Stochasticity, conflicts, and instability: Biological strategies for optimal growth in a complex environment

Dean Huang

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

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Paul A. Wiggins, Chair
Andrew H. Laszlo
Samu Taulu

Program Authorized to Offer Degree:
Physics

University of Washington

Abstract

Stochasticity, conflicts, and instability:
Biological strategies for optimal growth in a complex environment

Dean Huang

Chair of the Supervisory Committee:
Paul A. Wiggins
Physics and Bioengineering

Life is complicated. Even in the bacterium *Escherichia coli*, cell proliferation is dependent on the maintenance of over 2000 small-molecule metabolites, as well as the synthesis of more than 600 essential proteins. In addition to the sheer scale of this regulatory challenge, the regulatory processes themselves are stochastic in nature. In recent decades, biologists have made great progress in developing a functional map of cellular metabolism, but the dynamics and regulatory behavior of cellular processes remain largely opaque.

In this dissertation, I have aimed to develop a series of mathematical and experimental methods that form a foundational framework for investigating and characterizing cellular dynamics and robustness. I first use the behavior of exponential growth to establish a direct correspondence between stochastic and deterministic cell models, bridging the gap between experimental stochasticity and observed population demographics. Using this correspondence, I then introduce the method of lag-time analysis, which experimentally characterizes the *in vivo* dynamics of replication for an exponentially-growing bacterial population. We use the method to measure replication pauses down to the precision of seconds, replication fork velocity in units of base pairs per second, and temporal oscillations in fork velocity in three evolutionarily-divergent species. Next, I introduce the robustness-load trade-off model, which incorporates stochasticity and an asymmetric

fitness landscape to predict a lower limit for transcription of essential genes, metabolic load balancing between transcription and translation, and a generic overabundance of essential proteins.

In the final chapter, I describe some preliminary work on regulatory feedback dynamics. We predict that regulation is a strategy that the cell uses to maintain robustness, complementary to the overabundance strategy. We also find an oscillatory signature that agrees with the lag-time analysis results, and demonstrate a trade-off between feedback strength, speed of return to equilibrium, and network stability. Although this research is not yet complete, this chapter provides a road map for further analysis and experimental tests. My hope is that the emergent phenomena described in this dissertation provide a solid foundation for future work on cellular dynamics and regulation.

Table of Contents

Page

List of Figures viii

List of Tables xi

Chapter 1: Introduction 1
1.1 The consequences of stochasticity 5
1.2 Lag-time analysis 7
1.3 The Robustness-Load Trade-Off (RLTO) model 9
1.4 Regulatory feedback dynamics 14
1.5 Summary and motivation 16
1.6 Bibliography 19

Chapter 2: Characterizing stochastic cell cycle dynamics in exponential growth . . 23
2.1 Introduction 24
2.2 Results 27
2.2.1 Deterministic model 27
2.2.2 Application to cell cycle dynamics 32
2.2.3 Stochastic model 39
2.2.4 The exponential mean 42
2.2.5 Model correspondence 43
2.2.6 Implications for cell cycle phenomenology 46
2.3 Discussion 47
2.3.1 Applicability of the stochastic model 48
2.3.2 On the applicability of the exponential mean 48
2.3.3 On the significance of stochasticity 49
2.4 Acknowledgments 50
2.5 Supplemental derivations 51

2.5.1 Derivation of the rate equation in the deterministic model 51
2.5.2 Derivation of the solution in the deterministic model 51
2.5.3 Derivation of the PDF and CDF in the deterministic model 52
2.5.4 Derivation of the cumulative creation number in terms of $N(t)$ 53
2.5.5 Derivation of the solution in the stochastic model 54
2.5.6 The exponential mean of a very narrow distribution 56
2.5.7 Derivation of the consistency condition in the stochastic model 56
2.5.8 A generalization of the stochastic model 57
2.6 Bibliography 58

Chapter 3: The *in vivo* measurement of replication fork velocity and pausing by lag-time analysis 61
3.1 Introduction 62
3.2 Results 63
3.2.1 The bacterial cell cycle 63
3.2.2 Lag-time analysis 63
3.2.3 Determination of replisome pause durations 64
3.2.4 Determination of the fork velocity 66
3.2.5 Lag-time analysis reveals *V. cholerae* replication dynamics 66
3.2.6 The measurement of the duration of fast processes 68
3.2.7 The fork velocity is locus dependent 68
3.2.8 Bilateral symmetry supports a time-dependent mechanism 68
3.2.9 The replisome pauses briefly at rDNA in *B. subtilis* 70
3.2.10 Strong, head-on conflicts lead to long pauses 73
3.2.11 Slow retrograde replication in *B. subtilis* 74
3.2.12 Rapid late replication due to genomic asymmetry 75
3.2.13 Fork number determines velocities in *V. cholerae* 75
3.2.14 The fork velocity oscillates in *E. coli* 76
3.2.15 Fork velocity oscillations are observed in three organisms 80
3.3 Discussion 80
3.3.1 The significance of the fork velocity 83
3.3.2 Applications to eukaryotic cells 83
3.3.3 Importance of a model-independent approach 84

3.3.4 Slow growth increases noise . . . 85
3.3.5 Systematic error in datasets . . . 85
3.3.6 Multiple factors determine replisome dynamics . . . 86
3.3.7 dNTP pools regulate the fork velocity . . . 86
3.3.8 Fork-velocity oscillations are observed in divergent species . . . 87
3.3.9 Retrograde fork motion leads to slow replication velocities . . . 87
3.3.10 Replisome pausing . . . 88
3.3.11 Conclusion . . . 88
3.4 Methods . . . 89
3.4.1 Strains used in this study . . . 89
3.4.2 Introduction to marker-frequency analysis . . . 89
3.4.3 Stochastic simulations support the log-slope relation . . . 90
3.4.4 The exponential-mean duration . . . 92
3.4.5 Marker-frequency demography . . . 93
3.4.6 Stochasticity has a minimal effect on the marker frequency . . . 94
3.4.7 Time resolution . . . 94
3.4.8 Fork-velocity resolution . . . 95
3.5 Data and code availability . . . 95
3.6 Acknowledgments . . . 96
3.7 Supplementary tables . . . 97
3.7.1 Bacterial strains used in this study . . . 97
3.7.2 Datasets used in this study . . . 98
3.8 Supplementary methods and derivations . . . 100
3.8.1 Growth media and determination of growth phase . . . 100
3.8.2 Generation of marker frequency data for this study . . . 101
3.8.3 Marker frequency analysis . . . 101
3.8.4 Bilateral symmetry analysis . . . 108
3.8.5 Estimation of average fork number per cell cycle . . . 109
3.8.6 Statistically significant deviations of local fork velocity from the global mean . . . 111
3.9 Supplementary notes on the stochastic simulation . . . 112
3.9.1 Stochastic simulations method . . . 112
3.9.2 Stochastic simulations match the predictions of the log-slope . . . 113

3.9.3 Simulation models . . . 114
3.9.4 Simulation results . . . 116
3.9.5 Re-scaling simulation units to compare to measured data . . . 116
3.9.6 Movies of marker frequency dynamics approaching steady state growth 117
3.10 Supplementary figures and data tables . . . 117
3.10.1 Comparison with GC content . . . 117
3.10.2 *V. cholerae* WT on LB . . . 119
3.10.3 *V. cholerae* WT on M9 fructose . . . 120
3.10.4 *V. cholerae* MCH1 on LB . . . 121
3.10.5 *V. cholerae* MCH1 on M9 fructose . . . 122
3.10.6 *V. cholerae oriR4* on M9 fructose . . . 123
3.10.7 *B. subtilis oriC*-257° on S7 fumarate . . . 124
3.10.8 *B. subtilis oriC*-94° on S7 fumarate . . . 125
3.10.9 *B. subtilis oriN* on S7 fumarate . . . 126
3.10.10 *B. subtilis oriN*-257° on S7 fumarate . . . 128
3.10.11 *B. subtilis rrnIHG* inversion on MOPS glucose w/ Casamino Acids . 129
3.10.12 *B. subtilis rrnIHG* inversion on MOPS glucose – Minimal . . . 130
3.10.13 *E. coli* WT on LB . . . 131
3.10.14 *E. coli* WT on M9 glucose . . . 132
3.10.15 Flattening of marker frequency profile of *V. cholerae* MCH1 in LB . . 133
3.11 Supplementary discussion . . . 133
3.12 Bibliography . . . 135

Chapter 4: Noise robustness and metabolic load determine the principles of central dogma regulation . . . 141
4.1 Introduction . . . 142
4.2 Results . . . 143
4.2.1 Defining the RLTO Model . . . 143
4.2.2 The fitness landscape of a trade-off . . . 147
4.2.3 RLTO predicts protein overabundance . . . 147
4.2.4 RLTO predicts larger overabundance in bacteria . . . 150
4.2.5 Overabundance is a robust prediction . . . 152
4.2.6 RLTO predicts proteins are buffered to depletion . . . 152

4.2.7 Overabundance is observed in a range of experiments 153
4.2.8 RLTO predicts a one-message transcription threshold 154
4.2.9 A lower threshold is observed for message number 154
4.2.10 Prediction of the optimal load ratio 157
4.2.11 Translation efficiency is predicted to increase with transcription . . . 157
4.2.12 Message number also responds to message cost 159
4.2.13 RLTO predicts the yeast global regulatory response 161
4.2.14 Parameter-free prediction of proteome fraction 161
4.2.15 RLTO predicts proteome fractions in eukaryotic cells. 162
4.2.16 RLTO model predicts non-canonical noise scaling 162
4.2.17 Non-canonical noise scaling is observed in yeast 163
4.2.18 Prediction of noise from protein-message relation 163
4.3 Discussion 164
4.3.1 Understanding the rationale for overabundance 164
4.3.2 Implications of overabundance for inhibitors 167
4.3.3 Implications for non-essential genes 167
4.3.4 Load balancing 168
4.3.5 Comparisons to previous work 168
4.3.6 Implications of noise 169
4.3.7 The principles that govern central dogma regulation 169
4.4 Data availability 170
4.5 Acknowledgments 170
4.6 Supplemental analysis of the RLTO model 170
4.6.1 Detailed description of the noise model 170
4.6.2 The derivation of the RLTO growth rate 174
4.6.3 Message number and translation efficiency optimization 178
4.6.4 Modifications of the RLTO model for bacterial cells 179
4.6.5 Arrest probability 180
4.6.6 Discussion of *E. coli* essential genes below the one-message-rule threshold 181
4.6.7 Measurements of the load ratio 181
4.6.8 Increased protein load analysis 183
4.6.9 Analysis of translational limits. 186
4.6.10 Estimate of the message cost and metabolic load 187

4.6.11 Prediction of the proteome fraction . . . 189
4.7 Analysis of alternative models . . . 191
4.7.1 Threshold (RLTO) model . . . 191
4.7.2 Model 2: Slow-Growth model . . . 191
4.7.3 Model 3: Symmetric model . . . 192
4.7.4 Conclusions from fitness-landscape analysis . . . 192
4.8 Quantitation of central dogma parameters for the one-message-rule . . . 194
4.8.1 Selection of central dogma parameter estimates . . . 194
4.8.2 Quantitative estimates of central dogma parameters . . . 197
4.9 Supplemental analysis of Noise-Protein-Abundance Relation in Yeast . . . 198
4.9.1 Estimating *protein number* (μ_p) for the noise analysis . . . 198
4.9.2 Empirical models for yeast gene expression . . . 200
4.9.3 Supplemental analysis of gene expression noise . . . 203
4.10 Comments on the Hausser et al. analysis . . . 207
4.11 Bibliography . . . 209

Chapter 5: Metabolic homeostasis, oscillations, and instability . . . 216
5.1 Introduction . . . 216
5.1.1 Feedback-based regulatory control . . . 217
5.1.2 Background on dNTP metabolism . . . 218
5.1.3 Significance . . . 218
5.2 Aim 1: Modeling—Identify the determinants of regulatory oscillations and instability . . . 221
5.2.1 Sub-aim 1.1: What are the determinants of regulatory oscillations and instability? . . . 223
5.2.2 Sub-aim 1.2: Explore the putative role of small RNA and proteins in stabilizing regulation . . . 225
5.3 Aim 2: Test the dNTP-oscillation model by characterizing fork-velocity oscillations . . . 226
5.3.1 Sub-aim 2.1: Increase the temporal resolution and precision of fork velocity measurement . . . 229
5.3.2 Sub-aim 2.2: Test role of RNR in limiting fork velocity by inhibition . 229
5.3.3 Sub-aim 2.3: Test feedback-regulation model for oscillations . . . 230

5.4 Aim 3: Test transcriptional-regulation model by characterizing cell-cycle-dependent transcription . . . 232
5.4.1 Sub-aim 3.1: Test RNR-expression oscillation hypothesis by fluorescence imaging . . . 234
5.4.2 Sub-aim 3.2: Optimize cell synchronization protocol to test hypotheses using bulk assays . . . 235
5.4.3 Sub-aim 3.3: Test *nrdAB*-transcription oscillation hypothesis . . . 236
5.5 Aim 4: Test metabolite oscillation model by direct cell-cycle dependent measurement of levels . . . 238
5.5.1 Sub-aim 4.1: Test NTP-oscillation model by LC-MS . . . 238
5.5.2 Sub-aim 4.2: Test dNTP-oscillation model by LC-MS . . . 240
5.6 Outlook . . . 241
5.7 Bibliography . . . 242

Appendix A: An interbacterial DNA deaminase toxin directly mutagenizes surviving target populations . . . 249

List of Figures

Figure Number		Page
1.1	Visual representation of biochemical pathways.	2
1.2	Functional map of metabolism.	3
1.3	Correspondence between deterministic and stochastic model.	6
1.4	Fork velocity oscillations.	10
1.5	Asymmetric fitness landscape in the RLTO model.	12
1.6	Central dogma regulatory principles.	13
1.7	Feedback-based regulatory control.	15
1.8	Feedback models and solutions.	17
2.1	Schematic for positioning and timing of events during the *E. coli* cell cycle.	26
2.2	Creation and annihilation of cell quantities.	29
2.3	Cell Age PDF.	33
2.4	Numbers of cell quantities in an exponential culture.	36
2.5	Schematic of the stochastic model of the cell cycle.	40
2.6	Correspondence between deterministic and stochastic model.	45
3.1	Lag-time analysis.	65
3.2	Replication fork dynamics in *V. cholerae*.	67
3.3	*B. subtilis* fork dynamics and transcriptional conflicts.	72
3.4	Reducing fork number increases fork velocity.	77
3.5	Observed oscillations are consistent with a temporal mechanism.	79
3.6	Fork velocity oscillations.	81
3.7	Analysis of simulated data.	91
3.8	Schematic of stochastic simulation model.	115
3.9	Relative fork velocity and % GC skew as a function of position.	118
3.10	Marker frequency data for *V. cholerae* WT on LB.	119
3.11	Marker frequency data for *V. cholerae* WT on M9 fructose.	120
3.12	Marker frequency data for *V. cholerae* MCH1 on LB.	121

3.13 Marker frequency data for *V. cholerae* MCH1 on M9 fructose. 122
3.14 Marker frequency data and tabulated results for *V. cholerae oriR4* on M9 fructose. 123
3.15 Digitized marker frequency data and tabulated results for *B. subtilis oriC*-257° on S7 fumarate. 124
3.16 Digitized marker frequency data and tabulated results for *B. subtilis oriC*-94° on S7 fumarate. 125
3.17 Digitized marker frequency data and tabulated results for *B. subtilis oriN* on S7 fumarate, fit with two-slope model. 126
3.18 Digitized marker frequency data and tabulated results for *B. subtilis oriN* on S7 fumarate, fit with pause model. 127
3.19 Digitized marker frequency data and tabulated results for *B. subtilis oriN*-257° on S7 fumarate. 128
3.20 Digitized marker frequency data and tabulated results for *B. subtilis rrnIHG* inversion on MOPS glucose w/ Casamino Acids. 129
3.21 Digitized marker frequency data and tabulated results for *B. subtilis rrnIHG* inversion on MOPS glucose – Minimal. 130
3.22 Marker frequency data for *E. coli* WT on LB. 131
3.23 Marker frequency data for *V. cholerae* WT on M9 glucose. 132
3.24 Flattening of marker frequency profile of *V. cholerae* MCH1 in LB. 133

4.1 Features of the RLTO model. 145
4.2 RLTO model prediction of the fitness landscape. 149
4.3 RLTO prediction of overabundance. 151
4.4 The one-message-rule. 155
4.5 Load balancing for three model species. 158
4.6 RLTO prediction of message number. 160
4.7 Comparison of noise models in yeast. 165
4.8 Central dogma regulatory principles. 166
4.9 The protein abundance is approximately gamma distributed. 173
4.10 Cell cycle arrest probability. 185
4.11 Increased protein cost decreases optimal translation efficiency. 185
4.12 Exploring the mathematical mechanism of overabundance. 193
4.13 Transcription in three model organisms. 199
4.14 Fit to rescale fluorescence intensity to protein number. 205

4.15 Yeast noise fit against canonical noise model, with a noise floor. 205
5.1 Characteristics of regulatory feedback control. 219
5.2 Schematic and model of nucleotide metabolism in *E. coli*. 220
5.3 Homeostatic model results. 224
5.4 Measuring replication fork velocity by lag-time analysis. 227
5.5 Evidence of cell-cycle-dependent expression. 233
5.6 dNTP fluctuations. 240

List of Tables

Table Number Page

2.1 A summary of the model notation. 28
2.2 The effect of mutants on the doubling time T and C period duration C of an exponential culture. 46

3.1 Fork number and velocities under different growth conditions. 69
3.2 Velocity oscillation characteristics for different bacterial species and growth conditions. 82
3.3 Strains used in the study. 97
3.4 Datasets used in the study 99
3.5 Stochastic simulation results. 116

4.1 Summary of RLTO parameters. 148
4.2 Below-threshold essential genes identified in *E. coli*. 182
4.3 Central dogma parameters for three model organisms with detailed references. 198

Acknowledgments

Choosing to work with my advisor, **Paul Wiggins**, has been one of the best decisions I have ever made. Despite being way smarter than me, Paul always treated me as a scientific equal, giving me the space and opportunities I needed to grow and learn. I'll probably never have another boss who makes me laugh as much as he has over the past six years. I couldn't have asked for a more brilliant, patient, hilarious, supportive, and inspiring person to be my scientific role model, mentor, and friend.

Over the years, one of the pieces of advice that Paul gave to me most often was: "Don't let the perfect be the enemy of the good." Thus, any imperfections in this dissertation are simply a reflection of my respect for Paul's wisdom, whether he likes it or not. Thanks, Paul!

Next, I want to thank the only two other people who will likely ever read this dissertation in full: **Andrew Laszlo** and **Samu Taulu**. Andrew has been the source of many illuminating scientific conversations, and has provided a lot of useful advice, both for the preparation of this dissertation and for research in general. Without Samu's kindness and generosity with his time, I would likely still be struggling to find someone willing to read this gargantuan dissertation.

