.* total number of genes); whereas, the fitness cost of arrest of essential processes is very high.

2.5.6 Discussion: RLTO predicts larger overabundance in bacteria.

There are two distinctive features of bacterial cells that could affect the model predictions: (i) the translation efficiency is constant [23] and bacterial gene expression is subject to a large-magnitude noise floor that increases the noise for high-expression genes [10]. The optimization of message number at fixed translation efficiency does result in a slightly modified optimization condition for the message number (Sec. b.); however, the predicted overabundance is only subtly perturbed (Fig. 2.2C). In contrast, the noise floor increases the predicted overabundance, especially for high-expression proteins. As a result, the RLTO model predicts that the vast majority of bacterial proteins are expressed in significant overabundance. (See Fig. 2.2C.)

2.5.7 Discussion: RLTO predicts proteins are buffered to depletion.

A principle motivation for our analysis is the observation that many protein levels appear to be buffered. To explore the prediction of the RLTO model for protein depletion, we first

computed the optimal message numbers and translation efficiencies for a range of protein thresholds. To model the effect of protein depletion, we computed the change in growth rate as function of protein depletion (equivalent to a reduction of the translation efficiency relative to the optimum.) The growth rate is shown in Fig. 2.11 for the RLTO model with parameters representative of a bacterial cell. (See Sec. c.)

In general, the RLTO model predicts that protein numbers have very significant robustness (*i.e.* buffering) to protein depletion. This is especially true for low expression proteins that are predicted to have the largest overabundance. For these genes, even a ten-fold depletion leads to very subtle reductions in the growth rate. For a three-fold reduction in the growth rate, only the very highest-expression genes (*e.g.* ribosomal genes) are expected to lead to qualitative phenotypes.

2.5.8 Results: Detailed development of load balancing

a. Prediction of the optimal load ratio.

The two-stage amplification of the central dogma implies that the expression and noise levels can be controlled independently by the balance of transcription to translation. How does the cell achieve high and low gene expression optimally, and how does this strategy depend on the message cost?

To understand the optimization, we first define the load ratio R for a gene as the metabolic cost of translation relative to transcription:

$$R \equiv \frac{\mu_p}{\lambda \mu_m} = \frac{\varepsilon}{\lambda}. \quad (2.72)$$

In Sec. b., we show that the optimal load ratio is:

$$\hat{R} \equiv \frac{1}{\Lambda \ln 2} p_{\Gamma}(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \frac{1}{\delta \ln 2}), \quad (2.73)$$

where p_{Γ} is the PDF of the gamma distribution. The optimal load ratio is shown in Fig. 2.10C.

The dependence of the optimal load ratio $\hat{R}$ on Λ is extremely weak, but it is strongly dependent on message number. As a result, for low transcription genes ($\mu_m \ll 10$), the metabolic load is predicted to be dominated by transcription; whereas, for highly transcribed genes ($\mu_m \gg 10$), the metabolic load is dominated by translation. These predictions are robust since they are independent of the relative load Λ .

b. Measurements of the load ratio

Unfortunately, there is somewhat limited data to which to compare the model. The best source we found was Kafri *et al.* [44] who analyzed the differences in fitness between transcription and transcriptional-and-translation of a fluorescent protein driven by the pTDH3 promoter in yeast. This promoter is one of the strongest in yeast. Based on the RLTO model, we would predict this promoter to have a very high translation efficiency and therefore a large load ratio; however, the translation efficiency is much lower than one would predict based on a global analysis and likewise its load ratio is roughly unity, which based on the smaller than expected translation efficiency is broadly consistent with our expectations. A satisfactory test of this prediction will require larger-scale measurements that probe more representative genes.

2.5.9 Results: Translation efficiency is predicted to increase with transcription.

Now that we have defined the optimal load ratio (Eq. 2.73), the equation for optimal translation efficiency can be written concisely:

$$\hat{\varepsilon} = \lambda \hat{R} \quad \text{or} \quad \hat{E} = \Lambda \hat{R}, \quad (2.74)$$

where $\hat{R}$ depends weakly on the relative load Λ . The RLTO model predicts that optimal partitioning of amplification between transcription (gain μ_m) and translation (gain ε) has two important qualitative features: (i) As the message cost (λ) rises, the optimal translation efficiency increases in proportion. (ii) The optimal translation efficiency is also approximately proportional to message number ($\hat{\varepsilon} \propto \mu_m$). (See Fig. 2.4A.) Therefore, the RLTO model

predicts that low expression levels should be achieved with low levels of transcription and translation, whereas high expression genes are achieved with high levels of both. We call this relation between optimal transcription and translation *load balancing*.

2.5.10 Results: RLTO predicts that message number responds to message cost

We will first focus on analyzing the implications of the message cost dependence in Eq. 2.74. At a fixed load ratio, Eq. 2.74 clearly implies that the translation efficiency increases as the message cost λ increases; however, the message number (and load ratio) also respond to compensate to changes in λ . To probe the dependence on message cost in an experimentally relevant context, consider optimal message numbers in a reference condition (relative load Λ_0) relative to a second perturbed condition (relative load Λ). The predicted relation between the optimal messages numbers is shown in Fig. 2.12A. The resulting relation between the optimal message numbers is roughly linear on a log-log plot, predicting the approximate power-law relation:

$$\hat{\mu}_m(\Lambda) \propto \hat{\mu}_m(\Lambda_0)^\alpha, \quad (2.75)$$

describing a non-trivial global change in the regulatory program.

2.5.11 Results: RLTO predicts the yeast global regulatory response

To test the RLTO predictions, we compared the relative message numbers for yeast growing under phosphate depletion, which increases the message cost [44], to a reference condition [45]. As predicted, the relative transcriptome data was well described by a power law (Eq. 2.75) and the observed slope was smaller than one: $\hat{\alpha} = 0.837 \pm 0.001$, as predicted by the increased message cost. See Fig. 2.12B.

The observation of this large-scale regulatory change has an important implication: This response supports a nontrivial hypothesis that the RLTO model not only can predict how the cell is optimized in an evolutionary sense, but can predict global regulatory responses as well.

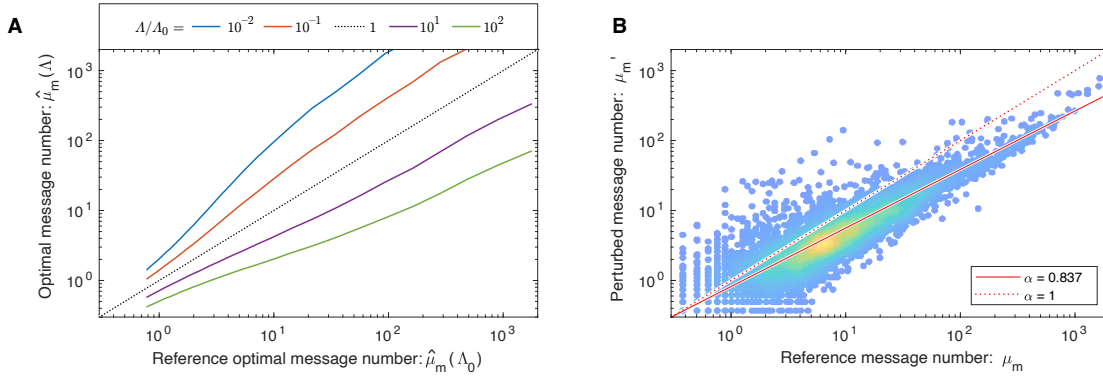


Figure 2.12: **The message cost affects transcription genome wide. Panel A: Message number decreases with increased relative load Λ .** The optimal message number responds to changes in the message cost. The RLTO model predicts an approximate power-law relation (linear on a log-log plot) between message numbers. **Panel B: A power-law relation is observed.** To test whether central dogma regulation would adapt dynamically as predicted, we analyzed the relation between the yeast transcriptome under reference conditions and phosphate depletion (perturbed), which increases the message cost [44]. (Data from Ref. [45].) As predicted by the RLTO model, a global change in regulation is observed, which generates a power-law relation with scaling exponent $\alpha = 0.837 \pm 0.01$. The observed exponent is smaller than one, as predicted by an increased relative load Λ .

Note that Metzl-Raz *et al.* [45] also explored conditions that increased the cost of protein; however, here the predictions of the model are ambiguous. The complication arises due to the observed decrease in size of the cells in the experimental condition, which decreases the total metabolic load N_0 . As discussed above, the relative load $\Lambda = \lambda/N_0$ is the key determinant in Eq. 2.75; however, even as the relative cost of transcription λ decreases in the experimental condition, the total metabolic load N_0 also decreases, making no clear prediction about how the relative load Λ changes.

2.5.12 Methods: Prediction of the proteome fraction

a. Parameter-free prediction of proteome fraction.

We now turn our focus to an analysis of the implications of *load balancing*: the message number dependence of the optimal translation efficiency (Eq. 2.74). The most direct test of this prediction is measuring the relation between proteome fraction and message number. The RLTO model predicts proteome fraction (Eq. 2.42):

$$\hat{\Phi}_p = \hat{E} \hat{\mu}_m \propto \hat{\mu}_m^2, \quad (2.76)$$

where μ_m is the observed message number and the optimal relative translation efficiency is predicted by Eq. 2.74. The proportionality is only approximate but gives important intuition for how protein number depends on message number in the RLTO model, in contrast to a constant-translation-efficiency model: $\Phi_p \propto \mu_m$. To compare these predictions to protein abundance measurements, we will renormalize the protein fraction to be defined relative to total protein number rather than N_0 . This renormalization eliminates the Λ dependence to result in a parameter-free prediction of the proteome fraction.

b. RLTO: proteome fraction

Starting from Eq. 2.104, clearly:

$$\hat{\mu}_p \propto \hat{\mu}_m p_{\Gamma}(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \quad (2.77)$$

which can be used to predict the proteome fraction (where we have restored the explicit gene i subscript):

$$\hat{\Phi}_{p,i} \equiv \frac{\hat{\mu}_{p,i}}{\sum_j \hat{\mu}_{p,j}}, \quad (2.78)$$

where the second subscript is the gene index. To predict the proteome fraction, we computed the proportionality constant C :

$$C \equiv \sum_i [\mu_{m,i} p_{\Gamma}(\mu_m | \frac{\mu_{m,i}}{\ln 2}, o \ln 2) |_{o=\hat{o}(\mu_{m,i})}], \quad (2.79)$$

where the message numbers μ_{mi} for gene i are the experimentally observed message numbers, the implicit o_i values are predicted by the RLTO model (Eq. 2.61) for message number μ_{mi} and the sum index i runs over all genes. The predicted optimal proteome fraction is:

$$\hat{\Phi}_{p,i} = C^{-1} \left[\mu_{m,i} p_{\Gamma}(\mu_{m,i} | \frac{\mu_{m,i}}{\ln 2}, o \ln 2) |_{o=\hat{o}(\mu_{m,i})} \right], \quad (2.80)$$

which generates the predicted solid curves shown in Fig. 2.4BCD.

c. Constant-translation-efficiency model: proteome fraction

For the constant translation efficiency model, we define the normalization:

$$C' \equiv \sum_i \mu_{m,i}, \quad (2.81)$$

and the predicted proteome fraction is:

$$\Phi'_{p,i} = C'^{-1} \mu_{m,i}, \quad (2.82)$$

which generates the predicted dotted curves shown in Fig. 2.4BCD.

d. Sources of experimental data for proteome fraction analysis

For *E. coli* data, the protein abundance data was generated by mass spec measurements and the message abundance data was from [23]. For the yeast data, the protein abundance data is measured by mass spec and message abundances are determined by [22]. For the mammalian

data, we used mouse data. The protein abundance data is measured by mass-spec and message abundances are determined by [24].

We estimated the message number μ_m as described in Sec. c.. For the mouse data, the study provided message lifetimes, the cell cycle duration and abundances in molecules per cell [24]. For the *E. coli* and yeast, the total number of proteins, messages *etc*, cell cycle duration and message lifetimes for each organism and their sources are described in Tab. 2.2.

2.5.13 Results: Load balancing is observed in eukaryotic cells

A non-trivial prediction of the RLTO Model is that translation efficiency and message number should be roughly proportional. Qualitatively, this strategy allows expression levels to be increased while distributing the added metabolic load between transcription, which reduces noise, and translation, which does not affect the noise. We predict the optimal translation efficiency versus message number which matches the observations in eukaryotic cells (Fig. 2.4BC). However, in *E. coli*, the translation efficiency and message number are *not* strongly correlated (Fig. 2.4D). Why does this organism appear not to load balance? We demonstrate that the observed translation efficiency is consistent with the RLTO model, augmented by a ribosome-per-message limit. (See Sec. 2.5.15.) Hausser *et al.* have proposed just such a limit, based on the ribosome footprint of mRNA molecules [11]. (See Sec. 2.5.17.) Although this augmented model is consistent with central dogma regulation in *E. coli*, it is not a complete rationale. This proposed translation-rate limit could be circumvented by increasing the lifetime of *E. coli* messages which would increase the translation efficiency. Why the message lifetime is as short as observed will require a more detailed *E. coli*-specific analysis.

2.5.14 Discussion: Relation between load balancing and previous results

Hausser *et al.* have previously performed a more limited analysis of the trade-off between metabolic load and gene-expression noise [11]. In this section, we will provide some more context into the differences between the two approaches.

Hausser *et al.* assume a symmetric (not an asymmetric) fitness landscape and consider only the metabolic cost of transcription (but not translation). Their model depends on two (not one) gene-specific parameters: an optimal protein number and a sensitivity, which defines the curvature of the fitness [11]. The authors maximize fitness with respect to the transcription rate (but not the translation rate) and the condition they derive depends on the two (not one) unknown, gene-specific parameters. As a result, this condition is not predictive of global regulatory trends without non-trivial, gene-specific measurements or assumptions about the unknown sensitivity.

The Hausser model assumes that the growth rate has the form:

$$k = k_0 - \frac{1}{2}|k''|(N_p - \mu_p)^2 - k_0\Lambda\mu_m, \quad (2.83)$$

where k is the growth rate, and we have rewritten the form of the fitness to better match our own definitions. Here k'' is the second derivative of the growth rate at the optimal protein number μ_p and N_p is the stochastic protein number. If we take the expectation with respect to the protein number, we get:

$$k = k_0 - \frac{1}{2}|k''|\sigma_p^2 - k_0\Lambda\mu_m, \quad (2.84)$$

and substituting the noise model for the variance of the protein number gives:

$$k = k_0 - \frac{1}{2}|k''|\mu_p^2\left(\frac{\ln 2}{\mu_m} + C_0\right) - k_0\Lambda\mu_m, \quad (2.85)$$

where C_0 is the noise floor, and we have assumed the mean protein number is optimal (μ_p). If we maximize the growth rate with respect to μ_m , we get the following condition on the optimal message number:

$$\hat{\mu}_m^2 = \frac{1}{2} \frac{|k''|}{k_0} \mu_p^2 \frac{\ln 2}{\Lambda}, \quad (2.86)$$

which depends on the unknown curvature k'' . To make global predictions about how transcription and translation are related, some added assumptions are necessary to describe how k'' scales with protein abundance.

To illustrate how this expression does not make explicit global predictions, let's consider a number of plausible possibilities. First, we will assume that k'' is independent of μ_p and on average all proteins are equally sensitive to changes in protein number. In this case, we find:

$$\mu_p \propto \hat{\mu}_m \sqrt{\Lambda}, \quad (2.87)$$

$$\hat{\varepsilon} \propto \sqrt{\Lambda}, \quad (2.88)$$

implying a constant translation efficiency which is inversely proportional to the square root of the relative load.

Alternatively, we can assume that $k'' \propto \mu_p^{-2}$ and, on average, the cell is equally sensitive to changes in the relative number of proteins (*i.e.* $\Delta p/\mu_p$), regardless of expression level. In this case,

$$\hat{\mu}_m \propto 1/\sqrt{\Lambda}, \quad (2.89)$$

$$\hat{\varepsilon} \propto \mu_p \sqrt{\Lambda}, \quad (2.90)$$

implying a constant message number, irrespective of expression level, and a translation efficiency that is proportional to expression level.

Finally, we will assume that $k'' \propto \mu_p^{-1}$, which is the intermediate case. Here:

$$\mu_p \propto \hat{\mu}_m^2 \Lambda, \quad (2.91)$$

$$\hat{\varepsilon} \propto \hat{\mu}_m \Lambda, \quad (2.92)$$

implying that translation efficiency should increase with message number, analogous to our prediction. It appears that Hausser *et al.* implicitly also favor this model, since they define their sensitivity parameter to include a power of protein number μ_p . They justify this assumption by arguing that since $\sigma_p^2 \propto \mu_p$, it makes sense to define the *sensitivity to noise* to include a factor of μ_p [11]. At best, this is somewhat fuzzy logic since, as we demonstrate in the paper, Eq. 2.92 implies that the protein variance does not scale $\sigma_p^2 \propto \mu_p$!

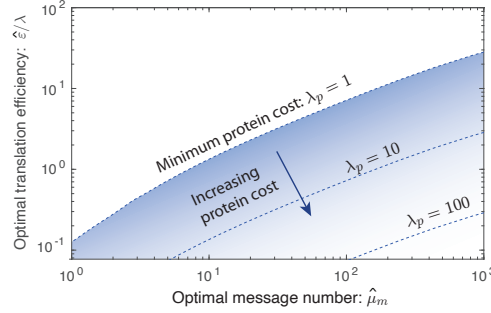


Figure 2.13: **Increased protein cost decreases optimal translation efficiency.** A protein cost of $\lambda_p = 1$ corresponds to the metabolic cost of protein synthesis only, and is the minimum protein cost. For larger protein costs, the optimal translation efficiency is lower. As a result, the $\lambda_p = 1$ curve represents an upper bound of the optimal translation efficiency.

The authors also propose a lower limit on the translation-transcription ratio; however, their limit is dependent on the noise floor, which only affects genes with the highest transcription rates in eukaryotic cells. The implementation of a more appropriate estimate of the noise, relevant for the vast majority of genes, does not lead to the same limit.

2.5.15 Methods: Analysis of translational limits & gene-specific load analysis

To explore the consequences of a protein-specific load, we can modify the metabolic load term in the growth rate equation (Eq. 3.62):

$$E\mu_m \rightarrow \lambda_p E\mu_m, \quad (2.93)$$

which includes an additional parameter: the protein cost λ_p , which is 1 if the fitness cost is equal to the metabolic load and greater than one if the cost is higher. We will also treat the metabolic load per message λ as a gene-specific parameter in this section only. The optimization can be repeated for this augmented model.

To analyze the effect of increased protein load, we modify Eq. 3.62:

$$\ln \frac{k}{k_0} = -(\Lambda + \lambda_p E)\mu_m - \frac{1}{\ln 2} \gamma\left(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}\right), \quad (2.94)$$

to contain the supplemental load factor λ_p which is unity if the only protein load is metabolic and $\lambda_p > 1$ if there is additional load (*e.g.* toxicity). The optimization conditions (Eqs. 2.55 and 2.56) become:

$$0 = -\frac{\lambda + \lambda_p \hat{\varepsilon}}{N_0} - \frac{1}{(\ln 2)^2} \gamma_{,1} \left(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon} \ln 2} \right), \quad (2.95)$$

$$0 = -\frac{\lambda_p \hat{\mu}_m}{N_0} + \frac{n_p}{\hat{\varepsilon}^2 (\ln 2)^2} \gamma_{,2} \left(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon} \ln 2} \right). \quad (2.96)$$

Using the same algebraic approach as before, we can derive the same optimal overabundance and load equations:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m} \gamma \left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{\delta} \ln 2} \right), \quad (2.97)$$

$$\hat{R} = \frac{1}{\Lambda \ln 2} p_{\Gamma}(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{\delta} \ln 2); \quad (2.98)$$

however, the relation between the load and the translation efficiency now has an extra factor: λ_p :

$$R = \frac{\lambda_p \varepsilon \mu_m}{\lambda \mu_m} = \frac{\lambda_p \varepsilon}{\lambda}, \quad (2.99)$$

representing the modified total load ratio.

2.5.16 Results: Increased protein-specific cost reduces the optimal translation efficiency.

The relation between the overabundance and message number is unchanged (Eq. 2.71). This result can be rationalized in the following way: The optimal overabundance is determined by the noise which is determined by message number only. This relation is unaffected by the added parameter λ_p . However, the optimal translation efficiency is affected:

$$\hat{\varepsilon} = \frac{\lambda}{\lambda_p} \hat{R}, \quad (2.100)$$

where $\hat{R}$ is the optimal load ratio, defined by Eq. 2.74. The optimal curves are shown in Fig. 2.13.

How do these added considerations affect the RLTO predictions? First, we consider message and protein length. What are the optimal translation efficiencies for two proteins,

one ten times the length of the other, at fixed protein number? In this case, we will assume that both the transcriptional cost (λ) as well as the translational cost (λ_p) increase tenfold. These increases cancel, resulting in the same optimal translation efficiency since it is only the relative cost of transcription to translation that is determinative of the translation efficiency.

Now consider a tenfold protein-specific increase in protein cost at fixed message cost and fixed protein number. The message number and translation efficiency would change by compensatory factors of 10:

$$\hat{\mu}_m \rightarrow 10 \cdot \hat{\mu}_m, \quad (2.101)$$

$$\hat{\varepsilon}_m \rightarrow \frac{1}{10} \cdot \hat{\varepsilon}_m, \quad (2.102)$$

to maintain the protein number.

Returning to our original motivation, we can understand how genes with a higher protein-to-message cost migrate downwards and rightwards off the optimal $\lambda_p = 1$ curve, predicting a cloud versus a narrow strip in proteome fraction measurements shown in Fig. 2.4. If the relative load Λ were directly measured, we would expect the predicted optimal translation efficiency curve for $\lambda_p = 1$ to lie at the top edge of the observed data cloud rather than the bisecting it. This bisection is the consequence of fitting an effective relative load parameter to the abundance data in the unaugmented RLTO model.

2.5.17 Discussion: Translation limits in *E. coli*

A critical assumption in the RLTO model to this point has been that the optimal central dogma parameters are realizable in the cell; however, translation can be limited by a number of different mechanisms. The superior performance of the constant- over the optimal-translation-efficiency model in *E. coli* (Fig. 2.4D) suggests that this assumption may not be satisfied for bacteria. How do translation limits affect the model phenomenology?

When considering possible limits on translation, there are two natural mechanisms: (i) ribosome-number limit, where the number of ribosomes in the cell limits translation and (ii) a ribosome-per-message limit, where the number of ribosomes per message is limiting.

Assuming the ribosome-number-limit mechanism, the original unconstrained optimization problem can be recast as a constrained optimization problem where the protein cost λ_p is reinterpreted as a Lagrange multiplier to constrain the number of proteins translated (*e.g.* [46]). In spite of this reformulation, we would still predict the same functional form for the coupling between the optimal translation efficiency and message number. (*I.e.* it is still optimal to have a higher translation efficiency for highly-expressed genes even if the total number of proteins is fixed.) Therefore, the ribosome-number-limit mechanism cannot be the rationale for the constant translation efficiency observed in *E. coli*.

Assuming the ribosome-per-message-limit mechanism, we limit the translation efficiency to a restricted range of values. If the unconstrained optimum lies above this range, the optimum is at the maximum limiting value. If the unconstrained optima for all genes lie above the realizable range, the model predicts a translation efficiency uncoupled from the message number, as observed. These predictions are consistent with the observed central dogma regulatory program in *E. coli*. In added support of this hypothesis, Hausser *et al.* have argued that *E. coli* translates close to just such a ribosome-per-message limit as a consequence of the finite ribosome complex footprint on a message [11].

2.5.18 Methods: Estimate of the message cost and metabolic load

We can estimate the message cost λ from the known total protein number for yeast and mammalian cells. (For *E. coli* this estimate is not possible since the protein cost is not determinative of the translation efficiency.)

The optimal translation efficiency for gene i is (Eq. 2.59):

$$\hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} p_{\Gamma}(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \quad (2.103)$$

and therefore the optimal protein number for gene i is:

$$\hat{\mu}_p = \hat{\mu}_m \hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} \hat{\mu}_m p_{\Gamma}(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2). \quad (2.104)$$

We define the normalization constant A (where we restore the explicit gene i subscript):

$$A = \sum_i \hat{\mu}_{m,i} \cdot \frac{1}{\Lambda \ln 2} p_{\Gamma}(\hat{\mu}_{m,i} | \frac{\hat{\mu}_{m,i}}{\ln 2}, \hat{o}_i \ln 2), \quad (2.105)$$

where we have restored the explicit gene i index running over all genes. Now, by summing Eq. 2.104, over all genes, we derive an expression for the total protein number N_p^{tot} in terms of the message cost λ and the normalization constant A :

$$N_p^{\text{tot}} = \lambda A. \quad (2.106)$$

Solving for the protein cost results in the estimate:

$$\hat{\lambda} = \frac{N_p^{\text{tot}}}{A}. \quad (2.107)$$

This message cost estimate $\hat{\lambda}$ can then be plugged into the metabolic load definition:

$$N_0 \equiv L_0 + \sum_i (\lambda + \varepsilon_i) \mu_{m,i}, \quad (2.108)$$

to estimate its size:

$$\hat{N}_0 \equiv L_0 + \hat{\lambda} N_m^{\text{tot}} + N_p^{\text{tot}}, \quad (2.109)$$

where we have ignored the non-protein and non-message contributions to the load ($L_0 = 0$).

a. Detailed protocol

We first estimate the message numbers, as described in Sec. c., from data. For each gene i , we set the optimal message number equal to the observed message number and then compute the optimal overabundance from the message number using Eq. 3.70. (Since the result is independent of the assumed Λ value, we set an arbitrary initial value of $\Lambda = 10^{-5}$.) We then use these single gene optimal message number and overabundances to compute A using Eq. 2.105. In Eqs. 2.107 and 2.109, we use the N_p^{tot} from Tab. 2.2. N_m^{tot} is computed by summing the estimated message numbers.

b. Estimate the message cost and metabolic load in yeast

In yeast, the estimates are:

$$A = 4.8 \times 10^5, \quad (2.110)$$

$$\hat{\lambda} = 1.0 \times 10^2, \quad (2.111)$$

$$\hat{N}_0 = 6.2 \times 10^7, \quad (2.112)$$

$$\hat{\Lambda} = 1.6 \times 10^{-6}, \quad (2.113)$$

where the data sources are described in detail in Sec. b..

c. Estimate the message cost and metabolic load in human cells

In human cells, the estimates are:

$$A = 4.3 \times 10^6, \quad (2.114)$$

$$\hat{\lambda} = 7.1 \times 10^2, \quad (2.115)$$

$$\hat{N}_0 = 2.4 \times 10^9, \quad (2.116)$$

$$\hat{\Lambda} = 2.9 \times 10^{-7}, \quad (2.117)$$

where the data sources are described in detail in Sec. c..

d. Estimate of the modified relative load in bacterial cells

In bacterial cells, we will assume a constant translation efficiency model. We therefore use the modified relative load formula (Eq. 3.64) to estimate Λ' . We will assume that the load is dominated by proteins and messages:

$$N_0 = \sum_i (\lambda + \varepsilon) \mu_{m,i} = (\lambda + \varepsilon) N_m, \quad (2.118)$$

where N_m is the total number of messages. We can then solve this equation for Λ' :

$$\hat{\Lambda}' = \frac{\lambda + \varepsilon}{N_0} = \frac{1}{N_m} \approx 10^{-5}, \quad (2.119)$$

based on the total message number estimate for *E. coli*. (See Tab. 2.2.)

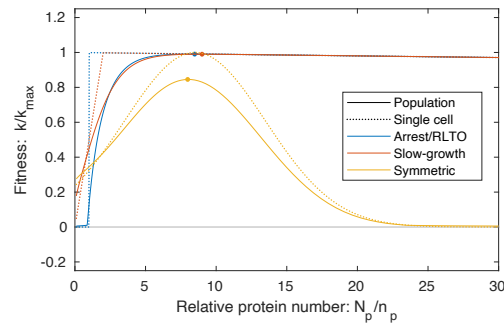


Figure 2.14: **Exploring the mathematical mechanism of overabundance.** Single-cell and population growth rate are compared for three different models: arrest (RLTO), slow-growth, and symmetric models. In the arrest model (RLTO), the growth rate goes to zero below threshold protein level n_p . In the slow-growth model, the growth rate transitions continuously to zero as the N_p is depleted below n_p . In both the arrest and slow-growth models, there is a small negative slope above the threshold corresponding to the metabolic load. In the symmetric model, the fitness cost is symmetric about the optimum. Both the threshold-like and slow-growth models are optimized at mean expression levels μ_p far exceeding the threshold level n_p . This is a consequence of the highly-asymmetric dependence of the fitness on protein number N_p . This leads to the phenomenon of protein overabundance. In contrast, the symmetric model is optimized in close proximity to its single-cell optimum.

2.5.19 Model robustness & exploring alternatives to RLTO

In this section, we investigate the phenomenology of three different single-cell growth rate functions to determine what model features result in overabundance. We consider an *arrest model* (the RLTO model), a *slow-growth model*, and a *symmetric model*.

a. Methods: Defining alternative models

In each case, we will assume that the protein number is described by a gamma distribution:

$$N_p \sim \Gamma\left(\frac{\mu_m}{\ln 2}, \varepsilon \ln 2\right). \quad (2.120)$$

We will assume the cell-cycle duration T is determined by this stochastic protein number N_p and then compute the population growth rate using Eq. 2.44 for a range of different message numbers μ_m . In each case, $\tau_0 = 1/N_0$, $N_0 = 10^5$, $\varepsilon = 30$, $n_p = \varepsilon \ln 2$. The mean expression level is $\mu_p = \mu_m \varepsilon$.

b. Model 1: Arrest (RLTO) model

The arrest (RLTO) model has cell cycle duration:

$$T = \tau_0 \begin{cases} \infty, & N_p < n_p \\ N_0 + N_p, & N_p > n_p \end{cases}, \quad (2.121)$$

where protein expression below threshold n_p results in growth arrest.

c. Model 2: Slow-Growth model

In the slow-growth model, we imagine two processes: (i) checkpoint process X and (ii) other processes. The cell will divide after whichever process finishes last. Other processes will finish after time predicted by the metabolic load, identical to the threshold model defined above. However, we model checkpoint process X as the completion of a fixed amount of activity in an irreversible process. We will therefore assume it will take a time inversely

proportional to the amount of enzyme X (N_p). The amount of activity is set by effective threshold n_p :

$$T = \tau_0 \max\{N_0 + N_p, \frac{2n_p N_0}{N_p}\} \quad (2.122)$$

such that n_p defines the level of protein required to make the growth rate half the metabolic limit.

Unlike the arrest model, cell growth slows but does not stop for $N_p < n_p$. This model will test whether the results of the RLTO model are an artifact of the assumed arrest-based slow growth.

d. Model 3: Symmetric model

For the symmetric model, we choose the model parameters such that the single-cell optimum was close to the other models: $n_0 = 8.5$, $\sigma_n = 5$. The cell-cycle duration is:

$$T = \tau_0 N_0 \exp\left(\frac{(N_p - n_0)^2}{2\sigma_n^2}\right), \quad (2.123)$$

such that the noise-free growth rate will be Gaussian is N_p .

2.5.20 Results: Overabundance is a robust prediction

The growth rates as a function of the mean expression level μ_p are shown in Fig. 2.14. The symmetric model has a population optimum in close proximity to its single-cell optimum, as we intuitively expect. However, both the arrest (RLTO) model and the slow-growth model have optima far above the threshold number n_p . We therefore conclude that it is fitness asymmetry rather than growth arrest that is responsible for the overabundance phenomenon.

Why doesn't growth arrest of a sub-population lead to a stronger effect than the same sub-population growing slowly? In Ref. [30], we showed that the population doubling time $\bar{T}$ can be understood as the exponential mean of the stochastic cell-cycle duration:

$$\bar{T} \equiv f^{-1}[\mathbb{E}_T f(T)], \quad (2.124)$$

Model organism	Growth condition	Doubling time:	Message lifetime:	Message recycling ratio:	Total number of			Average	
					messages /cell:	messages /cell-cycle:	proteins:	translation efficiency:	translation rate:
					$N_{m/c}^{\text{tot}}$	N_m^{tot}	N_p^{tot}	ε	$\beta_p \text{ (h}^{-1}\text{)}$
<i>Escherichia coli</i> (<i>E. coli</i>)	LB	30 min	2.5 min	12	7.8×10^3	9.4×10^4	3×10^6	22	530
	M9	[36] 90 min	[37] 2.5 min	36	[47] 2.4×10^3	8.6×10^4	[48] 3×10^6	24	580
<i>Saccharomyces cerevisiae</i> (<i>Yeast-haploid</i>)	YEPD	[36] 90 min	[37] 22 min	4	[47] 2.9×10^4	1×10^5	[48] 5×10^7	4×10^2	410
		[49]	[50]		[51]		[52]		
<i>Mus musculus</i> (<i>Mammalian-mouse</i>)	Tissue	27.5 h	15 h	1.8	1.7×10^5	3×10^5	3×10^9	1×10^4	660
		[24]	[24]		[24]	[24]	[24]		
<i>Homo sapiens</i> (<i>Human</i>)	Tissue	24 h	14 h	1.7	3.6×10^5	5×10^5	2×10^9	4×10^3	120
		[53]	[54]		[49]		[48]		

Table 2.2: **Central dogma parameters for three model organisms with detailed references.** Columns three through seven hold representative values for measured central-dogma parameters for the model organisms described in the paper. Each value is followed by a reference for its source.

where $\mathbb{E}_T$ is the expectation over the stochastic duration T and $f(t) \equiv \exp(-kt)$, where $k = \bar{T}^{-1} \ln 2$ is the population growth rate. Due to the functional form of $f(t)$, any long cell cycles are exponentially suppressed in their contribution to the exponential mean. Therefore, low-probability extremely-long-duration cell cycles only contribute to the growth rate by reducing the fraction of growing cells.