Beth Traxler has taken on many roles during my time here at UW: supervisory committee member, Graduate School Representative, lab-space provider, biology encyclopedia, bus-buddy, fellow dim sum enthusiast, and one of the wisest mentors I know. Her support has been critical for keeping our whole lab on the rails.

Working and interacting with **Jason Detwiler** has always been a delight. From the department orientation, to teaching E&M together, to my dissertation defense, Jason's welcoming demeanor and intellectual agility are characteristics I aspire to emulate.

I'm also grateful to the other members of my supervisory committee: **Armita Nourmohammad**, **Masha Baryaktar**, and **Miguel Morales**, for their time, insightful questions, and flexibility with scheduling.

External to my supervisory committee, there were a few professors here at the **University of Washington** that made a tremendous positive impact on my graduate career. It was a stroke of luck that I was assigned to teach upper-division lab classes with **David Pengra**, who has proven to be one of my strongest advocates and closest confidants in the department. He has taught me so much about physics, electronics, teaching, and life. Our conversations were thought-provoking and entertaining, and I deeply admire his compassion for everyone he interacts with. **Jiun-Haw Chu** and **Subramanian Ramachandran** were also a pleasure to teach with. Their willingness to answer my questions about their career paths helped me with forging my own. **Marcel den Nijs** was helpful in making sure that my progress through graduate school went on at a steady pace. I also appreciate that he had enough confidence in me to ask me to tutor for PHYS5XX—it was a very rewarding experience.

Professors aside, I would like to thank an underappreciated group of people in the UW physics department: **the staff**. Without them working behind the scenes doing administrative work and building maintenance, and also acting as the face of the department at the front desk and advising, absolutely nothing would get done in this department. From the very first time I visited UW, I knew that I would like our graduate program advisor, **Catherine Provost**. Her care for us students was evident from every interaction. Her guidance each step of the way and her generous supply of chocolate were both critical to my completion of graduate school. **Steven Troy** has by far the coolest office in the department—it houses the collection of all the physics lecture demos. I have enjoyed chatting with my fellow 3D-printing enthusiast about various clever demos, and I appreciate his infinite patience while we were coming up with demo ideas for my defense.

It takes a special kind of educator to have a lasting impact on a student long after the student has graduated from their institution. From **New York University**, a few have stood out as being especially critical in my development as a scientist. **Leif Ristroph** introduced me to the joys of experimental physics. I hope that I've inherited even a little of his kindness, creativity, and scientific ingenuity after working with him. **Charles Seife** once said to me, "You're going to have a *veeery* interesting career." This was after I explained that the reason I stuck a bunch of utensils in my soda and was jostling them around was that I don't like carbonated beverages—I wanted to make the drink go flat, so I added the utensils to increase the surface area and hence the increase the nucleation sites for the bubbles. My hope is to make Charles' prediction come true. **David Grier**'s infectious enthusiasm in my first semester of college was what made me first fall in love with physics. In my second semester, **Andrew Kent** gave me the first big confidence boost regarding my aptitude for physics by being such an effective teacher. **Andrew MacFadyen** taught me about the power of a computational approach to doing scientific research. **Paul Chaikin** taught by far the hardest class I ever took at NYU. The problems he came up with gave me a taste of theoretical research, where finding an answer might very well be impossible, and he also introduced us to a certain useful guide to semiconductor physics. **Bill LePage**, the undergraduate physics advisor, was a strong advocate for me and always had something interesting to say.

At NYU, I was extremely lucky to get a job tutoring at the **University Learning Center**. It was where I learned the power of teaching, and gained experience tutoring hundreds of students. I really appreciate **Soomie Han** for hiring me and for hiring so many of my favorite people. Soomie and **Caroline Cristal** were great bosses. They did an amazing job of creating a healthy work environment for all of us.

Three of my high school teachers from **Taipei American School** deserve a special shout-out for starting me on my scientific journey. **Peter Morgan** was my first scientific

role model. He taught me the importance of having both breadth and depth to my knowledge. I will never forget his organic molecule dances. **Nyoli Connor** taught my two-person IBHL Math class, and showed me that learning certain things just takes time. A lot of time. Our after-school calculus marathons helped me build up my stamina for tackling difficult math problems. I hope the duck song tradition still lives on. **Jude Clapper** was the first person to introduce me to scientific research and the camaraderie that comes with being in a lab.

Speaking of lab, I couldn't have asked for a more friendly, welcoming, and supportive group of colleagues (and friends) than the **Wiggins lab**. It's difficult to condense into a few lines the impact that **Isaac Shelby** had on my time in graduate school, so the following list of roles he played is not exhaustive: Mentor, friend, grad-school-commiseration expert, fellow musketeer (see Sec. AIMdyn), physics tutor, crossword buddy, and chess grandmaster. **Han Kyou (James) Choi** and I have had endless enjoyable debates about everything under the sun (and also about the sun itself!). His unconventional thinking and genuine compassion will never cease to *fascinate* me. I hope we both continue to grow and change each other's minds. **Dani Koch** and I like to say that our conversations tend towards the doom and gloom, but her personality and mien are the opposite of that. She brings some much needed sunshine to our windowless basement lab. **Kevin Cutler** has a special kind of energy (in the physics sense of the word) that I wish I could bottle up and keep around me at all times. **Teresa Lo** and I joined the lab at the same time, and she has helped me out more times than I can count. **Zeeshawn Kazi** is one of the best listeners I know, and I admire his effortless compassion.

The undergraduates in our lab have also been fantastic. I had a lot of fun chatting with **Brandon Sim** (my first protégé), **Joey Turnbull** (egg boi), **Nandor Marosan** (Maro-san!), **Maddie Walter** (wee-badger), **Ivy LeGassick** (erg-chair aficionado), and **Brighton Reed** (my final protégé).

Although many of my scientific collaborators are acknowledged in the ensuing chapters, I would like to thank the following for their special contributions: **Anna Johnson** and **Houra Merrikh** for working closely with me to obtain stellar sequencing results and to shore up my microbiology knowledge. **Marcos de Moraes**, **Fosheng Hsu**, **Brook Peterson**, and **Joseph Mougous** for including me in the DddA project and for trusting in my work despite my inexperience. Joseph in particular has taught me the importance of advocating for myself and having frank discussions about important aspects of science, such as authorship and journal selection. **Bryan Andrews** for his help with obtaining λ phage. **Amy Schaefer** for sharing her *Vibrio cholerae* expertise and her lab's chemicals.

My graduate school experience was made unique through my work experience at **AIMdyn, Inc.** Somewhere between academia and industry, it was the perfect training ground for gaining some experience outside the ivory tower. Working there taught me how transferable my PhD skills are. I also got the chance to lead the AIMdyn side of the Enabling Confidence project, which certainly enabled my confidence. I am extremely grateful to **William Redman** (see Sec. William Redman for more details) for referring me to the company. The greatest perk of working at AIMdyn was definitely meeting and interacting with **Maria Fonoberova**. She is a wonderful mentor and exactly the kind of boss I would like to be. She taught me many interpersonal skills that I know will help me tremendously with my career. It's hard not to be impressed by the genius of **Igor Mezić**, but even more impressive is his kindness. Despite having a relentless schedule, he always made time to hear us out whenever we had new research ideas. I also had the pleasure of working with **Marco Pravia**, who has an impressive intuitive feel for physics that I admire. One of the best team research experiences I've had so far has been with **Jordan Garrett** and **Isaac Shelby**. Whether we were being The Three Musketeers or The Three Stooges, the complementarity was perfect. I also really enjoyed working with and learning from **Chris DuPre**, **Ryan Mohr**, and **Zlatko Drmač**.

On the more personal side, I have been blessed with an incredibly strong and wide support network of people that I am lucky enough to call friends. I met many of these amazing people in grad school, and they have helped to fight off the Seattle gloom over the past six years. It is weird for **Wan Jin Yeo** to appear in the friends section, since she's almost certainly a long-lost sibling of mine. I still reminisce fondly about our many shenanigans, such as the Great Screen Door Heist of 2020, for which we are still on the run. $\hat{\vec{8}}$ for life. **Vasilis Niaouris**, my dissertation writing buddy, has been a constant source of entertainment. I hope we continue to build worlds together, both in the imagination and in real life. **Arnab Manna** (the diplomat) and **William Atienza** (the golden retriever) were an absolute blast to hang out with.

One of the underrated aspects of graduate school is the existence of sporadic hallway conversations and weekly teaching with lovely people. I appreciate all the conversations I had with **Chris Thomas** (biophysics buddy), **Tahiyat Rahman** (tenacity personified), **John Cenker** (pet raccoon), **Alex Kato** (wise office mate), **Felicia Tsai** (explainer of modern parlance), **Al Snow** ("I'm all" fed up!), **Emmett Hough** (a worthy heir of my first-year office desk), **Ella Henry** (the tea!!!), **Mikael Kovtun** (Mössbauer Mike), **Ryan Lanzetta** (NYC represent), **Dan Matthias** (the wisest itinerant salmon sage), and **Mike Smith** (best peer mentor). I also want to thank all my **students** (little rascals) for teaching me while I was teaching them.

I feel very fortunate that I have made lifelong friends whose support reaches across vast stretches of spacetime. Even if the frequency of meetups has been redshifted with the expansion of the universe, the strength of the bonds have not been broken. **William Redman**, goofball extraordinaire, super-rememberer, Snapchat iconoclast, chief curator of the number one Bojangles-sponsored art collective, ideal roomie, job referrer, pseudo-fam provider, connoisseur of my questionable inventions, co-conspirator, and dad of my favorite dog. How many bests could a best friend best if a best friend could best bests? To the victor,

the potatoes! My beloved ULC friends: **Chappel Sharrock** (birthday twin, telepath, master of cross products), **Kobi Dent** (bean counter, wordplay enthusiast, final boss in all games), **Mats Thijssen** (the best intellectual rival, you snus you lose, my favorite archnemesis), **Chish Malata** (the ultimate vinecyclopedia), **Anthony Kuo** (too many veggies!), and **EJ Kim** (VR visionary). Let's do another game day soon! Also a shout-out to the goons: **Ali Hassan**, **Shiloh Pitt**, **Kevin Zhuang**, and **Sanchit Chaturvedi**.

Maaz Ul-Haq was the cool older brother I always wish I had. Gone too soon, but not forgotten.

From before college: **Bethany Shieh** is the first friend I felt truly close with, and I am so glad she moved to Seattle. **Kevin Lin** is one of my oldest friends (since fifth grade!), and was a great roommate. **Grace Lee**, **Kenneth Chen**, **Morgan Chien-Hale**, **Daniel Hsieh**, **Penny Lin**, **Annie Mao**, **Valerie Lin**, and **Sara Chen** are the reason why my sense of humor is so immature. **Alysia Lo**, **Timothy Chen**, **Ray Chen**, **Grace Chen**, **Claire Peng**, **Eric Syu**, and **Rachel Tan** taught me the meaning of infinity.

Of course, if you follow the turtles all the way down to my fundamental base of support, you will end up finding my family. My **parents**, who have provided me with life, sustenance, a home, many unforgettable experiences, an incredible education, and unconditional love. They have given me my most useful icebreaker through their choice of names for me and my sisters: **Sean**, who taught me how to make friends, and **Jean**, who is the only one I can talk science with in the family. I am lucky that my cousins, **Annie Lin** and **Chanel Lin**, and my aunt **Sophie Fang** were stationed in Seattle. They were a piece of **Taiwan** away from Taiwan. My **grandmas** and the rest of my **extended family** have also helped me feel connected to my roots.

I have been the lucky-luckiest person alive to have met and fallen in love with **Nicole Hardman**. Her unflagging support, patience, sweetness, creativity, silliness, comfort, and love have kept me happy and healthy during the last few years of my PhD. We have built

up a lovely little family of our own, full of emotional support animals: **Parsley** (the babiest pig), **Nutmeg** (sunshine awoooo!), **Coco** (the softest rabbit in the West), and **Fig** (FICO score of 850). I look forward to our continued growth together!

Dedication

to exponential growth,
and all who make it possible

Chapter 1

Introduction

Life is complicated [1]. From filing taxes to navigating interpersonal relationships, we all know this to be true at an intuitive level. Moreover, beneath the surface of every living organism, even down to the smallest bacterium, there lies a rich, hidden world of unimaginable complexity. This is the world of biochemistry, where stochasticity, molecular competition, and regulation reign supreme.

For the uninitiated, the complexity of biochemistry is perhaps best conveyed by a diagram of cellular metabolism, as shown in Fig. 1.1 [2]. Metabolism is the set of all chemical reactions that sustain life. Each arrow in Fig. 1.1 represents a chemical reaction between different metabolites, the collection of molecules that life utilizes. Although the *functional map* of metabolism is well-established (see Fig. 1.2), the *dynamics* and *regulatory behavior* are not [2]. As physicists, we hope to better understand these metabolic dynamics by mathematically modeling them.

As stipulated by statistical mechanics and chemical kinetics theory, every one of these chemical reactions is fundamentally probabilistic [3]. Furthermore, every chemical reaction and environmental fluctuation causes a change in the concentrations of relevant metabolites, leading to a constantly shifting landscape of available molecules that life must navigate [3]. Regulatory pathways cause feedback, introducing another layer of self-referential complexity [3, 4]. These factors all combine to make the mathematical formalization of cellular dynamics a formidable task. Although the work done in this dissertation is far from solving that challenge, my hope is that it provides a few stepping stones to help pave the path towards it.

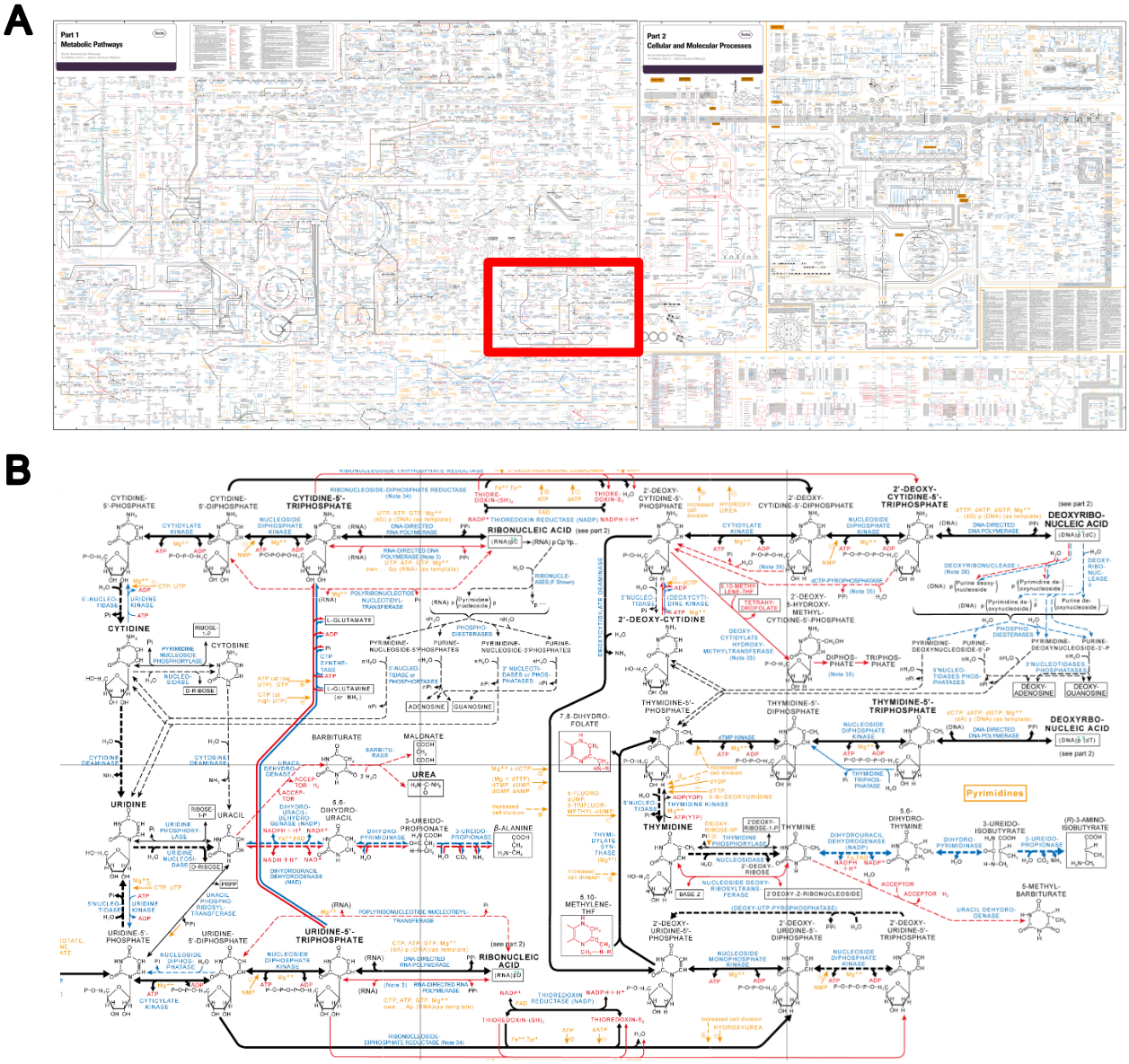


Figure 1.1: **Visual representation of biochemical pathways. Panel A: Complete chart in two parts.** These charts of all the known biochemical pathways used by life are designed by Gerhard Michal and distributed by Roche [2]. Each arrow represents a chemical reaction between different metabolites, which are represented with structural formulae and common names. The red box denotes the collection of pathways that are shown in Panel B (expanded for clarity and to demonstrate complexity). **Panel B: A subset of pathways associated with nucleotide metabolism.** This collection of pathways is responsible for pyrimidine metabolism (purine metabolism shown elsewhere).