2.6 Quantitation of central dogma parameters for one-message-rule

The RLTO model predicts the *one-message-rule* for the lower threshold on transcription for essential genes. In this section, we use transcriptome data from the literature to test this prediction. We first describe the sources of the data (Sec. 2.6.1), how the estimates are computed (Sec. 2.6.2), the results (Sec. 2.6.3) and discussion (Sec. 2.6.4).

2.6.1 Methods: Selection of central dogma parameter estimates

The estimates for central dogma model parameters come from two types of data: (i) quantitative measurement of cellular-scale parameters for each organism (total number of messages in the cell, cell cycle duration, *etc*) and (ii) genome-wide studies quantitative of mRNA and protein abundance.

For the cellular-scale central dogma parameters, we relied heavily on an online compilation of biological numbers: BioNumbers [55]. This resource provides a collection of curated quantitative estimates for biological numbers, as well as their original source. In the interest of conciseness, we have cited only the original source in the Tab. 2.2, although we are extremely grateful and supportive of the creators of the BioNumbers website for helping us very efficiently identify consensus estimates for the parameters of the central dogma parameters.

For the selection of genome-wide studies on abundance, we used many of the same resources cited in BioNumbers as well as studies selected by a previous study of a quantitative analysis of the central dogma: Hausser *et al.* [11].

a. E. coli data

Message lifetimes: The message lifetimes (and median lifetime) were taken from a recent transcriptome-wide study by Chen *et al.* [37]. These investigators measured the lifetime in both rapid (LB) and slow growth (M9).

Noise: Taniguchi *et al.* have performed a beautiful simultaneous study of the proteome and transcriptome with single-molecule sensitivity [10]. Although we use the noise analysis data from this study for our supplemental analysis of *E. coli* noise, it is not the source for our *E. coli* transcriptome data due to the extremely slow growth of the cells in this study (150 minute doubling time), which is not consistent with the growth conditions for the other sources of data.

mRNA abundance: Instead, we used data from the more recent Bartholomaeus *et al.* study [47], which characterizes the transcriptome in both rapid (LB) and slow growth (M9).

Total cellular message number. This study was chosen since it was the source of the BioNumbers estimates of cellular message number in *E. coli* (BNID 112795 [55]).

Doubling time: The source of the doubling times for rapid (LB) and slow (M9) growth of *E. coli* comes from Bernstein [36].

Essential gene classification. The classification of essential genes in *E. coli* comes from the construction of the Keio knockout collection from Baba *et al.* [56].

Protein number. The total protein number in *E. coli* came from Milo's recent review of this subject [48].

b. Yeast data

Message lifetimes: The message lifetimes (and median lifetime) were taken from Chia *et al.* [50].

Noise: The noise data was taken from the Newman *et al.* study, which used flow cytometry of a library of fluorescent fusions to characterize protein abundance with single-cell resolution [9].

mRNA abundance: The transcriptome data comes from the very recent Blevins *et al.* study [57].

Total cellular message number. There are a wide-range of estimates for the total cellular message number in yeast: 1.5×10^4 [58] (BNID 104312 [55]), 1.2×10^4 [59] (BNID 102988

[55]), 6.0×10^4 [60] (BNID 103023 [55]), 2.6×10^4 [51] (BNID 106763 [55]) and 3.0×10^4 [61]. We used the compromise value of 2.9×10^4 .

Doubling time: The doubling time was taken from [49].

Protein number. The total protein number in yeast comes from Futcher *et al.* [52].

Essential gene classification. The classification of essential genes in yeast comes from van Leeuwen *et al.* [20].

Proteome abundance data: The proteome abundance data came from two sources: flow cytometry of fluorescent fusions from Newman *et al.* [9] as well as mass-spec data from de Godoy *et al.* [62].

c. Human data

Message lifetimes: The message lifetimes (and median lifetime) were taken from Yang *et al.* [54] who reported a median half life of 10 h which corresponds to a lifetime of 14 h.

mRNA abundance: The transcriptome data comes from the data compiled by the Human Protein Atlas [63], which we averaged over tissue types.

Total cellular message number. The total cellular message number in human comes from Velculescu *et al.* [64] (BNID 104330 [55]).

Doubling time: The doubling time was taken from [53].

Protein number. The total protein number in human came from Milo's recent review of this subject [48].

Essential gene classification. The classification of essential genes in human comes from Wang *et al.* [21].

2.6.2 Methods: Quantitative estimates of central dogma parameters

a. Estimating the cellular message number: $\mu_{m/c}$

For each model organism (and condition), we found a consensus estimate from the literature for the total number of mRNA messages per cell $N_{m/c}^{\text{tot}}$. This number and its source are

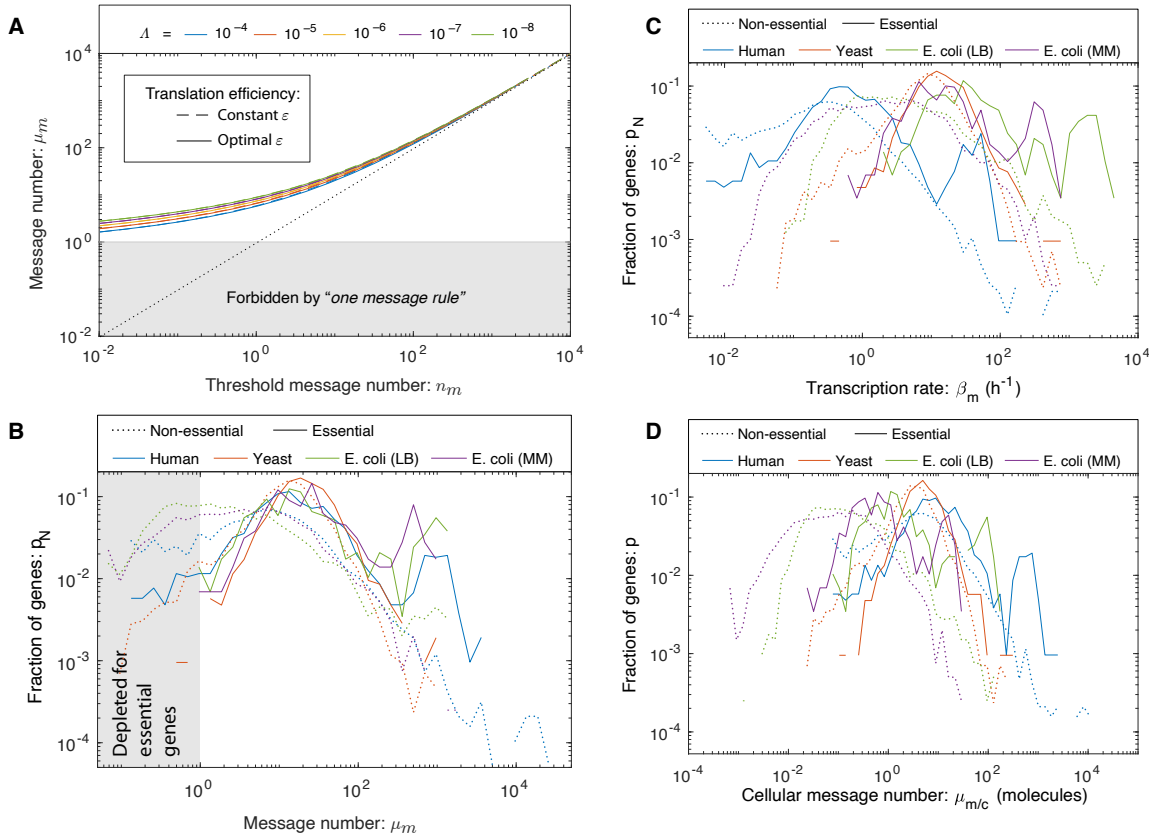


Figure 2.15: **The one message rule.** **Panel A: One-message-rule for essential genes.** For highly transcription genes (high μ_m), little compensation for noise is required and the optimal message number tracks with the threshold message number n_m . However, as the threshold message number approaches one ($n_m \rightarrow 1$), the noise is comparable to the mean, and the optimal message number μ_m increases to compensate for the noise. As a result, a lower threshold of roughly one message per cell-cycle is required for essential genes. This threshold is predicted for both fixed (dashed) and optimized translation efficiency (solid). The threshold is weakly dependent on relative load Λ . **Panel B: A one message threshold is observed in three evolutionarily-divergent organisms.** As predicted by the RLTO model, essential, but not nonessential genes, are observed to be expressed above a one message per cell-cycle threshold. All organisms have roughly similar distributions of message number for essential genes, which are not observed for message numbers below a couple per cell cycle. **Panel C: The distribution of gene transcription rate.** The typical transcription rate varies by two orders-of-magnitude between organisms. **Panel D: The distribution of gene cellular message number.** There is also a two-order-of-magnitude variation between typical cellular message numbers. No consistent lower threshold is observed for either statistic.

provided in Tab. 2.2. To estimate the number of messages corresponding to gene i , we re-scaled the un-normalized abundance level r_i :

$$N_{m/c,i} = N_{m/c}^{\text{tot}} \frac{r_i}{\sum_j r_j}, \quad (2.125)$$

where the sum over gene index j runs over all genes.

b. Estimating the transcription rate: β_m

To estimate the transcription rate for gene i , we start from the estimated cellular message number $N_{m/c,i}$ and use the canonical steady-state noise model prediction for the cellular message number:

$$N_{m/c,i} = \beta_{m,i} / \gamma_{m,i}, \quad (2.126)$$

where $\gamma_{m,i}$ is the message decay rate. Since gene-to-gene variation in message number is dominated by the transcription rate (*e.g* [37]), we estimate the decay rate as the inverse gene-median message lifetime:

$$\gamma_{m,i} = \tau_m^{-1}, \quad (2.127)$$

for which a consensus value was found from the literature. This number and its source are provided in Tab. 2.2. We then estimate the gene-specific transcription rate:

$$\beta_{m,i} = N_{m/c,i} / \tau_m. \quad (2.128)$$

c. Estimating the message number: μ_m

To estimate the message number of gene i , we use the predicted value from the canonical steady-state noise model (see a.):

$$N_{m,i} = T \beta_{m,i} = \frac{T}{\tau_m} N_{m/c,i}, \quad (2.129)$$

where T is the doubling time and $N_{m/c,i}$ is the cellular message number (Eq. 2.125).

2.6.3 Results: Histograms of central dogma transcriptional statistics

We generated histograms for each of the three transcriptional statistics: transcription rate β_m , cellular message number $\mu_{m/c}$, and message number μ_m . The histograms for transcription rate and cellular message number do not show a consistent lower limit (as predicted) and are shown in Fig. 2.15; however, the histogram for message number does show a consistent lower bound for the three model organisms and is shown in Fig. 2.15B.

2.6.4 Discussion: *E. coli* essential genes below the one-message-rule threshold

Since our own preferred model system is *E. coli*, we focus here. Our essential gene classification was based on the construction of the Keio knockout library [56]. By this classification, 10 essential genes were below threshold. (See Tab. 2.3.) Our first step was to determine what fraction of these genes were also classified as essential using transposon-based mutagenesis [65, 66]. Of the 10 initial candidates, only one gene, *ymfK*, was consistently classified as an essential gene in all three studies, and we estimate that its message number is just below the threshold ($\mu_m = 0.4$). *ymfK* is located in the lambdoid prophage element e14 and is annotated as a CI-like repressor which regulates lysis-lysogeny decision [67]. In λ phase, the CI repressor represses lytic genes to maintain the lysogenic state. A conserved function for *ymfK* is consistent with it being classified as essential, since its regulation would prevent cell lysis. However, since *ymfK* is a prophage gene, not a host gene, it is not clear that its expression should optimize host fitness, potentially at the expense of phage fitness. In summary, closer inspection of below-threshold essential genes supports the threshold hypothesis.

2.7 Analysis of gene-expression noise

This section provides a detailed development of gene expression noise. We continue the discussion of the model from Sec. a. that provided a self-contained development of the

Gene name	Message number:	Annotated function from Ecocyc	Essential (E)/Nonessential (N) Ref. [56], [65], [66]
<i>alsK</i>	μ_m 0.3	The <i>alsK</i> gene encodes a D-allose kinase. Its role in the degradation of D-allose is unclear; AlsK is not required for utilization of a D-allose carbon source; this effect may be due to the presence of other ambiguous sugar kinases within <i>E. coli</i> K-12.	E, N, N
<i>bcsB</i>	0.4	BcsB is encoded in a predicted operon together with <i>bcsA</i> , <i>bcsZ</i> and <i>bcsC</i> . In other organisms, these genes are involved in cellulose biosynthesis, a characteristic of the rdar (red, dry and rough) morphotype. However, the K-12 laboratory strain of <i>E. coli</i> does not show a rdar morphotype and does not produce cellulose.	E, N, N
<i>entD</i>	0.4	AcpS is the founding member of a 4'-phosphopantetheinyl (P-pant) transferase protein family that includes <i>E. coli</i> EntD, <i>E. coli</i> o195 protein, and <i>Bacillus subtilis</i> Sfp; family members share two conserved motifs but relatively low sequence identity overall.	E, N, N
<i>yafF</i>	0.4	No information about this protein was found by a literature search conducted on April 19, 2017.	E,-, N
<i>yagG</i>	0.6	<i>yagGH</i> is predicted to be a member of the XylR regulon; its products may mediate transport (YagG) and hydrolysis (YagH) of xylooligosaccharides; putative XylR and CRP binding sites are identified upstream of <i>yagGH</i> .	E,-, N
<i>yceQ</i>	0.2	No information about this protein was found by a literature search conducted on July 12, 2017.	E, E, N
<i>ydiL</i>	0.2	No information about this protein was found by a literature search conducted on April 7, 2017.	E, N, N
<i>yhhQ</i>	0.4	YhhQ is an inner membrane protein implicated in the uptake of queuosine (Q) precursors - 7-cyano-7-deazaguanine (<i>preQ0</i>) and 7-aminomethyl-7-deazaguanine (<i>preQ1</i>) - for Q salvage. Q-modified tRNA is absent in $\Delta queD$ and $\Delta queD \Delta yhhQ$ strains grown in minimal media with glycerol; Q-modified tRNA is detected when a $\Delta queD$ strain is grown in minimal media plus 10 nM <i>preQ0</i> or <i>preQ1</i> but is absent when a $\Delta queD \Delta yhhQ$ strain is grown under these conditions. <i>yhhQ</i> expressed from a plasmid restores the presence of Q-modified tRNA in a $\Delta queD \Delta yhhQ$ strain.	E,-, N
<i>yibJ</i>	0.3	No information about this protein was found by a literature search conducted on July 9, 2018.	E, N, N
<i>ymfK</i>	0.4	YmfK is a component of the relic lambdoid prophage $\phi 14$ and is likely the SOS-sensitive repressor. It is similar to the P34 gene of the <i>Shigella flexneri</i> bacteriophage SfV and belongs to the LexA group of SOS-response transcriptional repressors.	E, E, E

Table 2.3: **Below-threshold essential genes identified in *E. coli*.** This table describes the message numbers and annotations for essential genes that we estimated to have expression below the threshold of one message per cell cycle. However, in the final column, we show classifications from three different studies. Only one of the identified genes, *ymfK*, was consistently defined as essential.

noise models developed by others which are the input to the RLTO model. Secs. 2.7.1-2.7.6 describe the RLTO prediction of non-canonical noise scaling and the test of this model.

2.7.1 Results: RLTO model predicts non-canonical noise scaling

The predicted scaling of the optimal translation efficiency with message number has many important implications, including on the global characteristics of noise. Based both on theoretical and experimental evidence, it is widely claimed that gene-expression noise should be inversely proportional to protein abundance [10, 68]:

$$\text{CV}_p^2 \propto \mu_p^{-1}, \quad (2.130)$$

for low-expression proteins, as observed in *E. coli* [10]; however, the more fundamental prediction is that the noise is inversely proportional to the message number:

$$\mu_p = \mu_m \varepsilon, \quad (2.131)$$

$$\text{CV}_p^2 = \frac{\ln 2}{\mu_m}. \quad (2.132)$$

In *E. coli*, the translation efficiency is roughly constant (*i.e.* $\hat{\mu}_p \propto \hat{\mu}_m$, Fig. 2.4D) and therefore Eq. 2.132 is consistent with the canonical noise model (Eq. 2.130). However, in eukaryotes, the translation efficiency grows with message number (*i.e.* $\hat{\mu}_p \propto \hat{\mu}_m^2$, Fig. 2.4BC). If we substitute this proportionality into Eq. 2.132, we predict the non-canonical noise scaling:

$$\text{CV}_p^2 \propto \mu_p^{-1/2}, \quad (2.133)$$

for eukaryotic cells.

2.7.2 Methods: Analysis of gene expression noise

The quantitative model for gene expression noise includes multiple contributions:

$$\text{CV}_p^2 \approx \frac{1}{\mu_p} + \frac{\ln 2}{\mu_m} + c_0, \quad (2.134)$$

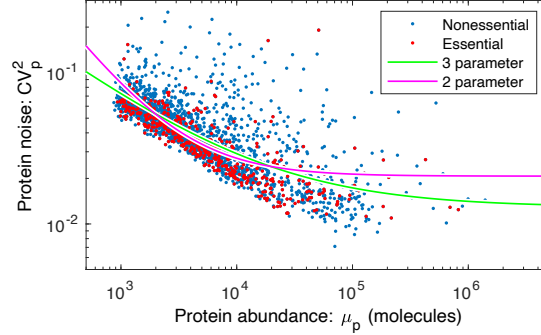


Figure 2.16: **Yeast noise fit against canonical noise model, with a noise floor.** Yeast noise data fit with the 2- (null hypothesis with μ_p^{-1} dependence) and 3- parameter (μ_p^a) models. The two-parameter model corresponds to the canonical noise model (Eq. 2.130) and fails to quantitatively fit the data.

where the first term can be understood to represent the Poisson noise from translation, the second term the Poisson noise from transcription, and the last term, c_0 , is called the *noise floor* and is believed to be caused by the cell-to-cell variation in metabolites, ribosomes, and polymerases *etc.* [69, 70].

In the main text of the paper, we have ignored the role of the noise floor in the analysis of noise in yeast. Unlike *E. coli*, where the noise floor is high ($CV_p^2 = 0.1$) and is determinative of the noise associated with almost all essential genes [10, 69, 70], in yeast the noise floor is much lower ($CV_p^2 = 0.01$) and therefore affects only genes with the highest expression.

In this section, we will consider models that include the noise floor, since its presence can make the noise scaling more difficult to interpret. To determine if the scaling of the noise is consistent with the canonical assumption that the noise is proportional to μ_p^{-1} for low expression, we will consider two competing empirical models for the noise (Fig. 2.16). In the null hypothesis, we will consider a model:

$$\eta_0(\mu_p; b, c) = \frac{b}{\mu_p} + c, \quad (2.135)$$

and an alternative hypothesis with an extra exponent parameter a :

$$\eta_1(\mu_p; a, b, c) = \frac{b}{\mu_p^a} + c. \quad (2.136)$$

We will assume that CV_p^2 is normally distributed about η with unknown variance σ_η^2 .

In this context, a maximum likelihood analysis is equivalent to least-squares analysis. Let the sum of the squares be defined:

$$S_I(\boldsymbol{\theta}) \equiv \sum_i [CV_{p,i}^2 - \eta_I(\mu_{p,i}; \boldsymbol{\theta})]^2, \quad (2.137)$$

for model I where $\boldsymbol{\theta}$ represents the parameter vector. The maximum likelihood parameters are

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} S_I(\boldsymbol{\theta}), \quad (2.138)$$

with residual norm:

$$\hat{S}_I = S_I(\hat{\boldsymbol{\theta}}). \quad (2.139)$$

To test the null hypothesis, we will use the canonical likelihood ratio test with the test statistic:

$$Z \equiv 2 \ln \frac{q_1}{q_0}, \quad (2.140)$$

where q_0 and q_1 are the likelihoods of the null and alternative hypotheses, respectively. Wilks' theorem states that Z has a chi-squared distribution of dimension equal to the difference of the dimension of the alternative and null hypotheses ($3 - 2 = 1$).

a. Hypothesis test I

In our first analysis, we will estimate the variance directly. We computed the mean-squared difference for successive CV_p^2 values, sorted by mean protein number μ_p . The variance estimator is

$$\hat{\sigma}_\eta^2 = \frac{1}{2} \langle (CV_{p,i}^2 - CV_{p,i+1}^2)^2 \rangle_i = 6.3 \times 10^{-4}, \quad (2.141)$$

where the brackets represent a standard empirical average over gene i for the μ_p -ordered gene CV_p^2 values. The test statistic can now be expressed in terms of the residual norms:

$$Z = (\hat{S}_1 - \hat{S}_2) / \hat{\sigma}_\eta^2, \quad (2.142)$$

$$= 3.3 \times 10^4, \quad (2.143)$$

which corresponds to a p-value far below machine precision. We can therefore reject the null hypothesis.

b. Hypothesis test II

In a more conservative approach, we can use maximum likelihood estimation to estimate the variance of each model independently as a model parameter. In this case, the test statistic can again be expressed in terms of the residual norms:

$$Z = N \ln \frac{\hat{S}_1}{\hat{S}_2}, \quad (2.144)$$

$$= 1.6 \times 10^2, \quad (2.145)$$

where N is the number of data points. (Details of derivation are in Sec. d.) In this case, the p-value can be computed assuming the Wilks' theorem (*i.e.* the chi-squared test):

$$p = 6 \times 10^{-36}, \quad (2.146)$$

again, strongly rejecting the null hypothesis.

c. Maximum likelihood estimates of the parameters

In the alternative hypothesis, the maximum likelihood estimate (MLE) of the empirical noise model (Eq. 2.136) parameters are (Fig. 2.16):

$$a = 0.57 \pm 0.02, \quad (2.147)$$

$$b = 3.0 \pm 0.5, \quad (2.148)$$

$$c = 0.013 \pm 0.001, \quad (2.149)$$

where the parameter uncertainty has been estimated using the Fisher Information in the usual way using the MLE estimate of the variance [**Rao1945**, **Cramer1946**].

d. Details: Statistical details MLE estimate of the variance

The minus-log-likelihood for the normal model I is:

$$h_I(\hat{\boldsymbol{\theta}}, \sigma^2) = \frac{N}{2} \ln 2\pi\sigma^2 + \frac{1}{2\sigma^2} \hat{S}_I, \quad (2.150)$$

where $\hat{S}_I$ is the least-square residual. We then minimize h_I with respect to the variance σ^2 :

$$\partial_{\sigma^2} h|_{\hat{\sigma}^2} = 0, \quad (2.151)$$

to solve for the MLE $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{N} \hat{S}_I. \quad (2.152)$$

Next we evaluate h at the variance estimator:

$$h_I(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \frac{N}{2} \left[\ln 2\pi \frac{\hat{S}_I}{N} + 1 \right]. \quad (2.153)$$

The test statistics can be written in terms of the h 's:

$$Z = 2h_0(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - 2h_1(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2), \quad (2.154)$$

$$= N \ln \frac{\hat{S}_0}{\hat{S}_1}, \quad (2.155)$$

which can be evaluated directly in terms of the residual norms for the null and alternative hypotheses.

2.7.3 Results: Non-canonical noise scaling is observed in yeast.

To test the RLTO model predictions for noise scaling, we reanalyze the dataset collected by Newman *et al.*, who performed a single-cell proteomic analysis of yeast by measuring the abundance of fluorescent fusions by flow cytometry [9]. Since the competing models (Eqs. 2.130 and 2.133) make different scaling predictions, we first apply a statistical test to determine whether the observed scaling is consistent with the canonical model (Eq. 2.130). We consider the null hypothesis of the canonical model (Eq. 2.135) and the alternative hypothesis with an unknown scaling exponent (Eq. 2.136). To test the models, we perform

a null hypothesis test. (A detailed description of the statistical analysis, which includes the contribution of the noise floor, is given in the Sec. 2.7.) We reject the null hypothesis with a p-value of $p = 6 \times 10^{-36}$. The observed scaling exponent is $\hat{a} = -0.57 \pm 0.02$, which is close to our predicted estimated exponent from the RLTO model ($-\frac{1}{2}$).

2.7.4 Methods: Parameter-free prediction of noise from protein-message relation.

By combining the noise model (Eq. 2.132) with a protein-message abundance relation, the relation between protein abundance and noise can be predicted without additional fitting parameters. (We call this prediction *parameter-free* since, although a parameter is fit when determining the protein-message abundance, once this relation has been established, no new parameters are fit in order to predict the noise.) To test this prediction, we will compare three competing models: (i) the RLTO model, (ii) an empirical protein-message abundance model, and (iii) the constant-translation-efficiency model.

a. Estimating protein number (μ_p) for the noise analysis

The protein abundance data for yeast grown in YEPD media and measured with flow cytometry fluorescence [9] were given in arbitrary units (AU). In order to convert from AU to protein number, the fluorescence values were rescaled by comparing with mass-spectrometry protein abundance data for yeast grown in YNB media [62]. Since the protein abundance from mass-spectrometry was given in terms of intensity, the intensity values were first rescaled by the total number of proteins in yeast, 5×10^7 . (See Sec. b.) The mass-spectrometry protein data was thresholded at 10 proteins, based on the assumption that the noise of the data for 10 and fewer proteins makes the data unreliable. Next, the log of the fluorescence protein abundance in AU as a function of the log of thresholded mass-spectrometry protein abundance was fit as a linear function with an assumed slope of 1 to find the offset, 3.9, (Fig. 2.17) which corresponds to a multiplicative scaling factor. We then used that offset value to

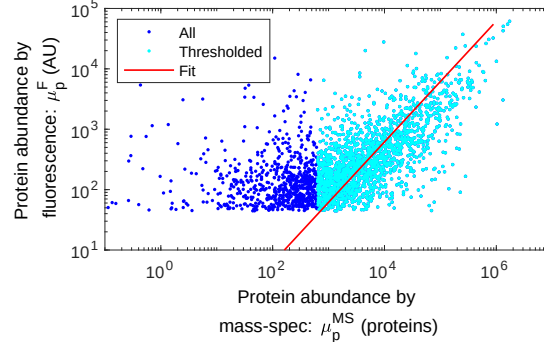


Figure 2.17: **Fit to rescale fluorescence intensity to protein number.** Protein abundance from flow cytometry fluorescence [9] as a function of mass-spectrometry scaled abundance [62]. The mass-spectrometry data was thresholded at 10 proteins, and then a linear fit was performed to find the multiplicative offset of 3.9, which was used to convert protein fluorescence AU to number.

rescale the fluorescence data from AU to protein number. We also compared to yeast grown in SD media [9] and found a similar offset result.

b. Empirical models for yeast gene expression

To generate the empirical model for protein number as a function of message number, we used protein abundance data from Newman *et al.* [9], re-scaled to estimate protein number (Sec. a.) and transcriptome data from Lahtvee *et al.* [71], re-scaled to estimate message number (Sec. c.).

c. Empirical model for protein number

We initially fit the empirical model for protein number,

$$\mu_p = C_0 \mu_m^{\alpha_0}, \quad (2.156)$$

to the data using a standard least-squares approach; however, the algorithm led to a very poor fit since it does not account for uncertainty in both independent and dependent variables. We therefore used an alternative approach [72], which assumes comparable error in both

variables. The model parameters are:

$$\alpha_0 = 2.1 \pm 0.04, \quad (2.157)$$

$$C_0 = 8.0 \pm 1.0, \quad (2.158)$$

where the uncertainties are the estimated standard errors. The result of the empirical model fit is shown in Fig. 2.18A, along with the constant-translation-efficiency model, and the RLTO model.

d. Empirical model for message number

For the prediction of the coefficient of variation, it is useful to invert Eq. 2.156 to generate a model for message number as a function of protein number:

$$\mu_m = C_0^{-1/\alpha_0} \mu_p^{1/\alpha_0}, \quad (2.159)$$

$$= C_1 \mu_p^{\alpha_1}, \quad (2.160)$$

where the last line defines two new parameters: a coefficient C_1 and an exponent α_1 . The resulting parameters and uncertainties are:

$$\alpha_1 \equiv 1/\alpha_0, \quad (2.161)$$

$$= 0.48 \pm 0.01, \quad (2.162)$$

$$C_1 \equiv C_0^{-1/\alpha_0}, \quad (2.163)$$

$$= 0.37 \pm 0.02, \quad (2.164)$$

where the uncertainties are the estimated standard errors.

e. Empirical model for translation efficiency

To generate an empirical model for translation efficiency, we started from the empirical model for protein number (Eq. 2.156), and then use Eq. 2.131 to relate protein number, message

number, and translation efficiency:

$$\varepsilon = \frac{\mu_p}{\mu_m}, \quad (2.165)$$

$$= C_0 \mu_m^{\alpha_0 - 1}, \quad (2.166)$$

$$= C_2 \mu_m^{\alpha_2}, \quad (2.167)$$

where the last line defines two new parameters: a coefficient C_2 and an exponent α_2 . The resulting parameters and uncertainties are:

$$\alpha_2 = \alpha_0 - 1, \quad (2.168)$$

$$= 1.07 \pm 0.04, \quad (2.169)$$

$$C_2 = C_0, \quad (2.170)$$

$$= 8.0 \pm 1.0, \quad (2.171)$$

where the uncertainties are the estimated standard errors.

f. Empirical model for the coefficient of variation

To generate an empirical model for the coefficient of variation, we started from the empirical model for message number (Eq. 2.160), and then substitute this into the statistical model prediction for CV_p^2 (Eq. 2.132):

$$CV_p^2 = \frac{\ln 2}{\mu_m}, \quad (2.172)$$

$$= C_0^{1/\alpha_0} \ln 2 \cdot \mu_p^{-1/\alpha_0}, \quad (2.173)$$

$$= C_3 \mu_p^{\alpha_3}, \quad (2.174)$$

where the last line defines two new parameters: a coefficient C_3 and an exponent α_3 . The resulting parameters and uncertainties are:

$$\alpha_3 \equiv -1/\alpha_0, \quad (2.175)$$

$$= -0.48 \pm 0.01, \quad (2.176)$$

$$C_3 \equiv C_0^{1/\alpha_0} \ln 2, \quad (2.177)$$

$$= 1.9 \pm 0.1, \quad (2.178)$$

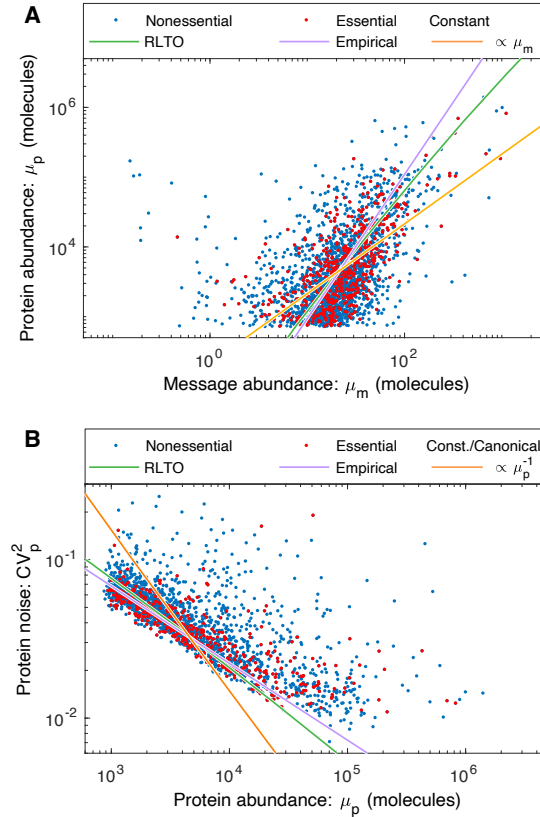


Figure 2.18: **Load balancing predicts the scaling of noise. Panel A: Three competing models for protein abundance in yeast.** The empirical model (purple) fits the slope and the y offset. The RLTO (green) and constant-translation-efficiency (orange) models fit a parameter corresponding to the y offset only. As discussed in the analysis of the proteome fraction, the RLTO model qualitatively captures the scaling of the protein abundance with message number better than the constant translation efficiency model; however, the predicted fit does not correspond to the optimal power law, which is represented by the empirical model. The protein abundance has a cutoff near 10^1 due to autofluorescence [9]. **Panel B: Predictions of the noise-protein abundance relation.** Using each competing protein abundance model, the noise-protein abundance relation can be predicted using Eq. 2.132. The canonical noise model (Eq. 2.130) fails to capture even the scaling of the noise. In contrast, both the RLTO and empirical models quantitatively predict both the scaling and magnitude of the noise. The empirical model has the highest performance, presumably due to its two-parameter fit to the protein abundance in Panel A. A fit accounting for the noise floor is shown in Fig. 2.16.

where the uncertainties are the estimated standard errors.

2.7.5 Results: Parameter-free prediction of noise-abundance in yeast

The fit of the competing protein-message abundance models are shown in Fig. 2.5A. Using each model, we can now predict the relation between protein abundance and noise without additional fitting parameters. The predictions of the three competing models are compared to the experimental data in Fig. 2.5B.

In both its ability to capture the protein abundance and predict the noise, the RLTO model vastly outperforms the constant-translation-efficiency model. The purely empirical model that best captures the protein abundance data, due to directly fitting both the y-offset and slope, also performs best in predicting the noise. It is important to emphasize that the prediction of the noise in all models is non-trivial since there are no free parameters fit, once the protein abundance relation is determined. We therefore conclude that the noise model (Eq. 2.132) quantitatively predicts the observed noise from the message number and that eukaryotic noise has non-canonical scaling due to load balancing.

2.7.6 Discussion: Implications of noise

What are the biological implications of gene expression noise? Many important proposals have been made, including bet-hedging strategies, the necessity of feedback in gene regulatory networks, *etc.* [8]. Our model suggests that robustness to noise fundamentally shapes the central dogma regulatory program. With respect to message number, the one-message-rule sets a lower bound on the transcription rate of essential genes. (See Fig. 2.6B.) With respect to protein expression, robustness to noise has two important implications: Protein overabundance significantly increases protein levels above what would be required in the absence of noise and therefore reshapes the metabolic budget. (See Fig. 2.6A.) Robustness to noise also gives rise to load balancing, the proportionality of the optimal transcription and translation rates. (See Fig. 2.6C.) Not only does robustness to noise affect central dogma

regulation, but there is an important reciprocal effect: Load balancing changes the global scaling relation between noise and protein abundance. (See Fig. 2.18B.)

2.8 Data Tables

2.8.1 Datasets for one-message-rule analysis.

Data S1: Organism: *S. cerevisiae* (yeast). Original source: [71], [57], [62] & [9]. Essential gene classification: [20]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. 2.6.

Data S2: Organism: *Homo sapiens* (human). Original sources: mRNA abundances: Human Protein Atlas [63]. Essential gene classification: [54]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. 2.6.

Data S3: Organism: *E. coli* grown in rich media (LB). Original sources: mRNA abundance: [47]. Essential gene classification: [56]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. 2.6.

Data S4: Organism: *E. coli* grown in minimal media (MM), Original sources: mRNA abundance: [47]. Essential gene classification: [56]. Processing: We merged these datasets. We added message number (messages per cell-cycle) as described in Sec. 2.6.

2.8.2 Datasets for load-balancing analysis.

Data S5: Load-balancing data for yeast cells. Organism: *S. cerevisiae* (yeast). Original sources: Protein abundance [22]. mRNA abundance: [71]. Processing: We merged these datasets. We added protein number and message number (messages per cell-cycle) as described in Sec. 2.6.

Data S6: Load-balancing data for mammalian cells. Organism: *Mus musculus* (mammalian). Original source: [24]. Processing: We added protein number and message number (messages per cell-cycle) as described in Sec. 2.6.

Data S7: Load-balancing data for *E. coli*. Organism: *E. coli*. Original source: [23]. Processing: We added protein number and message number (messages per cell-cycle) as described in Sec. 2.6.

2.8.3 Datasets for noise analysis.

Data S8: Noise data for yeast. Organism: *S. cerevisiae* (yeast). Original sources: Noise measurements: [9]. Essential gene classifications: [20]. Protein fluorescence data, rescaled by mass-spec data as described in Sec. a.: [9]. Processing: We merged these datasets. We added protein number as described in Sec. 2.6.

2.8.4 Datasets for one-message-rule exceptions.

Data S9: Exceptions to the one-message-rule for yeast cells. Organism: *S. cerevisiae* (yeast). Original sources: Message abundances: [71]. Essential gene classifications: [20].

Data S10: Exceptions to the one-message-rule for human cells. Organism: *H. sapiens* (human). Original sources: Message abundances: [63]. Essential gene classifications: [54].

Data S11: Exceptions to the one-message-rule for *E. coli* grown in rich media (LB). Organism: *E. coli*. Original sources: Message abundances: [47]. Essential gene classifications: [56].

2.9 Bibliography

- [1] T. W. Lo, H. J. Choi, D. Huang, and P. A. Wiggins, “Noise robustness and metabolic load determine the principles of central dogma regulation,” *Science Advances*, vol. 10, no. 34, eado3095, 2024. DOI: [10.1126/sciadv.ado3095](https://doi.org/10.1126/sciadv.ado3095).
- [2] E. Dekel and U. Alon, “Optimality and evolutionary tuning of the expression level of a protein,” *Nature*, vol. 436, no. 7050, pp. 588–92, Jul. 2005. DOI: [10.1038/nature03842](https://doi.org/10.1038/nature03842).
- [3] L. Keren *et al.*, “Massively parallel interrogation of the effects of gene expression levels on fitness,” *Cell*, vol. 166, no. 5, 1282–1294.e18, Aug. 2016. DOI: [10.1016/j.cell.2016.07.024](https://doi.org/10.1016/j.cell.2016.07.024).

- [4] J. M. Peters *et al.*, “A comprehensive, CRISPR-based functional analysis of essential genes in bacteria,” *Cell*, vol. 165, no. 6, pp. 1493–1506, Jun. 2016. DOI: [10.1016/j.cell.2016.05.003](https://doi.org/10.1016/j.cell.2016.05.003).
- [5] L. A. Gallagher, J. Bailey, and C. Manoil, “Ranking essential bacterial processes by speed of mutant death,” *Proc Natl Acad Sci U S A*, vol. 117, no. 30, pp. 18010–18017, Jul. 2020. DOI: [10.1073/pnas.2001507117](https://doi.org/10.1073/pnas.2001507117).
- [6] H. G. S. Joseph W. Lengeler Gerhart Drews, Ed., *Biology of the Prokaryotes*. Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998.
- [7] N. M. Belliveau *et al.*, “Fundamental limits on the rate of bacterial growth and their influence on proteomic composition,” *Cell Syst*, vol. 12, no. 9, 924–944.e2, Sep. 2021. DOI: [10.1016/j.cels.2021.06.002](https://doi.org/10.1016/j.cels.2021.06.002).
- [8] J. M. Raser and E. K. O’Shea, “Noise in gene expression: Origins, consequences, and control,” *Science*, vol. 309, no. 5743, pp. 2010–3, Sep. 2005. DOI: [10.1126/science.1105891](https://doi.org/10.1126/science.1105891).
- [9] J. R. S. Newman *et al.*, “Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise,” *Nature*, vol. 441, no. 7095, pp. 840–6, Jun. 2006. DOI: [10.1038/nature04785](https://doi.org/10.1038/nature04785).
- [10] Y. Taniguchi *et al.*, “Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells,” *Science*, vol. 329, no. 5991, pp. 533–8, Jul. 2010. DOI: [10.1126/science.1188308](https://doi.org/10.1126/science.1188308).
- [11] J. Hausser, A. Mayo, L. Keren, and U. Alon, “Central dogma rates and the trade-off between precision and economy in gene expression,” *Nat Commun*, vol. 10, no. 1, p. 68, Jan. 2019. DOI: [10.1038/s41467-018-07391-8](https://doi.org/10.1038/s41467-018-07391-8).
- [12] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–3, Aug. 1970. DOI: [10.1038/227561a0](https://doi.org/10.1038/227561a0).
- [13] J. Paulsson and M. Ehrenberg, “Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks,” *Phys Rev Lett*, vol. 84, no. 23, pp. 5447–50, Jun. 2000. DOI: [10.1103/PhysRevLett.84.5447](https://doi.org/10.1103/PhysRevLett.84.5447).
- [14] N. Friedman, L. Cai, and X. S. Xie, “Linking stochastic dynamics to population distribution: An analytical framework of gene expression,” *Phys Rev Lett*, vol. 97, no. 16, p. 168302, Oct. 2006. DOI: [10.1103/PhysRevLett.97.168302](https://doi.org/10.1103/PhysRevLett.97.168302).
- [15] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, “Interdependence of cell growth and gene expression: Origins and consequences,” *Science*, vol. 330, no. 6007, pp. 1099–102, Nov. 2010. DOI: [10.1126/science.1192588](https://doi.org/10.1126/science.1192588).
- [16] J. I. Steinfeld, J. S. Francisco, and W. L. Hase, *Chemical Kinetics and Dynamics*, 2nd. Prentice-Hall, 1999.

- [17] D. A. Charlebois, N. Abdennur, and M. Kaern, “Gene expression noise facilitates adaptation and drug resistance independently of mutation,” en, *Phys. Rev. Lett.*, vol. 107, no. 21, p. 218 101, Nov. 2011.
- [18] M. R. Silvis *et al.*, “Morphological and transcriptional responses to CRISPRi knockdown of essential genes in *Escherichia coli*,” *mBio*, vol. 12, no. 5, e0256121, Oct. 2021. DOI: [10.1128/mBio.02561-21](https://doi.org/10.1128/mBio.02561-21).
- [19] S. Donati *et al.*, “Multi-omics analysis of CRISPRi-knockdowns identifies mechanisms that buffer decreases of enzymes in *E. coli* metabolism,” *Cell Syst*, vol. 12, no. 1, 56–67.e6, Jan. 2021. DOI: [10.1016/j.cels.2020.10.011](https://doi.org/10.1016/j.cels.2020.10.011).
- [20] J. van Leeuwen *et al.*, “Systematic analysis of bypass suppression of essential genes,” *Mol Syst Biol*, vol. 16, no. 9, e9828, Sep. 2020. DOI: [10.15252/msb.20209828](https://doi.org/10.15252/msb.20209828).
- [21] T. Wang *et al.*, “Identification and characterization of essential genes in the human genome,” *Science*, vol. 350, no. 6264, pp. 1096–101, Nov. 2015. DOI: [10.1126/science.aac7041](https://doi.org/10.1126/science.aac7041).
- [22] S. Ghaemmaghani *et al.*, “Global analysis of protein expression in yeast,” *Nature*, vol. 425, no. 6959, pp. 737–41, Oct. 2003. DOI: [10.1038/nature02046](https://doi.org/10.1038/nature02046).
- [23] R. Balakrishnan *et al.*, “Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria,” *Science*, vol. 378, no. 6624, eabk2066, Dec. 2022. DOI: [10.1126/science.abk2066](https://doi.org/10.1126/science.abk2066).
- [24] B. Schwanhäusser *et al.*, “Global quantification of mammalian gene expression control,” *Nature*, vol. 473, no. 7347, pp. 337–42, May 2011. DOI: [10.1038/nature10098](https://doi.org/10.1038/nature10098).
- [25] H. Dourado, M. Mori, T. Hwa, and M. J. Lercher, “On the optimality of the enzyme-substrate relationship in bacteria,” en, *PLoS Biol.*, vol. 19, no. 10, e3001416, Oct. 2021.
- [26] M. Mori, S. Schink, D. W. Erickson, U. Gerland, and T. Hwa, “Quantifying the benefit of a proteome reserve in fluctuating environments,” *Nat Commun*, vol. 8, no. 1, p. 1225, Oct. 2017. DOI: [10.1038/s41467-017-01242-8](https://doi.org/10.1038/s41467-017-01242-8).
- [27] G. Lambert and E. Kussell, “Memory and fitness optimization of bacteria under fluctuating environments,” *PLoS Genet*, vol. 10, no. 9, e1004556, Sep. 2014. DOI: [10.1371/journal.pgen.1004556](https://doi.org/10.1371/journal.pgen.1004556).
- [28] J. Monod, A. M. Pappenheimer Jr, and G. Cohen-Bazire, “The kinetics of the biosynthesis of beta-galactosidase in *Escherichia coli* as a function of growth,” *Biochim Biophys Acta*, vol. 9, no. 6, pp. 648–60, Dec. 1952. DOI: [10.1016/0006-3002\(52\)90227-8](https://doi.org/10.1016/0006-3002(52)90227-8).
- [29] B. Bosch *et al.*, “Genome-wide gene expression tuning reveals diverse vulnerabilities of *M. tuberculosis*,” *Cell*, vol. 184, no. 17, 4579–4592.e24, Aug. 2021. DOI: [10.1016/j.cell.2021.06.033](https://doi.org/10.1016/j.cell.2021.06.033).

- [30] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, “Characterizing stochastic cell-cycle dynamics in exponential growth,” *Phys. Rev. E*, vol. 105, p. 014420, 1 Jan. 2022. DOI: [10.1103/PhysRevE.105.014420](https://doi.org/10.1103/PhysRevE.105.014420).
- [31] J. L. Hargrove and F. H. Schmidt, “The role of mRNA and protein stability in gene expression,” *FASEB J*, vol. 3, no. 12, pp. 2360–70, Oct. 1989. DOI: [10.1096/fasebj.3.12.2676679](https://doi.org/10.1096/fasebj.3.12.2676679).
- [32] A. L. Koch and H. R. Levy, “Protein turnover in growing cultures of *Escherichia coli*,” *J Biol Chem*, vol. 217, no. 2, pp. 947–57, Dec. 1955.
- [33] M. Martin-Perez and J. Villén, “Determinants and regulation of protein turnover in yeast,” *Cell Syst*, vol. 5, no. 3, 283–294.e5, Sep. 2017. DOI: [10.1016/j.cels.2017.08.008](https://doi.org/10.1016/j.cels.2017.08.008).
- [34] D. Gillespie, “A rigorous derivation of the chemical master equation,” *Physica A*, vol. 188, no. 1, pp. 404–425, 1992.
- [35] D. T. Gillespie, “Exact stochastic simulation of coupled chemical reactions,” *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977. DOI: [10.1021/j100540a008](https://doi.org/10.1021/j100540a008). eprint: <https://doi.org/10.1021/j100540a008>.
- [36] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen, “Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays,” *Proc Natl Acad Sci U S A*, vol. 99, no. 15, pp. 9697–702, Jul. 2002. DOI: [10.1073/pnas.112318199](https://doi.org/10.1073/pnas.112318199).
- [37] H. Chen, K. Shiroguchi, H. Ge, and X. S. Xie, “Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*,” *Mol Syst Biol*, vol. 11, no. 5, p. 808, May 2015. DOI: [10.15252/msb.20159000](https://doi.org/10.15252/msb.20159000).
- [38] E. W. Weisstein, *CRC Encyclopedia of Mathematics*. Chapman & Hall/CRC, 2009.
- [39] R. Phillips, J. Kondev, J. Theriot, and N. Orme, *Physical Biology of the Cell*. Garland Science, 2013, ISBN: 9780815344506.
- [40] M. Kafri, E. Metzl-Raz, F. Jonas, and N. Barkai, “Rethinking cell growth models,” *FEMS Yeast Res*, vol. 16, no. 7, Nov. 2016. DOI: [10.1093/femsyr/fow081](https://doi.org/10.1093/femsyr/fow081).
- [41] D. Huang, T. Lo, H. Merrikh, and **Paul A. Wiggins**, “Quantitative in vivo measurement of fork velocity in bacterial cells by deep sequencing,” *Nature Communications (In press.)*, 2023.
- [42] E. Levien, J. Min, J. Kondev, and A. Amir, “Non-genetic variability in microbial populations: Survival strategy or nuisance?” *Reports on Progress in Physics*, vol. 84, no. 11, p. 116601, Nov. 2021. DOI: [10.1088/1361-6633/ac2c92](https://doi.org/10.1088/1361-6633/ac2c92).
- [43] E. Powell, “Growth rate and generation time of bacteria, with special reference to continuous culture,” *Microbiology*, vol. 15, no. 492, 1956.

- [44] M. Kafri, E. Metzl-Raz, G. Jona, and N. Barkai, “The cost of protein production,” *Cell Rep*, vol. 14, no. 1, pp. 22–31, Jan. 2016. DOI: [10.1016/j.celrep.2015.12.015](https://doi.org/10.1016/j.celrep.2015.12.015).
- [45] E. Metzl-Raz, M. Kafri, G. Yaakov, and N. Barkai, “Gene transcription as a limiting factor in protein production and cell growth,” *G3 (Bethesda)*, vol. 10, no. 9, pp. 3229–3242, Sep. 2020. DOI: [10.1534/g3.120.401303](https://doi.org/10.1534/g3.120.401303).
- [46] G. Arfken, *Mathematical Methods for Physicists*, Third. San Diego: Academic Press, Inc., 1985.
- [47] A. Bartholomäus *et al.*, “Bacteria differently regulate mRNA abundance to specifically respond to various stresses,” *Philos Trans A Math Phys Eng Sci*, vol. 374, no. 2063, Mar. 2016. DOI: [10.1098/rsta.2015.0069](https://doi.org/10.1098/rsta.2015.0069).
- [48] R. Milo, “What is the total number of protein molecules per cell volume? A call to rethink some published values,” *Bioessays*, vol. 35, no. 12, pp. 1050–5, Dec. 2013. DOI: [10.1002/bies.201300066](https://doi.org/10.1002/bies.201300066).
- [49] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular Biology of the Cell*, 4th. Garland, 2002.
- [50] L. L. Chia and C. McLaughlin, “The half-life of mRNA in *Saccharomyces cerevisiae*,” *Mol Gen Genet*, vol. 170, no. 2, pp. 137–44, Feb. 1979. DOI: [10.1007/BF00337788](https://doi.org/10.1007/BF00337788).
- [51] V. Pelechano, S. Chávez, and J. E. Pérez-Ortín, “A complete set of nascent transcription rates for yeast genes,” *PLoS One*, vol. 5, no. 11, e15442, Nov. 2010. DOI: [10.1371/journal.pone.0015442](https://doi.org/10.1371/journal.pone.0015442).
- [52] B. Futcher, G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels, “A sampling of the yeast proteome,” *Mol Cell Biol*, vol. 19, no. 11, pp. 7357–68, Nov. 1999. DOI: [10.1128/MCB.19.11.7357](https://doi.org/10.1128/MCB.19.11.7357).
- [53] G. M. Cooper, *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates 2000, 2000, ISBN: 0-87893-106-6.
- [54] E. Yang *et al.*, “Decay rates of human mRNAs: Correlation with functional characteristics and sequence attributes,” *Genome Res*, vol. 13, no. 8, pp. 1863–72, Aug. 2003. DOI: [10.1101/gr.1272403](https://doi.org/10.1101/gr.1272403).
- [55] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, “Bionumbers—the database of key numbers in molecular and cell biology,” *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D750–3, Jan. 2010. DOI: [10.1093/nar/gkp889](https://doi.org/10.1093/nar/gkp889).
- [56] T. Baba, H.-C. Huan, K. Datsenko, B. L. Wanner, and H. Mori, “The applications of systematic in-frame, single-gene knockout mutant collection of *Escherichia coli* K-12,” *Methods Mol Biol*, vol. 416, pp. 183–94, 2008. DOI: [10.1007/978-1-59745-321-9_12](https://doi.org/10.1007/978-1-59745-321-9_12).
- [57] W. R. Blevins *et al.*, “Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker’s yeast,” *Sci Rep*, vol. 9, no. 1, p. 11005, Jul. 2019. DOI: [10.1038/s41598-019-47424-w](https://doi.org/10.1038/s41598-019-47424-w).

- [58] L. M. Hereford and M. Rosbash, “Number and distribution of polyadenylated RNA sequences in yeast,” *Cell*, vol. 10, no. 3, pp. 453–62, Mar. 1977. DOI: [10.1016/0092-8674\(77\)90032-0](https://doi.org/10.1016/0092-8674(77)90032-0).
- [59] T. von der Haar, “A quantitative estimation of the global translational activity in logarithmically growing yeast cells,” *BMC Syst Biol*, vol. 2, p. 87, Oct. 2008. DOI: [10.1186/1752-0509-2-87](https://doi.org/10.1186/1752-0509-2-87).
- [60] D. Zenklusen, D. R. Larson, and R. H. Singer, “Single-RNA counting reveals alternative modes of gene expression in yeast,” *Nat Struct Mol Biol*, vol. 15, no. 12, pp. 1263–71, Dec. 2008. DOI: [10.1038/nsmb.1514](https://doi.org/10.1038/nsmb.1514).
- [61] F. Miura *et al.*, “Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs,” *BMC Genomics*, vol. 9, p. 574, Nov. 2008. DOI: [10.1186/1471-2164-9-574](https://doi.org/10.1186/1471-2164-9-574).
- [62] L. M. F. de Godoy *et al.*, “Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast,” *Nature*, vol. 455, no. 7217, pp. 1251–4, Oct. 2008. DOI: [10.1038/nature07341](https://doi.org/10.1038/nature07341).
- [63] M. Uhlen *et al.*, “Proteomics. Tissue-based map of the human proteome,” *Science*, vol. 347, no. 6220, p. 1260419, Jan. 2015. DOI: [10.1126/science.1260419](https://doi.org/10.1126/science.1260419).
- [64] V. E. Velculescu *et al.*, “Analysis of human transcriptomes,” *Nat Genet*, vol. 23, no. 4, pp. 387–8, Dec. 1999. DOI: [10.1038/70487](https://doi.org/10.1038/70487).
- [65] S. Y. Gerdes *et al.*, “Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655,” *J Bacteriol*, vol. 185, no. 19, pp. 5673–84, Oct. 2003. DOI: [10.1128/JB.185.19.5673-5684.2003](https://doi.org/10.1128/JB.185.19.5673-5684.2003).
- [66] E. C. A. Goodall *et al.*, “The essential genome of *Escherichia coli* K-12,” *mBio*, vol. 9, no. 1, Feb. 2018. DOI: [10.1128/mBio.02096-17](https://doi.org/10.1128/mBio.02096-17).
- [67] P. Mehta, S. Casjens, and S. Krishnaswamy, “Analysis of the lambdoid prophage element e14 in the *E. coli* K-12 genome,” *BMC Microbiol*, vol. 4, p. 4, Jan. 2004. DOI: [10.1186/1471-2180-4-4](https://doi.org/10.1186/1471-2180-4-4).
- [68] A. Bar-Even *et al.*, “Noise in protein expression scales with natural protein abundance,” *Nat Genet*, vol. 38, no. 6, pp. 636–43, Jun. 2006. DOI: [10.1038/ng1807](https://doi.org/10.1038/ng1807).
- [69] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, “Stochastic gene expression in a single cell,” *Science*, vol. 297, no. 5584, pp. 1183–6, Aug. 2002. DOI: [10.1126/science.1070919](https://doi.org/10.1126/science.1070919).
- [70] P. S. Swain, M. B. Elowitz, and E. D. Siggia, “Intrinsic and extrinsic contributions to stochasticity in gene expression,” *Proc Natl Acad Sci U S A*, vol. 99, no. 20, pp. 12795–800, Oct. 2002. DOI: [10.1073/pnas.162041399](https://doi.org/10.1073/pnas.162041399).
- [71] P.-J. Lahtvee *et al.*, “Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast,” *Cell Syst*, vol. 4, no. 5, pp. 495–504.e5, May 2017. DOI: [10.1016/j.cels.2017.03.003](https://doi.org/10.1016/j.cels.2017.03.003).

- [72] K. H. Hellton and M. Thoresen, “The impact of measurement error on principal component analysis,” *Scandinavian Journal of Statistics*, vol. 41, no. 4, pp. 1051–1063, Apr. 2014. DOI: [10.1111/sjos.12083](https://doi.org/10.1111/sjos.12083).

Chapter 3

Protein overabundance is driven by growth robustness on a threshold-like fitness landscape

Originally published as: [1] H. J. Choi, T. W. Lo, K. J. Cutler, D. Huang, W. R. Will, and P. A. Wiggins, “Protein overabundance is driven by growth robustness,” *bioRxiv*, 2024. DOI: [10.1101/2024.08.14.607847](https://doi.org/10.1101/2024.08.14.607847). eprint: <https://www.biorxiv.org/content/early/2024/08/17/2024.08.14.607847.full.pdf>.

Author contributions: H.J.C., T.W.L., K.J.C., D.H., and P.A.W. conceived the research. H.J.C., W.R.W., and P.A.W. performed the experiments. H.J.C., T.W.L., K.J.C., and P.A.W. performed the analysis. H.J.C., T.W.L., D.H., and P.A.W. wrote the paper.

Abstract

Protein expression levels optimize cell fitness: Too low an expression level of essential proteins will slow growth by compromising essential processes; whereas overexpression slows growth by increasing the metabolic load. This trade-off naïvely predicts that cells maximize their fitness by sufficiency, expressing just enough of each essential protein for function. We test this prediction in the naturally-competent bacterium *Acinetobacter baylyi* by characterizing the proliferation dynamics of essential-gene knockouts at a single-cell scale (by imaging) as well as at a genome-wide scale. In these experiments, cells proliferate for multiple generations as target protein levels are diluted from their endogenous levels. This approach facilitates a proteome-scale analysis of the fitness landscape with respect to protein abundance. We find that most essential proteins are subject to a threshold-like fitness landscape: growth is independent of protein abundance above a critical threshold and

arrests below that threshold. We have recently analyzed the implications of this landscape for growth robustness. Confirming signature predictions of this model, we find that (i) roughly 70% of essential proteins are overabundant, (ii) overabundance increases as the expression level decreases and (iii) the lowest abundance proteins are in vast excess ($>10\times$) of what is required for growth in the typical cell. These results reveal that robustness plays a fundamental role in determining the expression levels of essential genes and that overabundance is a key mechanism for ensuring robust growth.

3.1 Introduction

Understanding the rationale for protein expression levels is a fundamental question in biology with broad implications for understanding cellular function [2]. Measured expression levels appear to be paradoxically both *optimal* and *overabundant*. For instance, repeated investigations support the idea that gene expression levels optimize cell fitness [3, 4]. Since the overall metabolic cost of protein expression is large [5, 6], fitness optimization would seem to imply that protein levels should satisfy a *Goldilocks condition*: Expression levels should be *just high enough* to achieve the required protein activity [7, 8]. However, a range of approaches suggest that many essential genes are expressed in vast excess of the levels required for function [8–10]. How can expression levels be at once optimal with respect to fitness as well as in excess of what is required for function?

The cell faces a complex regulatory challenge: Even in a bacterium, there are between five and six hundred essential proteins, each of which is required for growth [11]. How does the cell ensure the robust expression of each essential factor? We recently argued that the stochasticity of gene expression processes fundamentally shape the principles of central dogma regulation, including the optimality of protein overabundance [12]. Specifically, we proposed a quantitative model, the Robustness-Load Trade-Off (RLTO) model, which makes a parameter-free prediction of protein overabundance as a function of gene transcription level [12]. The optimality of overabundance can be understood as

the result of a highly-asymmetric fitness landscape: the fitness cost of essential protein underabundance, which causes the arrest of essential processes, is far greater than the fitness cost of essential protein overabundance, which leads to slow growth by increasing the metabolic load. However, critical model assumptions and predictions remain untested which is the motivation for the current study. Here, we will quantitatively measure the fitness landscape with respect to protein abundance and determine the level of overabundance for all essential proteins in the bacterium *Acinetobacter baylyi*.

3.2 Results

3.2.1 Natural competence facilitates knockout-depletion

To characterize the fitness landscape for essential gene expression, we must deplete the levels of essential proteins. Both degenon- and CRISPRi-based approaches have been applied; however, these approaches require careful characterization of protein levels [9, 13–16] and introduce significant cell-to-cell variation on top of the endogenous noise which further obscures the underlying fitness landscape [17]. To circumvent these difficulties, we will use an alternative approach: *knockout-depletion* in the naturally competent bacterium *A. baylyi* ADP1 [18, 19]. In this approach, cells are transformed with a *geneX::kan* knockout cassettes targeting essential gene X, carrying a kanamycin resistance allele Km^R . (See Fig. 3.1A.) Cells that are not transformed arrest immediately on selective media. The crux of the approach is that transformants remain transiently *geneX*⁺, due to the presence of already synthesized target protein X, even after the transcription of the target *geneX* stops. Growth can continue, diluting protein X abundance, as long as this residual abundance remain sufficient for function. The success of the knockout-depletion approach is dependent on the extremely high transformation efficiency of *A. baylyi*.

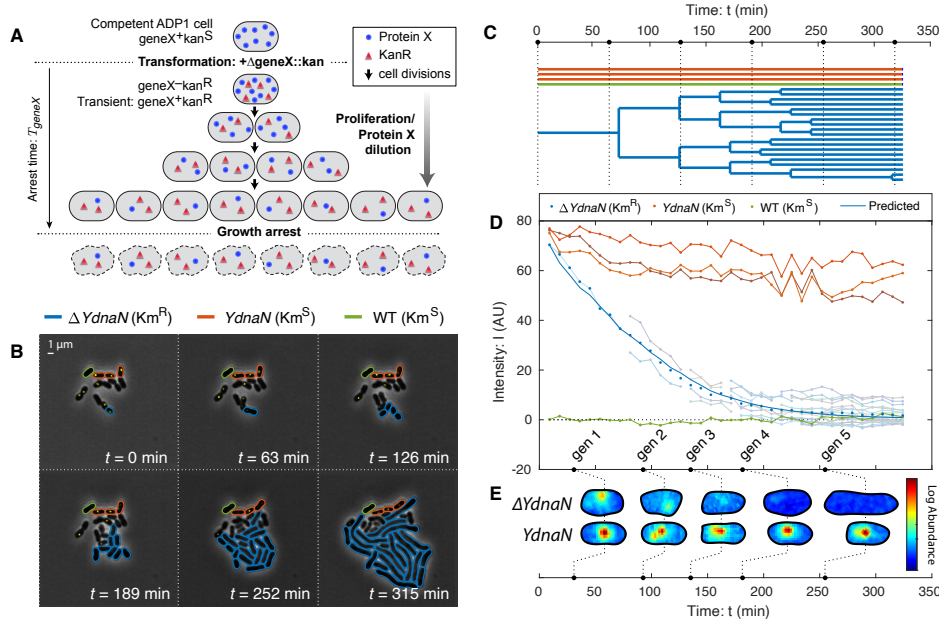


Figure 3.1: Knockout-depletion experiments. Panel A: Experimental schematic. Competent ADP1 cells are transformed with Δ geneX::kan. Untransformed cells arrest immediately on selective media. Transformed cells proliferate, but cease protein X expression (blue circles) while expressing Kan (red triangles). Existing protein X abundance is diluted as cells proliferate. For essential genes, cell growth continues until protein levels are diluted to the threshold level required for growth, after which growth arrests. **Panel B & C: Visualization of knockout depletion.** The fluorescent fusion YPet-dnaN to essential gene dnaN is knocked out at $t = 0$. Cell proliferation is visualized using phase-contrast microscopy while protein abundance is measured by fluorescence microscopy (yellow). Transformed cells (Δ YdnaN, blue) have a Km^R allele and can proliferate over several generations before arrest; however, untransformed cells (YdnaN, orange) and wild-type cells (WT, green) were both kanamycin sensitive and therefore arrested immediately. **Panel C: Lineage tree.** Black dotted lines represent time points shown in Panel B. **Panel D: Target protein is diluted by proliferation.** Protein concentration is measured by integrated fluorescence. Arrested YdnaN cells maintain protein abundance, whereas proliferating transformed cells (Δ YdnaN, blue) show growth-induced protein depletion. The protein concentration over all transformed progeny (blue points) are consistent with the dilution-model prediction (solid blue). **Panel E: Protein function is robust to dilution.** Representative single-cell images of transformed (Δ YdnaN) and untransformed (YdnaN) cells are shown for successive time points. The YPet-DnaN fusion shows punctate localization, consistent with function, even as protein abundance is depleted. No puncta are observed in the last generation and the cells form filaments, consistent with the replication arrest phenotype.

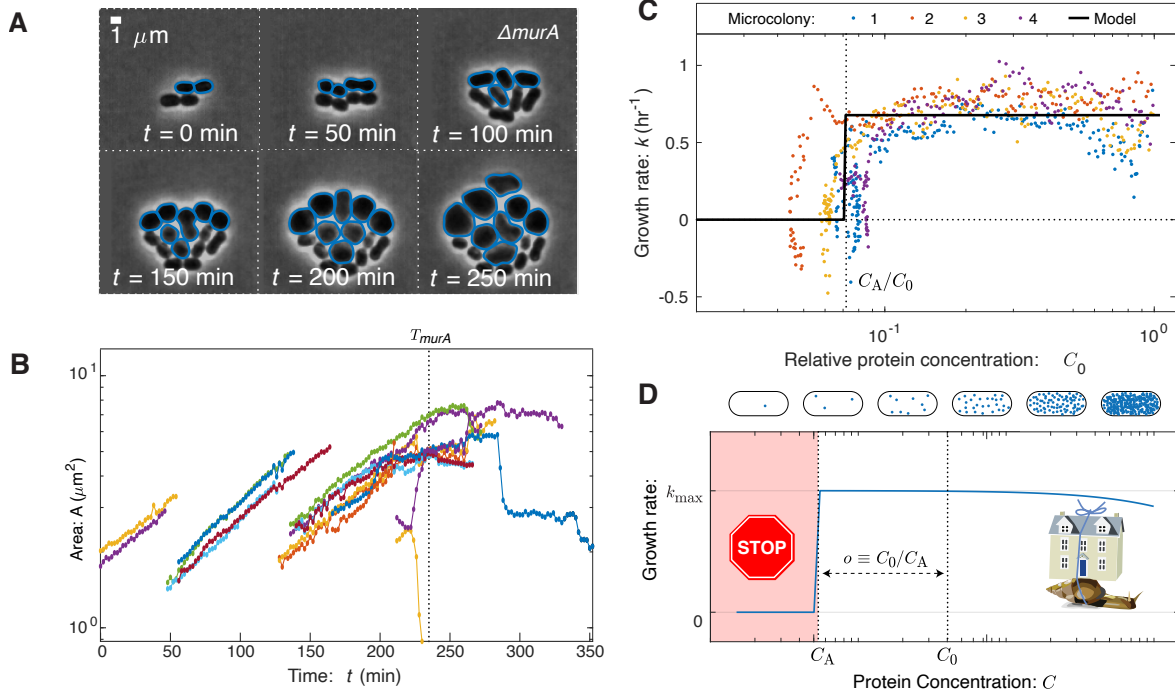


Figure 3.2: The fitness landscape. Panel A: Visualization of growth in a *murA* knockout. Essential gene *murA* is knocked out at $t = 0$ and cell proliferation is visualized by phase-contrast microscopy. Blue outlines represent the Omnipose cell segmentation. Cell proliferation continues for multiple generations after deletion. Furthermore, we observe transient growth of WT cells (without outline). These represent transient growth of WT resulting from the transformation of only one chromosome before division. Following cell division, the sister cell that lacks the *kan*-deletion cassettes can temporarily proliferate due to the residual KanR protein inherited from the parent.