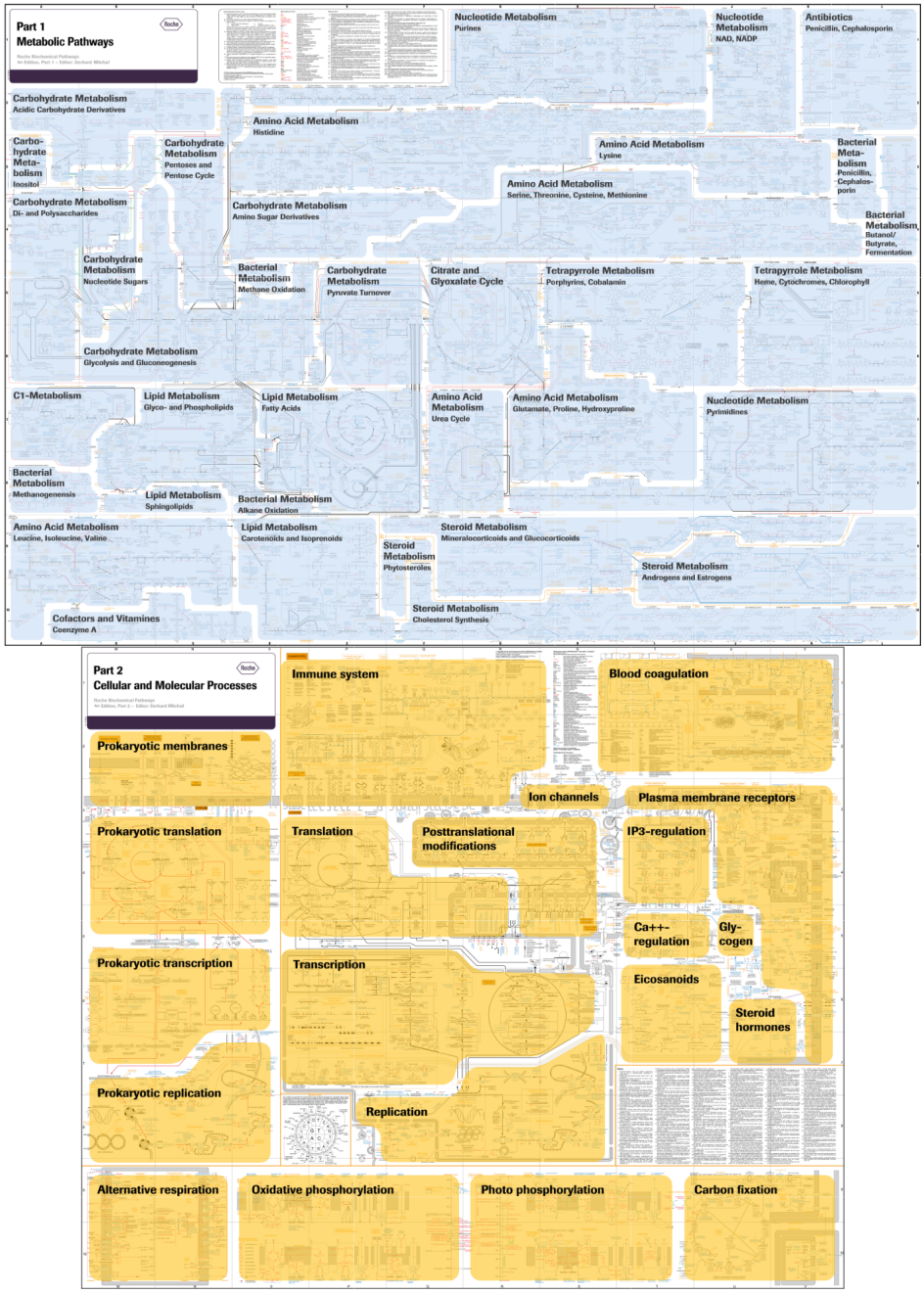


Figure 1.2: **Functional map of metabolism.** These charts of all the known biochemical pathways used by life are designed by Gerhard Michal and distributed by Roche [2]. The pathways from Fig. 1.1A, grouped by function. The upper chart shows the metabolic pathways. The lower chart shows the cellular and molecular processes.

In particular, I have aimed to develop a series of mathematical and experimental methods that form a foundational framework for investigating and characterizing cellular dynamics and robustness [5–7]. The development of these methods were driven specifically by the following scientific questions, which I hope to address in this dissertation: (i) How do we reconcile stochastic and deterministic models for different types of experiments, particularly during exponential growth? (ii) How can we experimentally measure cellular timing and the dynamics of molecular machines in living cells, particularly for DNA replication, one of the most fundamental processes for all living organisms? (iii) What strategies do living organisms use to maintain robustness under perpetually-changing and noisy conditions? (iv) How might metabolic regulation lead to unexpected cellular phenomena, and how can we mathematically model them?

Before we launch into the rest of this introduction, a few words about its structure might be beneficial: A side effect of interdisciplinary research is that the required background knowledge covers a lot of potentially disparate scientific fields. Furthermore, the relevant background for each research project often does not intersect, and a reader from a physics background might disagree with someone from a biology background on what constitutes common knowledge. Thus, rather than a single background chapter that lays out all knowledge needed to understand the results that appear in the later chapters, I have instead included the relevant background in each separate chapter. The purpose of this introduction is therefore to tie together these seemingly disjoint projects into a single coherent narrative, while also providing an overview of the results that can be expected in each chapter. The reader of this dissertation may find it helpful to think of this introductory chapter as a (potentially nonlinear) road map of the next 200+ pages. Some parts may be intriguing, others may be confusing (despite my best attempts at clarity), but my recommendation is to follow wherever one's curiosity leads, and to not treat the chapter layout as a prescribed way to read this dissertation.

1.1 The consequences of stochasticity

Given the inherently probabilistic nature of all cellular processes, one of our first goals was to reconcile this stochasticity with the deterministic (stochasticity-free) models that biologists have successfully used for decades to explain their experimental results [8, 9]. The prime example of a deterministic model is the original 1968 Cooper-Helmstetter model of the *Escherichia coli* cell cycle, which modeled the cell cycle as a series of precisely-timed, deterministic stages [8]. The deterministic assumption of the model has proven successful for modeling many experimental results, leading some to argue that stochasticity is not biologically significant [10]. Although there have been many methods developed since for characterizing cell cycle dynamics [9], none prior to our own work in Ref. [5] (described in Chapter 2) have incorporated the significant level of stochasticity seen in cell cycle timing. Since the steady-state behavior of populations given sufficient nutrients is exponential growth, we chose to focus on this stage of population growth to begin our investigations.

In Chapter 2, I describe how we exactly solve a set of deterministic and stochastic models that utilize the mathematical features of exponential growth [5]. We first solve the deterministic model to obtain the cellular demographics of an exponentially growing culture. We then solve the inverse problem: Given a set of observed demographics, how can cell cycle state timing be inferred based on the deterministic model? Next, we introduce and solve for the demographics of an exponential culture using the more realistic stochastic model, with cell-state lifetimes represented by probability distributions. We then define an exponential mean, which allows us to establish an exact correspondence between the demographics of the deterministic and stochastic models. In particular, the exponential means of state ages in the stochastic model correspond exactly to effective deterministic state ages, as shown in Fig. 1.3. This equivalence resolves the long-standing incongruity between the success of deterministic models and the fundamentally stochastic nature of cellular processes. Some nontrivial consequences of this exponential-mean equivalence in the context of interpreting experimental data are also discussed.

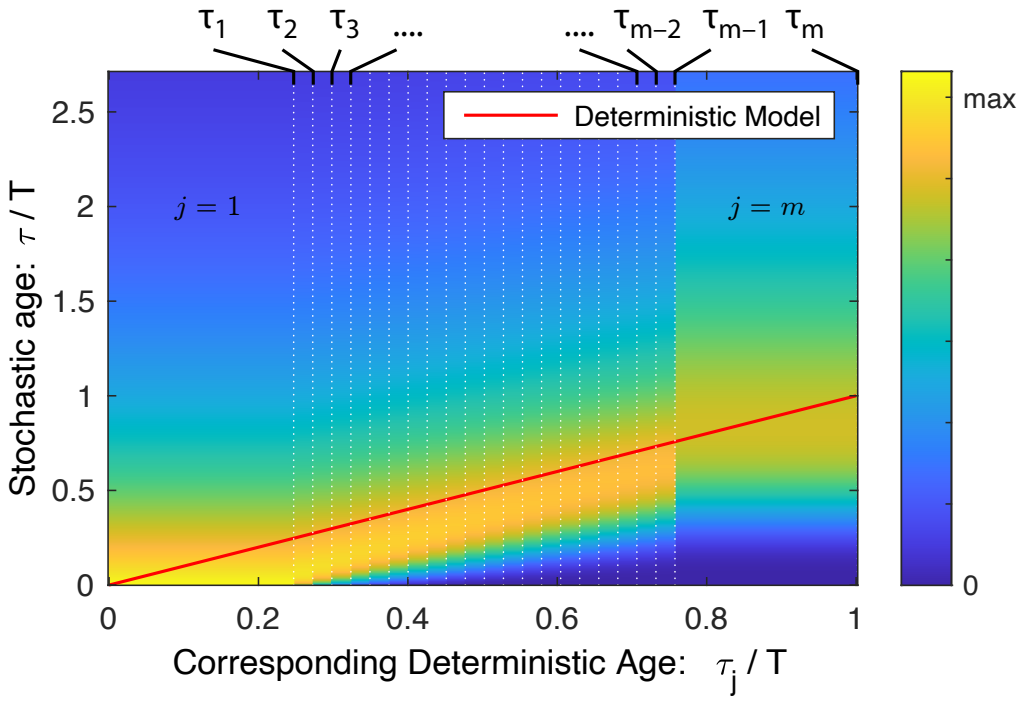


Figure 1.3: **Model correspondence.** Figure reproduced from Ref. [5]. The deterministic and stochastic models generate identical statistics in exponential growth once a suitable correspondence is defined between cell state j in the stochastic model and cell age τ in the deterministic model. State j in the stochastic model corresponds to age interval $(\overline{\tau}_{j-1}, \overline{\tau}_j]$ in the deterministic model, which is represented by the red line. In the stochastic model, the PDF of age as a function of state j is shown. Qualitatively, the age in the deterministic model tracks with the mode of the age in the stochastic model.

One key result from Chapter 2 is that, due to the exponential mean, it is not possible to distinguish between (i) slow overall population growth, and (ii) a fraction of the cell population arresting (ceasing growth), a common result of DNA replication conflicts [11–14]. This fact has implications for the interpretation of two distinct classes of experimental approaches in cellular biology: The first class consists of unsynchronized approaches, where a snapshot of an entire exponentially growing culture is analyzed to generate statistics (e.g., [15–17]). The second class consists of synchronized approaches, where the population is either first physiologically synchronized (e.g., [18]), investigated at a single-cell level, or synchronized via post-processing (e.g., [19, 20]). The results from these two classes can differ drastically if cell-to-cell variability is not taken into account using our stochastic approach. In Chapter 3, described in the next section, this nuanced interpretation plays a key role in understanding the data obtained from Next-Generation DNA Sequencing (NGS), an example of an unsynchronized approach [21]. The results in Chapter 2 also allow us to calculate the growth rates that are used later on in the Robustness-Load Trade-Off (RLTO) model of Chapter 4.

1.2 Lag-time analysis

DNA replication is one of the most fundamental processes for all living organisms. Every cell needs to make a complete copy of its genome, so that each of its daughter cells can have all the genetic information needed to survive. An equally fundamental process, transcription, is responsible for creating various RNA molecules, which are used either as templates for protein production or as necessary molecules for other cellular functions [22]. During replication, individual deoxynucleotide triphosphates (dNTPs) are incorporated into the DNA backbone by molecular machines called DNA polymerases, which are a family of enzymes responsible for DNA synthesis [22]. These DNA polymerases share the same DNA substrate as RNA polymerases, which are responsible for transcription. The production of both DNA and RNA is essential to proper cell function, but the molecular machines that produce them can have

antagonistic interactions as they travel along the shared DNA substrate [23, 24]. Add in a layer of complexity due to the stochastic nature of each step that the molecular machines takes, and it becomes clear why replication-transcription conflicts are still a highly active area of research [25–31].

Although the pausing of molecular motors has been experimentally investigated *in vitro* (not in living cells) [13, 14], an analogous experimental method *in vivo* (in living cells) has proven difficult [12]. In Ref. [6] (described in Chapter 3), we developed lag-time analysis, a method that measures *in vivo* replication pausing, down to the precision of seconds, and rate of DNA replication, in units of kilobases per second. Like the exponential mean described in Chapter 2, this approach uses features of exponential growth to provide information about individual cell states in an exponentially growing unsynchronized population (NGS marker frequency data) [5]. In particular, lag-time analysis uses exponential growth as a stopwatch to provide temporal replication dynamics and precise cell timing. Since replication is a key metabolic process, and one of the central hubs of the biochemical landscape [2], we believe this experimental method is a useful tool for revealing important metabolic regulatory dynamics (more details in Chapter 5).

In Chapter 3, I describe the approach of lag-time analysis, the various hypotheses we tested with it, and the quantitative results we obtained. To verify that our method was able to accurately recover known parameters, we implemented a stochastic Gillespie simulation [32]. We then used lag-time analysis to investigate three model bacterial systems: *Bacillus subtilis*, *Vibrio cholerae*, and *Escherichia coli*. For *B. subtilis*, we looked at replication-transcription conflicts in various growth conditions and genetic mutants, finding pauses down to the precision of seconds. For *V. cholerae*, a model bacterium with two separate chromosomes, we investigated both the effects of having multiple simultaneous replication forks on the overall speed of replication, and the locus- and time-dependent behavior of fork velocity. For the simplest model organism, *E. coli*, we find a clear oscillatory time-dependent signature in the fork velocity across various growth conditions. By the concept of lag time, we are able to directly compare all three evolutionarily-divergent species using the same real temporal units.

Comparing fork velocities over time, we find that all three organisms show the same signature temporal oscillations, each with their own species-dependent frequencies, consistent across different growth media, as shown in Fig. 1.4. We attempt to explain and mathematically model these temporal fork velocity oscillations in Chapter 5.

In the formulation of lag-time analysis, the results of Chapter 2 describing the effects of stochasticity were essential. Although lag-time analysis provides an experimental platform to investigate stochastic *in vivo* dynamics and replication conflicts, it does not theoretically explain how the cell remains robust under such circumstances. What strategies do cells use to survive the constantly changing metabolic landscape? We attempt to address this with the RLTO model in Chapter 4.

1.3 The Robustness-Load Trade-Off (RLTO) model

It is well-known that cells optimize their protein levels in order to maximize cell fitness (i.e., growth rate) [33, 34]. Every cell needs a bare minimum of various essential proteins in order to properly function [35]. Every additional protein produced takes up some of the cell's resources, which increases the overall metabolic load of the cell [36]. Naively, we might expect that after billions of years, cells have evolved a Goldilocks approach, where they produce slightly above the bare minimum of essential proteins in order to save on the metabolic costs of excess production. This approach of "living life on the edge" is incompatible with the highly stochastic nature of gene expression—any lessening of the protein levels due to statistical fluctuations would leave the cell without enough essential proteins to function. Since there are hundreds of essential proteins that each need to meet their threshold levels in order to maintain robust cell growth, stochasticity is almost guaranteed to cause growth arrest in the naive Goldilocks approach.

In Ref. [7], described in Chapter 4, we introduce the Robustness-Load Trade-Off (RLTO) model, which suggests alternative strategies for robustly optimizing cell fitness under stochasticity. In particular, we stipulate that the metabolic cost of additional proteins is

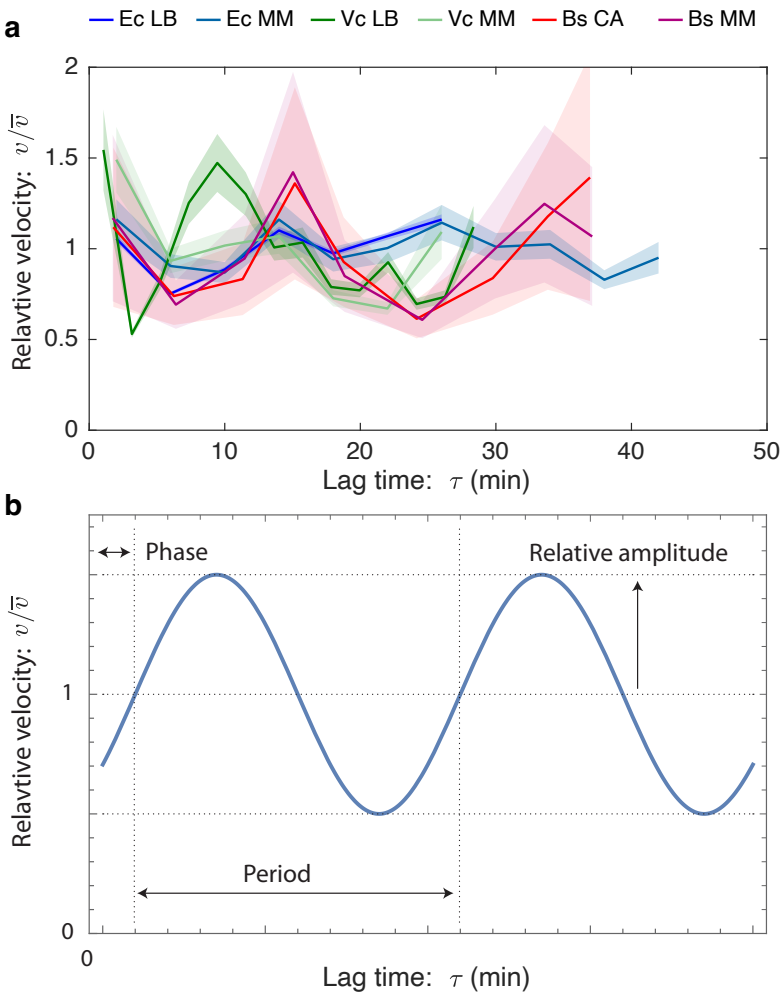


Figure 1.4: **Fork velocity oscillations.** Figure reproduced from Ref. [6]. **Panel a: Temporal velocity oscillations are observed in three bacterial species:** *E. coli* (Ec), *B. subtilis* (Bs), and *V. cholerae* (Vc). The fork velocity starts high before decaying rapidly and then recovering. Data are presented as mean values ± SEM. **Panel b: Oscillation characteristics.** The definition of the phase, amplitude, and period of the fork velocity oscillation.

low, leading to an asymmetric fitness landscape, as shown in Fig. 1.5 [37]. Combining the asymmetric fitness landscape with stochasticity allows us to generate quantitative predictions for various cellular phenomena. In particular, we find that (i) essential proteins are vastly overabundant, (ii) essential genes have a transcriptional lower limit of one message per cell cycle, and (iii) metabolic load is balanced between transcription and translation (the conversion of mRNA into proteins). See Fig. 1.6 for a schematic representation of these main results.

In Chapter 4, I first describe the mathematical details of the RLTO model. I then describe how we used the RLTO model to explain the observed overabundance of essential proteins. In essence, since the cost of additional proteins is low, it is advantageous for the cell to produce them in excess to avoid the growth arrest that accompanies sub-threshold expression. The amount of overabundance (number of actual proteins divided by the minimum threshold) for each protein is dependent on the threshold level, with low-threshold proteins having the highest overabundance and vice versa.

The existence of noise in the transcription process leads to the one-message-rule, which states that essential genes must be transcribed above a threshold of one message per cell cycle. We provide experimental evidence for this rule based on preexisting data. Finally, we demonstrate that load balancing occurs between transcription and translation in certain organisms. This provides the cell with two levers to adjust expression levels with.

Although we have discussed experimental evidence for many of the RLTO predictions, the overabundance prediction remains experimentally untested. As of this writing, our group is working on these experimental predictions and have found substantial evidence that the RLTO overabundance prediction is accurate [37]. For more details on these experiments, please see either our lab's future publications or my colleague Han Kyou (James) Choi's future dissertation.

One interesting feature of the preliminary results is that there are certain classes of genes that break from the RLTO model prediction of overabundance. I propose in Chapter 5 that

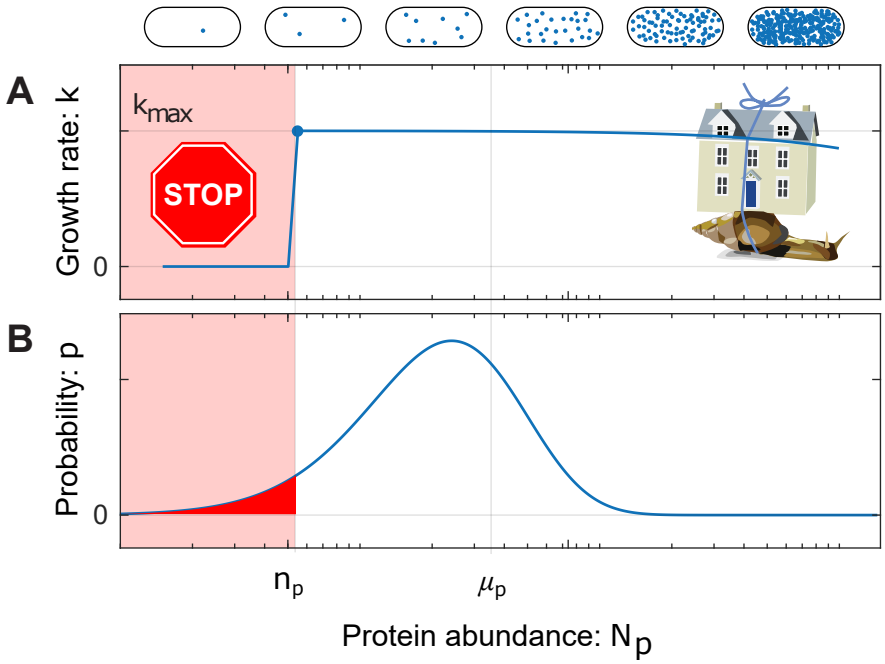


Figure 1.5: **Asymmetric fitness landscape in the RLTO model.** Figure reproduced from [37]. **Panel A: The fitness landscape is asymmetric in the RLTO model.** Cell fitness is modeled using the Robustness-Load Trade-Off model (RLTO). In the model, there is a metabolic cost of protein expression which 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 [7]. **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 [38].

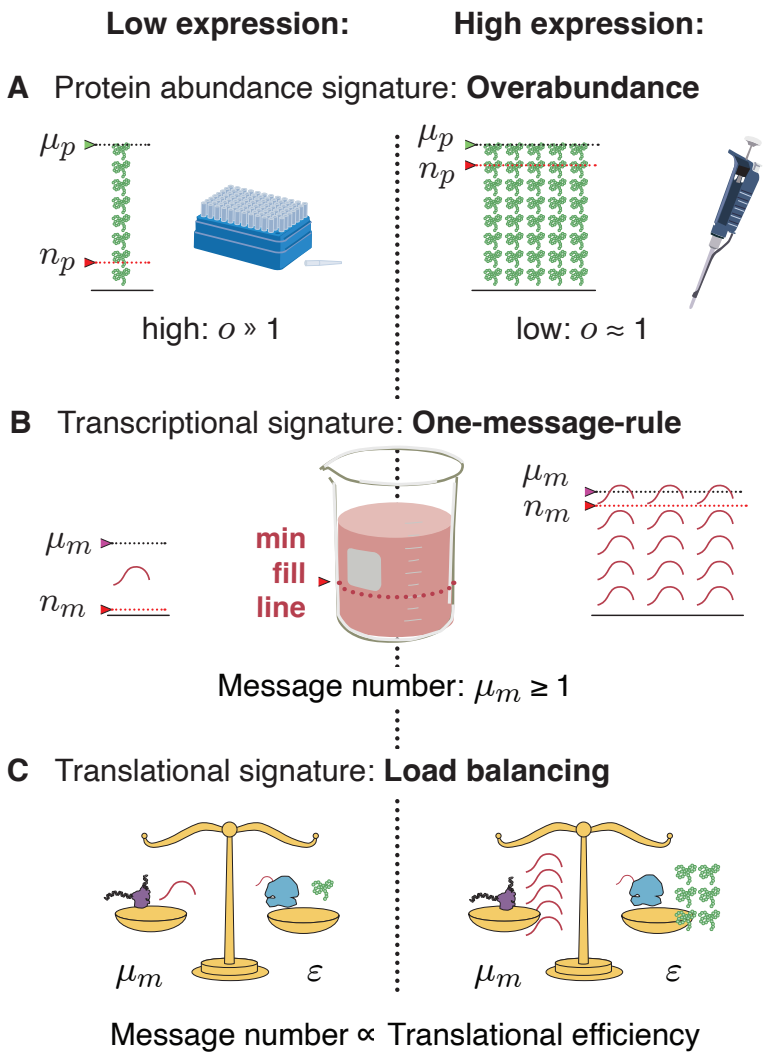


Figure 1.6: **Central dogma regulatory principles.** Figure reproduced from Ref. [7]. **Panel A: Overabundance.** Low-expression essential genes are expressed with high overabundance; whereas, high-expression essential genes are expressed with low overabundance. Lab supply analogy: Low-cost items that 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.

an alternative strategy for maintaining robustness, without overabundance, is the use of regulatory feedback.

1.4 Regulatory feedback dynamics

Imagine you're in the shower[1]. Although you turned the tap to the approximate location, the water is either freezing cold or it's scalding! You repeatedly adjust and overshoot, iteratively getting closer to that ideal temperature. This is an example of feedback-based regulatory control, depicted schematically in Fig. 1.7. We propose that this kind of regulatory feedback is an alternative strategy to the overabundance strategy (described in Chapter 4) for maintaining robustness for essential genes. How might this behavior arise in cellular metabolism?

As shown in Fig. 1.1, none of the metabolic pathways occur in isolation [2]. The metabolic network is deeply interconnected, with many regulatory loops incorporating feedback control. Although the decades-old field of chemical kinetics successfully describes independent chemical reactions, it fails to incorporate regulatory feedback [3]. An alternative approach called flux balance analysis attempts to expand the scope of kinetic theory by following the flow of metabolites through a metabolic network, but it only provides steady-state metabolite concentrations and also does not incorporate regulatory feedback [40]. Since we want to track the temporal metabolic dynamics as they vary throughout a cell cycle, we require additional mathematical machinery.

In Ref. [39], described in Chapter 5, we describe our attempt to mathematically model regulatory feedback dynamics, along with many proposed experiments for testing our models. We first focus on developing a set of minimal models for nucleotide metabolism (inspired by our work with DNA replication in Chapter 3). The enzyme ribonucleotide reductase (RNR) plays a major role in DNA synthesis, converting a set of metabolites into DNA-compatible counterparts, but RNR itself has multiple levels of regulation [41–47]. Our models expand

[1]Hopefully not with a printed copy of my dissertation!

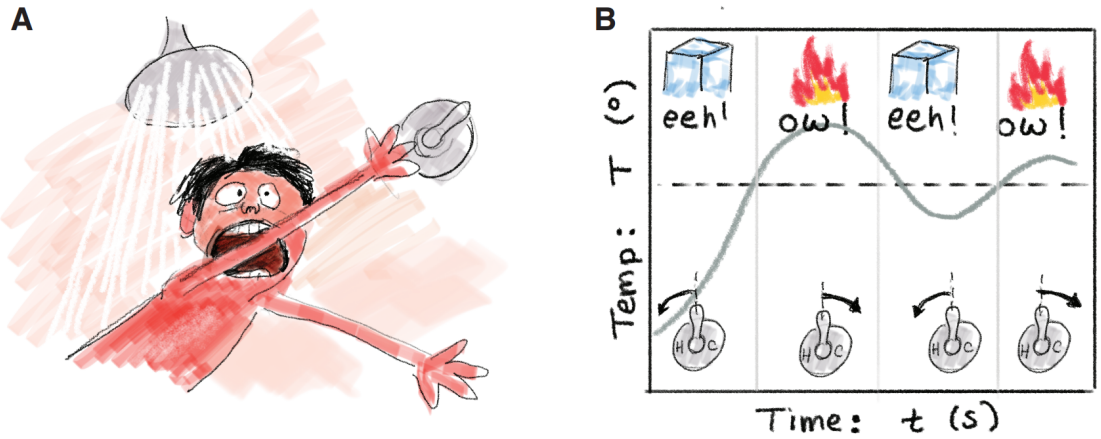


Figure 1.7: **Feedback-based regulatory control.** Figure reproduced from the proposal for the following NSF grant: NSF-Phys-2412326 [39]. **Panel A: Some feedback is required!** Most of us apply feedback-based regulatory control every day. Although I know the approximate position of the shower tap, I always need to adjust the tap to achieve the ideal temperature. **Panel B: Overshoot in regulation.** We apply negative feedback to control the water temperature. If the temperature is too high (low), we adjust the controls to decrease (increase) the temperature; however, we typically overcorrect due to the finite response time, leading to the phenomena of *overshoot* and oscillations. In a well-designed control scheme, these oscillations are quickly damped and the optimal conditions (e.g., temperature) are achieved.

on the basic theory of chemical kinetics by using coupled differential equations. We begin by solving the simplest model analytically (see Fig. 1.8Aa), which is equivalent in form to a damped harmonic oscillator. The solution demonstrates underdamped oscillatory behavior when there is strong feedback (tight regulation) and overdamped behavior when there is weak feedback (see Fig. 1.8B). These temporal oscillations match the experimental results for fork velocity that we found in Chapter 3.

The results also suggest that after a perturbation, strong feedback with overshoot leads to a faster return to homeostasis (metabolic equilibrium) than weak feedback with no overshoot. This runs counter to the naive expectation that evolution would have tuned all regulatory processes to turn on "just the right amount" to return exactly to equilibrium without overshoot. Is the best strategy instead to have the strongest feedback possible? The answer to this turns out to be quite nuanced, as evidenced by some of our computational tests, where we introduce greater complexity to our model (see Fig. 1.8Ab) and find that above a certain threshold, strong feedback actually leads to instability. More work still needs to be done to fully explain this result. In this chapter, we also provide multiple experimental approaches to test our hypotheses. I hope this can provide a road map for future research into this topic.

1.5 Summary and motivation

In the beginning of this introductory chapter, I laid out a set of questions that have guided the research in this dissertation. In revisiting and answering them here, I hope to paint a complete picture of what I have attempted to accomplish in my work with my colleagues.
(i) *How do we reconcile stochastic and deterministic models for different types of experiments, particularly during exponential growth?* In Chapter 2, I demonstrate that we can use the behavior of exponential growth to establish, in the form of an exponential mean, a direct correspondence between stochastic and deterministic models [5]. This equivalence bridges

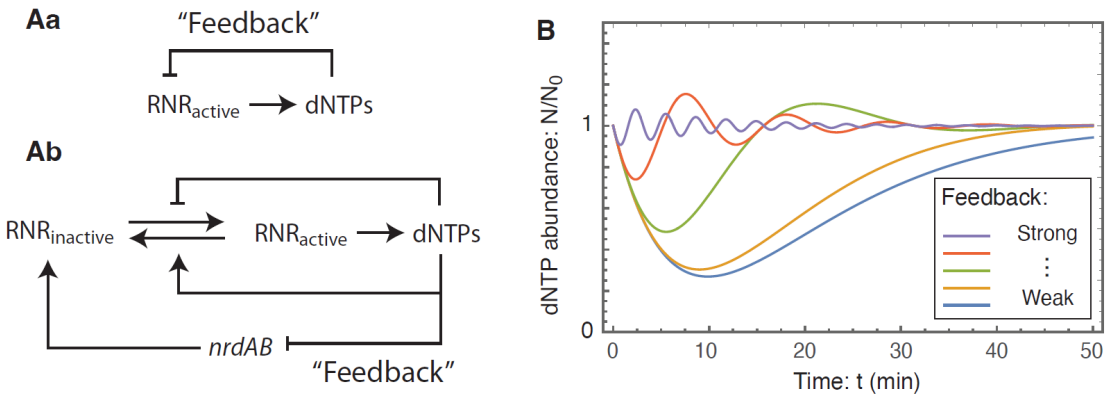


Figure 1.8: **Feedback models and solutions.** Figure reproduced from the proposal for the following NSF grant: NSF-Phys-2412326 [39]. **Panel A: Successive levels of complexity in the homeostatic model.** Panel Aa shows a minimal model with only two components (RNR and dNTPs) which is always stable, but can oscillate. Panel Ab shows a higher dimensional model which includes a more detailed and realistic mechanism of the homeostatic network, including transcription and translation, etc. **Panel B: Trade-off between overshoot, precision, and rapidity.** The oscillation period is controlled by the strength of the regulatory response. The homeostatic control can be overdamped (blue) to eliminate overshoot (i.e., oscillations); however, the size of dNTP pool depletion is larger and response time is slower.

the gap between experimental stochasticity and observed population demographics, allowing us to characterize quantities like growth rate and fractional cell pausing.

(ii) *How can we experimentally measure cellular timing and the dynamics of molecular machines in living cells, particularly for DNA replication, one of the most fundamental processes for all living organisms?* In Chapter 3, I introduce lag-time analysis, which uses the stochastic-deterministic correspondence to experimentally characterize the *in vivo* dynamics of replication for an exponentially-growing bacterial population [6]. We use the method to measure replication pauses down to the precision of seconds, and replication fork velocity in units of base pairs per second. We observe temporal oscillations in fork velocity in three evolutionarily-divergent species, which is the expected phenomenology for a strongly-regulated metabolic feedback pathway.

(iii) *What strategies do living organisms use to maintain robustness under perpetually-changing and noisy conditions?* In Chapter 4, I describe the robustness-load trade-off model, which incorporates stochasticity and an asymmetric fitness landscape to make quantitative predictions for various cellular phenomena [7]. We find a lower limit for transcription of essential genes, and we find that metabolic load is balanced between transcription and translation. Finally, we predict and observe an overabundance phenomenon for essential proteins, which we suggest is one major strategy for maintaining cellular robustness.

(iv) *How might metabolite regulation lead to unexpected cellular phenomena, and how can we mathematically model them?* In Chapter 5, I describe our work on regulatory feedback dynamics [39]. We propose that regulation is a strategy that the cell uses to maintain robustness, complementary to the overabundance strategy. The mathematical model that we propose exhibits the oscillatory signature, an unexpected cellular phenomenon, that we have observed with lag-time analysis. We find a trade-off between feedback strength, speed of return to equilibrium, and network stability. Although this research is still a work in progress, we propose a road map for further analysis and experimental tests.

My hope is that this dissertation provides a basis, however small, for humanity to further its understanding of life.

1.6 Bibliography

[1] Do you really need to see a reference for this?

[2] G. Michal, *Roche Biochemical Pathways Wall Charts by Gerhard Michal*, Jan. 2021. DOI: 10.5281/zenodo.4446230.

[3] D. L. Nelson and M. M. Cox, *Lehninger principles of biochemistry*, en, 7th ed. New York, NY: W.H. Freeman, 2017.

[4] J. A. van der Knaap and C. P. Verrijzer, "Undercover: Gene control by metabolites and metabolic enzymes," *Genes & Development*, vol. 30, no. 21, pp. 2345–2369, Nov. 2016. DOI: 10.1101/gad.289140.116.

[5] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014 420, 1 Jan. 2022. DOI: 10.1103/PhysRevE.105.014420.

[6] 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," *Nat Commun*, vol. 14, no. 1, p. 1762, Mar. 2023. DOI: 10.1038/s41467-023-37456-2.

[7] T. W. Lo, H. K. J. Choi, D. Huang, and P. A. Wiggins, "Noise robustness and metabolic load determine the principles of central dogma regulation," *preprint*, Oct. 2023. DOI: 10.1101/2023.10.20.563172.

[8] 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.

[9] L. Willis and K. C. Huang, "Sizing up the bacterial cell cycle," *Nat Rev Microbiol*, vol. 15, no. 10, pp. 606–620, Oct. 2017. DOI: 10.1038/nrmicro.2017.79.

[10] 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.

[11] H. Merrikh, Y. Zhang, A. D. Grossman, and J. D. Wang, "Replication-transcription conflicts in bacteria," *Nat Rev Microbiol*, vol. 10, no. 7, pp. 449–58, Jun. 2012. DOI: 10.1038/nrmicro2800.

[12] 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.

[13] K. Adelman and J. T. Lis, "Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans," *Nature Reviews Genetics*, vol. 13, no. 10, pp. 720–731, 2012. DOI: 10.1038/nrg3293.

[14] R. J. Davenport, G. J. Wuite, R. Landick, and C. Bustamante, "Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase," *Science*, vol. 287, no. 5462, pp. 2497–500, Mar. 2000. DOI: 10.1126/science.287.5462.2497.

[15] X. Wang, C. Possoz, and D. J. Sherratt, "Dancing around the divisome: Asymmetric chromosome segregation in *Escherichia coli*," *Genes Dev*, vol. 19, no. 19, pp. 2367–77, Oct. 2005. DOI: 10.1101/gad.345305.

[16] 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.

[17] C. J. Rudolph, A. L. Upton, A. Stockum, C. A. Nieduszynski, and R. G. Lloyd, "Avoiding chromosome pathology when replication forks collide," *Nature*, vol. 500, no. 7464, pp. 608–11, Aug. 2013. DOI: 10.1038/nature12312.

[18] 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.

[19] N. J. Kuwada, B. Traxler, and P. A. Wiggins, "Genome-scale quantitative characterization of bacterial protein localization dynamics throughout the cell cycle," *Mol Microbiol*, vol. 95, no. 1, pp. 64–79, Jan. 2015. DOI: 10.1111/mmi.12841.

[20] P. Wang *et al.*, "Robust growth of *Escherichia coli*," *Curr Biol*, vol. 20, no. 12, pp. 1099–103, Jun. 2010. DOI: 10.1016/j.cub.2010.04.045.

[21] E. R. Mardis, "Next-Generation DNA Sequencing Methods," *Annual Review of Genomics and Human Genetics*, vol. 9, no. Volume 9, 2008, pp. 387–402, 2008, ISSN: 1545-293X. DOI: https://doi.org/10.1146/annurev.genom.9.081307.164359.

[22] B. Alberts *et al.*, *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, 2015.

[23] R. Phillips, J. Kondev, J. Theriot, and N. Orme, *Physical Biology of the Cell*. Garland Science, 2013, ISBN: 9780815344506.

[24] I. Tinoco Jr and R. L. Gonzalez Jr, "Biological mechanisms, one molecule at a time," *Genes Dev*, vol. 25, no. 12, pp. 1205–31, Jun. 2011. DOI: 10.1101/gad.2050011.

[25] M. Elías-Arnanz and M. Salas, "Resolution of head-on collisions between the transcription machinery and bacteriophage phi29 DNA polymerase is dependent on RNA polymerase translocation," *EMBO J*, vol. 18, no. 20, pp. 5675–82, Oct. 1999. DOI: 10.1093/emboj/18.20.5675.