Panel B: Quantitative analysis of cell proliferation with single-cell resolution. Cell area (log scale) as a function of time for the *murA* deletion (colored lines). Each different line represents a single cell. The log-slope represents the single-cell growth rate. The vertical dotted line represents the arrest time at which transformant growth slows to growth arrest, a value we quantify for estimating overabundance. Untransformed cells are arrested throughout the experiment since the cells are Km^S .

Panel C: Growth rate as a function of protein depletion for ΔmurA . The growth rate (Eq. 3.2) is observed to obey a threshold-like-dependence on protein abundance (Eq. 3.1), transitioning between wild-type growth to arrest at the vertical dotted line. We define the critical dilution as $o \equiv C_0/C_A$ where C_A is the protein concentration at arrest. Cell-to-cell variation is observed in the arrest of four microcolonies that originate from distinct progenitor cells. The model represents a threshold-like model fit to the growth rate data. **Panel D: The fitness landscape is threshold-like.** Motivated by single-cell growth data, cell fitness is modeled using the Robustness-Load Trade-Off model (RLTO)[12].

3.2.2 Target proteins are depleted by dilution

A key untested assumption in the experimental design of the knockout-depletion approach is that target protein translation stops after transformation, and that the protein abundance is depleted by dilution. The model predicts that the protein concentration is:

$$C(t) = C_0 \cdot V_0/V(t), \quad (3.1)$$

where C_0 and V_0 are the concentration and volume of the progenitor cell at deletion and $V(t)$ is the total volume of the progeny. To test the predicted protein depletion hypothesis, we designed a knockout-depletion experiment to target a protein we had previously studied that can be visualized using a fluorescent fusion and whose localization is activity dependent: the essential replication gene *dnaN*, whose gene product is the β sliding clamp [20–22]. We constructed a N-terminal fluorescent fusion to *dnaN* using YPet in *A. baylyi* at the endogenous locus. The resulting mutant (YdnaN) had no measurable growth defect under our experimental conditions. We then knocked out the *YPet-dnaN* fusion, yielding $\Delta dnaN$, and characterized the protein levels by quantifying YPet-DnaN abundance by fluorescence. The experimentally measured fluorescence intensity is consistent with the dilution model (Eq. 3.1), as expected. (See Fig. 3.1D.) We therefore conclude that knockout-depletion experiments are consistent with the experimental design shown schematically in Fig. 3.1A.

3.2.3 Replication persists during DnaN depletion

A key subhypothesis of the overabundance model for transient growth is that target protein function continues as the target protein abundance is depleted. An alternative hypothesis for transient growth of the $\Delta dnaN$ strain is a high initial chromosomal copy-number that is partitioned between daughter cells, even after the replication process itself arrests due to target protein depletion [5, 23]. The imaging-based knockout-depletion experiment tests this hypothesis as well. The localization of DnaN is dependent on activity: During ongoing replication, DnaN is localized in puncta corresponding to replisomes, whereas in the absence of active replication, DnaN has diffuse localization [20–22, 24, 25]. During the

			Log Overabundance:			
Gene	Annotated gene function:	Message number: μ_m (mRNA molecules/cell cycle)	TFNseq	Imaging-based		(N_C, N_P)
			Replication $\log_{10} o$	Elongation $\log_{10} o$	Septation $\log_{10} o$	
<i>dnaA</i>	Regulation of replication initiation	30	0.02± 0.02	0.7 ± 0.1	0.0± 0.2	(4,4)
<i>dnaN</i>	Replication beta sliding clamp	49	1.5 ± 0.1	2.0 ± 3.0	1.4± 0.1	(134,8)
<i>ftsN</i>	Essential cell division/septation protein	20	2.6 ± 0.1	1.8 ± 0.2	0.6± 0.2	(19,5)
<i>murA</i>	Cell wall precursor synthesis	26	0.7 ± 0.5	1.1 ± 0.1	0.9± 0.2	(16,4)

Table 3.1: **Measured overabundance for sequencing- versus imaging-based approaches.** The overabundance was determined by both sequencing- and imaging-based approaches. For the imaging-based approach, we show two measurements based on different metrics for arrest: The first is based on the arrest of cell elongation, as defined by Eq. 3.3, and the second is based on the arrest of the septation process, as visualized by microscopy.

knockout-depletion experiment, we observed YPet-DnaN puncta persist as the targeted fusion was depleted (Fig. 3.1DE), consistent with replication activity after dilution. Only after the YPet-DnaN puncta disappear do the cells begin to adopt the $\Delta dnaN$ phenotype: cell filamentation (Fig. 3.1BE). We therefore conclude that function (replication) is robust to significant target protein (DnaN) dilution.

3.2.4 Many essential knockouts undergo transient growth

To understand the generic consequences of essential protein depletion, we used the imaging- based knockout-depletion experiments to explore essential genes with a range of functions. We initially targeted four essential genes: the replication initiation regulator gene *dnaA* (Movie S5-6), the beta-clamp gene *dnaN* (Movie S7-8), the cell-wall-synthesis gene *murA* (Movie S9-10), and septation-related gene *ftsN* (Movie S11-12), as well as a non-essential IS element with no phenotype as a negative control ([movie](#)). (Representative frame mosaic images and cytometry appear in Supplementary Material Sec. d..) In each case, transformants continued to proliferate through multiple cell-cycle durations [18] and are therefore consistent with the essential protein overabundance hypothesis. However, in Ref. [18], we were unable to perform a quantitative single-cell analysis of

these time-lapse experiments since existing segmentation packages failed to segment the observed morphologies [26]. We therefore developed the *Omnipose* package, which facilitated quantitative analysis of the growth dynamics with single-cell resolution [26]. (See Fig. 3.2A.)

3.2.5 The fitness landscape is threshold-like

A key input to the RLTO model is the fitness landscape (growth rate) as a function of protein abundance. Omnipose segmentation facilitates the measurement of single-cell growth rates from the time-lapse imaging experiments. We focus first on the single-cell areal growth rate:

$$k(t) = \frac{d}{dt} \ln A(t), \quad (3.2)$$

where $A(t)$ is the area of the cell at time t . This areal growth rate is more convenient than a cell-length based rate since we avoid the necessity of defining cell length for unusual cell morphologies like those observed in the $\Delta murA$ mutant. Fig. 3.2B shows representative knockout-depletion dynamics of cell area for the essential-gene target *murA*. The log slope remains constant for multiple generations, consistent with a constant growth rate, even as the gene targeted is depleted over multiple cell cycles. By combining the dilution model (Eq. 3.1) and the growth rate (Eq. 3.2), a single knockout-depletion measurement determines the growth rate for a range of protein abundances between wild-type abundance and those realized at growth arrest. This fitness landscape is shown for MurA protein in Fig. 3.2C. For all four mutants, the areal growth rate is roughly constant for multiple generations before undergoing a rapid transition to growth arrest.

3.2.6 Protein overabundance

We will define the overabundance as the ratio of protein concentration in wild-type cells (C_0) to the concentration at cell arrest (C_A):

$$o \equiv C_0/C_A, \quad (3.3)$$

as shown in Fig. 3.2D. (Supplementary Material Sec. 3.5.4 gives a detailed description of the inferred overabundance from single-cell data.) The measured overabundance for the four

mutants imaged by microscopy is summarized in Tab. 3.1, using three distinct metrics for growth: (i) replication derived from TFNseq data, and (ii) cell elongation and (iii) septation obtained through timelapse imaging. We conclude that for each gene, with the exception of *dnaA*, rapid growth continues after the knockout due to the vast overabundance of the target protein.

3.2.7 The RLTO model predicts protein overabundance

The RLTO model explicitly analyzes the trade-off between growth robustness to noise and metabolic load and predicts the optimal central-dogma regulatory principles [12]. Critically, the model incorporates the observed threshold-like dependence of growth rate on protein abundance (Fig. 3.2CD). The model quantitatively predicts protein overabundance with a signature feature: high-expression genes have low protein overabundance ($o \approx 1$) due to the high metabolic cost of increasing expression and low inherent noise of high expression genes; however, low-expression genes have high overabundance ($o \gg 1$) due to the low metabolic cost of increasing expression and the high inherent noise of low expression genes. (See Supplemental Material Sec. 3.5.7 for a more detailed description of the model.)

3.2.8 TFNseq determines overabundances genome-wide

To test the signature expression-dependent overabundance prediction of the RLTO model, we now transition to a genomic-scale analysis. The Manoil lab developed a TFNseq-approach to knockout-depletion experiments for targeting all genes simultaneously in *A. baylyi* [19]. In short: A genomic library was prepared and mutagenized using a transposon carrying the Km^R allele. The resulting DNA was then transformed into *A. baylyi*. The transformants were propagated on selective liquid media and fractions collected every two hours from which genomic DNA was extracted. The transposons were then mapped using Tn-seq to generate the relative abundance trajectory for each mutant [19]. (See Fig. 3.3AB.) We then analyzed each mutant trajectory statistically using three competing growth models: no-effect, sufficiency, and overabundance, using two successive null-hypothesis tests. (See

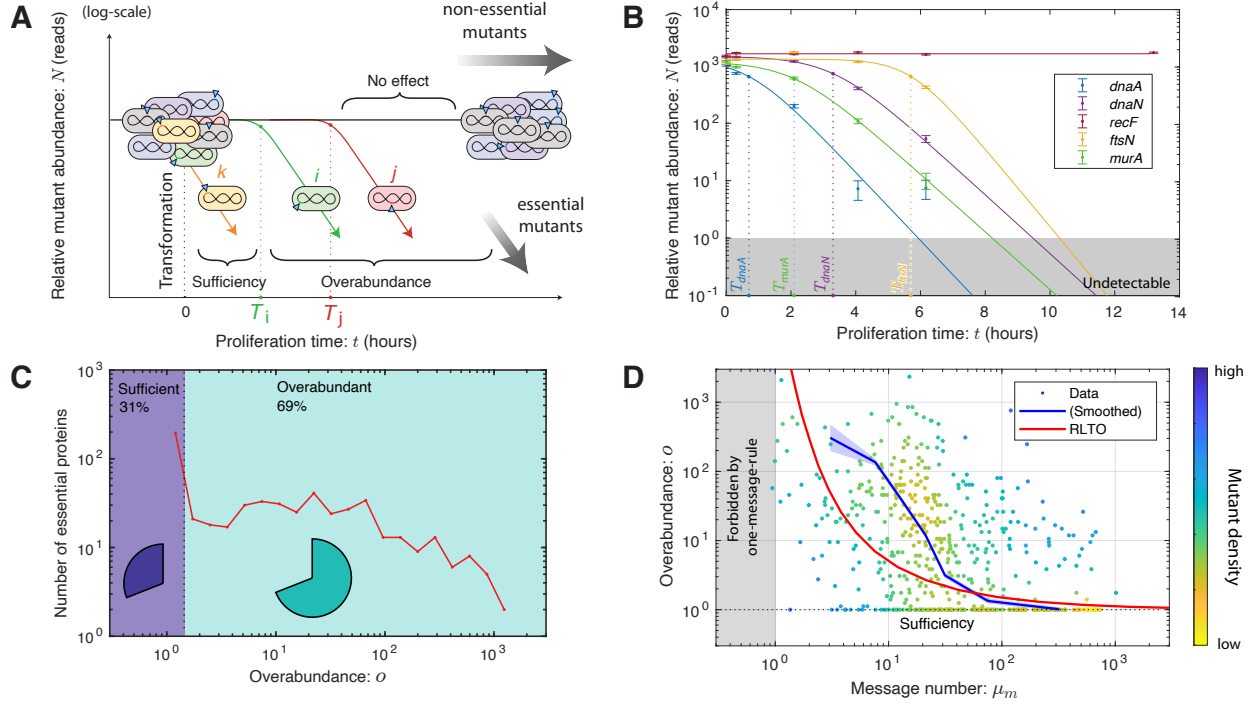


Figure 3.3: **A proteome-wide analysis of protein overabundance. Panel A: TFNseq schematic.** A poly-clonal library of knockout mutants is generated by the transformation of ADP1 with DNA mutagenized by transposon insertions. The library is proliferated on selective media and sequential fractions are collected. The relative-abundance trajectories of mutants are determined by mapping transposon insertion sites by sequencing. **Panel B: TFNseq-trajectory analyses for five mutant strains.** Each mutant trajectory is well fit by one of the three trajectory models. As expected, the no-effect model is selected for the non-essential gene *recF*. For the other four essential genes, the overabundance model is selected. The dotted line represents the arrest time for each mutant. **Panel C: Overabundance varies by orders of magnitude between essential proteins.** The protein overabundance is inferred from the arrest time using Eq. 3.4. Sufficient expression genes have overabundance $o = 1$, while overabundant genes vary from $o > 1$ to very large overabundance ($o > 100$). **Panel D: Overabundance is large for low-expression essential proteins.** The measured message-number-overabundance pairs are shown for essential genes (including estimated gene density.) The smoothed experimental data is shown in blue (with experimental uncertainty.) The RLTO model (red) predicts that overabundance grows rapidly as the transcription level is reduced. The RLTO model qualitatively captures the trend of the data (blue); however, it appears to underestimate the measured overabundance for intermediate expression genes.

Supplementary Material Sec. 3.5.5.) For each mutant i described by the overabundance model, the TFNseq experiment measures a growth arrest time T_i and the corresponding target protein overabundance:

$$o_i = \exp(k_0 T_i), \quad (3.4)$$

where k_0 is the wild-type growth rate. (See Supplementary Material Sec. 3.5.5.)

To test the consistency of this TFNseq approach with imaging-based knockout-depletion measurements, we focused first on the analysis of the mutants *dnaA*, *dnaN*, *ftsN*, and *murA*. As shown in Fig. 3.3B, the trajectories for *dnaA*, *murA*, *ftsN*, and *dnaN* show an unambiguous step-like change in growth dynamics: The no-effect trajectory model (null hypothesis) are rejected with p-values that are below machine precision, and the sufficiency trajectory model is also rejected with $p < 10^{-4}$ for all genes. In Tab. 3.1, we compare protein overabundances determined by imaging- and sequencing-based approaches. These numbers are qualitatively consistent. For instance, the single-cell analysis of *dnaA* mutant shows a nearly immediate phenotype by imaging (*i.e.* cell filamentation). (See Supplementary Sec. d.) Likewise, the TFNseq-approach finds an overabundance of 1.0, meaning that protein expression is sufficiency. On the other hand, all three of the other mutants (*murA*, *ftsN*, and *dnaN*) are found to have very large overabundances, and are roughly comparable. Finally, a representative non-essential gene (*e.g.* *recF*) shows no effect. These results support the use of the TFNseq approach to analyze protein overabundance genome wide.

3.2.9 Many essential proteins have vast overabundance

To determine the protein overabundance genome-wide, we analyzed the knockout-depletion trajectories for all genes in *A. baylyi*. (See Fig. 3.3BCD.) Our analysis showed that the vast majority (90%) of genes annotated as non-essential were classified as having *no effect* and 10% of non-essential genes had measurable growth defects. (See Supplementary Material Fig. 3.14.) The most severe growth defect in non-essential annotated genes were observed for the genes *gshA* and *rplI*. For essential genes, all mutants were observed to have growth

defects, as anticipated; however, only 31% of essential proteins were classified as *sufficient*, corresponding to an immediate change in growth rate. Notable genes in this category include ribosomal proteins RpsQ and RpsE, ribonucleotide reductase subunits NrdA and NrdB, and ATP synthase subunits AtpA and AtpD. However, as predicted by the RLTO model, the majority of essential proteins (69%), were classified as *overabundant*, meaning that they required significant dilution before a growth rate change was detected. Fig. 3.3D shows a histogram of essential gene overabundances.

3.2.10 Low-expression genes are highly overabundant

To understand the overall significance of overabundance in a typical biological process, we determined the median essential protein overabundance: 7-fold. To understand the significance of overabundance from the perspective of the metabolic load, we also determine the mean protein overabundance, weighted by the expression level: 1.6-fold. These two superficially-conflicting statistics emphasize a key predicted regulatory principle: overabundance is high for low-abundance proteins; however, it is close to unity for the high-abundance proteins, which constitute the dominant contribution to the metabolic load.

To explicitly test the predicted relation between protein expression and overabundance, we measured the relative abundance of mRNA messages by RNA-Seq for exponentially growing *A. baylyi* cells. (See Supplementary Material Sec. c..) We computed the message number (transcripts per gene per cell cycle) for each essential gene. (See Supplementary Material Sec. b..) Fig. 3.13 compares measured message numbers and overabundances for all essential genes with the prediction of the RLTO model.

As predicted, the data shows a clear trend of decreasing overabundance with increasing message number (Fig. 3.3D). To quantitatively capture this trend, we computed the mean log overabundance over windows of message number (blue curves) to compare the data cloud to the RLTO model predictions. With very few exceptions, high expression genes have extremely low overabundance. At the other extreme, low expression genes typically have large to very large overabundance as shown by the sharp up-turn of the blue curve

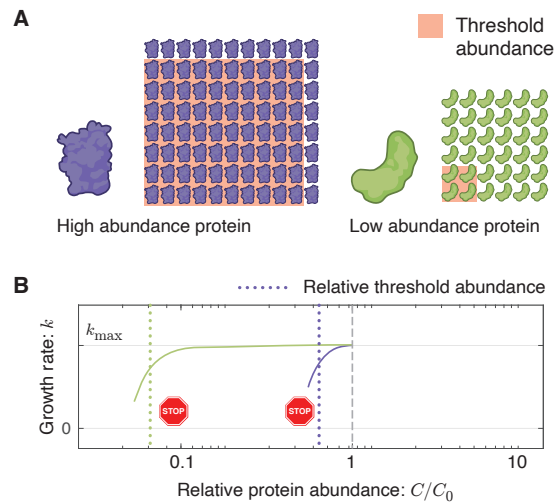


Figure 3.4: **How rate-limited kinetics shapes the fitness landscape. Panel A: Protein abundance and threshold.** Two essential protein species with different abundances are pictured schematically. The threshold abundance at which each protein becomes limiting is represented by the pink square and the total cellular abundance is represented by the protein array. **Panel B: Emergent fitness landscape.** A schematic model of the growth rate versus relative protein abundance is shown for the two protein species. The RLTO model predicts that low-abundance proteins (green) have high overabundance, which leads to significant insensitivity to protein depletion. High-abundance protein (purple) are predicted to have small overabundance leads to high sensitivity to protein dilution. The growth rate rapidly decreases with concentration once a species becomes limiting.

as the message number approaches the one-message-rule threshold, a lower threshold on transcription that we recently proposed [12].

3.3 Discussion

3.3.1 The shape of the fitness landscape

Despite some large-scale measurements [9, 10, 27, 28], fundamental questions remain about the structure of the fitness landscape and its rationale [8]. Our measurements reveal that most (69%) essential proteins show a step-like transition between wild-type and arrested

growth below a critical threshold protein abundance. Although asymmetric landscapes have been observed previously (*e.g.* [4, 27]), the knockout depletion approach is expected to yield more quantitative results. For instance, the use of either CRISPRi (*e.g.* [9]) or inducible promoters (*e.g.* [4]) significantly increases the cell-to-cell variation in protein abundance [17, 29], obscuring the features of the fitness landscape. The sharpness of the protein-abundance threshold is manifest in the single-cell analysis where the progeny begin from a common pool of protein in a single progenitor cell and are therefore not subject to noise (*e.g.* Fig. 3.2C).

3.3.2 The rationale for a threshold abundance

The observed threshold-like dependence can be rationalized in terms of chemical kinetics: If the protein target is not a rate-limiting reactant in an essential cellular process, then its depletion has no effect on the rate [12, 30]. See Fig. 3.4. We explicitly demonstrate protein function (*i.e.* replication) is robust to an order-of-magnitude depletion of replisome protein DnaN; however, for most proteins, we must infer this picture from the growth rate.

3.3.3 The rationale for overabundance

Rate-limiting kinetics does not in itself predict vast protein overabundance. The RLTO model predicts that this feature of the fitness landscape is a consequence of a balance between (i) the metabolic cost of protein expression, which favors minimizing protein abundance, and (ii) robustness to the noise in gene expression [31, 32]. The model predicts expression-dependent protein overabundance: large overabundance for low-abundance proteins and small overabundance for high-abundance proteins [12]. We show that this signature prediction is observed (Fig. 3.3D). In spite of predicting the genomic-scale trend, there are some significant outliers. We discuss their significance as well as evidence for the conservation of overabundance in Supplementary Material Sec. 3.5.1

3.3.4 Biological implications

Many important proposals have been made about the biological implications of noise [33]. Our work reveals that noise acts to inflate the optimal expression levels of low-expression proteins and, as a result, significantly increases the metabolic budget for protein, which constitutes 50-60% of the dry mass of the cell [5]. We believe this increased protein budget has cellular-scale implications. For instance, in stress response and stationary phase, the presence of a significant reservoir of overabundant protein provides critical resources, via protein catabolism, to facilitate the adaptation to changing conditions [34, 35]. Protein overabundance may have important implications for individual biological processes as well, including determining which proteins and cellular processes make attractive targets for small molecule inhibitors (*e.g.* antibiotics) [28]. Since overabundance defines the fold-depletion in protein activity required to achieve growth arrest, high-overabundance proteins are predicted to be extremely difficult targets for inhibition.

3.4 Conclusion

By combining imaging-, genomic-, and modeling-based approaches, we provide a both a quantitative measurement of the fitness landscape for all essential proteins as well as a clear qualitative and conceptual understanding of the rationale for the observed fitness landscape. The RLTO model fundamentally reshapes our understanding of the rationale for protein abundance. The model predicts, and experiments confirm, that low-abundance proteins are expressed in vast excess of what is required for growth. Despite the limitations of the experiments, the predicted trend is clearly resolved both at a genomic-scale, using sequencing-based approaches, as well as at the single-cell scale, as observed by microscopy. The rationale for the overabundance strategy is intuitive: Growth requires the robust expression of between five to six hundred distinct proteins. The cell contends with this extraordinary complex regulatory challenge by keeping all but the highest-abundance proteins in vast excess.

Data availability

We include source data files and sequencing data from RNA-Seq experiments to quantify transcription levels. NCBI Sequence Read Archive, <https://www.ncbi.nlm.nih.gov/sra> (accession no. PRJNA1173237).

Acknowledgments

The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, M. Cosentino-Lagomarsino, S. van Teeffelen, C. Manoil, L. Gallagher, J. Bailey, J. Mannik, G. Chure and S. Murray. H.J.C., T.W.L., D.H., and P.A.W. were supported by NIH grant R01-GM128191 and NSF grant GR046955. K.J.C. was supported by the Molecular Biophysics Training Program (NIH grant T32GM008268). W.R.W. was supported by NIH grant R01-AI150041.

3.5 Supplementary material

3.5.1 Supplementary Discussion

a. Discussion: Are non-essential proteins overabundant?

We have focused our analysis on essential genes in the model organism *A. baylyi* and demonstrated that most essential proteins are overabundant. To what extent is this mechanism generic to non-essential proteins? Several arguments support a generic applicability to non-essential genes. Our modeling suggests asymmetry rather than explicit growth arrest is the mathematical rationale for the optimality of overabundance [12]. We therefore predict that all proteins that increase cell fitness, not just essential proteins, will be overabundant. In addition, it is important to emphasize that the annotation of genes as *essential* is contextual. For instance, for *E. coli* proliferation on lactose, the gene *lacZ* is essential, although non-essential for other carbon sources. As a result, we predict that when expressed, LacZ should be overabundant, consistent with observation [36]. Finally,

the RLTO model also correctly predicts the balance between transcription and translation for all genes, not just essential genes, in eukaryotic cells, suggesting that it should generalize to nonessential genes as well [12].

b. Discussion: Limitations of knockout-depletion experiments.

In spite of the success of the RLTO model in predicting the genomic-scale overabundance trend, there are many significant outliers from this prediction. In considering their significance, it is important to emphasize the flaws both with the knockout-depletion experiments, as well as the RLTO model. With respect to the experiments, the mechanism of growth arrest plays an important role in determining which growth metric most accurately determines the arrest time. Consider the three arrest times measured for the septation-related essential gene *ftsN* in Tab. 3.1. Due to the absence of strict cell-cycle checkpoints in the bacterial cell, the arrest of the septation process does not immediately arrest cell elongation and replication [37]. Growth arrest is therefore detected first by the cell-number metric, directly dependent on septation, and later in the other two metrics.

c. Discussion: Limitations of the RLTO model.

Likewise, the RLTO model itself has some important limitations. For instance, the model assumes that the dominant contribution to the fitness cost of protein overabundance is metabolic load rather than toxicity [38]. We have already investigated the consequences of a toxicity-based increase in cost from a model perspective: The qualitative behavior of the model is unchanged; however, the optimal overabundance is reduced by toxicity [12]. Motivated by this prediction, we tested whether two classes of proteins, ATPases and enzymes [38], that are expected to exhibit toxicity, have lower overabundance. In Supplementary Material Sec. d., we demonstrate that this predicted trend is observed. Similarly, the low overabundance of DnaA also provides a second clue about a class of genes that is predicted to have low overabundance: *dnaA* is negatively autoregulated [39]. Tight regulation can reduce noise, and therefore we hypothesize that tight regulation, and auto-regulation in

particular [40], could therefore reduce the optimal overabundance [12]. In Supplementary Material Sec. e., we demonstrate that this predicted trend is also observed in the data. The putative importance both gene-product toxicity and gene regulation in determining the optimal overabundance emphasizes that the RLTO model describes only part of the biology that determines optimal expression levels.

d. Discussion: Is protein overabundance conserved?

To what extent is the overabundance of essential genes a conserved mechanism from bacteria, to single-cell eukaryotes, to multicellular organisms? As we emphasized above, CRISPRi protein depletions in a wide range of model organisms appear to be consistent with the overabundance hypothesis [9, 13–16]. Furthermore, we have demonstrated elsewhere that the RLTO model also predicts two other principles of central dogma function (the one-message-rule and load balancing in protein expression) that are observed in eukaryotic cells [12]. We therefore expect to observe the overabundance strategy in all organisms for low-expression genes [12].

3.5.2 *Acinetobacter baylyi* strains, manipulation, and culturing

Mutant strains were derived from *Acinetobacter baylyi* ADP1 (MAY101) (the gift of C. Manoil) [41]. Growth media were LB and M9, a minimal-succinate M9 medium [42], supplemented with 15 mM sodium succinate, 2 mM magnesium sulfate, 0.1 mM calcium chloride and 1–3 μ M ferrous sulfate (from sterile 5mM stock, made fresh at least once a month). For selective growth, media was supplemented with kanamycin at 20 μ g/mL. Cultures were grown at 30°C.

The strains used in the study are summarized in Tab. 3.2.

Short name:	Genotype:	Description:
wild-type	ADP1	Wild-type strain.
YdnaN	ADP1 <i>dnaN::YPet-dnaN</i>	The beta clamp (DnaN) is replaced by the fluorescent fusion <i>YPet-dnaN</i> at the endogenous locus. Stable
ΔIS	ADP1 <i>dnaN::YPet-dnaN</i> <i>ACIA0320-0321::kan</i>	This is a control strain where non-essential genes, corresponding to an IS element, are knocked out from the YdnaN strain. Even though the strain is stable, it is re-transformed in each knockout-depletion experiment. Transformed strain has a wild-type growth phenotype. Stable
$\Delta dnaA^\dagger$	ADP1 <i>dnaA::kan</i>	DnaA is an essential cell-cycle regulator. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type. Unstable
$\Delta dnaN^\dagger$	ADP1 <i>dnaN::kan</i>	The beta clamp (DnaN) is an essential component of the replisome. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type. Unstable
$\Delta YdnaN^\dagger$	ADP1 <i>YdnaN::kan</i>	The beta clamp (DnaN) is an essential component of the replisome. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is YdnaN (not wild-type). Unstable
$\Delta murA^\dagger$	ADP1 <i>murA::kan</i>	The gene product of <i>murA</i> is UDP-N-acetylglucosamine 1-carboxyvinyltransferase, an essential protein in synthesizing the precursors of cell wall synthesis. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type. Unstable
$\Delta ftsN$	ADP1 <i>ftsN::kan</i>	The gene product of <i>ftsN</i> is essential cell division protein FtsN. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type. Unstable

Table 3.2: **Summary of strains used in this study.** The *short name* describes the nomenclature of the strains as described in the text. $\dagger$ Strain re-created by transformations in each knockout-depletion experiment are not stable.

a. Methods: Construction of deletion mutations

We generated deletion mutants by transformation of linear DNA fragments, constructed by PCR using extension overlap [18]. A homologous overlap of ~ 2 kb flanking target genes was created that either directly joined (for marker-free deletions) or flanked a kanamycin resistance cassette (for kan-selectable deletions). Unmarked deletions were in-frame. Kan deletions were constructed from the *kan* gene from plasmid pACYC177 [43], in an orientation matching the deleted gene [18]. PCR reactions were performed using Q5 Polymerase (New England Biolabs) or Phusion HF polymerase (New England Biolabs) and DNA fragments were purified using Qiaquick columns (Qiagen) before transformation.

b. Methods: A. baylyi transformation protocol

DNA fragments were transformed into *A. baylyi* cultures prepared as follows. Cultures were grown overnight in minimal-succinate M9 media with 1 μ M ferrous sulfate. The culture was then back diluted 1:5 into fresh medium and grown one hour, shaking at 30°C. The DNA fragment was added at 1 μ g/mL, followed by incubation for 2.5 - 3 hours with shaking, and then plated on selective (for *kan*-deletion cassettes) or non-selective media (for marker-free cassettes). Marker-free deletion mutants were identified by screening single colonies by PCR using primers flanking targeted genes. Essential gene kan-marked deletion mutations were selected by plating on protective medium supplemented with 20 μ g/mL kanamycin. All unmarked and the marked non-essential deletion mutations were verified by PCR. For essential gene deletions, 0.1–1% of the cells were transformed, forming microcolonies of cells carrying the deletion.

c. Methods: Construction of YPet-dnaN fusion strain

In previous work in *Escherichia coli* and *Bacillus subtilis*, we visualized fluorescent fusions to the beta sliding clamp (*dnaN*) to study replication [20–22]. The DnaN protein imaging is a convenient tool for studying replication due to its relatively high abundance and the

change in its localization, from diffuse (non-replicating cells) to punctate (replicating cells), which serves as a convenient reporter of activity.

To construct a fluorescent fusion to the *A. baylyi* DnaN protein with a high probability of success, we used the exact same fluorescent protein and linker to that which R. Reyes-Lamothe had used to construct the *E. coli* fusion used in our previous work [25]. In this approach, we inserted the YPet-linker cassette at the 5' end of the gene. Since the transformation efficiency of *A. baylyi* is so high, we constructed a marker-free fusion. We screened colonies by both PCR and fluorescence localization. We then sequenced the mutant *YdnaN* strain to confirm that the desired construct was achieved. (We provide a supplemental file with the sequence.) Like the original *E. coli* strain, no growth phenotype is observed under experimental conditions.

3.5.3 Growth models for knockout-depletion experiments

To quantitatively analyze growth in knockout-depletion experiments, we define three nested growth models: (i) *No-Effect*, (ii) *Sufficiency*, and (iii) *Overabundance* models. In our statistical analysis, we will initially treat the No-Effect model as the null hypothesis and the Sufficiency model as the alternative hypothesis. If the null hypothesis is rejected, we will then adopt the Sufficiency model as the null hypothesis and adopt the Overabundance model as the alternative hypothesis.

a. *No-Effect model*

In the *No-effect model*, the mutant has no effect on the growth rate. The abundance in a log culture will therefore be:

$$N_N(t; N_0) = N_0 e^{k_0 t}, \quad (3.5)$$

where k_0 is the wild-type growth rate and N_0 is the abundance at $t = 0$.

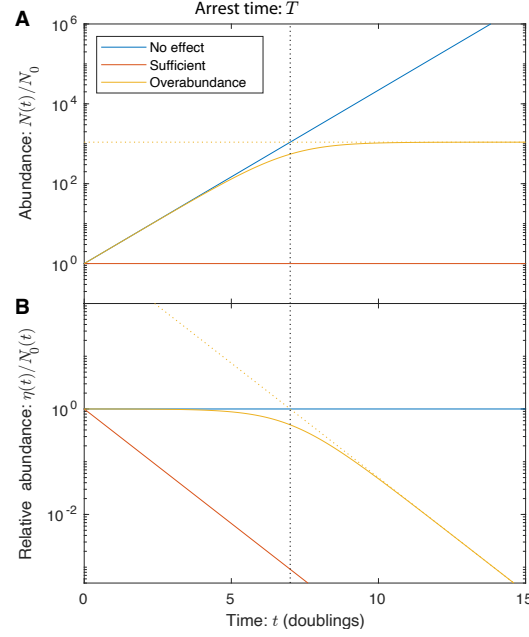


Figure 3.5: **Panel A: Mutant abundances for trajectory models.** Mutants described by the *No-effect model* (blue) grow at the wild-type growth rate. Mutants described by the *Sufficient trajectory model* (red) show an immediate change in growth rate after transformation. Mutants described by the *Overabundant trajectory model* (yellow) grow to the arrest time T (black dotted line) with the wild-type growth rate, before adopting a reduced growth rate of $k = 0$. **Panel B: Relative mutant abundances for trajectory models.** Same as above, but abundances are renormalized by wild-type growth.