[26] A. M. Deshpande and C. S. Newlon, "DNA replication fork pause sites dependent on transcription," *Science*, vol. 272, no. 5264, pp. 1030–3, May 1996. DOI: 10.1126/science.272.5264.1030.

[27] B. Liu and B. M. Alberts, "Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex," *Science*, vol. 267, no. 5201, pp. 1131–7, Feb. 1995. DOI: 10.1126/science.7855590.

[28] E. P. C. Rocha, "The organization of the bacterial genome," *Annu Rev Genet*, vol. 42, pp. 211–33, 2008. DOI: 10.1146/annurev.genet.42.110807.091653.

[29] E. V. Mirkin and S. M. Mirkin, "Replication fork stalling at natural impediments," *Microbiol Mol Biol Rev*, vol. 71, no. 1, pp. 13–35, Mar. 2007. DOI: 10.1128/MMBR.00030-06.

[30] N. Y. Yao and M. O'Donnell, "Replisome structure and conformational dynamics underlie fork progression past obstacles," *Curr Opin Cell Biol*, vol. 21, no. 3, pp. 336–43, Jun. 2009. DOI: 10.1016/j.ceb.2009.02.008.

[31] R. T. Pomerantz and M. O'Donnell, "The replisome uses mRNA as a primer after colliding with RNA polymerase," *Nature*, vol. 456, no. 7223, pp. 762–6, Dec. 2008. DOI: 10.1038/nature07527.

[32] 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. eprint: https://doi.org/10.1021/j100540a008.

[33] 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.

[34] 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.

[35] 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.

[36] H. G. S. Joseph W. Lengeler Gerhart Drews, Ed., *Biology of the Prokaryotes*. Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998.

[37] H. K. J. Choi, K. J. Cutler, D. Huang, T. W. Lo, W. R. Will, and P. A. Wiggins, "In preparation," 2024.

[38] 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.

[39] P. A. Wiggins, *Metabolic homeostasis, oscillations, and instability*, Proposal for National Science Foundation Physics of Living Systems Grant: NSF-Phys-2412326, May 2024.

[40] J. D. Orth, I. Thiele, and B. Palsson, "What is flux balance analysis?" *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, Mar. 2010. DOI: 10.1038/nbt.1614.

[41] 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.

[42] 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.

[43] 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. 14 496–501, Apr. 2004. DOI: 10.1074/jbc.M310142200.

[44] 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. 33 021–6, Oct. 2000. DOI: 10.1074/jbc.M005337200.

[45] 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.

[46] 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.

[47] 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.

Chapter 2

Characterizing stochastic cell cycle dynamics in exponential growth

Originally published as: [1] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014 420, 1 Jan. 2022. DOI: 10.1103/PhysRevE.105.014420.

Author contributions: D.H., T.L., H.M., and P.A.W. developed the approach. D.H., T.L., and P.A.W. wrote code, ran the model, and analyzed output data. D.H. and P.A.W. developed the mathematical theory and wrote the manuscript.

Abstract

Two powerful and complementary experimental approaches are commonly used to study the cell cycle and cell biology: One class of experiments characterizes the statistics (or demographics) of an unsynchronized exponentially-growing population, while the other captures cell cycle dynamics, either by time-lapse imaging of full cell cycles or in bulk experiments on synchronized populations. In this paper, we study the subtle relationship between observations in these two distinct experimental approaches. We begin with an existing model: a single-cell deterministic description of cell cycle dynamics where cell states (i.e., periods or phases) have precise lifetimes. We then generalize this description to a stochastic model in which the states have stochastic lifetimes, as described by arbitrary probability distribution functions. Our analyses of the demographics of an exponential culture reveal a simple and exact correspondence between the deterministic and stochastic models: The corresponding state ages in the deterministic model are equal to the exponential

mean of the age in the stochastic model. An important implication is therefore that the demographics of an exponential culture will be well-fit by a deterministic model even if the state timing is stochastic. Although we explore the implications of the models in the context of the *Escherichia coli* cell cycle, we expect both the models as well as the significance of the exponential-mean lifetimes to find many applications in the quantitative analysis of cell cycle dynamics in other biological systems.

2.1 Introduction

Methods to quantitatively characterize cell cycle dynamics have expanded dramatically [2] since the pioneering model of the *Escherichia coli* cell cycle described by Cooper and Helmstetter [3]. Their initial work represented the cell cycle as a deterministic process in which each step was precisely timed. Although these assumptions were almost certainly viewed as a matter of mathematical convenience, some later readers have interpreted the experimental success of this model as evidence that stochasticity in the cell cycle has little biological significance [4]. Some later authors have relaxed some of these assumptions and found that the predictions are in fact robust to the model details [4], but none have yet reanalyzed these dynamics in the context of the significant level of stochasticity observed in cell cycle timing (e.g., [5, 6]). In this paper, we study a class of stochastic models that can be solved exactly, even in the strong stochasticity limit, and we explore their phenomenology.

One fundamental difficulty with reconciling the quantitative analyses of the cell cycle is the existence of two distinct classes of experiments: In *unsynchronized approaches*, an exponential culture is analyzed and the number of cells at time t is used to generate statistics defined with respect to cell number [3]. Examples of this approach are snapshot imaging (e.g., [7]), flow cytometry (e.g., [8]), and many deep-sequencing based approaches (e.g., [9]). We contrast these with *synchronized approaches* in which cells of a known state in the cell cycle progression are analyzed. Examples of this approach are the use of any of the previously described methods on cells which are first synchronized using a baby machine (e.g., [10]).

Time-lapse imaging of full cell cycles (e.g., [11]), including the use of devices like the mother machine (e.g., [5]), can also be used to generate data for synchronized analyses. Although it might naively seem that averaging with respect to these two population ensembles are equivalent, they are not.

To demonstrate the subtlety of interpreting the data from an exponential culture, consider the probability of observing the Z ring, the ring-shaped protein complex that forms in *E. coli* at midcell and drives the process of septation (or cytokinesis) [12]. If the cell cycle has duration T and the Z ring has lifetime $\delta\tau_{\mathrm{Z}}$, one might naively assume the probability of observing the Z ring is:

$$p_{\mathrm{Z}} = \delta\tau_z/T. \tag{2.1}$$

See Fig. 2.1. Although this is true in the synchronized population, in an exponential culture the probability is 30% lower as a direct consequence of the relative abundance of cells by age. (The exact degree to which this is reduced depends on the ratio of τ_z/T as discussed below.) Why? The number of new-born cells is twice the abundance of cells at the end of the cell cycle when the Z ring forms. Although this seems like a trivial book-keeping annoyance, when we consider the stochastic model, this effect has consequential implications for timing throughout the cell cycle, including on the growth rate.

In Sec. 2.2.1, we will first revisit the existing *deterministic model*, where all events in the cell cycle are precisely timed. In this model, we will represent the fundamental state of the cell as an age τ and compute the statistics of cell age in an exponential culture. To make contact with observables, we then apply these results to describe the demographics of the *E. coli* cell cycle in Sec. 2.2.2. In Sec. 2.2.3, we consider a *stochastic model*, where the cell cycle is represented as discrete sequential states $j = 1...m$, each with a stochastic lifetime $\tau_{\delta j}$. Although this model cannot fully capture all the complexities of the cell cycle, it is analytically tractable and we can exactly compute expressions for all the same statistics as the deterministic model. The relation between the deterministic and stochastic model statistics is at this point opaque. In Sec. 2.2.4, we define an *exponential mean*, which is a mean biased toward younger cells that are overabundant in exponential culture. In Sec. 2.2.5,

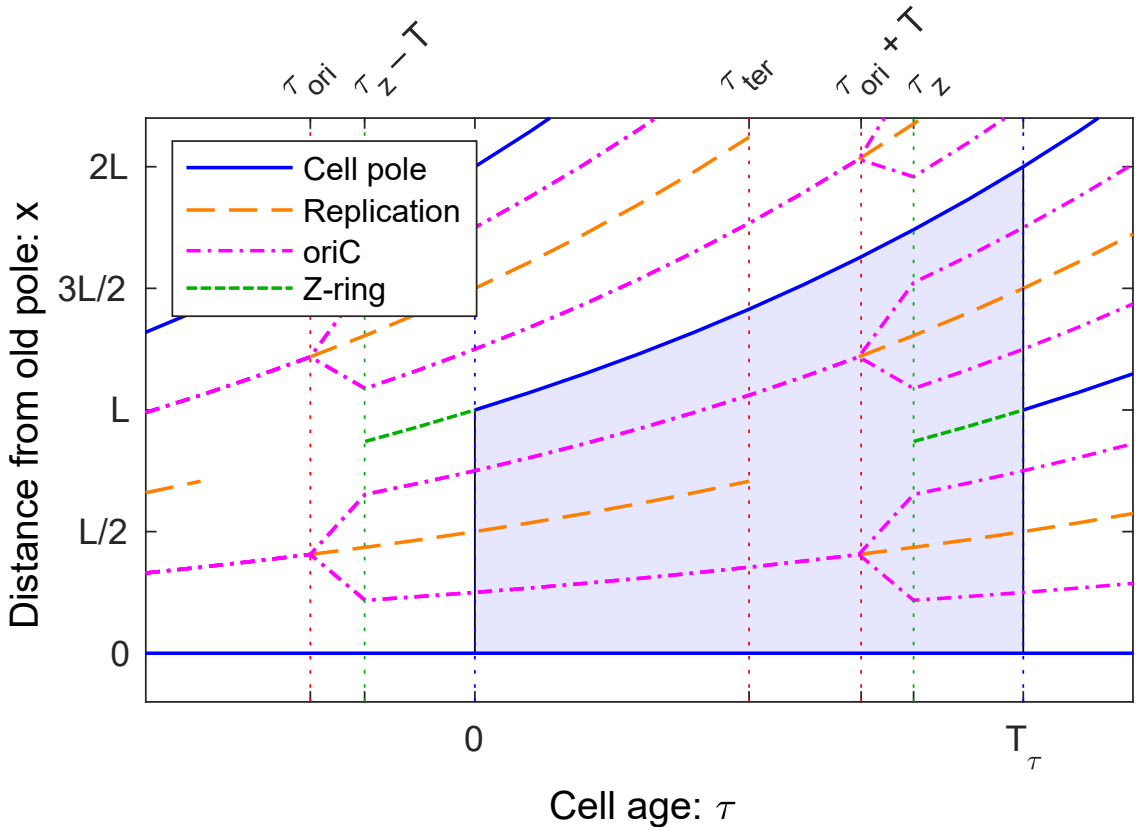


Figure 2.1: **Schematic for positioning and timing of events during the *E. coli* cell cycle.** In the deterministic model, the duration of the cell cycle as well as the timing of all events in the cell cycle are precise (i.e., deterministic). Furthermore, we shall assume the positioning of all complexes and the length of the cell are all deterministic as well. The shaded region represents one complete cell cycle. We have also annotated the positioning and timing of a number of cell cycle events: (i) Cell poles appear as the consequence of septation and do not disappear. (ii) Replication of the chromosome starts when the origin ($oriC$) is replicated and ends when the terminus is replicated. The origin is replicated before the start of the cell cycle. (iii) New origins are created and move from the quarter-cell positions to the eighth-cell positions after replication. (iv) The Z ring, which drives septation at midcell, assembles and disassembles at the end of the cell cycle.

we demonstrate that the predictions of the stochastic and deterministic models are in fact identical if the deterministic state ages τ_j are equal to the exponential-mean stochastic state ages $\overline{\tau}_j$. Finally, in Sec. 2.2.5, we consider a number of simple biological examples to underline both the mathematical behavior of the exponential mean as well as its biological implications. In the interest of brevity, we will discuss experimental support for this model elsewhere [13].

2.2 Results

In this section, we will derive the expressions for a large number of statistics relevant for describing an exponential culture. We will first derive expressions for the statistics in the deterministic model and then the stochastic model. In Tab. 2.1, we provide a summary of the notation.

2.2.1 Deterministic model

In the deterministic model, we will consider cells that are born with age $\tau = 0$ and divide deterministically at age $\tau = T_\tau$. By *cell age* τ, we mean a continuous cell state variable representing cell cycle progression, not *aging* in the context of reduced cell fitness over time [14].

2.2.1.1 Definition of the deterministic model

In the deterministic model, cell state is described by a continuous variable, cell age τ, and therefore the population is described in terms of a number density with respect to age τ at time t: $n_\tau(t)$. Age τ is defined on the interval $[0, T_\tau]$ with $\tau = 0$ corresponding to cell birth and T_τ corresponding to cell division. Let the cumulative creation number, $N_\tau^+(t)$, be the cumulative number of cells that have entered state (i.e., age) τ and the cumulative annihilation number, $N_\tau^-(t)$, be the cumulative number of cells that have transitioned out of state τ. The naming of the cumulative creation and annihilation numbers was motivated in relation to the creation and annihilation operators from quantum field theory. See Fig. 2.2.

Table 2.1: **A summary of the model notation.** Note that since the deterministic model is described in terms of a continuous state, the age τ, the number of cells in that state is represented as a *number density* n; whereas in the stochastic model, the cell state is represented by an integer j and, therefore, the number of cells in that state is represented by a *number* N_j. The symbol $\sim$ means that a random variable is *distributed like*.

Variable	**Meaning**
t	Experimental time
$N(t)$	Total number of cells at time t.
$N_{\text{obj}}(t)$	Number of objects 'obj'.
k	Culture growth rate
T	Culture mass doubling time
	Deterministic model
τ	Cell age $0 \leq \tau \leq T_\tau$
T_τ	Deterministic duration of the cell cycle
$N_\tau^+(t)$	Cumulative number of cells that have entered state τ (creation number)
$N_\tau^-(t)$	Cumulative number of cells that have transitioned out of state τ (annihilation number)
$n_\tau(t)$	Number density with respect to cell age τ
	Stochastic model
$\tau_{\delta j} \sim p_{\delta j}(\cdot)$	Stochastic lifetime of state j
$\tau_j \sim p_j(\cdot)$	Stochastic age of state j
$T_\tau \sim p(\cdot)$	Stochastic duration of the cell cycle
$N_j(t)$	Number of cells in state j
$N_j^+(t)$	Cumulative number of cells that have arrived to state j over all time
$N_j^-(t)$	Cumulative number of cells that have departed from state j over all time

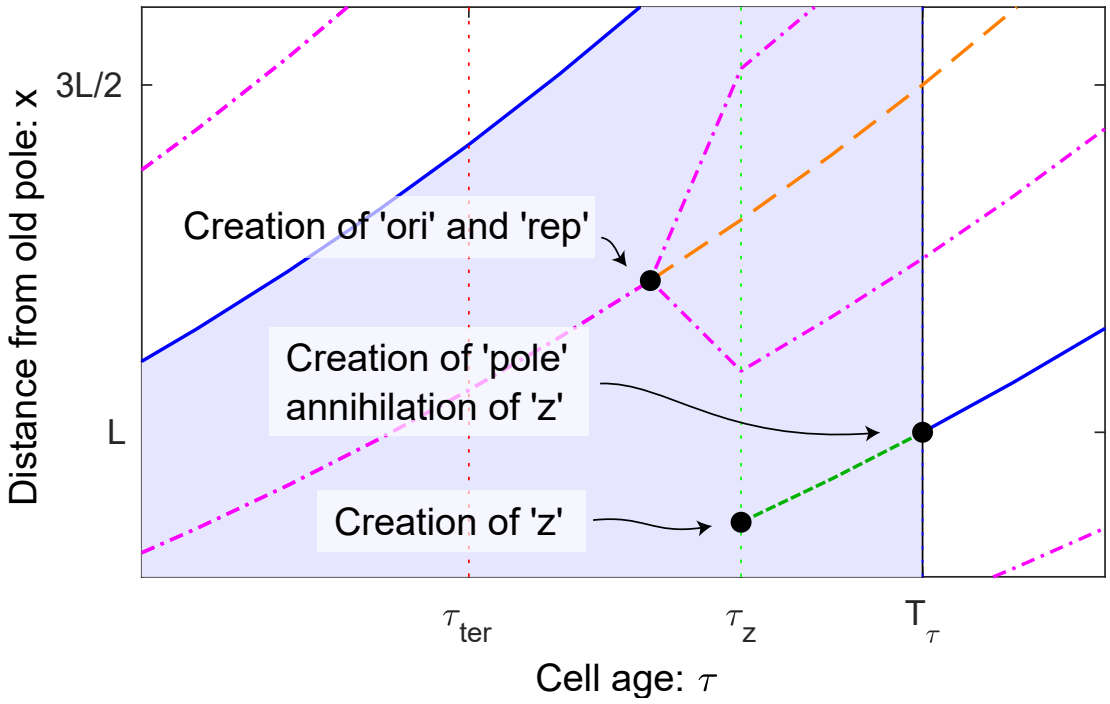


Figure 2.2: **Creation and annihilation of cell quantities.** The statistics of different types of quantities will have different statistical properties. *Transient quantities* undergo both creation and annihilation events. The Z ring and replisome are natural examples of transients since they assemble and then disassemble. In contrast, both cell poles and genetic loci are *perpetual quantities* that only undergo creation but never annihilation.

In the deterministic model, where the cell state τ is continuous (not discrete), the cumulative creation and annihilation numbers are identical and are related to the number density by the equation:

$$N_\tau^+(t) = N_\tau^-(t) = \int_0^t \mathrm{d}t_1 \; n_\tau(t_1). \tag{2.2}$$

It is convenient to define these cumulative numbers in addition to the number density, since they will be a powerful tool for computing some observable quantities. To describe the dynamics, we can write an equation describing the number of cells entering the infinitesimal age interval $[\tau, \tau + \delta\tau]$ in the infinitesimal time interval $[t, t + \delta t]$:

$$\delta t \, \delta\tau \, \dot{n}_\tau(t) = \delta t \, \dot{N}_\tau^+(t) - \delta t \, \dot{N}_{\tau+\delta\tau}^-(t), \tag{2.3}$$

where $\dot{A} \equiv \partial_t A$ and the first term on the RHS is the number of cells entering state τ and the second term represents the cells leaving state $\tau + \delta\tau$. Eq. 2.3 can then be rewritten:

$$\dot{n}_\tau(t) = -\partial_\tau \dot{N}_\tau^+(t), \tag{2.4}$$

except at division, where some care is required. Now consider the process of cell division explicitly: The division process can be understood as the annihilation of a cell in state $\tau = T_\tau$ and the creation of two new-born cells in state $\tau = 0$:

$$\dot{N}_\tau^+(t) = \begin{cases} 2n_{T_\tau}(t), & \tau = 0 \\ n_\tau(t), & \tau > 0 \end{cases} \tag{2.5}$$

$$\dot{N}_\tau^-(t) = n_\tau(t). \tag{2.6}$$

Substituting Eq. 2.5 into Eq. 2.4 gives a single piecewise rate equation in terms of the number density $n_\tau(t)$:

$$\dot{n}_\tau(t) = -\begin{cases} 2n'_{T_\tau}(t), & \tau = 0 \\ n'_\tau(t), & \tau > 0 \end{cases}, \tag{2.7}$$

where $A' \equiv \partial_\tau A$. Eq. 2.7 completely describes the cell cycle dynamics in the deterministic model. The details of the derivation are given in Supplementary Sec. 2.5.1.