For modeling the TFNseq trajectories, it is the relative abundance that is measured and we therefore normalize by wild-type growth of the culture, resulting in the relative abundance:

$$\eta_N(t; \eta_0) = \eta_0, \quad (3.6)$$

where η_0 represents the initial relative abundance. (The relative abundance of the No-effect model is independent of t .) Both the abundance N_N and relative abundance η_N are plotted in Fig. 3.5. Both models depend on a single model parameter and are therefore dimension 1.

b. Sufficiency model

In the *Sufficiency model*, we model the effect of the mutant as immediate. The cell number is assumed to grow at a new unknown rate:

$$N_S(t; N_0, k) = N_0 e^{kt}, \quad (3.7)$$

where k is the new growth rate and N_0 is the number of the mutants at $t = 0$. For modeling the TFNseq trajectories, it is the relative abundance that is measured, and we therefore normalize by wild-type growth of the culture, resulting in the relative abundance:

$$\eta_S(t; \eta_0, \Delta k) = \eta_0 e^{-\Delta k t}, \quad (3.8)$$

where $\Delta k \equiv k_0 - k$ is the growth rate reduction of the mutant relative to the wild-type growth rate. Both the abundance N_S and relative abundance η_S are plotted in Fig. 3.5. Both models depend on two model parameters and are therefore dimension 2. Note that we might naïvely expect $k = 0$ for essential genes; however, we expect some transient growth due to residual protein levels, and these transients will dominate the fit.

c. Overabundance model

In the *Overabundance model*, we model the effect of the mutant with a delayed arrest time, T : the transient growth duration as protein dilutes to the threshold level. For short times, the mutant growth with a wild-type rate:

$$N_O = N_0 e^{k_0 t}, \quad (3.9)$$

however, at long times we expect growth with a new unknown growth rate k :

$$N_O = N'_0 e^{kt}. \quad (3.10)$$

We initially attempted to use a piecewise function to join these two limits; however, the sparsity of the data and discontinuous slope at the boundary appeared to give rise to

fitting artifacts. In addition, the cell-to-cell variation in protein expression smooths out the transition time. To fix these shortcomings, we adopted an empirical formula with the correct limits, but with a smooth transition at $t = T$:

$$N_O(t; N_0, k, T) = N_0 e^{k_0 t} \frac{e^{\Delta k T} + 1}{e^{\Delta k T} + e^{\Delta k t}}, \quad (3.11)$$

where $\Delta k \equiv k_0 - k$ is the loss in growth rate due to the mutation. Modeling the TFNseq trajectories, it is the relative abundance that is measured, and we therefore normalize by wild-type growth of the culture, resulting in the relative abundance:

$$\eta_O(t; \eta_0, k, T) = \eta_0 \frac{e^{\Delta k T} + 1}{e^{\Delta k T} + e^{\Delta k t}}, \quad (3.12)$$

where $\Delta k \equiv k_0 - k$ is the growth rate reduction of the mutant relative to the wild-type growth rate. Both the abundance N_O and relative abundance η_O are plotted in Fig. 3.5. Both models depend on three model parameters and are therefore dimension 3. Note that we might naïvely expect $k = 0$ for essential genes; however, we expect some transient growth due to residual protein levels, and these transients will dominate the fit.

3.5.4 Imaging-based knockout-depletion experiments

a. Methods: Experimental protocol

For single-cell imaging-based analyses, cells were imaged proliferating in M9 media supplemented with 2% low-melt agarose, and in most cases, kanamycin at 20 $\mu\text{g}/\text{mL}$.

Cell preparation for knockout-depletion experiments.

The transformation protocol described above was modified as follows: after the 2.5-3 hr incubation with DNA, cells were immediately spotted on selective media pads for imaging. In the knockout-depletion experiments, cells are transformed with knockout cassettes which recombine into the genome, resulting in Km^R knockout strains. If transformed cells are transferred to Km^+ media too quickly, the competent cells do not have sufficient time to integrate the *kan* cassette before growth arrest. If cells are transferred too late, essential proteins are depleted before imaging begins. How do we know transformants after 2.5-3

hr outgrowth are at their initial stages of transient growth? With the 2.5-3 h outgrowth period, many cells still grow slowly (compared to log phase growth) for 10-15 min consistent with the expression of the kanamycin phosphotransferase (the gene product of the *kan* gene) not having reached a sufficiently high level to achieve a resistance phenotype. Furthermore, a significant number of heterogenic progenitors were observed. The presence of these heterogenic progenitor cells is consistent with the 2.5 h outgrowth period representing the typical recombination time for transformants. (See Sec. d. for a discussion of heterogenic progenitors.)

Sample/slide preparation. Thin pads were fabricated by melting the agarose (Invitrogen UltraPureTM LMP Agarose) and casting it between two slides with two layers of lab tape used as a shim to set the height. After the pad solidified (roughly 10 min), the top slide was carefully removed, and a razor blade was used to trim the pad to form a small square that could be covered with a #1.5 coverslip. For *E. coli* imaging, we typically use a pad that matches the size of the coverslip; however, *for A. baylyi* imaging, we trim the pad so it is less than 1 cm in width. This added space allows aerobic growth to continue over multiple hours. Finally, the coverslip is sealed using a hot glue gun.

Microscopy. The samples were imaged using a Nikon Eclipse Ti microscope in phase contrast and fluorescence. We imaged through a Nikon 60 $\times$ 1.4 NA Phase contrast objective onto a sCMOS camera (Andor Neo). An environmental chamber maintained the sample at 30°C during imaging. For phase imaging, a frame rate of 1 frame / 2 min was used; however, for combined phase and fluorescence imaging we reduced the frame rate to 1 frame / 3 min and 1 frame / 9 min to help reduce bleaching and phototoxicity. (The slowest frame rate was used to resolve the dim YPet-DnaN foci as the protein levels were depleted.) Typically, multiple (~ 10) fields of view were captured simultaneously in each experiment. For fluorescence-based analysis, we mixed in wild-type cells, in addition to fluorescent-fusion cells (1:2), to determine the autofluorescence levels in each experiment.

Image processing (cell segmentation) pipeline. Cell images were processed using the *OmniSegger* package [44] by running the `processExp` command with default settings.

Most of the analysis described in the paper was performed from the `clist.mat` files generated for each dataset.

b. Methods: Cytometry data analyses

Imaging-based analysis for protein overabundance was carried out by assessing the transient cell area growth and septation. The three different single-cell analysis approaches are explained below: *protein abundance*, *area*, and *number analysis*.

Accessing imaging-based cell cytometry data Most of the analysis described in the paper was performed using the `clist.mat` files generated for each dataset by the *SuperSegger-Omnipose* package. In particular, the `data3D` field provides time-dependent cell descriptors for each cell in each frame, including `Rod Length`, `Area`, `Fluor1 sum`, and `Fluor1 background`. These descriptors were the input for our analyses. To characterize cell progeny area of fluorescence, we would generate cell lineage trees and cell progeny IDs using the `getFamily` command and then sum fluorescence or area over all progeny as a function of time. For instance, this data is shown in Fig. 3.6.

Protein abundance analysis

To test the hypothesis that the targeted protein is depleted while protein-associated function continues for multiple generations, we visualized YPet-DnaN abundance and localization after the protein was knocked out as described in the paper. In short, we constructed a fluorescent fusion at the endogenous locus to make the YdnaN strain (Sec. c.), in which the endogenous *dnaN* was replaced by the fusion gene *YPet-dnaN*. In the knockout-depletion experiment, we knocked out the *YPet-dnaN* gene with the *kan* cassette to form *YPet-dnaN::kan*.

To test the protein dilution hypothesis, we measured total progeny fluorescence (the proxy for protein abundance of YPet-DnaN) as a function of time, as the cell progeny proliferated. The dilution model predicts that the protein abundance should scale with the total progeny area like:

$$C(t) = C(0) \frac{A_0}{A(t)}, \quad (3.13)$$

where $C(t)$ is the protein concentration at time t , C_0 is the abundance at time $t = 0$, A_0 is the progenitor area at time $t = 0$, and $A(t)$ is the total area of the progeny at time t . In the context of the fusion experiments, the observable is fusion fluorescence, equivalent to an intensity scaling of:

$$I(t) = I(0) \frac{A_0}{A(t)}, \quad (3.14)$$

where $I(t)$ and $I(0)$ are the average pixel intensity of the progeny at time t and the progenitor at $t = 0$. Both area A and intensity I are time-dependent quantities available in the `clist.mat` file. (See Sec. b..)

Several successive improvements in the experimental design and analysis were required to test the dilution hypothesis. (i) We initially attempted to image cells at the same frame rate as our phase contrast experiment (1 frame/2 min); however, to resolve YPet-DnaN foci after protein depletion, we had to significantly increase the exposure time of the fluorescence images and decrease the frame rate to avoid phototoxicity and bleaching. Although the predicted scaling (Eq. 3.14) was immediately observable in the data without corrections at short times, more care was required to observe the depletion at long times. (ii) First, we background subtracted to account for the background fluorescence level, computed as the average intensity in each frame outside the cell masks. This correction significantly improved the agreement with Eq. 3.14 at intermediate times, but did not yet account for cellular autofluorescence. (iii) Next, we analyzed a mixture of wild-type and *YdnaN* cells, using the intensity of the wild-type cell in the same microcolony for the background subtraction. This method led to good agreement with Eq. 3.14 even at long times (Fig. 3.1).

Why was a mixture of wild-type and *YdnaN* cells preferable to imaging the two strains independently? A detailed analysis of single cell intensities revealed that wild-type cells in close proximity to *YdnaN* cells in the microcolony had higher pixel intensity, due to the diffuse halo created by the bright *YdnaN* cells. The use of wild-type cells in the same field of view helped correct for the diffuse fluorescent light necessary for the analysis of protein abundance at large depletion times. Cell fluorescence intensities at $t = 0$ are used to differentiate between wild-type and *YdnaN* cells.

Areal growth analysis

In this section, we develop the statistical model for the analysis of cell-area based growth assays to determine both the model parameters and the statistical uncertainty of parameters on a per-experiment basis. We provide this development for completeness; **however, cell-to-cell variation will dominate the reported errors.**

Statistical procedure. For the imaging-based analyses, we define the following statistical procedure: For the analysis of essential genes, we will fix the asymptotic growth rate $k = 0$. Therefore, the Sufficiency model is now considered the null hypothesis since it is the lowest dimensional model. The first alternative hypothesis is the No-effect model, where the wild-type growth rate k_0 is fit in each analysis. If the Sufficiency model is rejected, we then adopt the No-effect model as the new null hypothesis and adopt the Overabundance model as the new alternative hypothesis.

Areal growth models. This growth metric is sensitive to cell elongation (rather than septation). Let $A(t)$ be the observed area of all cells sharing a single progenitor cell. For the areal growth model, we substitute cell area $A(t)$ for the abundance $N(t)$ and A_0 for N_0 in Eqs. 3.5, 3.7, and 3.11. The models are:

$$\ln A_S(t; A_0) = \ln A_0, \quad (3.15)$$

$$\ln A_N(t; A_0, k_0) = \ln A_0 + k_0 t, \quad (3.16)$$

$$\begin{aligned} \ln A_O(t; A_0, k_0, T) &= \ln A_0 + k_0 t + \dots \\ &\quad + \ln \frac{e^{k_0 T} + 1}{e^{k_0 T} + e^{k_0 t}}, \end{aligned} \quad (3.17)$$

where we have substituted $k = 0$.

Statistical model for areal growth analysis. We will model the error associated with determining the area of the cells as proportional to cell number or area:

$$\sigma_A \propto A(t). \quad (3.18)$$

This model is consistent with many mechanisms. Rather than fitting a model with a variable error, it is more convenient to introduce a new variable, a , with constant error:

$$a(t) \equiv \ln A(t). \quad (3.19)$$

Since $da = dA/A$, then $\sigma_a = \sigma_A/A$ which leads to an analysis with constant error.

The Shannon information (minus log likelihood) for the log area in frame i is:

$$h(a_i|\boldsymbol{\theta}) = \frac{1}{2} \ln 2\pi\sigma_a^2 + \frac{1}{2\sigma_a^2} [a_i - \mu_a(t_i; \boldsymbol{\theta})]^2, \quad (3.20)$$

where $\boldsymbol{\theta}$ represents the parameter vector, μ_a is the time-dependent mean log area defined by the growth models (Eqs. 3.15-3.17). For a time series with $i = 1 \dots N$ frames, the total Shannon information is:

$$h(\{a_{i=1 \dots N}\}|\boldsymbol{\theta}) = \frac{N}{2} \ln 2\pi\sigma_a^2 + \frac{1}{2\sigma_a^2} S^2, \quad (3.21)$$

which can be formulated as a least-squares minimization where:

$$\Delta a_i \equiv a_i - \mu_a(t_i; \boldsymbol{\theta}), \quad (3.22)$$

$$S^2(\boldsymbol{\theta}) = \sum_{i=1}^N \Delta a_i^2, \quad (3.23)$$

where i is the frame index.

Estimate of error for areal growth analysis. We will statistically estimate the relative area uncertainty (σ_a) from the wild-type growth data. The expression for the MLE for σ_a^2 is:

$$\hat{\sigma}_{a,\text{MLE}}^2 = \frac{1}{N} S^2(\hat{k}_0, \hat{a}_0), \quad (3.24)$$

where Eq. 3.23 is evaluated at the MLE values of the other parameters for the No-effect model. There is one additional improvement to this estimate which is straight forward to implement. It is well known that Eq. 3.24 is biased from below. We can construct an unbiased estimator by correcting for the complexity of the model for the mean (dimension two) [45]:

$$\hat{\sigma}_a^2 = \frac{1}{N-2} S^2(\hat{k}_0, \hat{a}_0), \quad (3.25)$$

which we will use for our variance estimator. Note that if only a single mean were fit, the prefactor would be $(N - 1)^{-1}$ accounting for the one model dimension; however, since we fit both the slope and the offset, the prefactor is $(N - 2)^{-1}$ accounting for the two model dimensions [45, 46].

From the wild-type growth data, the unbiased estimator for the error for log area (Eq. 3.25) is:

$$\sigma_a = 1.5 \times 10^{-3}, \quad (3.26)$$

or alternatively, this result can be stated in a more intuitive form: There is a 0.15% error in the cell area.

Application to observed data. To determine the model parameters (Eq. 3.71), we will minimize the Shannon information (Eq. 3.38) numerically, by a least-squares minimization of Eq. 3.22. We estimate the Fisher information using the resulting Jacobian from the least-squares minimization:

$$\hat{I} \equiv \frac{1}{\sigma_a^2} J J^T, \quad (3.27)$$

where the Jacobian matrices J are contracted over the frame index and σ_a is given by Eq. 3.26. The parameter uncertainties are then estimated from the Fisher information (Eq. 3.72). Although Eq. 3.72 accounts for the statistical uncertainty in the parameters, it does not account for the cell-to-cell variation. We found that this cell-to-cell variation was dominant. We therefore cite this cell-to-cell variation-based uncertainty. For the p-value calculations (Eq. 3.75), we compute the test statistic λ (Eq. 3.73) from the differences between residual norms for the null and alternative hypotheses:

$$\lambda = \frac{1}{2\sigma_a^2} (S_0^2 - S_1^2), \quad (3.28)$$

where σ_a is given by Eq. 3.26, and the residual norms for model I (the null (0) or the alternative (1) hypotheses) are defined in Eq. 3.23.

c. Cell-number growth analysis

In this section, we develop the statistical model for the analysis of cell-number based growth assays to determine both the model parameters and the statistical uncertainty of parameters on a per-experiment basis. We provide this development for completeness; **however, cell-to-cell variation will dominate the reported errors.**

Statistical procedure. For the imaging-based analyses, we define the following statistical procedure: For the analysis of essential genes, we will fix the asymptotic growth rate $k = 0$. Therefore, the Sufficiency model is now considered the null hypothesis since it is the lowest dimensional model. The first alternative hypothesis is the No-effect model where the wild-type growth rate k_0 is fit in each analysis. If the Sufficiency model is rejected, we then adopt the No-effect model as the new null hypothesis and adopt the Overabundance model as the new alternative hypothesis.

Cell-number growth models. For the cell-number growth model, we use Eqs. 3.5, 3.7, and 3.11. The statistical models depend on the growth rates as function of time for model I , which we define as:

$$k_I = \frac{\partial}{\partial t} \ln N_I(t; \theta_I), \quad (3.29)$$

where N_I is the cell abundance in model I at time t . The growth rates for the respective models are:

$$k_N(t; k_0) = k_0, \quad (3.30)$$

$$k_S(t) = 0, \quad (3.31)$$

$$k_O(t; k_0, T) = k_0 \cdot [1 + e^{k_0(t-T)}]^{-1}, \quad (3.32)$$

where Eq. 3.32 interpolates between the initial growth rate k_0 and final growth rate $k = 0$ at time T .

Deriving the Shannon information. Consider an experiment in which images are taken with a high frame rate, where the time duration between frames is δt . Let the frame number be denoted $I = 1 \dots m$ and the number of cells in each frame N_I . Let the model for

cell growth be formulated such that the growth rate at time t_I is:

$$k_I = k(t_I; \theta), \quad (3.33)$$

where θ represents a parameter vector. In this analysis, we will model cell division as a Markovian process where:

$$\dot{N} = kN, \quad (3.34)$$

which is to say that we will ignore the internal state of cells. For instance, at time t , cells have the same rate of division, irrespective of cell age.

In this model, the number of cell divisions n_I that occur over the short time interval δt is Poisson distributed:

$$q(n_I | \mu_I) = \frac{\mu_I^{n_I}}{n_I!} e^{-\mu_I}, \quad (3.35)$$

where

$$\mu_I \equiv \delta t N_I k(t_I; \theta), \quad (3.36)$$

is the mean number of divisions.

We now compute the Shannon information associated with the entire experiment:

$$h(\{N_I\}_{I=1\dots m} | \theta) = - \sum_{I=1}^m \ln q(n_I | \mu_I). \quad (3.37)$$

Substituting Eqs. 3.35 and 3.36, the equation is simplified to:

$$h = \sum_{I=1}^m \delta t N_I k_I - \sum_{I \in \text{Div}} n_I \ln \delta t N_I k_I + \sum_{I \in \text{Div}} \ln n_I!, \quad (3.38)$$

where Div represents the frames immediately preceding division. For instance, if there is one cell at frame 5 and two cells at frame 6, $\text{Div} = \{5\}$.

Application to observed data. To determine the model parameters (Eq. 3.71), we will minimize the Shannon information (Eq. 3.38) numerically, and determine the Hessian at the optimal parameter values to estimate the Fisher information:

$$\hat{I}_{ij} = H_{ij}, \quad (3.39)$$

where H is the Hessian matrix. The parameter uncertainties are then estimated from the Fisher information (Eq. 3.72). Although Eq. 3.72 accounts for the statistical uncertainty in the parameters, it does not account for the cell-to-cell variation. We found that this cell-to-cell variation was dominant. We therefore cite this cell-to-cell variation-based uncertainty. For the p-value calculations (Eq. 3.75), we compute the test statistic λ (Eq. 3.73) from the differences in the Shannon information (Eq. 3.38).

d. Results: Imaging-based analyses

Some progenitors have heterogenic progeny

Heterogenic progenitors are progenitor cells that are observed to have progeny with two distinct heritable phenotypes: the Km^{R} knockout phenotypes and the Km^{S} wild-type phenotype. For instance, in the ΔmurA knockout-depletion experiments, progenitors were observed with one daughter whose progeny proliferated for multiple generations on Km^+ media before lysing, the knockout phenotype, and whose other daughters proliferated for a short period but maintained wild-type morphology. The maintenance of the wild-type morphology suggested that the cells were $\text{murA}^+ \text{Km}^{\text{S}}$. How were these cells able to proliferate while other Km^{S} cells immediately arrested?

We hypothesize that since both cells had the same progenitor, recombination occurred in the mother cell, after the *murA* gene was replicated, leading to one wild-type chromosome and one ΔmurA chromosome. The transient growth of the wild-type cells was the result of overabundance of the *kan* gene product APH(3')II being expressed before cell division in the original mother cell.

Heterogenic progenitor cells appeared frequently for *dnaN* knockout-depletion experiments, presumably because of the location of *dnaN* in the immediate vicinity of the origin, resulting in early replication. In these experiments, an additional test of the heterogenic progenitor hypothesis was possible due to the fluorescent labeling of the target protein. Cells that arrested early with the wild-type morphology showed no protein

Gene:	Area				Cell-number			
	Model:	Growth rate: k (hr ⁻¹)	Arrest time: T (hr)	OA: $\log_{10} o$	Model	Growth rate: k (hr ⁻¹)	Arrest time: T (hr)	OA: $\log_{10} o$
<i>IS</i> (WT)	NE	0.925 ± 0.005	NA	NA	NE	1.04 ± 0.14	NA	NA
<i>dnaA</i>	OA	1.25 ± 0.02	1.2 ± 0.1	0.7 ± 0.1	S	1.04	0.0 ± 0.3	0.0 ± 0.2
<i>dnaN</i>	OA	1.02 ± 0.05	4.5 ± 7.7	2.0 ± 3.0	OA	0.88 ± 0.07	3.8 ± 0.1	1.4 ± 0.1
<i>ftsN</i>	OA	0.78 ± 0.06	5.2 ± 0.3	1.8 ± 0.2	OA	1.12 ± 0.25	1.3 ± 0.4	0.6 ± 0.2
<i>murA</i>	OA	0.70 ± 0.08	3.6 ± 0.4	1.1 ± 0.1	OA	0.96 ± 0.24	2.0 ± 0.3	0.9 ± 0.2

Table 3.3: **Detailed results from fitting imaging-based knockout-depletion experiments.** The table summarizes the analysis of cell proliferation by two complementary metrics: area and cell-number analyses. These two metrics depend on distinct cellular processes: Growth in cell area is dependent on cell elongation, whereas the proliferation of cell number is dependent on the septation process. The estimated standard error is provided for parameter fits. For each gene and analysis, different statistical models were fit to the data: OA (overabundance model), S (sufficient model), and NE (no-effect model), as indicated in the table. The model selected for each gene and metric is shown in the “Model” column. The estimated standard error is provided for parameter fits.

depletion; whereas cells that displayed the mutant phenotype (filamentation) showed depleted YPet-DnaN levels.

Wild-type imaging-based analyses

We analyzed two different strains with wild-type growth phenotypes: wild-type cells (*Acinetobacter baylyi* ADP1) and ACIA0320-0321::*kan*.

IS::*kan*. To generate a reference wild-type growth phenotype, we choose a non-essential gene with no reported phenotype, genes ACIA0320-0321, corresponding to an IS element. The deletion was performed on the YdnaN strain, which shows no growth phenotype under the experimental conditions. We will abbreviate this strain ΔIS . We constructed this deletion and measured its growth relative to wild-type on Km^r media, and no growth phenotype was observed. However, even though this strain can be stably maintained (since ACIA0320-0321 is non-essential), we transformed this cassette using the same protocol in knockout-depletion experiments. As expected, a comparable number of transformants were observed using this construct to those targeting essential genes.

A typical transformant from a knockout-depletion experiment targeting *IS* is shown in Fig. 3.6 for which six generations of growth are captured. Both the areal (cell-elongation-dependent) and cell-number (septation-dependent) analyses are consistent with the null hypothesis, the *No-effect model*, as expected. The growth rate was observed to be $k = 0.925 \pm 0.005 \text{ hr}^{-1}$ for the areal analysis and $k = 1.04 \pm 0.14 \text{ hr}^{-1}$ for the cell-number analysis.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. 3.6. Panel A shows a frame mosaic. The cells in this dataset show the log-phase growth phenotype of wild-type cells. Both cell number and area show exponential growth. The step-like growth of the cell number reflects the desynchronization of cell division events of the ancestors for a single progenitor.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected in favor of the No-effect model for both the area and cell-number analysis (both p-values under machine precision). The growth rate was observed to be $k = 1.04 \pm 0.14 \text{ hr}^{-1}$ for the areal analysis and $k = 0.925 \pm 0.005 \text{ hr}^{-1}$ for the cell-number analysis.

***dnaA* imaging-based analysis**

Annotated gene function. DnaA is an essential regulator of the cell cycle and DNA replication initiation in particular.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. 3.7. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype, cell filamentation, without undergoing significant growth-induced protein dilution. As a result, the cell number, shown in Panel B, is constant since no divisions are observed. However, as shown in Panel C, cell elongation continues for roughly 100 min before it begins to arrest. We interpret the metric that shows the earliest arrest to define the overabundance. In this case, since septation is not observed again after transformation, DnaA abundance is consistent with the Sufficiency model.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected in favor of the *Overabundance model* (p-value under machine precision) for the areal analysis. The initial

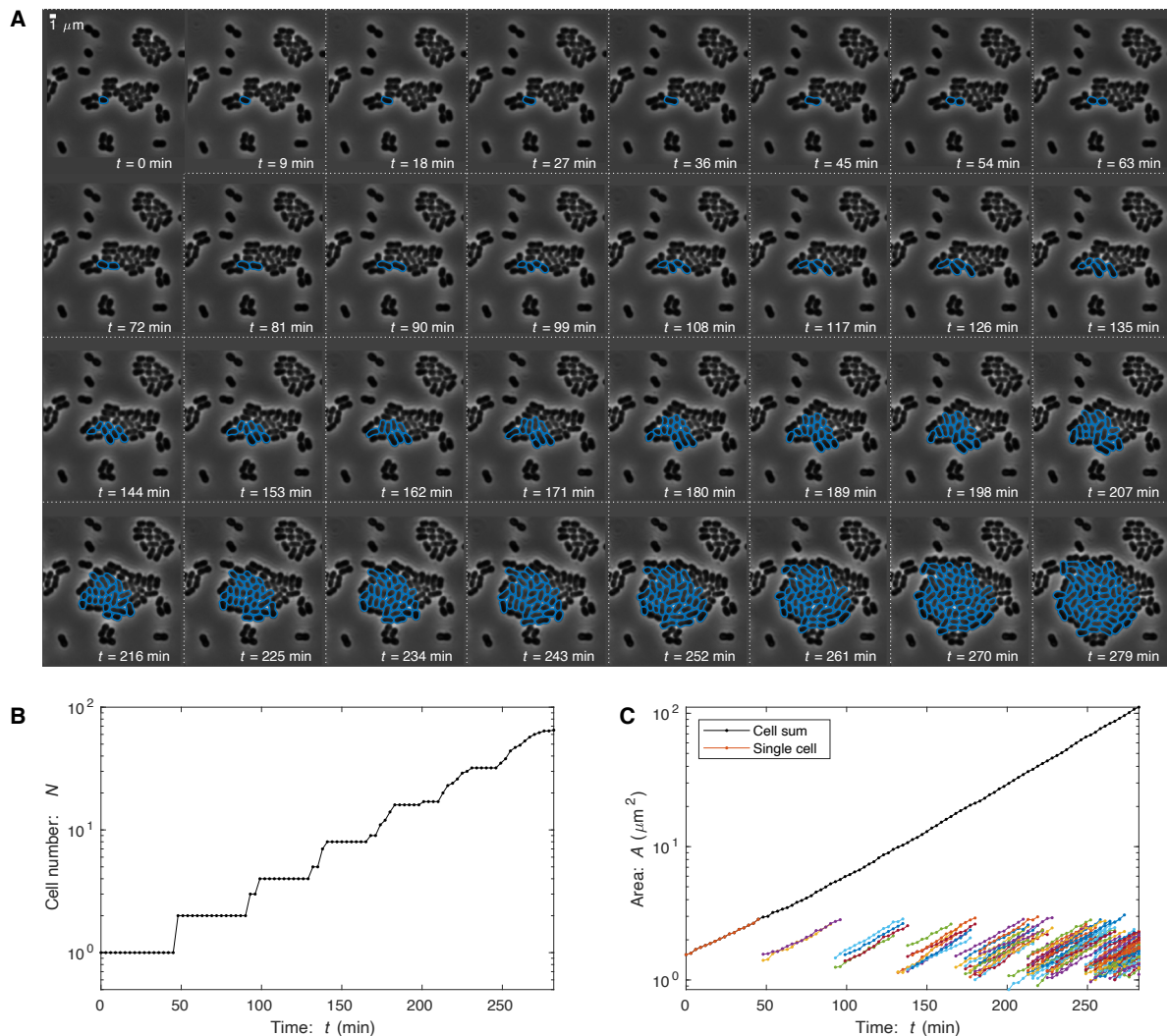


Figure 3.6: **Knockout-depletion experiment: IS element (Non-essential).** **Panel A: Frame mosaic.** In the knockout-depletion experiment, the majority of cells are not transformed and immediately arrest on media supplemented with kanamycin. The lone transformant (*IS::kan* (Km^R), blue) proliferates normally. Cells were segmented using OmniSegger for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. **Panel C: Progeny area.** Total progeny-cell area as a function of time. Total cell area is plotted with the black-dotted line, while individual cell areas are plotted with color.

growth rate was observed to be $k = 1.25 \pm 0.02 \text{ hr}^{-1}$ with an arrest time of $T = 1.24 \pm 0.10 \text{ hr}$. In case of the cell number analysis, we fail to reject the null hypothesis (*Sufficiency model*), indicating that there is no statistical significance to support the alternative hypothesis *No-effect model* ($p = 1.0$). We used the ΔIS wild-type growth rate ($k = 0.925 \pm 0.005 \text{ hr}^{-1}$) to fit the arrest time: $T = 0.0 \pm 0.3 \text{ hr}$.

***dnaN* imaging-based analysis**

Annotated gene function. The gene product of *dnaN* is the β sliding clamp (DnaN), which is an essential component of the replisome complex.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. 3.8. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype, cell filamentation, at about 220 min, after multiple rounds of cell division. As a result, the cell number, shown in Panel B, plateaus shortly after the filamentation is observed since the filamentation is a consequence of the failure of the cells to efficiently septate. However, as shown in Panel C, cell elongation continues, although slowing slightly, throughout the experiment. In this case, since arrest is observed first with respect to septation, we use the arrest of this process to define overabundance.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected for both the area ($p = 8.9 \times 10^{-140}$) and cell-number analysis ($p = 6.0 \times 10^{-19}$). The initial growth rate was observed to be $k = 1.02 \pm 0.05 \text{ hr}^{-1}$ with an arrest time of $T = 4.5 \pm 7.7 \text{ hr}$ for the areal analysis. For cell-number analysis, the initial growth rate was observed to be $k = 0.88 \pm 0.07 \text{ hr}^{-1}$ with an arrest time of $T = 3.8 \pm 0.1 \text{ hr}$.

***ftsN* imaging-based analysis**

Annotated gene function. The gene product of *ftsN* is essential cell division protein FtsN.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. 3.9. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype: the failure to septate, at roughly 150 minutes, after several rounds of division. As a result, the cell number, shown in Panel B, plateaus shortly after 150 min as a consequence

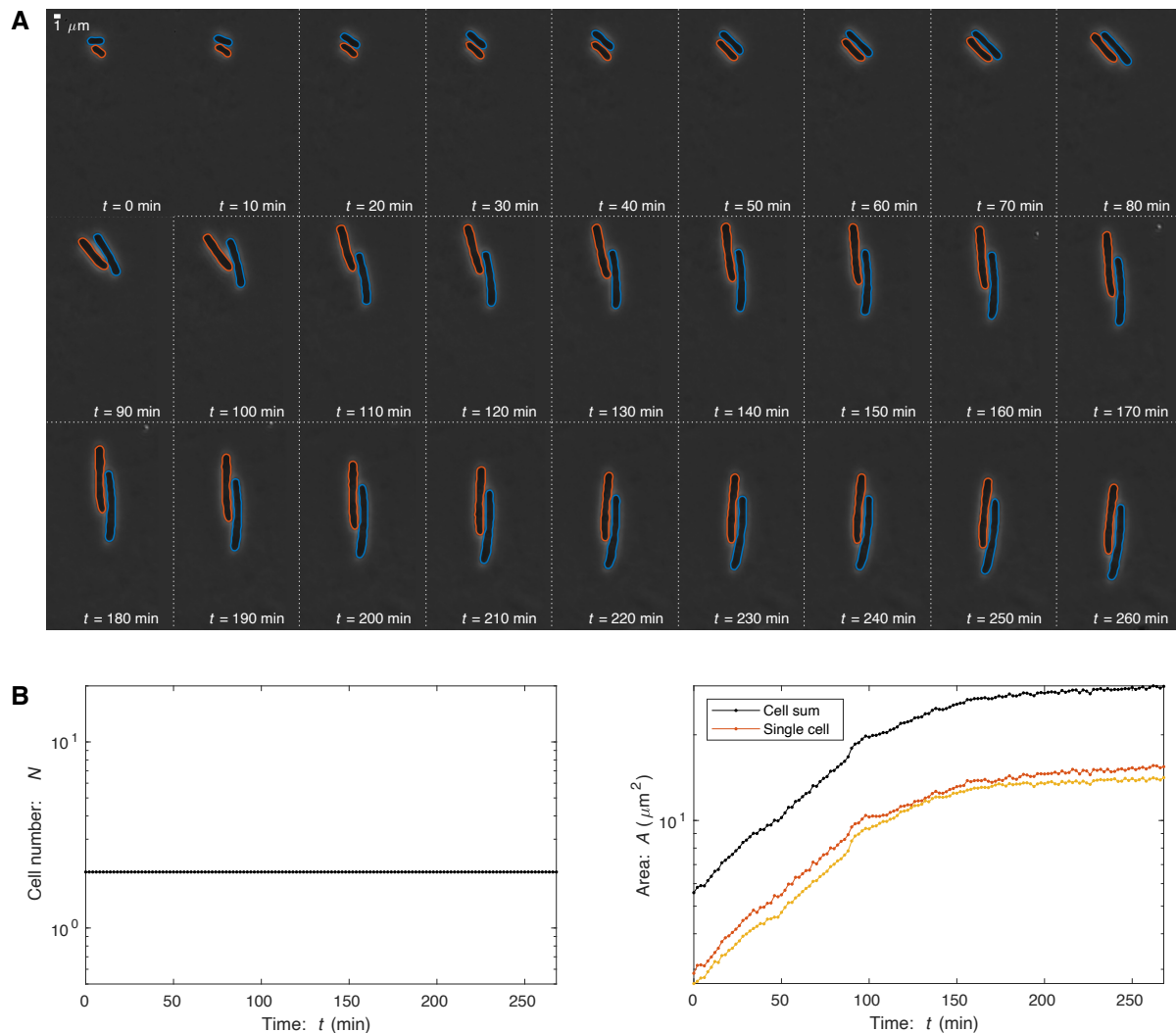


Figure 3.7: **Knockout-depletion experiment: $\Delta dnaA$.** **Panel A: Frame mosaic.** Two transformants ($dnaA::kan(Km^R)$, blue, orange) proliferate. DnaA is an essential regulator of replication initiation. Its depletion leads to a failure of the chromosome to replicate, and therefore results in cell filamentation. Cells were segmented using OmniSegger for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, cells fail to divide, consistent with DnaA expression being sufficient rather than overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. In spite of the arrest of septation/division, cell areal elongation persists for roughly 120 minutes.