2.2.1.2 Solution to the deterministic model

In steady-state growth, we can assume the total number of cells is:

$$N(t) = N_0 \exp(kt), \tag{2.8}$$

where k is the growth rate that is determined by solving the rate equation (Eq. 2.7), as detailed in Supplementary Sec. 2.5.2. It will often be convenient to rewrite the equations in terms of the doubling time:

$$T \equiv k^{-1} \ln 2, \tag{2.9}$$

rather than the growth rate k. Eq. 2.7 evaluated at $\tau = 0$ gives a consistency condition between doubling time T and the duration of the cell cycle T_τ:

$$T_\tau = T, \tag{2.10}$$

which is to say that the doubling time is equal to the duration of the cell cycle, as one would naively expect. In steady-state growth, one can compute the number density of cells, which is:

$$n_\tau(t) = n_0 \exp[k(t - \tau)], \tag{2.11}$$

where n_τ is the density with respect to cell age τ and n_0 is a constant determined by the initial cell number. The details of the derivations for the solution and the consistency condition are given in Supplementary Sec. 2.5.2.

2.2.1.3 Statistics of the deterministic model

The solution of Eq. 2.7 can be used to compute the probability (PDF) and cumulative (CDF) distribution functions with respect to cell age:

$$f_\tau(\tau) = 2ke^{-k\tau}, \tag{2.12}$$

$$F_\tau(\tau) = 2(1 - e^{-k\tau}). \tag{2.13}$$

The details of the derivation are given in Supplementary Sec. 2.5.3. Eq. 2.12 implies that in an exponential culture, there is an enrichment of young cells which decays exponentially with age τ. See Fig. 2.3.

Note that the canonical observable in an exponential culture is number as a function of time, rather than abundances relative to the total number of cells $N(t)$. However, we shall write each expression as the prefactor of $N(t)$ and therefore the prefactor can be interpreted as the abundance relative to cell number $N(t)$.

The cumulative creation number is:

$$N_\tau^+(t) = 2e^{-k\tau}N(t). \tag{2.14}$$

The details of the derivation are given in Supplementary Sec. 2.5.4. The number of cells younger than age τ is:

$$N_{<\tau}(t) = 2(1 - e^{-k\tau})\, N(t), \tag{2.15}$$

and the number of cells older than age τ is:

$$N_{>\tau}(t) = (2e^{-k\tau} - 1)\, N(t). \tag{2.16}$$

Finally, the number of cells in a state defined by the age range $\tau_1 < \tau < \tau_2$ is:

$$N_{[\tau_1,\tau_2]}(t) = N_{>\tau_1}(t) - N_{>\tau_2}(t), \tag{2.17}$$

where the two terms on the right hand side are defined in Eq. 2.16.

2.2.2 Application to cell cycle dynamics

In this section, we demonstrate how to apply these results in the context of the *E. coli* cell cycle dynamics shown schematically in Fig. 2.1. These formulae can be applied either to predict the numbers from the known replication timing or to infer timing from the observed numbers in an exponential culture.

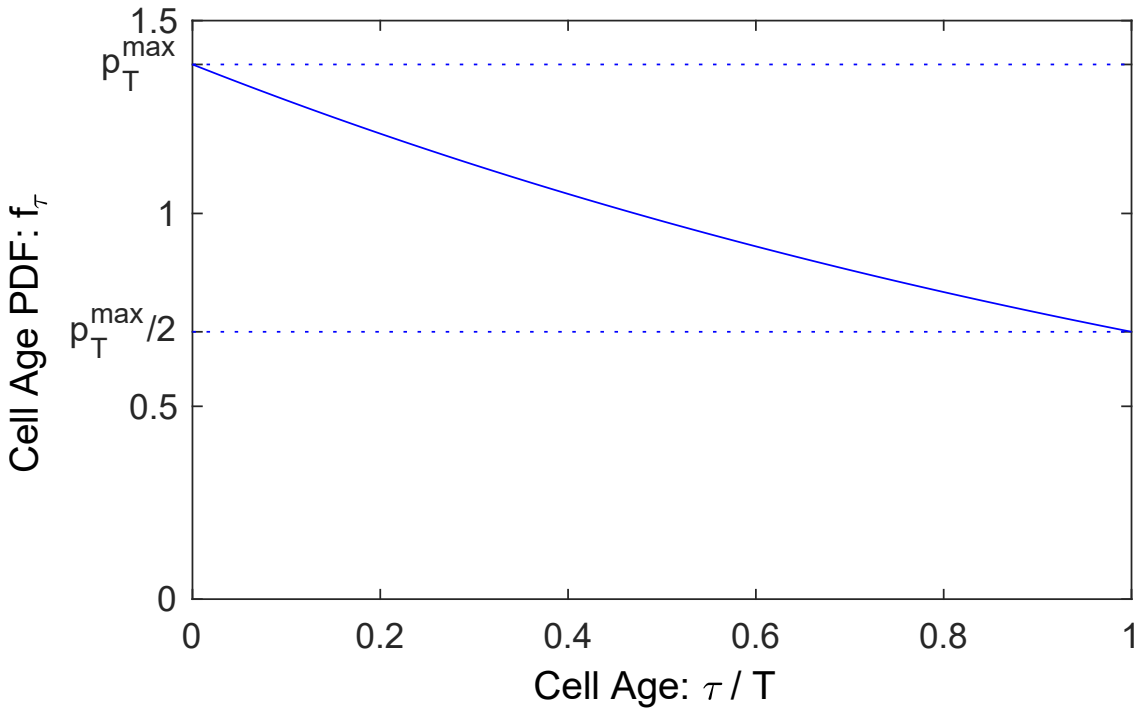


Figure 2.3: **Cell Age PDF.** In an exponential culture, there is an enrichment in young cells relative to old cells. The relative number of cells decays exponentially with cell age τ.

2.2.2.1 Z ring

The Z ring is an ultra-structural complex responsible for the process of bacterial cytokinesis (or septation) in which the cell envelope contracts at midcell forming a septum that closes to form the new poles of the nascent daughter cells [15]. The assembly, dynamics and disassembly of this structure is easily visualized using a wide range of fluorescent fusions in live cells or immunofluorescence in fixed cells [16].

The Z ring is an example of a transient complex, therefore we need to use N (as opposed to N^+). Furthermore, it assembles at $\tau = \tau_{\mathrm{Z}}$ and disassembles at the end of the cell cycle. The number of Z rings is therefore equal to the number of cells older than τ_{Z}:

$$N_{\mathrm{Z}}(t) \quad = \quad N_{>\tau_Z}(t) = (2e^{-k\tau_Z} - 1)\, N(t). \tag{2.18}$$

It is interesting to consider the limit as $\delta\tau_{\mathrm{Z}} \equiv T_\tau - \tau_{\mathrm{Z}}$ is small relative to the cell cycle duration T_τ in order to compare this to our intuitive guess (Eq. 2.1):

$$p_Z = \tfrac{N_Z}{N} \quad \approx \quad \tfrac{\delta\tau_Z}{T} \ln 2. \tag{2.19}$$

Since $\ln 2 \approx 0.69$, this is roughly 30% smaller than our naive estimate due to depletion of older cells in an exponential culture (Fig. 2.3).

2.2.2.2 Cell poles

Although the number of cell poles is twice the number of cells, it is useful to consider this example more formally. Unlike the Z ring which is transient, the poles are perpetual: Once the state is created, it is never annihilated (Fig. 2.2). In this context, we can use the cumulative creation number N^+. Note that we are immediately presented with a conundrum: Are two poles formed at the end of the cell cycle ($\tau = T$) or is one pole created at birth ($\tau = 0$)? Both approaches give the same number:

$$N_{\mathrm{pole}}(t) = 2N_T^+ = N_0^+ = 2N(t), \tag{2.20}$$

which is twice the number of cells, just as one intuitively expects.

2.2.2.3 DNA loci

The numbers of DNA loci can be observed by a number of different approaches: Modern deep sequencing methods allow a replication profile (i.e., the DNA copy number) of all loci to be measured in a single experiment (e.g., [9]). However, population-level analysis of the relative copy numbers of loci long predate this modern approach [3]. The single-cell dynamics of loci can also be observed: Imaging-based approaches, such as Fluorescence In Situ Hybridization and Fluorescent Repressor Operator Systems (or closely related approaches), can be used to visualize the numbers of segregated loci in single cells [17, 18].

There are multiple equivalent approaches to computing the numbers of genetic locus ℓ. First consider the slow-growth limit where both initiation and termination occur within the current cell cycle [3]. Assume the locus of interest is replicated at time τ_ℓ. The number of copies per cell is one before replication and two after replication. We can therefore write:

$$\begin{aligned} N_\ell(t) &= 1 \times N_{<\tau_\ell}(t) + 2 \times N_{>\tau_\ell}(t), && (2.21) \\ &= e^{-k(\tau_\ell - T)} N(t), && (2.22) \\ &= N_0 e^{k(t+T-\tau_\ell)}, && (2.23) \end{aligned}$$

in agreement with previous results [19, 20]. Unlike transient quantities (e.g., the number of Z rings), the form of Eq. 2.23 implies that the number of genetic loci can be understood as a temporal shift of $N(t)$ by $T_\tau - \tau_\ell$ to shorter times, as illustrated in Fig. 2.4.

The cancellation between the non-exponential terms between $N_{>\tau_\ell}$ and $N_{<\tau_\ell}$ in Eq. 2.22 may seem incidental, but from another perspective it is intuitive: The mathematical reason for the non-exponential terms in the prefactors of Eqs. 2.15-2.16 is *annihilation* (i.e., the reduction in the number of cells of a particular age τ due to aging). DNA loci correspond to a perpetual state: Once a locus state is created (i.e., replicated) it does not annihilate (i.e., *transition* into another *state*). To compute the number of genetic loci, we can therefore use the cumulative creation number N^+ formula (as opposed to N which is reduced by

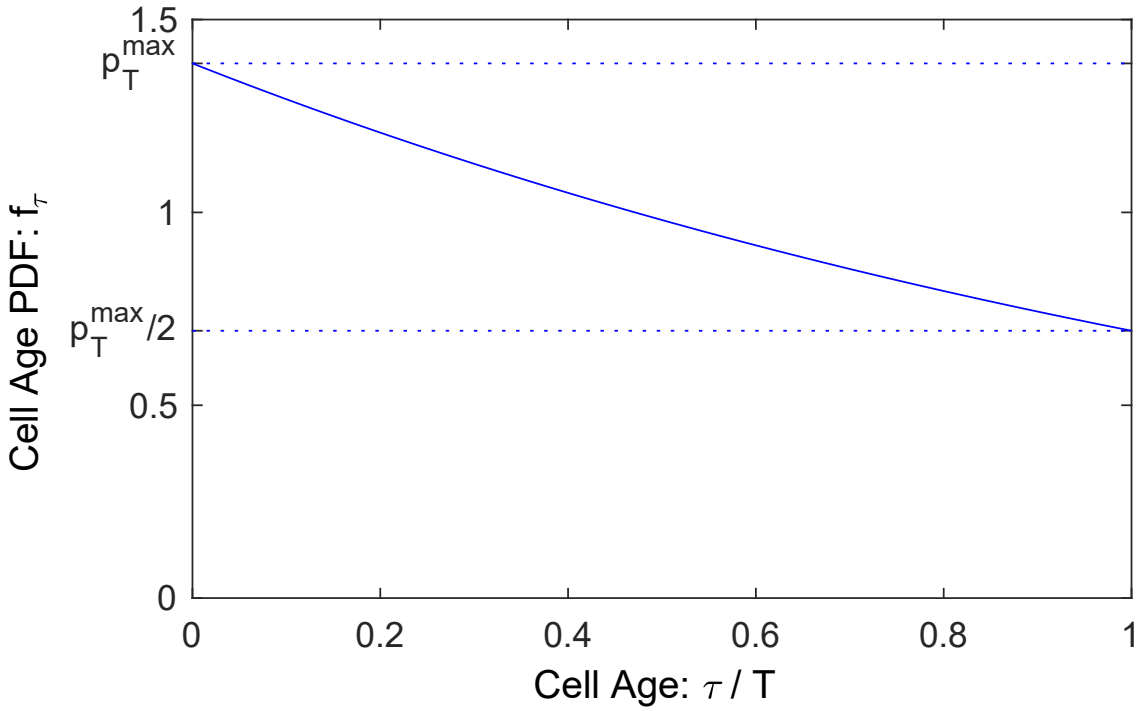


Figure 2.4: **Numbers in an exponential culture.** The numbers of all quantities grow exponentially with the same growth rate. For perpetual quantities (e.g., cell number, cell poles, DNA loci) the relative timing of the creation of a quantity can be inferred by the temporal offset of the $N_X(t)$ curve relative to the cell number curve $N(t)$. In contrast, transient quantities, like the number of Z rings, also grow exponentially, but their offset cannot directly be interpreted as a time.

annihilation). This more direct approach yields the same result as Eq. 2.22:

$$N_\ell(t) \quad = \quad N^+_{\tau_\ell}(t) = e^{-k(\tau_\ell - T)} N(t), \tag{2.24}$$

but is applicable for fast growth where replication initiates before the cell cycle begins (i.e., $\tau_\ell < 0$), as illustrated in the cell cycle schematic in Fig. 2.3.

2.2.2.4 B, C, and D period

Traditionally, the bacterial cell cycle is described by three periods: The B period is defined as the period between birth and replication initiation. The C period is defined as the cell cycle period during which replication occurs: I.e., after replication initiation and before termination [3]. The D period is defined as the period between replication termination and cell division [3]. There are multiple approaches to characterizing relative abundance of cells by period. A traditional approach is to infer this information from the relative abundance of the origin, terminus and number of cells [3]. However, more recent single-cell approaches can visualize the replication process itself in live cells [21, 22].

The relation between the durations of these periods and the locus number (Eq. 2.24) are:

$$B \equiv \tau_{ori} - 0 \quad = T \log_2 2N/N_{ori}, \tag{2.25}$$

$$C \equiv \tau_{ter} - \tau_{ori} = T \log_2 N_{ori}/N_{ter}, \tag{2.26}$$

$$D \equiv T - \tau_{ter} \quad = T \log_2 N_{ter}/N, \tag{2.27}$$

where N, N_{ori}, and N_{ter} are the number of cells, origins, and termini in the exponential culture (not per cell), which has previously been reported [3, 4]. Note that if replication initiates before the start of the cell cycle, $B = \tau_{ori} < 0$.

2.2.2.5 Replication

Finally, let us consider the replisomes and the replication process itself. A traditional population-level approach to determining the number of replicating cells is to infer it from the

relative origin, terminus, and cell abundances. However, many single-cell and live single-cell approaches exist today as well. For instance, fluorescent fusions to core replisome components that localize during replication can be used to characterize the number of replicating cells [21, 23, 24].

Like the Z ring, replication is a transient state; however, there is a significant subtlety here: Do we count (i) replicating cells, (ii) individual replication processes consisting of replisome-pairs, or (iii) individual replisomes?

First let us consider the number of replisome-pairs. Since the replication process can span the overlap between two successive cell cycles, it is most convenient to use differences in the cumulative creation number N^+. In fact, we can express the number of replisome-pairs concisely in terms of *oriC* and *ter*:

$$N_{\text{rep}}(t) = N_{ori} - N_{ter}, \tag{2.28}$$

and the number of individual replisomes will be twice the number of pairs. N_{ori} and N_{ter} are computed using Eq. 2.24.

For the number of replicating cells, we consider three different cases. First consider a case where the replication cycle is internal to the cell cycle. In this case, we have:

$$N_{\text{rep cell}}(t) = N_{[\tau_{ori},\tau_{ter}]} = N_{>\tau_{ori}} - N_{>\tau_{ter}} \tag{2.29}$$

which can be evaluated using Eq. 2.17. If the replication process overlaps by a single cell cycle but replication rounds do not overlap:

$$N_{\text{rep cell}}(t) = N_{>T+\tau_{ori}} + N_{<\tau_{ter}}, \tag{2.30}$$

where τ_{ori} is negative in this context, as noted in Sec. 2.2.2.4. Finally, if the rounds of replication overlap:

$$N_{\text{rep cell}}(t) = N(t), \tag{2.31}$$

and all cells are replicating in the deterministic model.

2.2.3 Stochastic model

An important complication of a more realistic model for cell cycle dynamics is stochasticity (i.e., randomness) in the lifetime of the states of the cell cycle. We will incorporate this stochasticity by dividing the cell cycle into m discrete states through which the cell must transition sequentially. This model is shown schematically in Fig. 2.5. The lifetime of each state $\tau_{\delta j}$ will be described by an arbitrary lifetime PDF, $p_{\delta j}(\cdot)$, for the jth state. It is important to note that this sequential-state stochastic model is not general enough to be an accurate representation of the bacterial cell cycle; however, it is sufficient to explore a number of interesting stochasticity-related phenomena and is exactly solvable.

2.2.3.1 Definition of the stochastic model

In our analysis, we will use the rate equation approach, rather than a master equation approach, since we are interested in the steady-state behavior of the model in the large cell number limit where the relative size of the fluctuations are vanishingly small.