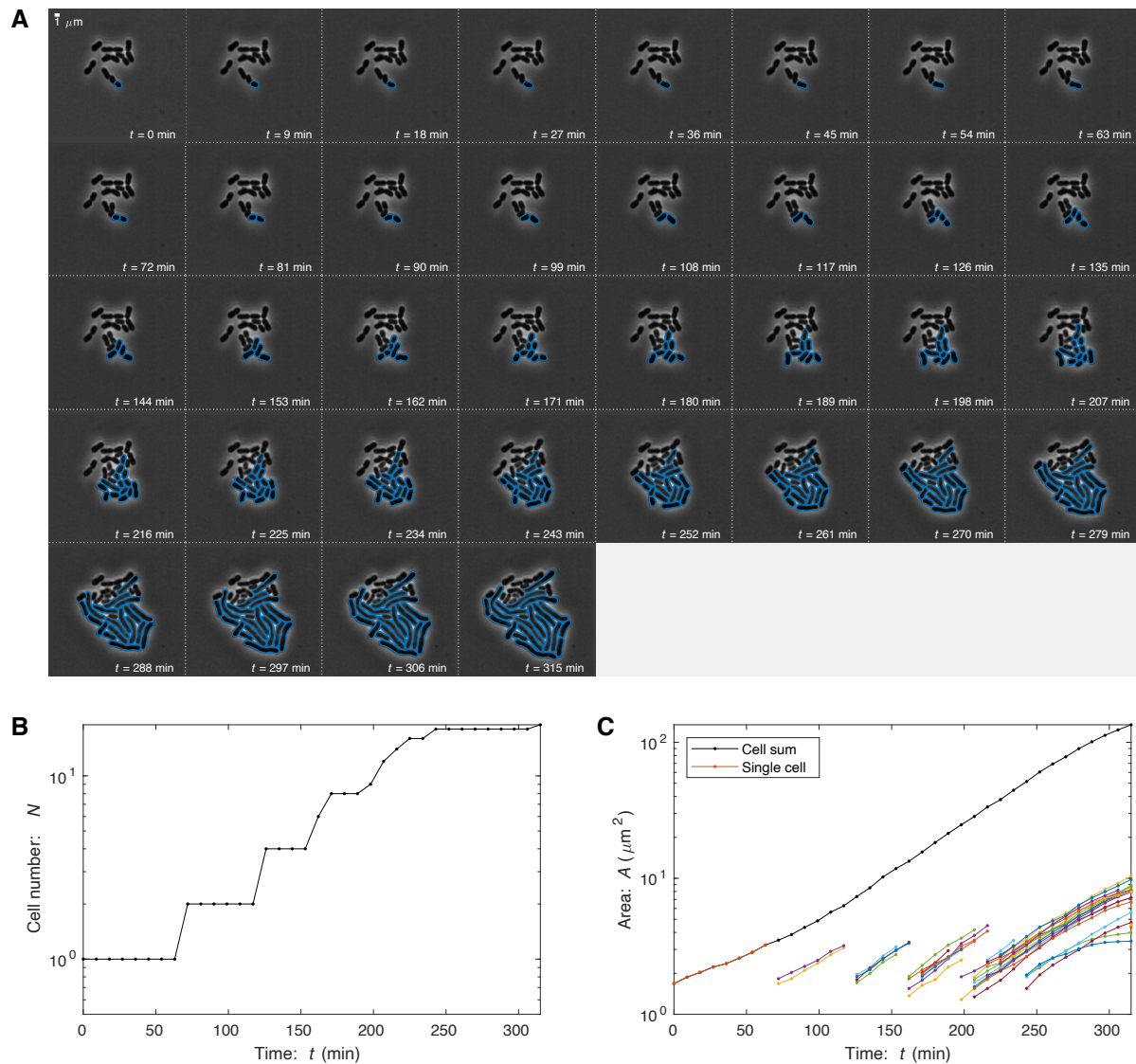


Figure 3.8: **Knockout-depletion experiment: $\Delta dnaN$.** **Panel A: Frame mosaic.** One transformant ($dnaA::kan(Km^R)$, blue) proliferates. The frame mosaic shows a typical imaging-based knockout-depletion experiment. DnaN is the sliding beta clamp, an essential DNA replication protein and a core component of the replisome. Its depletion leads to a failure of the chromosome to replicate and therefore results in cell filamentation. Cells were segmented using OmniSegger for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, normal growth persists for roughly 240 min, consistent with DnaN expression being overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. The areal elongation dynamics persists even are cell division arrests.

of the failure of the cells to efficiently septate. However, as shown in Panel C, cell elongation continues, although slowing slightly, to roughly 220 min.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected for both the area (p-value under machine precision) and cell-number analysis ($p = 1.7 \times 10^{-7}$). The initial growth rate was observed to be $k = 0.78 \pm 0.06 \text{ hr}^{-1}$ with an arrest time of $T = 5.2 \pm 0.3 \text{ hr}$ for the areal analysis. For cell-number analysis, the initial growth rate was observed to be $k = 1.12 \pm 0.25 \text{ hr}^{-1}$ with an arrest time of $T = 1.3 \pm 0.4 \text{ hr}$.

***murA* imaging-based analysis**

Annotated gene function. The gene product of *murA* is UDP-N-acetylglucosamine 1-carboxyvinyltransferase, an essential protein in synthesizing the precursors of cell wall synthesis.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. 3.10. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype: the loss of cell wall integrity, and therefore first the loss of wild-type cell morphology and then cell lysis. Cells begin to lose their wild-type morphology at roughly 120 min, after multiple rounds of cell division. As a result, the cell number, shown in Panel B, plateaus shortly after 150 min as a consequence of the failure of the cells to efficiently septate. However, as shown in Panel C, cell elongation continues, although slowing slightly, to roughly 200 min.

Supplemental approach. For this analysis, we did not want to explicitly model cell lysis. Therefore, in our fitting of the cell-number and areal growth curves, we locked the individual cell area at the last value taken immediately preceding lysis. Similarly, we treated cells that had lysed as arrested, not absent. (This fitting-refined data is *not* shown in Fig. 3.10. The resulting refined data for Panels B and C plateau rather than decrease after growth arrest.)

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected for both the area (p-value under machine precision) and cell-number analysis ($p = 1.2 \times 10^{-6}$). The initial growth rate was observed to be $k = 0.70 \pm 0.08 \text{ hr}^{-1}$ with an arrest time of $T = 3.6 \pm 0.4$

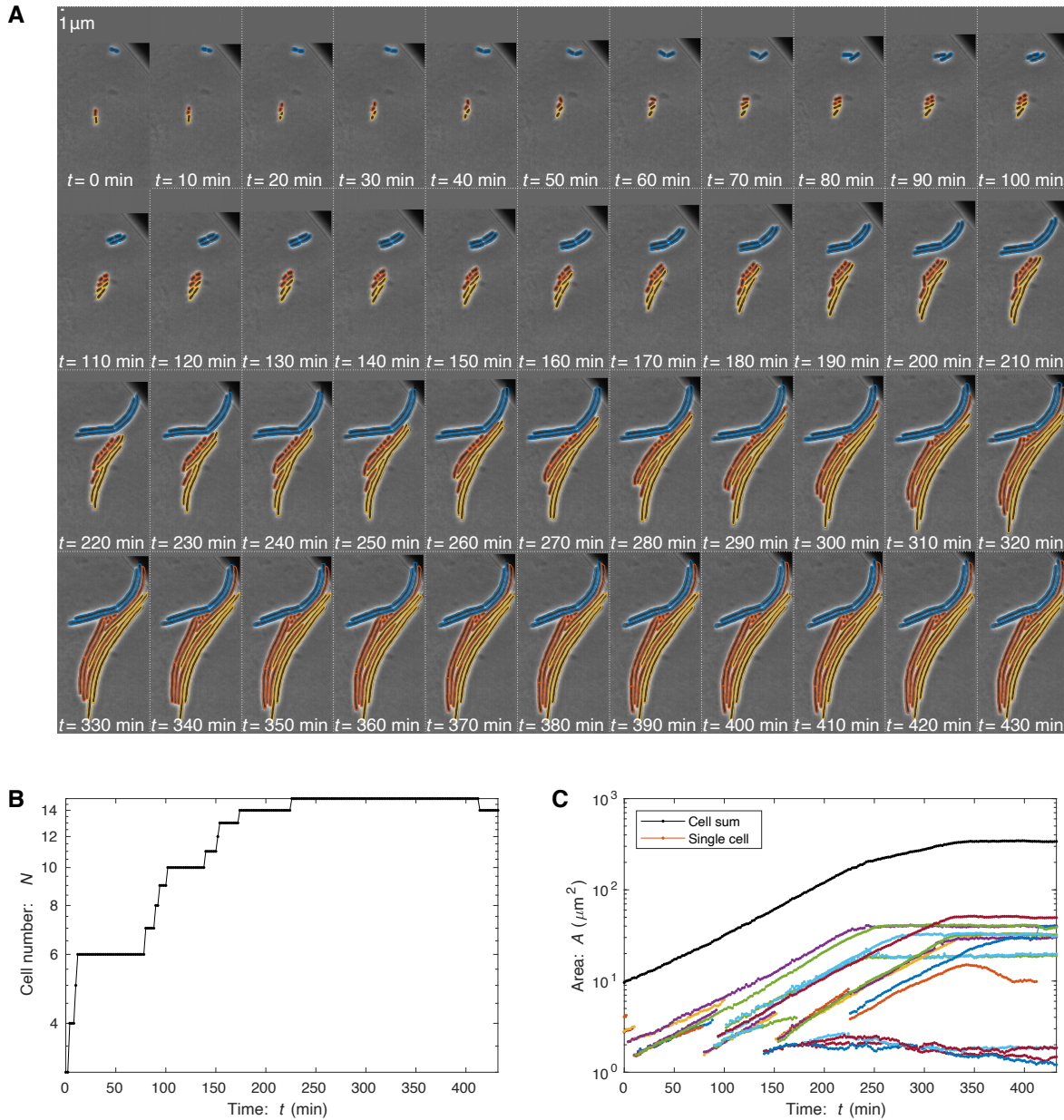


Figure 3.9: **Knockout-depletion experiment: $\Delta ftsN$.** **Panel A: Frame mosaic.** Three transformants ($ftsN::kan(Km^R)$, blue, yellow, orange) proliferate. FtsN is an essential cell division protein. Its depletion leads to a failure of the cells to septate. Cells were segmented using OmniSegger for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, normal growth persists for roughly 150 min, consistent with FtsN expression being overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. The areal elongation persists even after cell division arrests.

hr for the areal analysis. For cell-number analysis, the initial growth rate was observed to be $k = 0.97 \pm 0.24 \text{ hr}^{-1}$ with an arrest time of $T = 2.0 \pm 0.3 \text{ hr}$.

3.5.5 Statistical analysis of TFNseq trajectories

a. Methods: Time correction

Since the mutants transition from lag phase to log phase after transformation, we used a log phase equivalent time for the TFNseq-approach analysis. The corrected sampling times (t_s) are estimated from the number of doublings (D_s) for the non-essential mutants obtained from TFNseq experiment([19]):

$$t_s = D_s * [\text{doubling time}]. \quad (3.40)$$

For our experiment, the doubling time for ADP1 in M9 at 30°C is 37 min.

b. Methods: Defining the likelihood

We assume that deep-sequencing is well modeled by a Poisson process for which the probability mass function is:

$$p(n|\mu) = \frac{\mu^n}{n!} e^{-\mu}, \quad (3.41)$$

where n is the number of reads and μ is the mean-number parameter. For large n , we use the normal-distribution approximation:

$$p(n|\mu) \approx \frac{1}{\sqrt{2\pi\mu}} \exp \left[-\frac{(n-\mu)^2}{2\mu} \right]. \quad (3.42)$$

The total likelihood for sequential observations $n_{1...m}$ at time $t_{1...m}$ is therefore:

$$q(n_{1...m}|\theta_I) = \prod_{i=1}^m p(n_i|\mu)|_{\mu=N_I(t_i|\theta_I)}, \quad (3.43)$$

where N_I is one of the trajectory models and θ_I is the parameter vector for model I . The Shannon information is:

$$h(n_{1...m}|\theta_I) \equiv -\ln q(n_{1...m}|\theta_I), \quad (3.44)$$

$$= -\sum_{i=1}^m \ln p(n_i|\mu)|_{\mu=N_I(t_i|\theta_I)}. \quad (3.45)$$

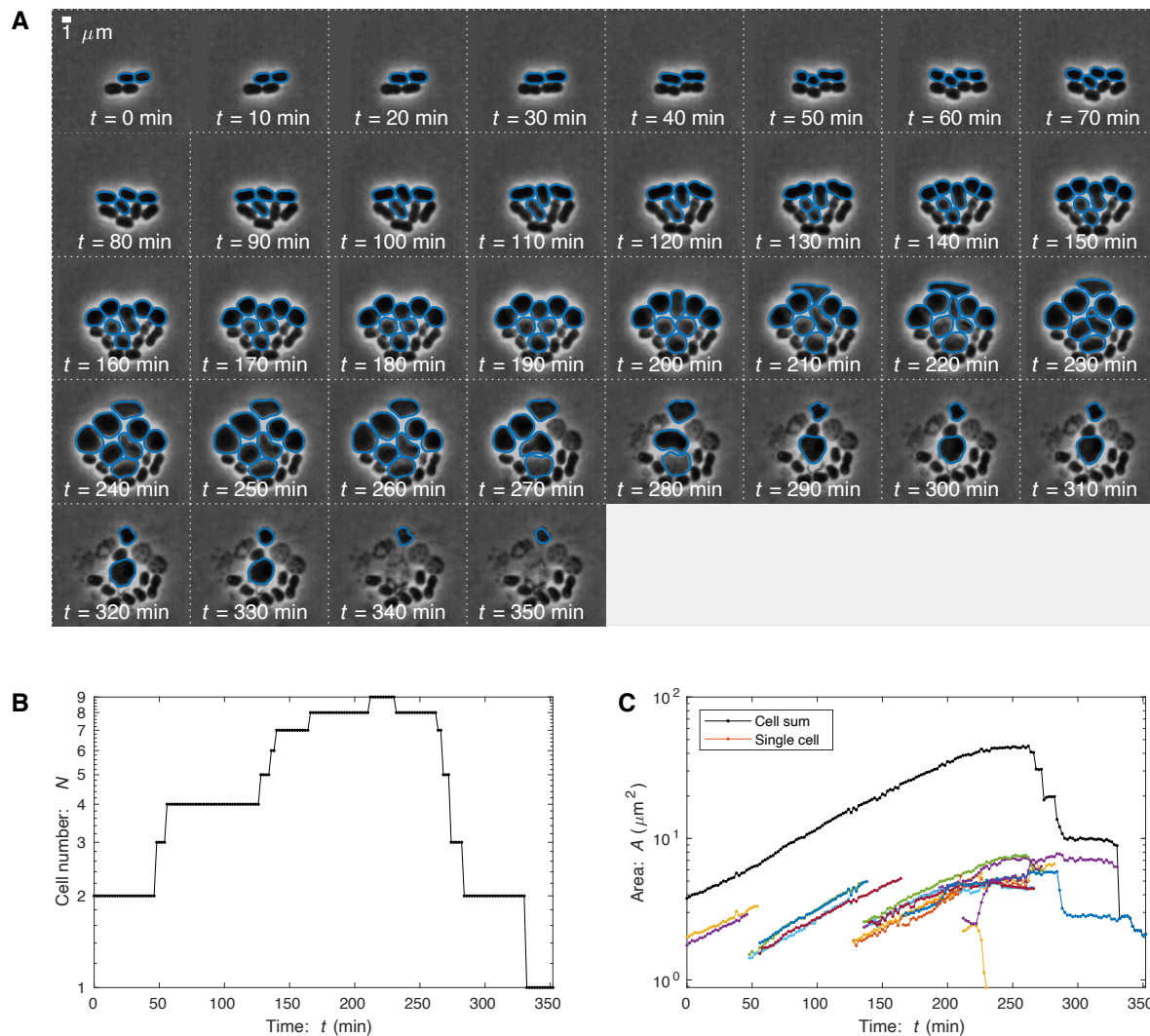


Figure 3.10: **Knockout-depletion experiment: $\Delta murA$.** **Panel A: Frame mosaic.** Two transformants (*murA::kan(Km^R)*, blue) proliferate. MurA is an essential enzyme responsible for cell wall precursor synthesis. Its depletion leads to the loss of cell wall integrity, and therefore first the loss of wild-type cell morphology and then cell lysis. Cells were segmented using OmniSegger for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, normal growth persists for roughly 200 min, consistent with MurA expression being overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. The areal elongation dynamics are largely consistent with the cell number dynamics: Normal growth persists for roughly 200 min.

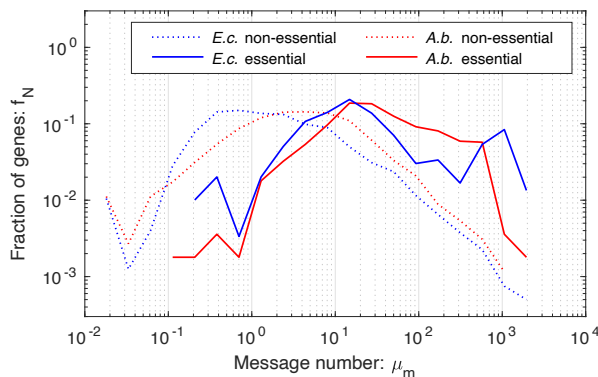


Figure 3.11: **Message number distributions for essential and non-essential genes in *E. coli* and *A. baylyi*.** Nearly all *A. baylyi* essential genes are expressed above the one-message-per-cell-cycle threshold. This distribution of both non-essential and essential genes in *A. baylyi* is qualitatively similar to that in *E. coli*, as predicted [12].

c. Methods: Analysis of overabundance for different Gene Ontologies (GO)

To classify genes, the gene ontology classifications and terms summarized in Tab. 3.4 were used.

d. Results: Toxicity reduces overabundance.

A second key assumption in the RLTO model is that the metabolic cost of transcription and translation are the dominant fitness costs of protein overabundance (*i.e.* there is no toxicity) [12]. To explore the potential role of toxicity, we generated groups of essential ATPases and enzymes, hypothesizing that these proteins would have higher cost due to excessive activity when overabundant, and a group of DNA-Binding Proteins (DBP), which we hypothesized would have low cost when overabundant. We find that the median overabundance for ATPase genes is 2-fold, and for enzymes more generally 5-fold, compared to 7-fold for all essential genes and 13-fold for DBP. These results are consistent with the hypothesis that toxicity, and in particular ATPase activity, is also a key determinant of overabundance. (See Fig. 3.13B.)

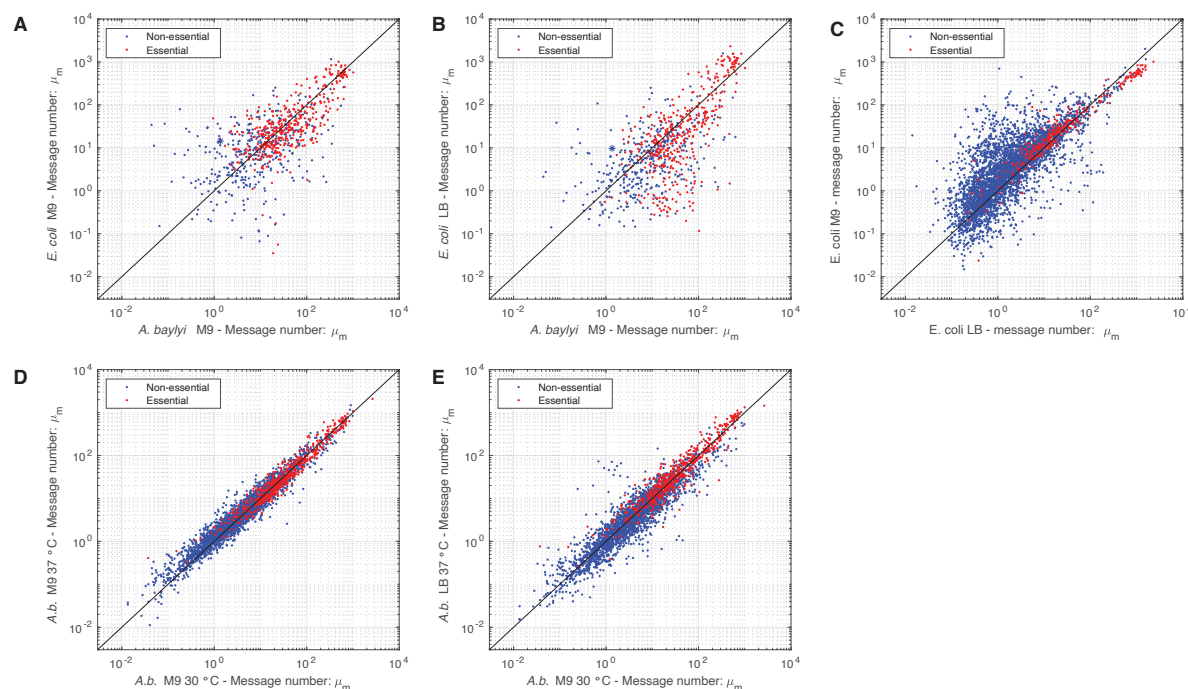


Figure 3.12: Transcriptome comparisons. Panel A: *E. coli* on M9 versus *A. baylyi* on M9. Panel B: *E. coli* on LB versus *A. baylyi* on M9. Panel C: *E. coli* on M9 versus on LB. Panel D: *A. baylyi* on M9 at 37°C versus at 30°C. Panel E: *A. baylyi* on LB at 37°C versus on M9 at 30°C. Throughout, there is broad consistency between the expression levels (message number) of genes, both between organisms and between conditions. These observations suggest a consistent overall transcriptional program governs gene expression both between organisms and growth conditions.

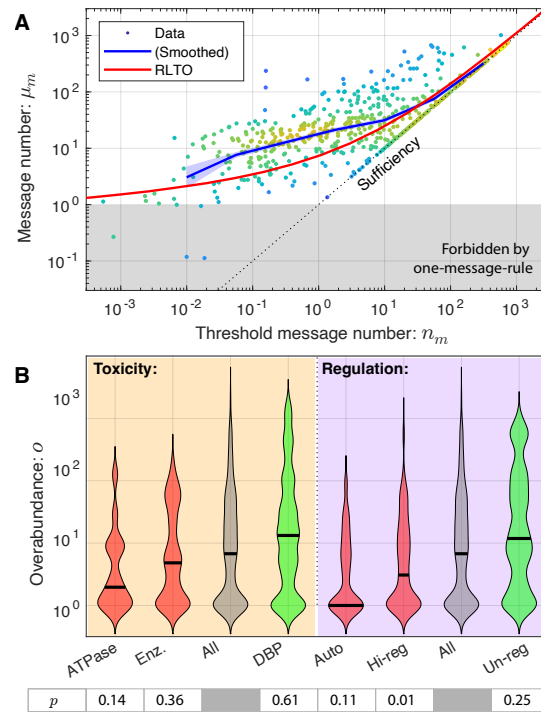


Figure 3.13: **Panel A: Threshold and message number.** Alternatively, the message number μ_m is shown as a function of the inferred threshold message number: $n_m \equiv \mu_m/o$. Although it is μ_m and o that are most directly measurable, the quantity n_m is a more intuitive quantity from a modeling perspective since μ_m is optimized to maximize fitness at fixed n_m in the RLTO model. **Panel B: Toxicity and regulation are determinants of overabundance.** We compared the overabundance measurements for six essential gene subgroups to determine whether toxicity and regulation could affect overabundance. Red groups were predicted to decrease overabundance while green groups were expected to increase it. The p-values for the consistency of each distribution with the all gene group is shown below each category. As hypothesized, the data is consistent with both toxicity and regulation decreasing overabundance.

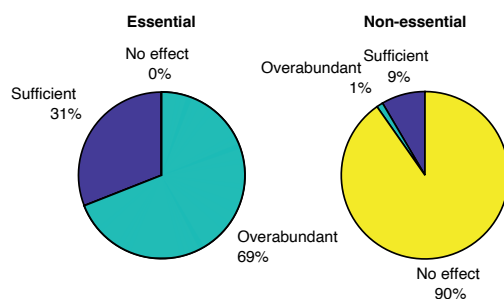


Figure 3.14: **Proteome-wide analysis of proliferation dynamics.** For genes classified as essential, 31% were best fit by the sufficiency model, while 69% were best fit by the overabundance model. For genes classified as non-essential, 90% were best fit by the no-effect model, while 10% showed a detectable reduction in growth rate.

e. Results: Regulation reduces overabundance.

Instead, we adopted a hypothesis-driven approach and attempted to construct subgroups of essential genes that violate the underlying assumptions used to formulate the RLTO model. A key assumption is that gene expression noise is a consequence of the message number only and is otherwise independent of regulation [12]. Precise control of expression could lead to a reduction in the optimal overabundance. To explore the regulatory hypothesis, we generated three lists of essential genes: autoregulatory, highly regulated (top 10% of genes ranked by number of regulators), and un-regulated. If regulation can obviate the need for overabundance, we would expect lower median overabundances in both regulated groups and potentially higher overabundances for the un-regulated group. Consistent with this hypothesis, we find that the median overabundance for autoregulatory genes is 1-fold and for highly-regulated, 3-fold, compared with 7-fold for all essential genes, and 12-fold for un-regulated genes, strongly supporting the hypothesis that tight regulation could reduce the need for overabundance. (See Fig. 3.13B.)

We analyzed only genes in *A. baylyi* that had homologues in *E. coli*. The *E. coli* classifications were downloaded from EcoCyc database [47].

Class	Terms
GO:0006260	DNA replication
GO:0051301	Cell division
GO:0008610	Lipid biosynthetic process
GO:0009252	Peptidoglycan biosynthetic process
GO:0008643	Carbohydrate transport
GO:0006355	Regulation of DNA-templated transcription
GO:0003824	Catalytic activity
GO:0003677	DNA binding

Table 3.4: **Gene ontology classifications and terms.** A summary of the gene ontology classifications and terms used in the study.

f. Analysis of overabundance for different gene regulatory controls

To investigate the effect of transcriptional regulation in determining protein overabundance, we assumed that the regulatory network in *A. baylyi* is roughly equivalent to that in *E. coli* which has been much more extensively studied. We used the EcoCyc database [47] to generate a list for each gene *i* of the list of direct regulators. For each gene, we counted the direct regulators of each gene, then ranked the genes in terms of regulator number, and finally we defined the top 10% of the genes as *highly regulated*. We also generated a list of genes directly inhibited by each gene. If a gene directly inhibited itself, we defined the gene as *autoregulatory*.

3.5.6 RNA-Seq analysis of transcription

a. Methods: RNA-Seq protocol

RNA extraction

ADP1 RNA was harvested through methods developed by Culviner et al. [48] Total RNA was harvested by mixing 1ml of *A. baylyi* (0.5 OD) with 110ul of ice-cold stop solution (95% ethanol and 5% acid-buffered phenol) and spinning in a tabletop centrifuge for 30 s at 13,000 rpm. The supernatant was flash-frozen and stored at -80°C until RNA extraction is ready. To start RNA extraction, 1ml of heated 65°C was added to the sample. The mixture was

shaken at 65°C for 10 min and flash-frozen at -80°C for at least 10 min. The pellets were thawed at room temperature and spun at top speed in a benchtop centrifuge at 4°C for 5 min. The supernatant was collected and added to 400 µl of 100% ethanol. The mixture was passed through DirectZol spin column (Zymo). The column was washed twice with RNA prewash buffer and once with RNA wash buffer (Zymo). RNA was eluted from the column with 90 µl diethyl pyrocarbonate (DEPC)-H₂O. Genomic DNA was removed with 4 µl of Turbo DNase I (Invitrogen) and supplemented with 10µl of 10x Turbo DNase I buffer to a final volume of 100µl. The solution was heated to 37°C for 40 min. Then RNA was diluted with 100 µl DEPC-H₂O, extracted with 200 µl buffered acid phenol-chloroform, followed by ethanol precipitation at -80°C for 4 h with 20 µl of 3 M sodium acetate (NaOAc), 2 µl GlycoBlue (Invitrogen), and 600 µl ice-cold ethanol. To pellet RNA, the samples were centrifuged at 4°C for 30 min at 21,000 × g. The pellets were washed twice with 500 µl of ice-cold 70% ethanol, followed by centrifugation at 4°C for 5 min. RNA pellets were then air dried and resuspended in 50 µl DEPC-H₂O. The yield and integrity of RNA was verified with NanoDrop spectrophotometer, and by running 50 ng of total RNA on a Novex 6% Tris-buffered EDTA (TBE)-urea polyacrylamide gel (Invitrogen).

rRNA Depletion

rRNA was depleted through the DIY method developed by Culviner et al. [48] as well. We used their 21 biotinylated oligonucleotides for *E. coli*. The selected biotinylated oligonucleotides were synthesized by IDT and resuspended to 100 µM in TE buffer (Qiagen). An oligonucleotide mixture was made by mixing equal volumes of each 16S and 23S primers and double volumes of 5S primers. The pooled mixture was diluted with DEPC treated H₂O based on the total RNA, using their Excel-based calculator. Using the Excel-based calculator, the calculated volume of Dynabeads MyOne streptavidin C1 beads (ThermoFisher) were washed three times in equal volume of 1x B&W buffer, resuspended in 30 µl of 2× B&W buffer and supplemented with 1µl of SUPERase-In RNase inhibitor (ThermoFisher). The beads were set aside in room temperature until the probes were ready to be pulled down. To collect rRNA, 2 to 3µg total RNA and 1 µl of the diluted biotinylated probe mix were

combined on ice into a final annealing reaction mixture of 1xSSC and 500 μ M EDTA. All the appropriate volumes were computed using the Excel-based calculator. The RNA and probe mixture was incubated at 70°C for 5 min, and slowly cooled to 25°C at a rate of 1°C per 30 s. The annealed mixture was then added to 30 μ l of beads that were resuspended in 2 $\times$ B&W buffer. The mixture was mixed by pipetting and vortexing at medium speed, and followed by incubating for 5 min at room temperature. The reaction mixtures were then vortexed, and incubated at 50°C for 5 min. To pull down the biotinylated probes, the reaction mixtures were placed immediately placed on the magnetic rack. The supernatant was carefully pipetted, placed on ice, and diluted to 200 μ l in DEPC-H₂O. The RNA was purified through ethanol precipitation with 20 μ l of 3 M NaOAc, 2 μ l GlycoBlue (Invitrogen), and 600 μ l ice-cold ethanol at -20°C for at least 1 h. To pellet RNA, the samples were centrifuged at 4°C for 30 min at 21,000 $\times$ g. The pellets were washed twice with 500 μ l of ice-cold 70% ethanol, followed by centrifugation at 4°C for 5 min. RNA pellets were then air dried and resuspended in 10 μ l DEPC-H₂O. The yield and rRNA depletion effectiveness was verified with NanoDrop spectrophotometer, and by running 50 ng of total RNA on a Novex 6% Tris-buffered EDTA (TBE)-urea polyacrylamide gel (Invitrogen). The yield and integrity of the library was checked by running the samples in qPCR using NEBNext Library Quant Kit for Illumina(NEB) and the Bioanalyzer.