Let $N_j(t)$ be the number of cells in state j, the cumulative creation number, $N_j^+(t)$, and the cumulative annihilation number, $N_j^-(t)$, be the total number of cells to have arrived and departed from state j over all time, respectively. The state dynamics is therefore described by the following rate equation:

$$\dot{N}_j = \dot{N}_j^+ - \dot{N}_j^-. \tag{2.32}$$

In this model, cells move sequentially through the m states before the final state ($j = m$) transitions to the initial state ($j = 1$) as two cells:

$$\dot{N}_j^+ = \begin{cases} \dot{N}_{j-1}^-, & j > 1, \\ 2\dot{N}_m^-, & j = 1. \end{cases} \tag{2.33}$$

Each state j has a PDF of lifetimes $p_{\delta j}(t)$ and therefore the relation between state j arrivals and departures is given by:

$$\dot{N}_j^- \;=\; p_{\delta j} \otimes \dot{N}_j^+, \tag{2.34}$$

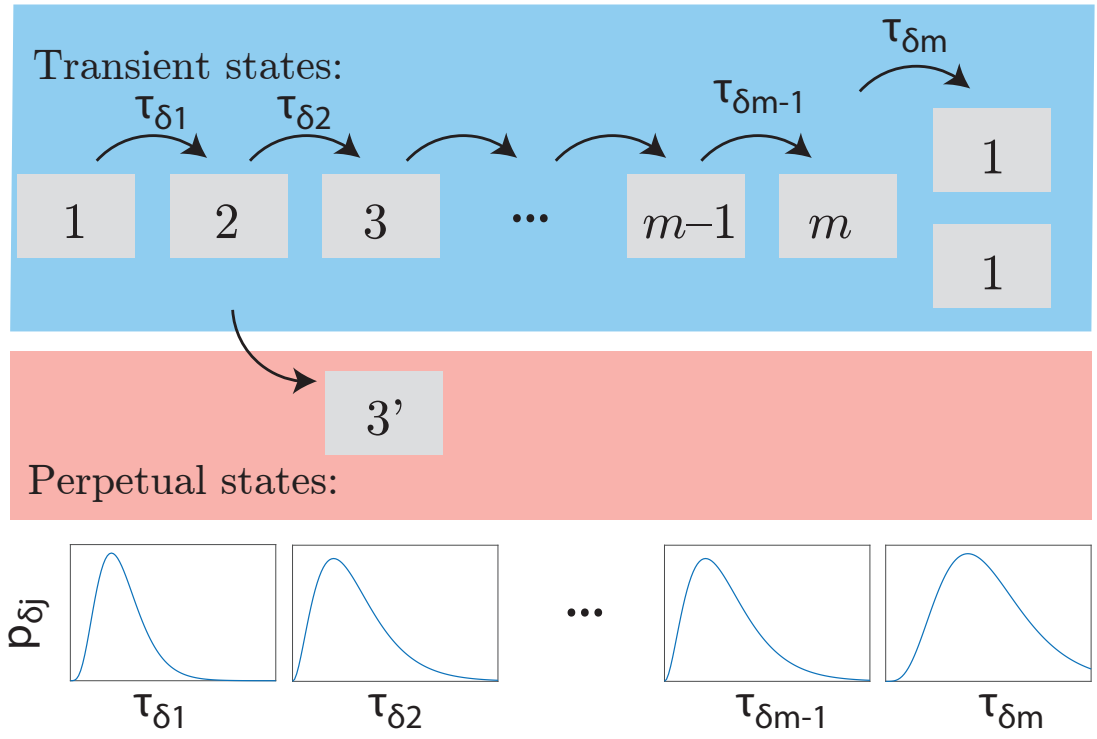


Figure 2.5: **Schematic of the stochastic model of the cell cycle.** We represent the cell cycle as a series of sequential states $j = 1...m$. After the mth state, two daughter cells in state 1 are born. The PDF of state lifetimes, $p_{\delta j}$, is distinct for each state j. Although most quantities of interest are transient, meaning that the states are *created* when the cells enter and then *annihilated* when the cells exit, we also consider perpetual quantities that are created but not annihilated (e.g., DNA loci).

where $\otimes$ is the convolution:

$$A \otimes B(t) \equiv \int_0^\infty \mathrm{d}t' \; A(t')\, B(t - t'). \tag{2.35}$$

Eqs. 2.32-2.34 completely specify the stochastic model.

2.2.3.2 Solution to the stochastic model

We will work in the steady-state growth limit as before. It is most convenient to work in terms of the Laplace transforms of the rate equations (Eqs. 2.32-2.34). The Laplace transform is defined:

$$\tilde{A}(\lambda) \equiv \int_0^\infty \mathrm{d}t \; A(t)\, e^{-\lambda t}, \tag{2.36}$$

where the tilde denotes the transformation from time t to Laplace conjugate λ.

The transformed representation is convenient since the ordinary differential equations become algebraic equations in terms of the transform quantities and the convolutions become products of transforms (e.g., [25]). Of particular importance in what follows is the relation between the PDF of lifetimes of individual state j, $p_{\delta j}(t)$, and the PDF of the age of the cell at the transition out of state j, $p_j(t)$:

$$\tilde{p}_j = \prod_{i=1}^{j} \tilde{p}_{\delta i}. \tag{2.37}$$

A detailed derivation of the solution to the rate equations (Eqs. 2.32-2.34) is given in Supplementary Sec. 2.5.5.

For steady-state exponential growth, the consistency condition that relates the growth rate k to the PDF of cell cycle durations $p(t) \equiv p_m(t)$ can be written concisely in terms of the Laplace transform:

$$1 = 2\,\tilde{p}(k), \tag{2.38}$$

an equation that is well known [26, 27]. This consistency condition is equivalent to Eq. 2.10 in the deterministic model, although the mathematical equivalence between these two relations is opaque for the moment.

2.2.3.3 The statistics of the stochastic model

In an exponential culture, the cell numbers are given by the expressions:

$$\text{Creation:} \quad N_j^+(t) = 2\tilde{p}_{j-1}N(t), \tag{2.39}$$

$$\text{In state} \leq j: \quad N_{\leq j}(t) = 2[1-\tilde{p}_j(k)]N(t), \tag{2.40}$$

$$\text{In state} > j: \quad N_{>j}(t) = [2\tilde{p}_j(k)-1]N(t), \tag{2.41}$$

$$\text{In state } j: \quad N_j(t) = 2[\tilde{p}_{j-1}(k)-\tilde{p}_j(k)]N(t), \tag{2.42}$$

which have a similar structure to the dynamics of the deterministic model (Eqs. 2.14-2.17), but are dependent on the Laplace transforms of the state lifetime PDFs. Intuitively, these Laplace transforms give rise to an effective mean time.

2.2.4 The exponential mean

To understand the biological significance of the Laplace transform of the lifetime PDF, consider the generalized f-mean (or Kolmogorov mean) where the random variable t is first transformed by function g, an arithmetic mean is performed, and then the inverse function is applied to generate a generalized expectation [28]:

$$\bar{t}[g] \equiv g^{-1}(\mathbb{E}_t g(t)), \tag{2.43}$$

where $\mathbb{E}_t$ is the arithmetic expectation over random variable t. Both the harmonic mean and geometric mean are special cases of this more general formulation. The Laplace transform of the lifetime and age PDFs can be reinterpreted as the expectation of $g(t) = \exp(-kt)$ and therefore we can generate the f-means:

$$\overline{\tau}_{\delta j}(k) \equiv -\tfrac{1}{k}\ln\tilde{p}_{\delta j}(k), \tag{2.44}$$

$$\overline{\tau}_j(k) \equiv -\tfrac{1}{k}\ln\tilde{p}_j(k), \tag{2.45}$$

which can be understood as the exponential-mean of the lifetime and age of state j respectively.

Before returning to our model, we will explore the behavior of the exponential mean. Consider the special case of a distribution that is very narrow relative to the growth rate. In this case:

$$\bar{t}(k) = \mathbb{E}_t t - \tfrac{1}{2} k \sigma_t^2 + ..., \tag{2.46}$$

where the exponential mean is equal to the mean to the order of the variance (σ_t^2) times the growth rate k. Details of the derivation are given in Supplementary Sec. 2.5.6. Short-lived states and states with small lifetime-variance will therefore have exponential means equal to the mean. More generally, the Jensen inequality always guarantees the exponential mean is less than or equal to the mean:

$$\bar{t}(k) \leq \mathbb{E}_t t, \tag{2.47}$$

since the function $g(t)$ is convex [29].

Finally, let us consider the consequences of a very wide distribution of lifetimes. Consider a state j in which fraction ϵ of cells arrest ($\tau_{\delta j} \to \infty$) while the remaining cells have exponential-mean lifetime $\overline{\tau}_{\delta j,0}$. Using Eq. 2.45, it is straightforward to compute the exponential-mean lifetime:

$$\overline{\tau}_{\delta j} = \overline{\tau}_{\delta j,0} + T \log_2 \tfrac{1}{1-\epsilon}, \tag{2.48}$$

where the second term acts to extend the lifetime by a positive multiple of the doubling time T. Although the arrested cells do lengthen the exponential-mean lifetime, it remains finite. Eq. 2.48 is a useful approximation anytime some fraction of the cells have a lifetime much longer than the doubling time even if all lifetimes are finite.

2.2.5 Model correspondence

To determine the differences between the deterministic and stochastic models, we eliminate the Laplace-transformed lifetime PDFs $\tilde{p}_{\delta j}$ in favor of the exponential-mean lifetimes $\overline{\tau}_{\delta j}$ using their definition (Eq. 2.45). First consider the consistency condition for exponential growth (Eq. 2.38). The convolution theorem ensures that the exponential-mean lifetimes of successive states add to generate the age of state j (e.g., the natural logarithm of Eq. 2.37).

Eq. 2.38 can now be rewritten as a relation between the exponential-mean cell-cycle duration and the doubling time:

$$\overline{T}_\tau \equiv \sum_{i=1}^{m} \overline{\tau}_{\delta i} = T, \tag{2.49}$$

which is now intuitively equivalent to the consistency condition in the deterministic model (Eq. 2.10). Details of the derivation are given in Supplementary Sec. 2.5.7.

Now consider the expressions for state number in the stochastic model (Eqs. 2.39-2.41). When the deterministic age τ is evaluated at exponential-mean stochastic age $\overline{\tau}_j$, the numbers are identical in the two models:

$$\text{Entered } j: \quad N_j^+(t) = N_\tau^+(t)\big|_{\tau=\overline{\tau}_{j-1}}, \tag{2.50}$$

$$\text{In state} \leq j: \quad N_{\leq j}(t) = N_{\leq \tau}(t)\big|_{\tau=\overline{\tau}_j}, \tag{2.51}$$

$$\text{In state} > j: \quad N_{>j}(t) = N_{>\tau}(t)\big|_{\tau=\overline{\tau}_j}, \tag{2.52}$$

$$\text{In state } j: \quad N_j(t) = N_{[\tau_{j-1},\tau_j]}(t). \tag{2.53}$$

We therefore conclude that the statistics of the deterministic and stochastic models are identical in an exponential culture for models with the same growth rate k, once a correspondence has been established between states j and ages τ. In the deterministic model, state j corresponds to times $\tau \in [\overline{\tau}_{j-1}, \overline{\tau}_j]$ where $\overline{\tau}_0 \equiv 0$ and $\overline{\tau}_m = T$. This correspondence is illustrated schematically in Fig. 2.6. Since we demonstrated a correspondence between the models, almost all the application discussed in Sec. 2.2.2 generalize by replacing the deterministic time τ with the corresponding exponential mean $\overline{\tau}$. The exceptions are the replicating cell statistics $N_{\text{rep cell}}$. In this case, the implicit assumption that states are sequential cannot always be implemented in the stochastic model. For instance, consider the re-initiation of replication at the quarter cell positions late in the cell cycle in rapidly proliferating cells, as illustrated in Fig. 2.1. If these events were modeled as independent, you will first see one replisome initiate at one quarter cell position and then the other. In this case one must compute the exponential means using the order statistics (e.g., [30]).

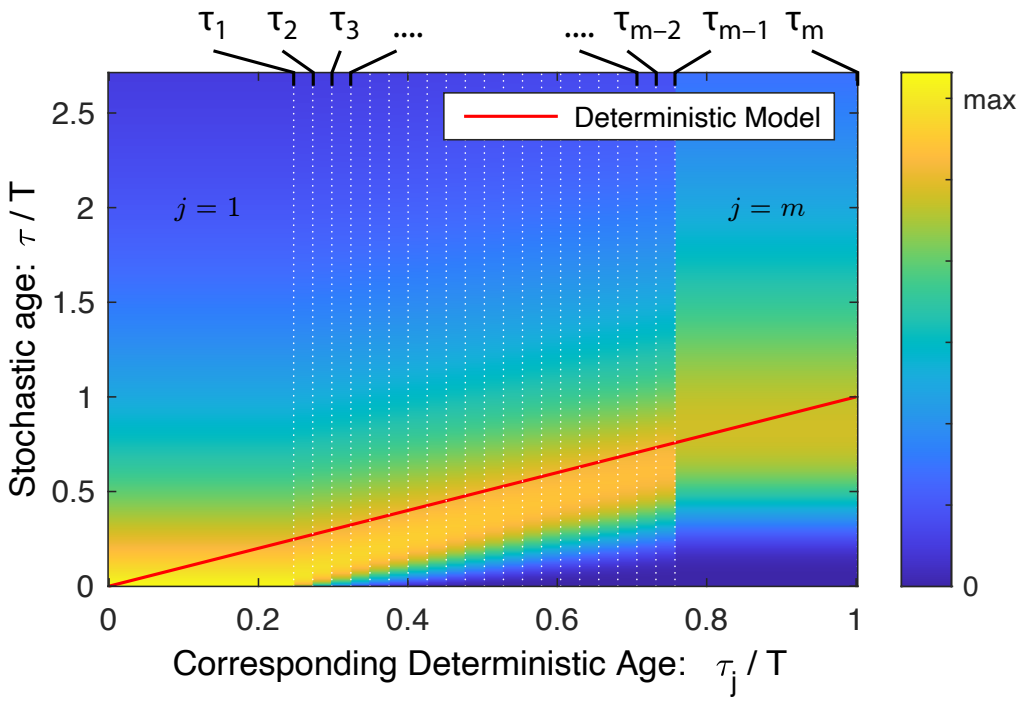


Figure 2.6: **Model correspondence.** The deterministic and stochastic models generate identical statistics in exponential growth once a suitable correspondence is defined between cell state j in the stochastic model and cell age τ in the deterministic model. State j in the stochastic model corresponds to age interval $(\overline{\tau}_{j-1}, \overline{\tau}_j]$ in the deterministic model, which is represented by the red line. In the stochastic model, the PDF of age as a function of state j is shown. Qualitatively, the age in the deterministic model tracks with the mode of the age in the stochastic model.

2.2.6 Implications for cell cycle phenomenology

To explore the nontrivial consequences of stochasticity in timing, consider an example motivated by replication conflicts [31]: By visualizing the replisome dynamics using single-molecule microscopy, we have recently reported that transcription leads to pervasive replisome instability [23]. To what extent should conflict-induced pauses in replication have been detectable in the classic analyses of unsynchronized cell populations?

Consider a simplified model in which an experiment probes the difference between the wildtype strain W and two mutant strains. The wildtype W grows with deterministic C period C_{W} and deterministic cell cycle duration T_{W}. In mutant strain A(rrest), a small fraction ϵ of cells arrest during replication (i.e., C period) and never complete the cell cycle, whereas non-arrested cells are identical to wildtype cells. In mutant strain S(low), the replication process is $1+\epsilon'$ times slower, but the B and D periods are identical to wildtype. Using Eq. 2.48, one can compute the C period duration C and doubling time T. To lowest order in ϵ, the inferred cell cycle durations and C period are presented in Tab. 2.2.

Table 2.2: The effect of mutants on the doubling time T and C period duration C of an exponential culture.

Strain	Doubling Time: T	C period: C
Wildtype	T_{W}	C_{W}
Mutant A	$T_{\mathrm{W}} + \frac{\epsilon}{\ln 2} T_{\mathrm{W}}$	$C_{\mathrm{W}} + \frac{\epsilon}{\ln 2} T_{\mathrm{W}}$
Mutant S	$T_{\mathrm{W}} + \epsilon' C_{\mathrm{W}}$	$C_{\mathrm{W}} + \epsilon' C_{\mathrm{W}}$

At an intuitive level, one aspect of the prediction is easy to understand: In both mutants the C period is lengthened, as one might naively expect since this is the replication period of the cell cycle. Furthermore, the doubling time increases by the same amount as the C period increases. But there is an aspect of this prediction which is perhaps less intuitive: One might naively expect to observe a more dramatic consequence of replication arrest, like a large buildup of C period cells, but the consequences are indistinguishable from a slowdown

in an exponential culture. In both mutants, there is a slight lengthening of the inferred C period, even though the slowdown is caused by replication arrest in the context of the A mutant. Although this prediction is not new in a qualitative sense, it concisely illustrates how the statistics of the exponential culture mask two mechanistically distinct phenomena.

The statistics of an exponential culture can also generate distinctions where seemingly none exist. Consider a more realistic model in which the duration of the D period is stochastic, has a non-zero width, and is *identical* for all three strains. The more rapid growth rate of the wildtype strain implies that its effective D period is shorter than for mutants cells:

$$D_{\mathrm{S}}, D_{\mathrm{A}} > D_{\mathrm{W}}, \tag{2.54}$$

even though the distributions of the D period durations are identical in all three strains. (To understand how this occurs, see the second term on the RHS of Eq. 2.46.) In most cases this effect should be subtle, but for large changes in growth rate, these changes could be quite significant and can clearly complicate the interpretation of effective period lifetimes in an exponential culture.

2.3 Discussion

In this paper, we provide a detailed analysis of both deterministic and stochastic models of the cell cycle. In Sec. 2.2.1, we solved the deterministic model in which the cell-aging and division processes are precisely timed and determined the demographics (i.e., statistics) of an exponential culture. Given a set of observed demographics, Sec. 2.2.2 provides a detailed road map for how to infer cell cycle state timing in the context of the deterministic model. In Sec. 2.2.3, we solved the more realistic stochastic model in which the lifetimes of sequential states are stochastic and again we determined the demographics of an exponential culture. By defining an exponential mean in Sec. 2.2.4, we demonstrated that the statistics of the two models were equivalent in Sec. 2.2.5. The effective lifetime of states in the deterministic model is the exponential mean of the lifetimes in the stochastic model. That is to say that the

exponential-mean lifetimes are the *sufficient statistics* of the model (e.g., [30]): Knowledge of only these lifetime statistics predicts the demographics of the exponential culture; therefore, inference on exponential-culture demographics infers only the exponential means, rather than the underlying lifetime distributions themselves. Finally in Sec. 2.2.6, we discussed some of the limitations of the exponential-mean lifetimes in resolving the underlying biological mechanisms.

2.3.1 Applicability of the stochastic model

Is the stochastic model sufficiently complex to capture all the relevant cell cycle phenomenology in *E. coli* and other bacterial systems? Like the deterministic model before it, the stochastic model is an idealized model that is simple enough to be tractable analytically, but complex enough to capture some important phenomenology. There are a number of shortcomings of this model but perhaps the most significant is that there is *no memory* beyond the cell state index j. As a consequence, it makes predictions at variance with some observed phenomenology: For instance, the stochastic model must predict that successive cell cycle durations are uncorrelated; however, these correlations are observed [5, 32]. (We briefly consider the implications of a more general model in Supplementary Sec. 2.5.8.) Another important limitation of the stochastic model is that cell divisions are symmetric, which is a good approximation in *E. coli*, but these types of stochastic models can easily be extended to the general asymmetric division case (e.g., [27]).

2.3.2 On the applicability of the exponential mean

Although the definition of the exponential mean was motivated by the correspondence between the deterministic and stochastic models, it almost certainly has much greater applicability to other more complicated scenarios. For instance, our own numerical experiments using more complex models suggest that the relation between the effective lifetime of the states and the exponential-mean lifetime appears to be more robust than the assumptions of the stochastic model might imply. Since the key mechanism for generating

bias toward short times is steady-state exponential growth, we expect the exponential mean of wait times to be the determinative statistic in more general models, as demonstrated in Supplementary Sec. 2.5.8. As such, the exponential-mean lifetime could be a powerful observable to bridge timescales between single-cell and culture phenomenology in two different contexts: (i) in experiments probing cell cycle dynamics at the single-cell level and (ii) in complex numerical simulations that are too slow and too memory intensive to simulate in the long time limit.