Library prep and sequencing

The RNA library was prepared with NEBNext® Multiplex Oligos for Illumina(NEB) and NEBNext ultra II RNA Library Prep Kit for Illumina(NEB). For the library prep protocol, we followed section 4 of the kit's provided protocol: Protocols for use with Purified mRNA or rRNA Depleted RNA. The quality of the final library was verified by running the samples on high sensitivity Bioanalyzer chip. The samples were pooled to a final concentration of 8.5nM, and were sequenced with NextSeq 150 cycle kit.

b. Methods: Computation of message number

To estimate the message number for gene i , defined as the total number of mRNA molecules transcribed per cell cycle, from the RNA-Seq data, we use the approach we described earlier [12]. Let the relative number of reads for gene i be r_i :

$$r_i = \frac{N_i}{N_{\text{tot}}}, \quad (3.46)$$

where N_i is the reads per kilobase (rpk) for gene i and N_{tot} is the rpk for all genes. We apply two different re-scalings: First we re-scale the relative message abundance to reflect the cellular abundance of the message, and then we scale this number by the ratio of cell cycle duration to mRNA lifetime to estimate the number of times a gene is transcribed per cell cycle. For *A. baylyi*, we use the same scaling factor as *E. coli*:

$$\mu_{m,i} = 9.4 \times 10^4 \cdot r_i, \quad (3.47)$$

where $\mu_{m,i}$ is the estimated message number (number of mRNA molecules transcribed per cell cycle).

To check the consistency of this estimate, we generated histograms for message number for essential and non-essential genes, and compared them to the histograms for *E. coli*. We expect the distribution of essential message numbers to abut 1 message per cell cycle, while non-essential genes can be expressed at significantly lower levels. The observed distribution are consistent with this expectation. (See Fig. 3.11.)

c. Results: Comparison of A. baylyi and E. coli gene expression

Knockout-depletion experiments are not tractable in *E. coli* and many other model systems. It is therefore difficult to directly test the overabundance hypothesis in these other systems. However, it is possible to determine if *E. coli* expression patterns are consistent with overabundance.

If overabundance were specific to *A. baylyi*, we would expect to see higher relative transcription of lower abundance essential genes in *A. baylyi*, where overabundance is large,

relative to *E. coli* if its expression levels were sufficient. Fig. 3.12 compares the message number between homologues in the two organisms and between growth conditions within a particular organism for all genes.

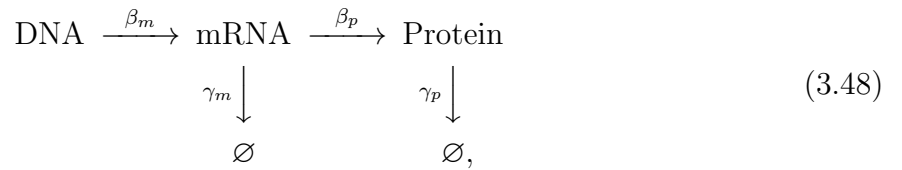
3.5.7 Robustness Load Trade-Off (RLTO) Model

We have provided a detailed description of the Robustness Load Trade-Off (RLTO) Model in Ref. [12]; however, in the interest of making this paper self-contained, we provide a concise summary of key elements and results from that paper in this supplementary section.

a. Methods: Detailed description of the noise model

Stochastic kinetic model for the central dogma.

The canonical steady-state noise model for the central dogma describes multiple steps in the gene expression process [31, 32, 49]: Transcription generates mRNA messages. These messages are then translated to synthesize the protein gene products [50]. Both mRNA and protein are subject to degradation and dilution [51]. At the single cell level, each of these processes are stochastic. We will model these processes with the stochastic kinetic scheme [50]:



where β_m is the transcription rate (s^{-1}), β_p is the translation rate (s^{-1}), γ_m is the message degradation rate (s^{-1}), and γ_p is the protein effective degradation rate (s^{-1}). The message lifetime is $T_m \equiv \gamma_m^{-1}$. For most proteins in the context of rapid growth, dilution is the dominant mechanism of protein depletion and therefore γ_p is approximately the growth rate [49, 52, 53]: $\gamma_p = T^{-1} \ln 2$, where T is the doubling time.

Statistical model for protein abundance.

Consistent with previous reports [31, 32], we find that the distribution of protein number per cell (at cell birth) was described by a gamma distribution [12]:

$$N_p \sim \Gamma(a, \theta), \quad (3.49)$$

where N_p is the protein number at cell birth and Γ is the gamma distribution, which is parameterized by a scale parameter θ and a shape parameter a . (See Sec. a..) We refer to this distribution as the *canonical steady-state noise model*; The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [32]: The scale parameter:

$$\theta = \varepsilon \ln 2, \quad (3.50)$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \quad (3.51)$$

where β_p is the translation rate and γ_m is the message degradation rate. ε is understood as the mean number of proteins translated from each message transcribed. The shape parameter a can also be expressed in terms of the kinetic parameters [32]:

$$a = \frac{\beta_m}{\gamma_p}; \quad (3.52)$$

however, we will find it more convenient to express the scale parameter in terms of the cell-cycle message number:

$$\mu_m \equiv \beta_m T = a \ln 2, \quad (3.53)$$

which can be interpreted as the mean number of messages transcribed per cell cycle. Forthwith, we will abbreviate this quantity *message number* in the interest of brevity.

b. Methods: Summary of the RLTO model fitness model

Metabolic load in the RLTO model

To produce a minimal model to study the trade-off between robustness and metabolic load, we must consider both the metabolic cost of transcription and translation. We will write that the metabolic load (in protein equivalents) associated with gene i is:

$$\delta N_i = \lambda \mu_{m,i} + \mu_{p,i}, \quad (3.54)$$

where λ is the message cost, the metabolic load associated with an mRNA molecule relative to a single protein molecule of the gene product.

$$\ln \frac{k}{k_0} = -\frac{(\lambda + \varepsilon_i) \mu_{m,i}}{N_0}. \quad (3.55)$$

This equation has an intuitive interpretation: growth slows in proportion to the relative added metabolic load. In resource allocation models [54], the capacity of the cell for growth can increase as protein sectors increase in size. In our context, this does not occur since we consider the uncoordinated changes in the levels of single proteins. *I.e.* we assume some other protein of factor is rate limiting. See the detailed discussion in Ref. [12].

Growth rate with stochastic arrest As discussed in Ref. [12], we idealize the slow growth associated with essential proteins falling below threshold as growth arrest. This arrest model has phenomenology consistent with more detailed and realistic models where cells experience a significant growth slowdown rather than true growth arrest [12].

In the idealized growth arrest model, if all essential proteins are above threshold, the cell cycle duration τ is determined by the metabolic load predictions (Eq. 3.55); however, if any essential protein is below threshold, the cell cycle duration is infinite. The probability mass function for the cycle-cycle duration T interpreted as a random variable is therefore:

$$p_T(t) = \begin{cases} P_+, & t = \tau \\ (1 - P_+), & t \rightarrow \infty \end{cases}. \quad (3.56)$$

As we show in Ref. [12], the growth rate can be computed exactly:

$$k = \tau^{-1} \ln(2P_+). \quad (3.57)$$

As expected, the growth rate goes down as the probability of growth P_+ decreases, stopping completely at $P_+ = \frac{1}{2}$. We can then compute the ratio of the growth with (k) and without arrest (k_0):

$$\frac{k}{k_0} = 1 + \frac{1}{\ln 2} \ln P_+, \quad (3.58)$$

where k_0 is computed by evaluating Eq. 3.57 at $P_+ = 1$.

RLTO growth rate

In the RLTO model, we will assume the probability of growth is the probability that all essential protein numbers are above threshold. We will further assume that each protein number is independent, and therefore:

$$P_+ = \prod_{i \in \mathcal{E}} \Pr\{N_{p,i} > n_{p,i}\}, \quad (3.59)$$

where $\mathcal{E}$ is the set of essential genes. Clearly, this assumption of independence fails in the context of polycistronic messages. We will discuss the significance of this feature of bacterial cells elsewhere, but we will ignore it in the current context. As we will discuss, the probability of arrest of any protein i to be above threshold is extremely small. It is therefore convenient to work in terms of the CDFs, which are very close to zero:

$$\ln P_+ \approx - \sum_{i \in \mathcal{E}} \gamma^-\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon_i \ln 2}\right), \quad (3.60)$$

where γ^- is the regularized lower incomplete gamma function (Eq. 3.81) and represents the probability of arrest.

Single-gene equation By summing the fitness losses from the metabolic load and cell arrest (Eqs. 3.55, 3.58, and 3.60), we can write an expression for the growth rate including contributions from essential gene i :

$$\ln \frac{k}{k_0} = -\frac{\lambda + \varepsilon_i}{N_0} \mu_{m,i} - \frac{1}{\ln 2} \gamma^-\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon_i \ln 2}\right), \quad (3.61)$$

where the first term on the RHS represents the fitness loss due to the metabolic load and the second term represents the fitness loss due to stochastic cell arrest due to protein i falling below threshold.

Optimization of transcription for bacteria

The growth rate is:

$$\ln \frac{k}{k_0} = -(\Lambda + \frac{\varepsilon}{N_0})\mu_m - \frac{1}{\ln 2} \gamma^-(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}), \quad (3.62)$$

where γ^- is the regularized lower incomplete gamma function (Eq. 3.81), which is the CDF of the gamma distribution and represents the probability of arrest due to gene i . For bacteria, we consider the special case of optimizing the message number only at fixed translation efficiency [12, 49]. To determine the optimal transcription level, we set the partial derivative of Eq. 3.62 with respect to μ_m to zero. The optimum message number $\hat{\mu}_m$ satisfies the equation:

$$\frac{(\lambda+\varepsilon) \ln 2}{N_0} = -[\partial_{\hat{\mu}_m} \gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\delta}}. \quad (3.63)$$

We define the relative load:

$$\Lambda \equiv \frac{(\lambda+\varepsilon)}{N_0}, \quad (3.64)$$

and substitute this into the optimum message number equation:

$$\Lambda \ln 2 = -[\partial_{\hat{\mu}_m} \gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\delta}}, \quad (3.65)$$

which is solved numerically.

Estimate of the relative load in bacterial cells

In bacterial cells, we will assume a constant translation efficiency model. We therefore use the modified relative load formula (Eq. 3.64) to estimate Λ . We will assume that the load is dominated by proteins and messages:

$$N_0 = \sum_i (\lambda + \varepsilon) \mu_{m,i} = (\lambda + \varepsilon) N_m, \quad (3.66)$$

where N_m is the total number of messages. We can then solve this equation for Λ :

$$\hat{\Lambda} = \frac{\lambda+\varepsilon}{N_0} = \frac{1}{N_m} \approx 10^{-5}, \quad (3.67)$$

based on the total message number estimate for *E. coli* [12].

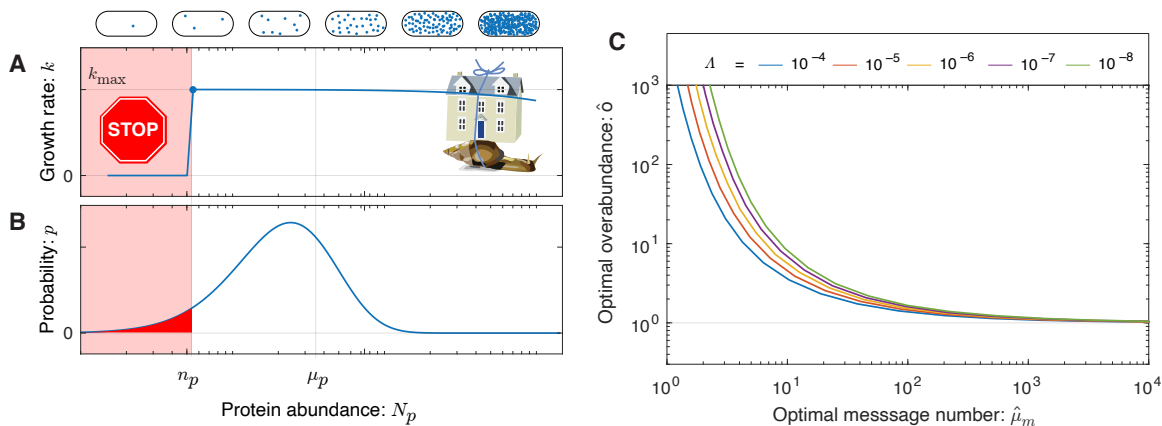


Figure 3.15: **Panel A: The fitness landscape is asymmetric in the RLTO model.** Motivated by single-cell growth data, cell fitness is modeled using the Robustness-Load Trade-Off model (RLTO). In the model, there is a metabolic cost of protein expression which favors low expression; however, growth arrests for protein number N_p smaller than the threshold level n_p (red). The relative metabolic cost of overabundance is small relative to the cost of growth arrest due to the large number of proteins synthesized, resulting in a highly asymmetric fitness landscape [12]. **Panel B: The gene expression process is stochastic.** There is significant cell-to-cell variation in protein abundance (N_p) around the mean level (μ_p). Even for mean expression levels significantly above the threshold level n_p , some cells fall below threshold (red). The distribution in protein number is modeled using a gamma distribution [49]. **Panel C: Overabundance is predicted to optimize cell fitness.** The asymmetry of the fitness landscape drives the optimal protein expression level to be overabundant ($\mu_p > n_p$). The RLTO model makes a quantitative prediction of the optimal overabundance ($\hat{o} \equiv \hat{\mu}_p/n_p$) as a function of the message number μ_m and a global parameter, the relative load $\Lambda \approx 10^{-5}$ (red curve). Overabundance is predicted to be extremely high ($\hat{o} \gg 1$) for low expression genes ($\mu_m \approx 1$) and much closer to sufficiency ($\hat{o} \approx 1$) for high expression genes ($\mu_m \gg 10$). Although the optimal overabundance depends on the relative load Λ , its qualitative dependence is unchanged over orders of magnitude in variation of the parameter.

c. Results: The fitness landscape of the RLTO model is highly asymmetric

In the RLTO model, the fitness landscape for a single cell is determined by an asymmetric fitness landscape: protein underabundance is extremely costly due to the risk of growth arrest, while the cost of protein overabundance is only associated with an increase in metabolic load. (See Fig. 3.15A.) Naïvely, this tradeoff predicts that the cell maximizes its fitness by simply expressing just above the minimum protein threshold for function [8]. However, achieving growth robustness at a population level is nontrivial. Gene expression is stochastic [33], leading to significant cell-to-cell variation in protein numbers, which we model with a gamma distribution (Fig. 3.15B) [31, 32]. Therefore, the strong asymmetry of the fitness landscape predicts protein overabundance.

d. Results: The RLTO model predicts overabundance is optimal for low-expression proteins

The optimal regulatory program for transcription and translation (μ_m and ε values) can be predicted analytically. The values are determined by a single global parameter, the relative load Λ , and the gene-specific threshold number n_p . The threshold number is not directly observable experimentally; instead we predict the optimal overabundance o , defined as the ratio of the mean protein number to the threshold number:

$$o \equiv \mu_p / n_p. \quad (3.68)$$

As shown in [12], by taking partial derivatives of the relative growth rate (Eq. 3.62) with respect to message number and translation efficiency, respectively, we can define the optimal overabundance:

$$\hat{o} \equiv \frac{\hat{\mu}_p}{n_p} = \frac{\hat{\varepsilon} \hat{\mu}_m}{n_p}, \quad (3.69)$$

in the large multiplicity limit where the overall metabolic load is much smaller than the metabolic load for a single gene: $N_0 \gg (\lambda + \hat{\varepsilon}) \hat{\mu}_m$. The optimal overabundance can be rewritten to find the optimization condition for message number:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m} \gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}\right). \quad (3.70)$$

As seen in Fig. 3.15C, the RLTO model generically predicts that for a range of relative loads, the optimal protein fraction is overabundant ($o > 1$); however, overabundance is not uniform for all proteins, but rather depends on transcription. For highly-transcribed genes ($\mu_m \gg 1$), the overabundance is predicted to be quite small ($o \approx 1$); however, for lowly-transcribed genes (message numbers approaching unity), the overabundance is predicted to be extremely high ($o \gg 1$).

e. Discussion: Does the detailed form of the fitness landscape affect RLTO predictions?

It is important to emphasize that the detailed mathematical form of the RLTO model is not essential to generate the predicted phenomenology. For instance, changes in the functional form of the protein expression noise, the metabolic load, or a more realistic model of the fitness landscape do not significantly change the predictions of the model. It is the strong asymmetry of the fitness landscape that is required to predict protein overabundance [12].

3.5.8 Methods: Statistical procedures

In this section, we provide a summary of statistical approaches that are common to the analyses in the paper.

a. Maximum Likelihood Estimation

The maximum likelihood (*i.e.* minimum information) estimates (MLE) of the parameters are defined:

$$\hat{\theta}^i = \arg \min_{\theta^i} h(\text{data}|\theta^i). \quad (3.71)$$

In all instances, these optimizations are performed numerically, either by direct minimization of the Shannon information (h), or for normal models, by least-squares minimization.

b. Parametric uncertainty estimates

To estimate the parameter uncertainty in the analysis of datasets, we use the Cramer-Rao bound to estimate of the uncertainty from the Fisher information [45]:

$$\sigma_{\theta^i} = \sqrt{[\hat{I}^{-1}]^{ii}}, \quad (3.72)$$

where σ_{θ^i} is the estimate of the standard error for parameter θ^i , $\hat{I}$ is the estimator of the Fisher information, and $[\hat{I}^{-1}]^{ii}$ is the ii component of the inverse Fisher information matrix. For each statistical model, we describe how the Fisher information is estimated in detail (Hessian or Jacobian *etc*).

c. Null-hypothesis-testing approach

For null-hypothesis testing, we define two sequential null-hypothesis tests of nested statistical models. If the initial null hypothesis is rejected, we then interpret the initial alternative hypothesis as the updated null hypothesis and adopt the remaining model as the alternative hypothesis. For each test, we will use a Likelihood Ratio Test (LRT) where we define the test statistic λ in terms of the Shannon information:

$$\lambda = h_0 - h_1, \quad (3.73)$$

where h_0 and h_1 are the Shannon information for the null and alternative hypotheses respectively. We will assume the Wilks' theorem: *I.e.* the test statistic Λ under the null hypothesis will have a chi-squared distribution [55, 56]:

$$\Lambda \sim \frac{1}{2}\chi_{\Delta K}^2, \quad (3.74)$$

where the degrees-of-freedom $\Delta K = 1$ is equal to the difference between the dimension of the alternative and null models. (The factor of 1/2 appears in this equation, since the test statistic is defined by the Shannon information difference rather than the deviance [45].) The p-value can then be computed:

$$p = \Pr\{\Lambda > \lambda\} = \gamma^+(\frac{1}{2}\Delta K, \lambda), \quad (3.75)$$

where γ^+ is the upper regularized incomplete gamma function (Eq. 3.91), $\Delta K = 1$ is the difference in model dimensions, and λ is the test statistic [45].

3.5.9 Distributions and conventions

a. Gamma distribution conventions

There are a number of conflicting conventions for the gamma function and distribution arguments. We will use those defined on Wikipedia and the CRC Encyclopedia of Mathematics [57].

The gamma distributed random variable X will be written:

$$X \sim \Gamma(a, \theta), \quad (3.76)$$

where a is the shape parameter and θ is the scale parameter. The PDF of the distribution is:

$$p_X(x|a, \theta) \equiv \frac{x^{a-1}}{\theta^a \Gamma(a)} e^{-x/\theta}, \quad (3.77)$$

where $\Gamma(a)$ is the gamma function. The CDF is therefore:

$$P_X(x|a, \theta) \equiv \Pr\{X < x|a, \theta\}, \quad (3.78)$$

$$= \int_0^x dx' p_\Gamma(x'|a, \theta), \quad (3.79)$$

$$= \int_0^{x/\theta} dx'' \frac{x''^{a-1}}{\Gamma(a)} e^{-x''}, \quad (3.80)$$

$$= \gamma^-(a, x/\theta), \quad (3.81)$$

where γ^- is the regularized lower incomplete gamma function. The survival function is:

$$\Pr\{X > x|a, \theta\} = 1 - P_X(x|a, \theta), \quad (3.82)$$

$$= \int_{x/\theta}^\infty dx'' \frac{x''^{a-1}}{\Gamma(a)} e^{-x''}, \quad (3.83)$$

$$= \gamma^+(a, x/\theta), \quad (3.84)$$

where γ^+ is the regularized upper incomplete gamma function.

b. Chi-squared distribution conventions

In statistical null hypothesis testing, the chi-squared distribution arises in the context of the Likelihood Ratio Test (LRT). Let Y be distributed like a chi-squared with k degrees of freedom:

$$Y \sim \chi_k^2, \quad (3.85)$$

where the PDF is:

$$p_Y(y|k) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} e^{-y/2}, \quad (3.86)$$

where Γ is the gamma function. The CDF is therefore:

$$P_Y(y|k) \equiv \Pr\{Y < y|k\}, \quad (3.87)$$

$$= \int_0^y dy' p_Y(y'|k), \quad (3.88)$$

$$= \int_0^x dx' p_X(x'|\frac{k}{2}, 2), \quad (3.89)$$

$$= \gamma^-(\frac{k}{2}, \frac{y}{2}), \quad (3.90)$$

where p_X is the PDF of the gamma distribution (Eq. 3.77) and γ^- is the regularized lower incomplete gamma function. The survival function is:

$$\Pr\{Y > y|k\} = \gamma^+(\frac{k}{2}, \frac{y}{2}), \quad (3.91)$$

where γ^+ is the regularized upper incomplete gamma function.

3.5.10 Description of supplementary data

a. Data Tables

Data S1: Overabundance for all genes as measured by TFNseq analysis. The original TFNseq data was previously generated by the Manoil lab [19]. Format: Open Document Format (xlsx).

Data S2: A list of essential genes ranked by overabundance. Format: Open Document Format (xlsx).

Data S3: Representative single-cell imaging-based cell cytometry data for wild-type *A. baylyi* proliferating on minimal media (Km^-) from a single progenitor cell (Sec. d.). Format: Open Document Format (xlsx).

Data S4: Representative single-cell imaging-based cell cytometry data for *A. baylyi* ΔIS proliferating on minimal media (Km^+) from a single progenitor cell in a knockout-depletion experiment (Sec. d.). Format: Open Document Format (xlsx).

Data S5: Representative single-cell imaging-based cell cytometry data for *A. baylyi* $\Delta dnaA$ proliferating on minimal media (Km^+) from a single progenitor cell in a knockout-depletion experiment (Sec. d.). Format: Open Document Format (xlsx).

Data S6: Representative single-cell imaging-based cell cytometry data for *A. baylyi* $\Delta dnaN$ proliferating on minimal media (Km^+) from a single progenitor cell in a knockout-depletion experiment (Sec. d.). Format: Open Document Format (xlsx).

Data S7: Representative single-cell imaging-based cell cytometry data for *A. baylyi* $\Delta murA$ proliferating on minimal media (Km^+) from a single progenitor cell in a knockout-depletion experiment (Sec. d.). Format: Open Document Format (xlsx).

Data S8: Representative single-cell imaging-based cell cytometry data for *A. baylyi* $\Delta ftsN$ proliferating on minimal media (Km^+) from a single progenitor cell in a knockout-depletion experiment (Sec. d.). Format: Open Document Format (xlsx).

b. Annotated sequences

Data S9: The annotated sequence of the DnaN fluorescent fusion *YPet-dnaN*. Format: Genbank file.

c. Supplemental movies

Movie S1: Wild-type *A. baylyi* proliferating on minimal media (Km^-). Frame rate: 1 frame/2 min. (Sec. d.) Raw images.

Movie S2: Wild-type *A. baylyi* proliferating on minimal media (Km^-). Frame rate: 1 frame/2 min. (Sec. d.) Annotated/segmented images.

Movie S3: *A. baylyi* ΔIS proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/3 min. (Sec. d..) Raw images.

Movie S4: *A. baylyi* ΔIS proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/3 min. (Sec. d..) Annotated/segmented images.

Movie S5: *A. baylyi* $\Delta dnaA$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. d..) Raw images.

Movie S6: *A. baylyi* $\Delta dnaA$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. d..) Annotated/segmented images.

Movie S7: *A. baylyi* $\Delta dnaN$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/9 min. (Sec. d..) Raw images.

Movie S8: *A. baylyi* $\Delta dnaN$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/9 min. (Sec. d..) Annotated/segmented images.

Movie S9: *A. baylyi* $\Delta murA$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. d..) Raw images.

Movie S10: *A. baylyi* $\Delta murA$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. d..) Annotated/segmented images.

Movie S11: *A. baylyi* $\Delta ftsN$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. d..) Raw images.

Movie S12: *A. baylyi* $\Delta ftsN$ proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. d..) Annotated/segmented images.

3.6 Bibliography

- [1] H. J. Choi, T. W. Lo, K. J. Cutler, D. Huang, W. R. Will, and P. A. Wiggins, “Protein overabundance is driven by growth robustness,” *bioRxiv*, 2024. DOI: [10.1101/2024.08.14.607847](https://doi.org/10.1101/2024.08.14.607847). eprint: <https://www.biorxiv.org/content/early/2024/08/17/2024.08.14.607847.full.pdf>.
- [2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular Biology of the Cell*, 4th. Garland, 2002.
- [3] E. Dekel and U. Alon, “Optimality and evolutionary tuning of the expression level of a protein,” *Nature*, vol. 436, no. 7050, pp. 588–92, Jul. 2005. DOI: [10.1038/nature03842](https://doi.org/10.1038/nature03842).

- [4] J.-B. Lallanne, D. J. Parker, and G.-W. Li, “Spurious regulatory connections dictate the expression-fitness landscape of translation factors,” *Mol Syst Biol*, vol. 17, no. 4, e10302, Apr. 2021. DOI: [10.15252/msb.202110302](https://doi.org/10.15252/msb.202110302).
- [5] H. G. S. Joseph W. Lengeler Gerhart Drews, Ed., *Biology of the Prokaryotes*. Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998.
- [6] M. Kafri, E. Metzl-Raz, F. Jonas, and N. Barkai, “Rethinking cell growth models,” *FEMS Yeast Res*, vol. 16, no. 7, Nov. 2016. DOI: [10.1093/femsyr/fow081](https://doi.org/10.1093/femsyr/fow081).
- [7] J. Hausser, A. Mayo, L. Keren, and U. Alon, “Central dogma rates and the trade-off between precision and economy in gene expression,” *Nat Commun*, vol. 10, no. 1, p. 68, Jan. 2019. DOI: [10.1038/s41467-018-07391-8](https://doi.org/10.1038/s41467-018-07391-8).
- [8] N. M. Belliveau *et al.*, “Fundamental limits on the rate of bacterial growth and their influence on proteomic composition,” *Cell Syst*, vol. 12, no. 9, 924–944.e2, Sep. 2021. DOI: [10.1016/j.cels.2021.06.002](https://doi.org/10.1016/j.cels.2021.06.002).
- [9] J. M. Peters *et al.*, “A comprehensive, CRISPR-based functional analysis of essential genes in bacteria,” *Cell*, vol. 165, no. 6, pp. 1493–1506, Jun. 2016. DOI: [10.1016/j.cell.2016.05.003](https://doi.org/10.1016/j.cell.2016.05.003).
- [10] S. Donati *et al.*, “Multi-omics analysis of CRISPRi-knockdowns identifies mechanisms that buffer decreases of enzymes in *E. coli* metabolism,” *Cell Syst*, vol. 12, no. 1, 56–67.e6, Jan. 2021. DOI: [10.1016/j.cels.2020.10.011](https://doi.org/10.1016/j.cels.2020.10.011).
- [11] T. Baba, H.-C. Huan, K. Datsenko, B. L. Wanner, and H. Mori, “The applications of systematic in-frame, single-gene knockout mutant collection of *Escherichia coli* K-12,” *Methods Mol Biol*, vol. 416, pp. 183–94, 2008. DOI: [10.1007/978-1-59745-321-9_12](https://doi.org/10.1007/978-1-59745-321-9_12).
- [12] T. W. Lo, H. J. Choi, D. Huang, and P. A. Wiggins, “Noise robustness and metabolic load determine the principles of central dogma regulation,” *Science Advances*, vol. 10, no. 34, Aug. 2024, ISSN: 2375-2548. DOI: [10.1126/sciadv.ado3095](https://doi.org/10.1126/sciadv.ado3095).
- [13] K. E. McGinness, T. A. Baker, and R. T. Sauer, “Engineering controllable protein degradation,” *Mol Cell*, vol. 22, no. 5, pp. 701–7, Jun. 2006. DOI: [10.1016/j.molcel.2006.04.027](https://doi.org/10.1016/j.molcel.2006.04.027).
- [14] J. H. Davis, T. A. Baker, and R. T. Sauer, “Small-molecule control of protein degradation using split adaptors,” *ACS Chem Biol*, vol. 6, no. 11, pp. 1205–13, Nov. 2011. DOI: [10.1021/cb2001389](https://doi.org/10.1021/cb2001389).
- [15] D. E. Cameron and J. J. Collins, “Tunable protein degradation in bacteria,” *Nat Biotechnol*, vol. 32, no. 12, pp. 1276–81, Dec. 2014. DOI: [10.1038/nbt.3053](https://doi.org/10.1038/nbt.3053).
- [16] X. Liu *et al.*, “High-throughput CRISPRi phenotyping identifies new essential genes in *Streptococcus pneumoniae*,” *Molecular Systems Biology*, vol. 13, no. 5, May 2017, ISSN: 1744-4292. DOI: [10.15252/msb.20167449](https://doi.org/10.15252/msb.20167449).

- [17] S. N. J. Franks, R. Heon-Roberts, and B. J. Ryan, “CRISPRi: A way to integrate iPSC-derived neuronal models,” *Biochem Soc Trans*, vol. 52, no. 2, pp. 539–551, Apr. 2024. DOI: [10.1042/BST20230190](https://doi.org/10.1042/BST20230190).
- [18] J. Bailey, J. Cass, J. Gasper, N.-D. Ngo, P. Wiggins, and C. Manoil, “Essential gene deletions producing gigantic bacteria,” *PLoS Genet*, vol. 15, no. 6, e1008195, Jun. 2019. DOI: [10.1371/journal.pgen.1008195](https://doi.org/10.1371/journal.pgen.1008195).
- [19] L. A. Gallagher, J. Bailey, and C. Manoil, “Ranking essential bacterial processes by speed of mutant death,” *Proc Natl Acad Sci U S A*, vol. 117, no. 30, pp. 18010–18017, Jul. 2020. DOI: [10.1073/pnas.2001507117](https://doi.org/10.1073/pnas.2001507117).
- [20] S. M. Mangiameli, C. N. Merrikh, P. A. Wiggins, and H. Merrikh, “Transcription leads to pervasive replisome instability in bacteria,” *Elife*, vol. 6, Jan. 2017. DOI: [10.7554/eLife.19848](https://doi.org/10.7554/eLife.19848).
- [21] S. M. Mangiameli, B. T. Veit, H. Merrikh, and P. A. Wiggins, “The replisomes remain spatially proximal throughout the cell cycle in bacteria,” *PLOS Genetics*, vol. 13, no. 1, e1006582, Jan. 2017, ISSN: 1553-7404. DOI: [10.1371/journal.pgen.1006582](https://doi.org/10.1371/journal.pgen.1006582).
- [22] S. M. Mangiameli, J. A. Cass, H. Merrikh, and P. A. Wiggins, “The bacterial replisome has factory-like localization,” *Current Genetics*, vol. 64, no. 5, pp. 1029–1036, Oct. 2018, ISSN: 0172-8083. DOI: [10.1007/s00294-018-0830-z](https://doi.org/10.1007/s00294-018-0830-z).
- [23] G. M. Cooper, *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates 2000, 2000, ISBN: 0-87893-106-6.
- [24] K. P. Lemon and A. D. Grossman, “Localization of bacterial dna polymerase: Evidence for a factory model of replication,” *Science*, vol. 282, no. 5393, pp. 1516–9, Nov. 1998. DOI: [10.1126/science.282.5393.1516](https://doi.org/10.1126/science.282.5393.1516).
- [25] R. Reyes-Lamothe, D. J. Sherratt, and M. C. Leake, “Stoichiometry and architecture of active DNA replication machinery in *Escherichia coli*,” *Science*, vol. 328, no. 5977, pp. 498–501, Apr. 2010. DOI: [10.1126/science.1185757](https://doi.org/10.1126/science.1185757).
- [26] K. J. Cutler *et al.*, “Omnipose: A high-precision morphology-independent solution for bacterial cell segmentation,” *Nat Methods*, vol. 19, no. 11, pp. 1438–1448, Nov. 2022. DOI: [10.1038/s41592-022-01639-4](https://doi.org/10.1038/s41592-022-01639-4).
- [27] L. Keren *et al.*, “Massively parallel interrogation of the effects of gene expression levels on fitness,” *Cell*, vol. 166, no. 5, 1282–1294.e18, Aug. 2016. DOI: [10.1016/j.cell.2016.07.024](https://doi.org/10.1016/j.cell.2016.07.024).
- [28] B. Bosch *et al.*, “Genome-wide gene expression tuning reveals diverse vulnerabilities of *M. tuberculosis*,” *Cell*, vol. 184, no. 17, 4579–4592.e24, Aug. 2021. DOI: [10.1016/j.cell.2021.06.033](https://doi.org/10.1016/j.cell.2021.06.033).
- [29] A. Novick and M. Weiner, “Enzyme induction as an all-or-none phenomenon,” *Proc Natl Acad Sci U S A*, vol. 43, no. 7, pp. 553–66, Jul. 1957. DOI: [10.1073/pnas.43.7.553](https://doi.org/10.1073/pnas.43.7.553).

- [30] D. L. Nelson and M. M. Cox, *Lehninger principles of biochemistry*, en, 7th ed. New York, NY: W.H. Freeman, 2017.
- [31] J. Paulsson and M. Ehrenberg, “Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks,” *Phys Rev Lett*, vol. 84, no. 23, pp. 5447–50, Jun. 2000. DOI: [10.1103/PhysRevLett.84.5447](https://doi.org/10.1103/PhysRevLett.84.5447).
- [32] N. Friedman, L. Cai, and X. S. Xie, “Linking stochastic dynamics to population distribution: An analytical framework of gene expression,” *Phys Rev Lett*, vol. 97, no. 16, p. 168302, Oct. 2006. DOI: [10.1103/PhysRevLett.97.168302](https://doi.org/10.1103/PhysRevLett.97.168302).
- [33] J. M. Raser and E. K. O’Shea, “Noise in gene expression: Origins, consequences, and control,” *Science*, vol. 309, no. 5743, pp. 2010–3, Sep. 2005. DOI: [10.1126/science.1105891](https://doi.org/10.1126/science.1105891).
- [34] D. Weichart, N. Querfurth, M. Dreger, and R. Hengge-Aronis, “Global role for ClpP-containing proteases in stationary-phase adaptation of *Escherichia coli*,” *J Bacteriol*, vol. 185, no. 1, pp. 115–25, Jan. 2003. DOI: [10.1128/JB.185.1.115-125.2003](https://doi.org/10.1128/JB.185.1.115-125.2003).
- [35] A. L. Goldberg and A. C. St John, “Intracellular protein degradation in mammalian and bacterial cells: Part 2,” *Annu Rev Biochem*, vol. 45, pp. 747–803, 1976. DOI: [10.1146/annurev.bi.45.070176.003531](https://doi.org/10.1146/annurev.bi.45.070176.003531).
- [36] G. Lambert and E. Kussell, “Memory and fitness optimization of bacteria under fluctuating environments,” *PLoS Genet*, vol. 10, no. 9, e1004556, Sep. 2014. DOI: [10.1371/journal.pgen.1004556](https://doi.org/10.1371/journal.pgen.1004556).
- [37] S. Autret, A. Levine, I. B. Holland, and S. J. S  ror, “Cell cycle checkpoints in bacteria,” *Biochimie*, vol. 79, no. 9-10, pp. 549–54, Oct. 1997. DOI: [10.1016/s0300-9084\(97\)82002-0](https://doi.org/10.1016/s0300-9084(97)82002-0).
- [38] B. Bolognesi and B. Lehner, “Reaching the limit,” *Elife*, vol. 7, Aug. 2018. DOI: [10.7554/eLife.39804](https://doi.org/10.7554/eLife.39804).
- [39] I. P. Menikpurage, K. Woo, and P. E. Mera, “Transcriptional activity of the bacterial replication initiator DnaA,” *Front Microbiol*, vol. 12, p. 662317, 2021. DOI: [10.3389/fmicb.2021.662317](https://doi.org/10.3389/fmicb.2021.662317).
- [40] D. Nevozhay, R. M. Adams, K. F. Murphy, K. Josic, and G. Bal  zsi, “Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression,” *Proc Natl Acad Sci U S A*, vol. 106, no. 13, pp. 5123–8, Mar. 2009. DOI: [10.1073/pnas.0809901106](https://doi.org/10.1073/pnas.0809901106).
- [41] V. Barbe *et al.*, “Unique features revealed by the genome sequence of *acinetobacter* sp. adp1, a versatile and naturally transformation competent bacterium,” *Nucleic Acids Res*, vol. 32, no. 19, pp. 5766–79, 2004. DOI: [10.1093/nar/gkh910](https://doi.org/10.1093/nar/gkh910).
- [42] J. Miller, *Experiments in Molecular Genetics* (Bacterial genetics - E.: Coli). Cold Spring Harbor Laboratory, 1972, ISBN: 9780879691066.

- [43] A. C. Chang and S. N. Cohen, “Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid,” *J Bacteriol*, vol. 134, no. 3, pp. 1141–56, Jun. 1978. DOI: [10.1128/jb.134.3.1141-1156.1978](https://doi.org/10.1128/jb.134.3.1141-1156.1978).
- [44] T. W. Lo, H. K. J. Choi, D. Huang, and P. A. Wiggins, “The one-message-per-cell-cycle rule: A conserved minimum transcription level for essential genes,” *bioRxiv*, 2023. DOI: [10.1101/2023.07.06.548020](https://doi.org/10.1101/2023.07.06.548020). eprint: <https://www.biorxiv.org/content/early/2023/07/07/2023.07.06.548020.1.full.pdf>.
- [45] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. London, England: Chapman & Hall, 1974.
- [46] K. P. Burnham and D. R. Anderson, “Multimodel inference: Understanding AIC and BIC in model selection,” *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, 2004.
- [47] P. D. Karp *et al.*, “The EcoCyc database (2023),” *EcoSal Plus*, vol. 11, no. 1, eesp-0002–2023, 2023. DOI: [10.1128/ecosalplus.esp-0002-2023](https://doi.org/10.1128/ecosalplus.esp-0002-2023). eprint: <https://journals.asm.org/doi/pdf/10.1128/ecosalplus.esp-0002-2023>.
- [48] P. H. Culviner, C. K. Guegler, and M. T. Laub, “A simple, cost-effective, and robust method for rRNA depletion in RNA-Sequencing studies,” *mBio*, vol. 11, no. 2, V. S. Cooper, Ed., Apr. 2020, ISSN: 2150-7511. DOI: [10.1128/mbio.00010-20](https://doi.org/10.1128/mbio.00010-20).
- [49] Y. Taniguchi *et al.*, “Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells,” *Science*, vol. 329, no. 5991, pp. 533–8, Jul. 2010. DOI: [10.1126/science.1188308](https://doi.org/10.1126/science.1188308).
- [50] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–3, Aug. 1970. DOI: [10.1038/227561a0](https://doi.org/10.1038/227561a0).
- [51] J. L. Hargrove and F. H. Schmidt, “The role of mRNA and protein stability in gene expression,” *FASEB J*, vol. 3, no. 12, pp. 2360–70, Oct. 1989. DOI: [10.1096/fasebj.3.12.2676679](https://doi.org/10.1096/fasebj.3.12.2676679).
- [52] A. L. Koch and H. R. Levy, “Protein turnover in growing cultures of *Escherichia coli*,” *J Biol Chem*, vol. 217, no. 2, pp. 947–57, Dec. 1955.
- [53] M. Martin-Perez and J. Villén, “Determinants and regulation of protein turnover in yeast,” *Cell Syst*, vol. 5, no. 3, 283–294.e5, Sep. 2017. DOI: [10.1016/j.cels.2017.08.008](https://doi.org/10.1016/j.cels.2017.08.008).
- [54] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, “Interdependence of cell growth and gene expression: Origins and consequences,” *Science*, vol. 330, no. 6007, pp. 1099–102, Nov. 2010. DOI: [10.1126/science.1192588](https://doi.org/10.1126/science.1192588).
- [55] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *The Annals of Mathematical Statistics.*, vol. 9, no. 1, pp. 60–62, 1938.

- [56] G. Casella and R. Berger, *Statistical Inference*. Duxbury Resource Center, Jun. 2001, ISBN: 0534243126.
- [57] E. W. Weisstein, *CRC Encyclopedia of Mathematics*. Chapman & Hall/CRC, 2009.

Chapter 4

Measuring metabolic oscillations

Originally published as: [1] D. Huang and P. A. Wiggins, “Metabolic homeostasis, oscillations, and instability,” NSF grant proposal, Jan. 2025.

Author contributions: D.H. and P.A.W. conceived the project and wrote the original grant proposal. H.J.C. performed experiments and expanded the project.

Abstract

Cellular homeostasis relies on regulatory mechanisms to maintain optimal protein and metabolite levels. Although many interactions involved in homeostatic regulations have been studied, there has been little investigation on temporal oscillations of metabolites. Here we investigate the phenomenon of metabolic oscillations, with a focus on deoxynucleotide triphosphate (dNTP) synthesis. High-temporal-resolution measurements, such as replication fork velocity derived from next-generation sequencing and lag-time analysis, reveal oscillations in fork velocity across multiple bacterial species for various conditions. This chapter also discusses experimental strategies for achieving high-precision cell-cycle synchronization, proposing periodic temperature oscillations as a minimally perturbative method. Together, these approaches aim to study the biophysical limits of homeostatic regulation and the mechanisms driving metabolic oscillations, providing new insights into the dynamics of cellular regulatory networks

4.1 *Introduction*

In the previous two chapters, I described the protein optimization strategy against intrinsic noise. However, the RLTO model does not fully capture all the strategies and methods of regulatory optimization. Unlike the assumption in the RLTO model, protein fitness landscape may have sharp peak with a narrow width. This can arise from additional fitness cost, such as toxicity, beyond metabolic load. Proteins with high toxicity would have fitness landscape where optimal expression levels are within a tight range. For such fitness a landscape, it would be more beneficial to have homeostatic protein abundance to keep the protein levels in a narrow functional range. This raises a question: How do cells regulate cellular contents to achieve homeostasis? Mechanistically, cellular homeostasis relies on complex networks of regulatory feedback which have been studied in many systems, but existing studies overlook a key phenomenon of regulatory feedback: temporal oscillations of metabolites.

In the chapter, we explore a different regulatory principle to overabundance: oscillatory behavior of feedback regulations. Here I will outline a future project consisting of two experimental approaches designed to help answer several biophysical questions: i) What are the limits of homeostatic regulatory precision achievable by the cell? (ii) What new insights into cellular regulatory mechanisms do these limits provide? (iii) What mechanism is responsible for the observed metabolic oscillations? By investigating these questions, I aim to deepen our knowledge on the cellular regulatory dynamics that give rise to robust and optimal homeostasis.

4.2 *Testing the dNTP-oscillation model by characterizing fork-velocity oscillations.*

4.2.1 Metabolic pathway of dNTP

dNTP synthesis is a highly-regulated pathway of central importance tied to DNA replication. The complex nucleotide synthesis pathway involves multiple layers of regulatory control. For *E. coli* under aerobic conditions, the protein products of genes *nrdAB* form the enzyme RiboNucleotide Reductase (RNR). RNR reduces nucleoside diphosphates (NDPs) to form deoxynucleoside diphosphates (dNDPs), which are the key precursor for DNA synthesis during DNA replication [2]. Because dNTP pool levels are rate-limiting for the DNA replication [3, 4] and dNTP synthesis is rate-limited by RNR activity[3], the regulation of RNR is thought to play a key role in determining replication and therefore cell-cycle dynamics [5]. The RNR is regulated in two distinct mechanisms: (i) allosteric regulation of the enzyme itself [6–10], (ii) transcriptional regulation [5].

4.2.2 Motivation

dNTP synthesis is a highly-regulated pathway of central importance tied to DNA replication. The complex nucleotide synthesis pathway involve multiple layers regulatory control. For *E. coli* under aerobic conditions, the protein product of genes *nrdAB* form enzyme RiboNucleotide Reductase (RNR). RNR reduces nucleoside diphosphates (NDPs) to form deoxynucleoside diphosphates (dNDPs), which are the key precursor for DNA synthesis during DNA replication [2]. Because dNTP pool levels are rate-limiting the DNA replication [3, 4] and dNTP synthesis is rate-limited by RNR activity[3], the regulation of RNR is thought to play a key role in determining replication and therefore cell-cycle dynamics [5]. The RNR regulated in two distict mechanisms: (i) allosteric regulation of the enzyme itself [6–10], (ii) transcriptional regulation [5].

Observation and characterization of the regulatory oscillations in this model system hold critical importance, as their discovery would suggest that oscillatory dynamics are

a ubiquitous feature of metabolic networks, even if it may be much more challenging to characterize the mechanism or the dynamics of the other pathways. Furthermore, observation of oscillatory dynamics in multiple divergent species for diverse growth conditions would suggest that regulatory oscillation may be a generic and fundamental feature of homeostasis of biological systems.

4.2.3 Preliminary work

There are two existing pieces of evidence supporting the presence of dNTP oscillations. First, the measurement of mutation rate as a function of locus position relative to the origin of has a symmetric wave-like dependence [11, 12]. Because elevation of dNTP pool levels leads to increases in mutation rates, it has been proposed that the wave-like dependence of mutation rate could be a result of oscillation of dNTP levels during replication [11, 12].

A second line of evidence comes from experiments performed by Huang et al. as well as Bhat et al, who made *in vivo* measurements of replication fork velocity [13, 14]. Furthermore, analysis revealed that the oscillations in fork velocity are time-dependent rather than locus-dependent [13]. This time-dependent oscillations of replication fork velocity were observed in different growth conditions for *E. coli* as well as *Vibrio cholerae* and *Bacillus subtilis*. Because the rate of replication is nucleotide limited, this evidence supports our model: the presence of temporal oscillations in dNTP pool results in temporal oscillations in fork velocity.

4.2.4 Approach

So, how can high-temporal-resolution measurements of dNTP levels be acquired to observe and characterize the temporal oscillations? DNA replication rate appears to be nucleotide-limited *in vivo*: decreases in dNTP pool levels leads to decreased replication rate, and increases in dNTP pool levels leads to faster replication speed. Recently, Huang et al.[13] developed a method for measuring *in vivo* replication fork velocity, *i.e.* the elongation rate, measured in bp/s, called lag-time analysis. In short: (i) cells are grown in

asynchronous culture and harvested in early log-phase growth (See Fig. ??A). (ii) Genomic DNA is harvested and prepared for deep sequencing (See Fig. ??B). (iii) The marker frequency is measured and analyzed to obtain rate of replication:

$$\tau(\ell) = \gamma^{-1} \ln \frac{N(0)}{N(\ell)}, \quad (4.1)$$

where γ is the culture growth rate and $N(\ell)$ is the marker frequency at a locus at position ℓ and $N(0)$ is the abundance of the origin of replication [13, 15–18]. One of the advantages of this approach is that no physiological perturbation is needed to acquire measurements. This is because it relies on the steady state of the asynchronous log-phase population as opposed to synchronization of the cells.

By employing lag-time analysis to measure replication fork velocity as a reporter for dNTP pool levels, we can indirectly observe temporal oscillations. To probe the oscillatory dynamics of metabolites, we will perturb nucleotide synthesis pathways—such as through controlled overexpression or suppression of *nrdAB*—and monitor resulting changes in fork velocity oscillations. Through these approaches, we hope to understand the dynamics of temporal oscillations in highly regulated systems.

4.3 *Cell-cycle dependent measurements through cell synchronization*

4.3.1 Motivation

Although lag-time analysis would provide us insight into temporal oscillations of fork velocity, the measurements are not a direct reporters of dNTP pool levels. So, how can we make direct measurements of the cell-cycle dynamics of different components involved in regulation of metabolites? There are a wide range of biochemical assays available, however, the majority of these assays are limited to bulk measurements. In order to use the various biochemical assays for our cell-cycle dependent observations, we would need to synchronize the cells with sufficient precision.

There are two classes of synchronization approaches that are available: (i) **baby-cell-enrichment-based** approaches or *baby machines* use a number of different approaches to filter and harvest newborn cells [19–22]. One of the advantages of this approach is that it does not interfere with cell physiology. However, due to the stochasticity in the duration of the B period (*i.e.* the time between cell birth and replication initiation) [23], this approach may result in poorer synchronization. The second class of synchronization is (ii) *arrest-based* approaches in which cell cycle is arrested at initiation and then released. This would allow the cells to begin C period synchronously. These approaches include the treatment of bacterial cells with serine hydroxamate (SHX) [24] and the use of a number of temperature-sensitive alleles (*dnaC2* and *dnaC28* [25], as well as *dnaA46* [26]). The advantage of this approach over the *baby-cell-enrichment-based* (or the *baby machines*) is that it has higher precision for synchronization as the process directly synchronizes the C period. However, the downside to this approach is that cell-cycle arrest can significantly perturb cell physiology.

4.3.2 Approach

Many conventional experiments utilizing synchronization of temperature sensitive alleles require holding cell cultures at nonpermissive temperatures for extended periods ($> 60\text{min}$) [27]. Although this method will synchronize the cellular process associated with the temperature sensitive allele, other cellular processes often remain unsynchronized, resulting in a physiologically asynchronous and perturbed sample. To achieve cell synchrony with minimal physiological disturbance, we will utilize one of the most powerful tools in physics, periodic driving. The idea is simple: instead of arresting initiation of replication for a long period, grow the cell culture in periodic cycle between arrest and releasing conditions. For example, a culture of temperature sensitive mutant, *dnaC2*, would be grown in an oscillating temperature between 30° (permissive temperature) for 40 min and 36° (non-permissive temperature) for 10min. The goal of this approach is to minimize physiological perturbation by avoiding cells being held in an arrested state for extended periods. Instead, they are

repeatedly released and progressively synchronized over multiple cycles, allowing gradual synchronization. This would provide a new method of synchronizing cells for that have minimal stochasticity and disturbance.

4.3.3 Preliminary work

We have preliminary synchronization data from fluorescence microscopy imaging of *dnaC2* with a *YPet-dnaN* fusion, which allows us to visualize initiation of DNA replication during temperature down shift. When the culture is held at 36°C sufficiently, the cells lose YPet-DnaN foci, a consequence of completing DNA replication and the inability to initiate the next round of replication (Fig. 4.1A). When temperature drops down to 30°, permissive temperature, foci are formed as the cells are allowed to initiate replication. Foci formation is synchronized, as the majority of cells show foci within a short period (<10min) after the temperature downshift.

4.3.4 Future work

Despite some promising results in synchronization attempt, current experimental setup has limitations. Over time we see that the cells continue to elongate without division (see Fig. 4.1 A). Previous study by Maisnier-Patin *et al.* on the impact of *dnaC2(ts)* allele on replication arrest showed that 18% of the *dnaC2(ts)* mutants grown in nonpermissive temperature were not able to finish replication, implying the continuous DnaC activity is required to complete replication with certainty [28]. Therefore, we suspect extended growth in nonpermissive temperature leading to elongated cells without division. To circumvent this effect, we would need to turn to a different temperature sensitive mutant such as strains with *dnaA46(ts)* allele. Unlike *dnaC2(ts)*, *dnaA46(ts)* mutants are able to finish ongoing rounds of replication even during growth in nonpermissive temperatures. Therefore, this allele would serve as a better alternative to achieve successful synchronization.

There are other complications with the current experiment: (i) only a couple rounds of temperature cycles can be applied before the cells are overpopulated on the slide. This can

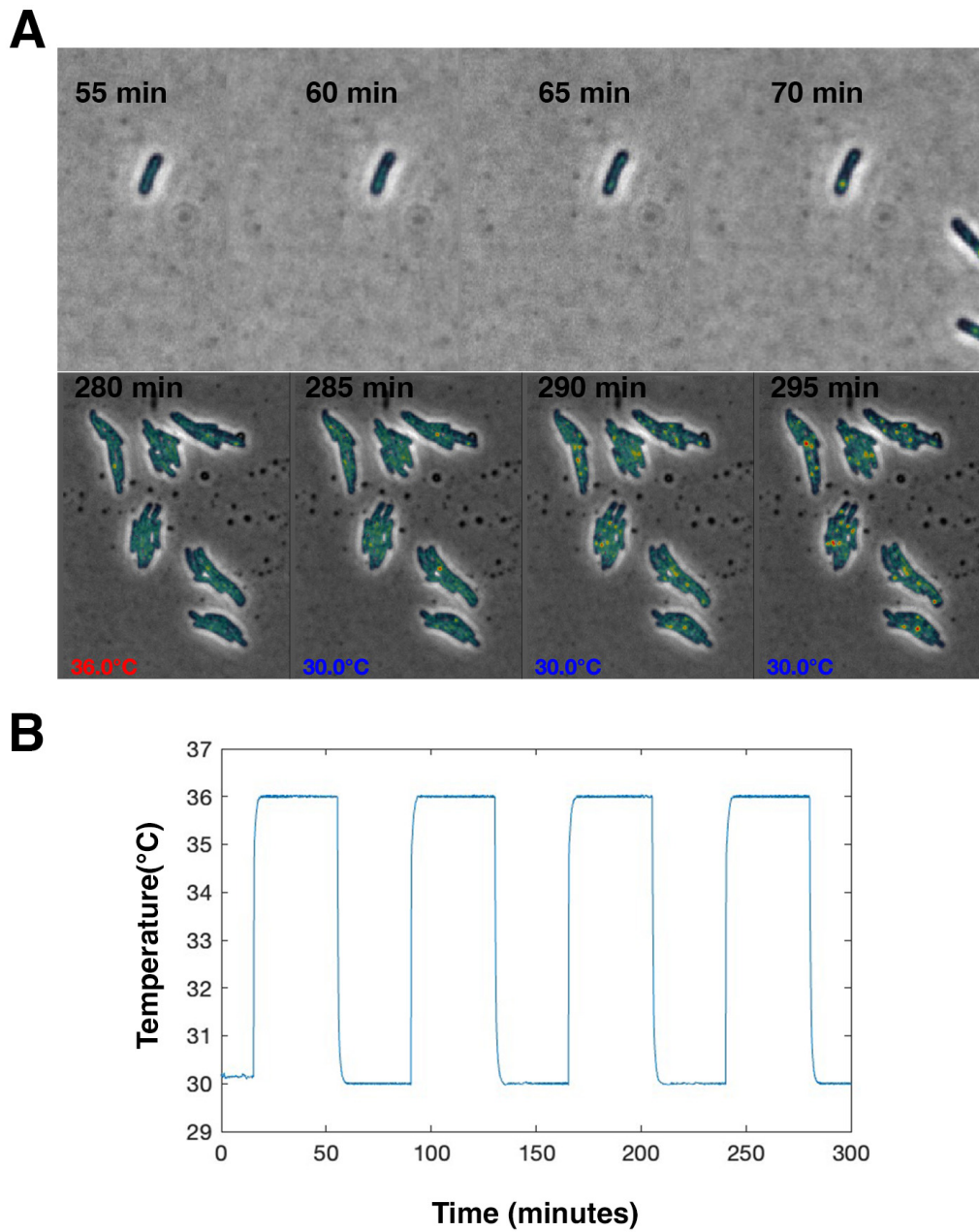


Figure 4.1: **Preliminary results for attempt to synchronize temperature sensitive mutants(*dnaC2*) through temperature oscillation** Panel A: **Timelapse imaging of *dnaC2(ts)* mutant with *YPet-dnaN* fusion.** The false color shows the intensity of the fluorescence within the cells. As the initiation of replication is released from the temperature downshift, the images show formation of YPet-dnaN foci, indicating that replication has been initiated. We also see that puncta are formed at similar time point within 10min of temperature down shift for majority of the cells. **Panel B: Temperature oscillation driving synchrony of initiation of replication.** The temperature is oscillated between 36°C for 40 minutes and 30°C for 35 minutes

be an issue because it may require a couple of temperature cycles before the entire culture is synchronized. (ii) It is difficult to match the driving temperature period with the culture's doubling time, which could lead to imprecise synchrony. (iii) Current experiment is limited to microscopy imaging. Ultimately, we would like to perform biochemical assays, such as LC-MS, to characterize the dNTP pool over time.

To address these complications, implementing a chemostat system with integrated temperature controls could offer a more robust synchronization of the culture. By using a Peltier device coupled with a precision controller, oscillating temperature cycles can be maintained indefinitely, eliminating the overpopulation issue (i) and enabling extended synchronization periods. Furthermore, the chemostat's adjustable dilution rate allows fine-tuning of the culture's doubling time (ii), ensuring precise alignment between the temperature oscillation period and the cell-division cycle for optimal synchrony. Finally, the continuous harvesting capability (iii) permits large-scale cell collection for biochemical assays like LC-MS, overcoming the microscopy limitation and enabling dynamic dNTP pool characterization. This approach not only resolves the experimental constraints of microscopy but also enhances flexibility in measurement tools.

4.4 Bibliography

- [1] D. Huang and P. A. Wiggins, "Metabolic homeostasis, oscillations, and instability," NSF grant proposal, Jan. 2025.
- [2] N. C. Brown and P. Reichard, "Role of effector binding in allosteric control of ribonucleoside diphosphate reductase," *J Mol Biol*, vol. 46, no. 1, pp. 39–55, Nov. 1969. DOI: [10.1016/0022-2836\(69\)90056-4](https://doi.org/10.1016/0022-2836(69)90056-4).
- [3] J. Nordman and A. Wright, "The relationship between dNTP pool levels and mutagenesis in an *Escherichia coli* NDP kinase mutant," *Proc Natl Acad Sci U S A*, vol. 105, no. 29, pp. 10 197–202, Jul. 2008. DOI: [10.1073/pnas.0802816105](https://doi.org/10.1073/pnas.0802816105).
- [4] M. Zhu *et al.*, "Manipulating the bacterial cell cycle and cell size by titrating the expression of ribonucleotide reductase," *mBio*, vol. 8, no. 6, Nov. 2017. DOI: [10.1128/mBio.01741-17](https://doi.org/10.1128/mBio.01741-17).

- [5] S. Gon, J. E. Camara, H. K. Klungsøyr, E. Crooke, K. Skarstad, and J. Beckwith, “A novel regulatory mechanism couples deoxyribonucleotide synthesis and DNA replication in *Escherichia coli*,” *EMBO J*, vol. 25, no. 5, pp. 1137–47, Mar. 2006. DOI: [10.1038/sj.emboj.7600990](https://doi.org/10.1038/sj.emboj.7600990).
- [6] P. L. Birgander, A. Kasrayan, and B.-M. Sjöberg, “Mutant R1 proteins from *Escherichia coli* class Ia ribonucleotide reductase with altered responses to dATP inhibition,” *J Biol Chem*, vol. 279, no. 15, pp. 14496–501, Apr. 2004. DOI: [10.1074/jbc.M310142200](https://doi.org/10.1074/jbc.M310142200).
- [7] P. Reichard, R. Eliasson, R. Ingemarson, and L. Thelander, “Cross-talk between the allosteric effector-binding sites in mouse ribonucleotide reductase,” *J Biol Chem*, vol. 275, no. 42, pp. 33021–6, Oct. 2000. DOI: [10.1074/jbc.M005337200](https://doi.org/10.1074/jbc.M005337200).
- [8] K.-M. Larsson, A. Jordan, R. Eliasson, P. Reichard, D. T. Logan, and P. Nordlund, “Structural mechanism of allosteric substrate specificity regulation in a ribonucleotide reductase,” *Nat Struct Mol Biol*, vol. 11, no. 11, pp. 1142–9, Nov. 2004. DOI: [10.1038/nsmb838](https://doi.org/10.1038/nsmb838).
- [9] H. Xu, C. Faber, T. Uchiki, J. W. Fairman, J. Racca, and C. Dealwis, “Structures of eukaryotic ribonucleotide reductase I provide insights into dNTP regulation,” *Proc Natl Acad Sci U S A*, vol. 103, no. 11, pp. 4022–7, Mar. 2006. DOI: [10.1073/pnas.0600443103](https://doi.org/10.1073/pnas.0600443103).
- [10] P. Reichard, “Ribonucleotide reductases: Substrate specificity by allostery,” *Biochem Biophys Res Commun*, vol. 396, no. 1, pp. 19–23, May 2010. DOI: [10.1016/j.bbrc.2010.02.108](https://doi.org/10.1016/j.bbrc.2010.02.108).
- [11] B. A. Niccum, H. Lee, W. MohammedIsmail, H. Tang, and P. L. Foster, “The symmetrical wave pattern of base-pair substitution rates across the *Escherichia coli* chromosome has multiple causes,” *mBio*, vol. 10, no. 4, Jul. 2019. DOI: [10.1128/mBio.01226-19](https://doi.org/10.1128/mBio.01226-19).
- [12] M. M. Dillon, W. Sung, M. Lynch, and V. S. Cooper, “Periodic variation of mutation rates in bacterial genomes associated with replication timing,” *mBio*, vol. 9, no. 4, Aug. 2018. DOI: [10.1128/mBio.01371-18](https://doi.org/10.1128/mBio.01371-18).
- [13] D. Huang, A. E. Johnson, B. S. Sim, T. W. Lo, H. Merrikh, and P. A. Wiggins, “The in vivo measurement of replication fork velocity and pausing by lag-time analysis,” *Nature Communications*, vol. 14, no. 1, 2023. DOI: [10.1038/s41467-023-37456-2](https://doi.org/10.1038/s41467-023-37456-2).
- [14] D. Bhat, S. Hauf, C. Plessy, Y. Yokobayashi, and S. Pigolotti, “Speed variations of bacterial replisomes,” *Elife*, vol. 11, Jul. 2022. DOI: [10.7554/eLife.75884](https://doi.org/10.7554/eLife.75884).
- [15] S. Cooper and C. E. Helmstetter, “Chromosome replication and the division cycle of *Escherichia coli* B/r,” *J Mol Biol*, vol. 31, no. 3, pp. 519–40, Feb. 1968. DOI: [10.1016/0022-2836\(68\)90425-7](https://doi.org/10.1016/0022-2836(68)90425-7).

- [16] G. Churchward and H. Bremer, "Determination of deoxyribonucleic acid replication time in exponentially growing *Escherichia coli* B/r," *J Bacteriol*, vol. 130, no. 3, pp. 1206–13, Jun. 1977. DOI: [10.1128/jb.130.3.1206-1213.1977](https://doi.org/10.1128/jb.130.3.1206-1213.1977).
- [17] H. Bremer and G. Churchward, "An examination of the Cooper-Helmstetter theory of dna replication in bacteria and its underlying assumptions," *J Theor Biol*, vol. 69, no. 4, pp. 645–54, Dec. 1977. DOI: [10.1016/0022-5193\(77\)90373-3](https://doi.org/10.1016/0022-5193(77)90373-3).
- [18] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014420, 1 Jan. 2022. DOI: [10.1103/PhysRevE.105.014420](https://doi.org/10.1103/PhysRevE.105.014420).
- [19] C. E. Helmstetter and D. J. Cummings, "Bacterial synchronization by selection of cells at division," *Proc Natl Acad Sci U S A*, vol. 50, pp. 767–74, Oct. 1963. DOI: [10.1073/pnas.50.4.767](https://doi.org/10.1073/pnas.50.4.767).
- [20] C. E. Helmstetter and D. J. Cummings, "An improved method for the selection of bacterial cells at division," *Biochim Biophys Acta*, vol. 82, pp. 608–10, Mar. 1964. DOI: [10.1016/0304-4165\(64\)90453-2](https://doi.org/10.1016/0304-4165(64)90453-2).
- [21] C. E. Helmstetter, C. Eenhuis, P. Theisen, J. Grimwade, and A. C. Leonard, "Improved bacterial baby machine: Application to *Escherichia coli* K-12," *J Bacteriol*, vol. 174, no. 11, pp. 3445–9, Jun. 1992. DOI: [10.1128/jb.174.11.3445-3449.1992](https://doi.org/10.1128/jb.174.11.3445-3449.1992).
- [22] D. Bates, J. Epstein, E. Boye, K. Fahrner, H. Berg, and N. Kleckner, "The *Escherichia coli* baby cell column: A novel cell synchronization method provides new insight into the bacterial cell cycle," *Mol Microbiol*, vol. 57, no. 2, pp. 380–91, Jul. 2005. DOI: [10.1111/j.1365-2958.2005.04693.x](https://doi.org/10.1111/j.1365-2958.2005.04693.x).
- [23] S. M. Mangiameli, B. T. Veit, H. Merrikh, and P. A. Wiggins, "The replisomes remain spatially proximal throughout the cell cycle in bacteria," *PLoS Genet*, vol. 13, no. 1, e1006582, Jan. 2017. DOI: [10.1371/journal.pgen.1006582](https://doi.org/10.1371/journal.pgen.1006582).
- [24] D. J. Ferullo, D. L. Cooper, H. R. Moore, and S. T. Lovett, "Cell cycle synchronization of *Escherichia coli* using the stringent response, with fluorescence labeling assays for DNA content and replication," *Methods*, vol. 48, no. 1, pp. 8–13, May 2009. DOI: [10.1016/j.ymeth.2009.02.010](https://doi.org/10.1016/j.ymeth.2009.02.010).
- [25] H. L. Withers and R. Bernander, "Characterization of *dnaC2* and *dnaC28* mutants by flow cytometry," *J Bacteriol*, vol. 180, no. 7, pp. 1624–31, Apr. 1998. DOI: [10.1128/JB.180.7.1624-1631.1998](https://doi.org/10.1128/JB.180.7.1624-1631.1998).
- [26] J. A. Wechsler and J. D. Gross, "*Escherichia coli* mutants temperature-sensitive for DNA synthesis," *Mol Gen Genet*, vol. 113, no. 3, pp. 273–84, 1971. DOI: [10.1007/BF00339547](https://doi.org/10.1007/BF00339547).

- [27] R. Yoshida, K. Korogi, Q. Wu, S. Ozaki, and T. Katayama, “Primosomal protein *prc* rescues replication initiation stress by bypassing the *dnaa*-*dnab* interaction step for *dnab* helicase loading at *oriC*,” *eLife*, vol. 13, A. Aguilera and D. Ron, Eds., RP103340, May 2025, ISSN: 2050-084X. DOI: [10.7554/eLife.103340](https://doi.org/10.7554/eLife.103340).
- [28] S. Maisnier-Patin, K. Nordström, and S. Dasgupta, “Replication arrests during a single round of replication of the *escherichia coli* chromosome in the absence of DnaC activity,” en, *Mol. Microbiol.*, vol. 42, no. 5, pp. 1371–1382, Dec. 2001.