We should note that although we believe our interpretation of the doubling time as an exponential mean (Eq. 2.49) is novel, it has already been appreciated in two important respects: (i) From a computational perspective, the Laplace-transform formulation (Eq. 2.38) of Eq. 2.49 has long been known [26]. (ii) From a qualitative perspective, biologists have long understood the consequences of the exponential-mean lifetime on cell growth rate: I.e., the doubling time T is "an average" of the cell cycle duration T_τ; however, a small arrested subpopulation, for whom $T_\tau \to \infty$, slows but does not stop growth. There is also physical precedent for this type of mean: Intriguingly, it emerges in the context of non-equilibrium statistical mechanics [33], although what connection this has to our cellular dynamics is opaque.

2.3.3 On the significance of stochasticity

How does stochasticity affect biological function? Experimentally, we have long known that although the statistics of an exponentially growing population are well described by the deterministic model [3], nontrivial stochasticity in cell cycle timing is observed [5, 6]. It is therefore tempting to conclude, based on the literature and perhaps even our own results, that stochasticity is either *small* or simply *does not* significantly affect biological function.

Our own conclusions are much more nuanced. Although our results guarantee that the deterministic model fits the exponential-culture demographic data just as well as the stochastic model, we have demonstrated that the stochasticity in timing is hidden in plain sight. The distribution of state lifetimes determine the exponential means. Therefore, the

success of the deterministic model should not be interpreted as evidence against stochasticity or against its importance, but rather it indicates that only the exponential-mean state lifetimes are determinative parameters in the model for the demographics of an exponential culture.

Perhaps more than anything else, the exact correspondence between the deterministic and stochastic models emphasizes the need for synchronized single-cell measurements: In Sec. 2.2.6, we illustrated (i) how similarities in the effective duration of the C period obscures distinct biological mechanisms as well as (ii) how differences in the effective D period could belie an identical mechanism.

At a mechanistic level, stochasticity plays a central role in many processes. For instance, the mechanism that restarts replication will prevent the existence of a *fat tail* on the distribution of C periods [23, 31, 34]. Although the existence of the fat tail—i.e., a small number of cells with very long C periods—does not *break* the correspondence with the deterministic model, it does increase the exponential-mean C period, which in turn decreases the growth rate. (E.g., see Tab. 2.2.) Since the growth rate is decreased, there is a strong selective pressure to reduce *stochasticity*. This argument predicts the existence of biological mechanisms to reduce stochasticity, as are already known in many contexts (e.g., replication restart). In fact, the subtle signature of stochasticity suggests an interesting hypothesis: a significant number of mutants that are currently known to reduce growth rate may in fact generate this phenotype by increasing the level of stochasticity in the cell cycle duration. Single-cell experimental analysis must play a central role in understanding these phenomena.

2.4 Acknowledgments

We acknowledge advice and comments from M. Cosentino-Lagomarsino, S. Iyer-Biswas, P. Levine, J. Mittler, R. Phillips, M. Transtrum, B. Traxler, I.M. Shelby, and H.K. Choi. This work was supported by NIH grant R01-GM128191.

2.5 Supplemental derivations

2.5.1 Derivation of the rate equation in the deterministic model

To obtain the cumulative creation and annihilation numbers in terms of the number density (Eq. 2.2), we integrate the number density at fixed age τ over all time t to obtain the cumulative number of cells that have ever entered (creation) or left (annihilation) age τ. They are equivalent due to the continuous nature of the deterministic model. If states were discrete, as in the stochastic model, then the cumulative creation and annihilation numbers would differ by the number of cells currently in the state τ.

To obtain Eq. 2.4 from Eq. 2.3, we divide both sides of Eq. 2.3 by $\delta t\, \delta\tau$ and replace $\dot{N}^{-}_{\tau+\delta\tau}(t)$ with the equivalent $\dot{N}^{+}_{\tau+\delta\tau}(t)$, which leaves:

$$\dot{n}_\tau(t) = -\frac{\dot{N}^{+}_{\tau+\delta\tau}(t) - \dot{N}^{+}_{\tau}(t)}{\delta\tau}. \tag{2.55}$$

Taking the limit as $\delta\tau$ goes to 0 and using the definition of a derivative, we are left with:

$$\dot{n}_\tau(t) = -\partial_\tau \dot{N}^{+}_{\tau}(t), \tag{2.56}$$

which is Eq. 2.4 in the main text. Eqs. 2.5-2.6 follow from taking the partial time derivative of Eq. 2.2 and taking into account the consistency condition:

$$n_0(t) = 2n_{T_\tau}(t). \tag{2.57}$$

Conceptually, this consistency condition describes how cell division at age T_τ leads to twice as many daughter cells of age $\tau = 0$.

2.5.2 Derivation of the solution in the deterministic model

In the deterministic model, we can assume that steady-state growth of the population is represented by an exponentially increasing time dependence factor, e^{kt}, with a constant unknown growth rate k. This assumption holds in the long time limit, since only the fastest growing mode remains in exponential growth, while all others (smaller k) are diluted out.

We thus stipulate that in the deterministic model, all cellular quantities must grow with this same time dependence. The number density is then a solution of the form:

$$n_\tau(t) = n_\tau(0)\, e^{kt}, \tag{2.58}$$

where $n_\tau(0)$ represents the initial age distribution at $t = 0$. Plugging this into Eq. 2.7 for the $\tau > 0$ case yields:

$$k\, n_\tau(t) = -\tfrac{\partial}{\partial \tau}(n_\tau(t)). \tag{2.59}$$

This can then be integrated to yield the solution:

$$\begin{aligned} n_\tau(t) &= n_0(t)\, e^{-k\tau}, & (2.60)\\ &= n_0\, e^{kt} e^{-k\tau}, & (2.61) \end{aligned}$$

where n_0 is a constant determined by the initial cell number. This equation appears in the main text as Eq. 2.11. To satisfy the $\tau = 0$ case of Eq. 2.7, we must use the consistency condition (Eq. 2.57):

$$\begin{aligned} n_0(t) &= 2n_{T_\tau}(t), & (2.62)\\ n_0 e^{kt} &= 2n_0 e^{kt} e^{-kT_\tau}. & (2.63) \end{aligned}$$

Dividing both sides by $n_0 e^{kt}$ and solving for T_τ gives:

$$T_\tau = k^{-1} \ln 2. \tag{2.64}$$

Therefore, the doubling time defined in Eq. 2.9 is equivalent to the cell cycle duration, as one would naively expect. Furthermore, this equation relates the growth rate k, a population measure, to the cell cycle duration T_τ, a single-cell measure.

2.5.3 Derivation of the PDF and CDF in the deterministic model

To obtain the probability distribution function, $f_\tau(\tau)$, with respect to cell age (Eq. 2.12), we must normalize the number density at any fixed time t:

$$f_\tau(\tau) = \frac{n(\tau)}{\int_0^{T_\tau} n(\tau)\mathrm{d}\tau}, \tag{2.65}$$

where $n(\tau) = n(\tau = 0)\, e^{-k\tau}$, which is just Eq. 2.11 with the fixed t factor absorbed into $n(\tau = 0)$. Evaluating the integral and replacing $n(\tau)$ with the expanded form yields:

$$f_\tau(\tau) = \frac{n(0)\, e^{-k\tau}}{-n(0)\frac{1}{k}(e^{-kT_\tau} - 1)}, \tag{2.66}$$

$$= \frac{ke^{-k\tau}}{1 - e^{-kT_\tau}}. \tag{2.67}$$

Now recall that $T_\tau = T \equiv k^{-1} \ln 2$, from Eqs. 2.9-2.10. Plugging this T_τ into Eq. 2.67 gives Eq. 2.12:

$$f_\tau(\tau) = 2ke^{-k\tau}. \tag{2.68}$$

To obtain the CDF, integrate with respect to τ from 0 to τ:

$$F_\tau(\tau) = \int_0^\tau 2ke^{-k\tau'} \mathrm{d}\tau' \tag{2.69}$$

$$= 2(1 - e^{-k\tau}). \tag{2.70}$$

2.5.4 Derivation of the cumulative creation number in terms of $N(t)$

Consider the cumulative creation number Eq. 2.2 evaluated at $\tau = 0$:

$$N_0^+(t) = 2N(t), \tag{2.71}$$

which is double the current number of cells N. The factor of two arises due to the cumulative nature of $N_\tau^+(t)$. To understand this intuitively, consider the total number of cells in each generation:

$$N_0^+(t) = (1 + \tfrac{1}{2} + \tfrac{1}{4} + ...)N(t), \tag{2.72}$$

which is a geometric series and can be summed to $2N$, matching Eq. 2.71. We can now multiply by the τ-dependence term, $e^{-k\tau}$, to obtain:

$$N_\tau^+(t) = 2e^{-k\tau}N(t). \tag{2.73}$$

More formally, we can use direct integration of Eq. 2.11:

$$\begin{aligned} N_\tau^+(t) &= \int n_\tau(t')\mathrm{d}t', && (2.74)\\ &= \tfrac{1}{k} n_0 e^{-k\tau} e^{kt} + c, && (2.75)\\ N_0^+(t) &= \tfrac{1}{k} n_0 e^{kt} + c, && (2.76) \end{aligned}$$

where c is an integration constant. We now integrate $n_\tau(t)$ over all τ to obtain the total number of cells at time t:

$$\begin{aligned} N(t) &= \int_0^{T_\tau} n_{\tau'}(t)\mathrm{d}\tau', && (2.77)\\ &= -\tfrac{1}{k} n_0 e^{kt}\big(e^{-kT_\tau} - 1\big), && (2.78)\\ &= \tfrac{1}{2k} n_0 e^{kt}. && (2.79) \end{aligned}$$

Combining this with the consistency condition Eq. 2.71, we get:

$$N_0^+(t) = \tfrac{1}{k} n_0 e^{kt}. \quad (2.80)$$

Setting this equal to Eq. 2.76 allows us to set the integration constant $c = 0$. Eq. 2.75 then becomes Eq. 2.73 from above, which is Eq. 2.14 from the main text.

2.5.5 Derivation of the solution in the stochastic model

Clearly, Eqs. 2.33 and 2.34 can be combined recursively to generate a relation between the number of cells entering states 1 and j. First let us define the state-transition time PDF p_j, describing the total time taken to transition from birth through state j:

$$\begin{aligned} \tilde{p}_j &\equiv \prod_{i=1}^{j} \tilde{p}_{\delta i}, && (2.81)\\ \tilde{p} &\equiv \tilde{p}_m && (2.82) \end{aligned}$$

where p is the lifetime PDF for the entire cell cycle. We then write an expression of the number arriving in state j:

$$\tilde{N}_j^+ = \tilde{p}_{j-1}\tilde{N}_1^+. \quad (2.83)$$

As before, using this same condition at the end of the cell cycle gives rise to a consistency condition:

$$\tilde{N}_1^+ = 2\tilde{p}\tilde{N}_1^+. \tag{2.84}$$

It follows that in steady-state exponential growth, the growth rate k must correspond to the solution to the equation:

$$1 = 2\tilde{p}(k), \tag{2.85}$$

an equation that is well known [26].

Let $N_{\leq j}$ be the total number of cells in states $i = 1...j$. The dynamics of this quantity has a simple form due to the telescoping form of the dynamics equations (Eqs. 2.32-2.34) where the number entering the ith state exactly cancel the number leaving the $i-1$th state:

$$\begin{aligned} \tilde{N}_{\leq j} &= \tilde{N}_1^+ - \tilde{N}_{j+1}^+, & (2.86)\\ &= (1-\tilde{p}_j)\tilde{N}_1^+. & (2.87) \end{aligned}$$

To determine the overall normalization, we can sum up the cells in all states and set that sum equal to the total number of cells $N(t)$:

$$\tilde{N} = \tilde{N}_{\leq m} = \tfrac{1}{2}\tilde{N}_1^+. \tag{2.88}$$

From $\tilde{N}_{\leq j}$, we can compute the number in individual states:

$$\begin{aligned} \tilde{N}_j &= \tilde{N}_{\leq j} - \tilde{N}_{\leq j-1}, & (2.89)\\ &= (\tilde{p}_{j-1} - \tilde{p}_j)\tilde{N}_1^+. & (2.90) \end{aligned}$$

In the long time limit, the fastest growing mode dominates the solution and therefore:

$$\begin{aligned} N(t) &= N(0)\ e^{-kt}, & (2.91)\\ N_j(t) &= 2[\tilde{p}_{j-1}(k) - \tilde{p}_j(k)]N(t), & (2.92)\\ N_{\leq j}(t) &= 2[1-\tilde{p}_j(k)]N(t), & (2.93)\\ N_j^+(t) &= 2\tilde{p}_{j-1}N(t), & (2.94) \end{aligned}$$

which also appear in the main text.

2.5.6 The exponential mean of a very narrow distribution

To obtain Eq. 2.46, we begin with the exponential mean:

$$\bar{t}(k) = -\tfrac{1}{k} \ln \mathbb{E}_t[\exp(-kt)], \tag{2.95}$$

which is obtained by using the function $g(t) = \exp(-kt)$ in the generalized expectation equation (Eq. 2.43). We then use a series expansion of the exponential term, $e^x = 1 + x + \frac{1}{2}x^2 + ...$, which yields:

$$\begin{aligned} \bar{t}(k) &= -\tfrac{1}{k} \ln \big(\mathbb{E}_t\big[1 - kt + \tfrac{1}{2}k^2t^2\big] + ...\big), & (2.96)\\ &= -\tfrac{1}{k} \ln \big(1 - k\mathbb{E}_t t + \tfrac{1}{2}k^2\mathbb{E}_t[t^2] + ...\big). & (2.97) \end{aligned}$$

We then use the series expansion $\ln(1 + x) = x - \frac{1}{2}x^2$, keeping only second order terms, since the distribution is very narrow:

$$\begin{aligned} \bar{t}(k) = -\tfrac{1}{k}\big(&- k\mathbb{E}_t t + \tfrac{1}{2}k^2\mathbb{E}_t[t^2] \\ &- \tfrac{1}{2}\big[-k\mathbb{E}_t t + \cancel{\tfrac{1}{2}k^2\mathbb{E}_t[t^2]}\big]^2 + ...\big) & (2.98)\\ = -\tfrac{1}{k}\big(&-k\mathbb{E}_t t + \tfrac{1}{2}k^2\big[\mathbb{E}_t[t^2] - (\mathbb{E}_t t)^2\big] + ...\big) & (2.99) \end{aligned}$$

Using the definition of the variance, $\sigma_t^2 = \mathbb{E}_t[t^2] - (\mathbb{E}_t t)^2$, we obtain Eq. 2.46:

$$\bar{t}(k) = \mathbb{E}_t t - \tfrac{1}{2}k\sigma_t^2 + ... \tag{2.100}$$

2.5.7 Derivation of the consistency condition in the stochastic model

In the deterministic model, Eq. 2.10 is a consistency condition that describes the naive expectation that the duration of the cell cycle is equal to the doubling time of the population:

$$T_\tau = T. \tag{2.101}$$

In the stochastic model, the consistency condition in terms of the Laplace transform is given by Eq. 2.38:

$$1 = 2\tilde{p}(k). \tag{2.102}$$

However, the mathematical equivalence is opaque for the moment. To make the equivalence clear, we use the exponential mean Eq. 2.45:

$$\overline{\tau}_j(k) \equiv -\tfrac{1}{k} \ln \tilde{p}_j(k), \tag{2.103}$$

along with the relation between the PDF of lifetimes of individual state j and the PDF of times taken to transition from birth through state j:

$$\tilde{p}_j = \prod_{i=1}^{j} \tilde{p}_{\delta i}, \tag{2.104}$$

which is Eq. 2.37 in the main text. Combining these two equations and letting j be the final state m, we obtain the exponential mean of the stochastic cell cycle duration:

$$\begin{aligned} \overline{T}_\tau &= -\tfrac{1}{k} \ln \left(\prod_{i=1}^{m} \tilde{p}_{\delta_i} \right), & (2.105) \\ &= \sum_{i=1}^{m} -\tfrac{1}{k} \ln \tilde{p}_{\delta_i} = \sum_{i=1}^{m} \overline{\tau}_{\delta_i}. & (2.106) \end{aligned}$$

If we use the consistency condition Eq. 2.38, we can also write:

$$\begin{aligned} \overline{T}_\tau &= -\tfrac{1}{k} \ln \tilde{p}(k), & (2.107) \\ &= k^{-1} \ln 2 = T, & (2.108) \end{aligned}$$

where the second equality came from the definition of the doubling time (Eq. 2.9 in the main text). We thus recover Eq. 2.49 from the main text:

$$\overline{T}_\tau \equiv \sum_{i=1}^{m} \overline{\tau}_{\delta i} = T, \tag{2.109}$$

which corresponds to the consistency condition in the deterministic model, Eq. 2.10.

2.5.8 A generalization of the stochastic model

Like the deterministic model before it, it seems almost certain that the phenomenology of the stochastic model is more general than some of the assumptions made to motivate and

derive it. In particular, the qualitative mechanism that makes the exponential-mean of state lifetime the determinative statistic would seem to depend only on the exponential enrichment of young cells in an exponential culture and not on the details of the sequential state structure of the stochastic model. We therefore offer a slightly more general derivation below.

In the generalized model, assume only that state or object j is created with wait time distribution p' relative to the birth of a new cell and assume steady-state growth at rate k. Under these assumptions, $\tilde{p}'$ replaces $\tilde{p}_{j-1}$ in Eqs. 2.83 and 2.94, even if k is not determined by Eq. 2.85 due to memory effects. Therefore, most of our results generalize in this new model if the suitable PDFs for the wait times replace the p_j's.

2.6 Bibliography

[1] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014 420, 1 Jan. 2022. DOI: 10.1103/PhysRevE.105.014420.

[2] L. Willis and K. C. Huang, "Sizing up the bacterial cell cycle," *Nat Rev Microbiol*, vol. 15, no. 10, pp. 606–620, Oct. 2017. DOI: 10.1038/nrmicro.2017.79.

[3] 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.

[4] 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.

[5] P. Wang *et al.*, "Robust growth of *Escherichia coli*," *Curr Biol*, vol. 20, no. 12, pp. 1099–103, Jun. 2010. DOI: 10.1016/j.cub.2010.04.045.

[6] L. Robert, M. Hoffmann, N. Krell, S. Aymerich, J. Robert, and M. Doumic, "Division in *Escherichia coli* is triggered by a size-sensing rather than a timing mechanism," *BMC Biol*, vol. 12, p. 17, Feb. 2014. DOI: 10.1186/1741-7007-12-17.

[7] X. Wang, C. Possoz, and D. J. Sherratt, "Dancing around the divisome: Asymmetric chromosome segregation in *Escherichia coli*," *Genes Dev*, vol. 19, no. 19, pp. 2367–77, Oct. 2005. DOI: 10.1101/gad.345305.

[8] 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.

[9] C. J. Rudolph, A. L. Upton, A. Stockum, C. A. Nieduszynski, and R. G. Lloyd, "Avoiding chromosome pathology when replication forks collide," *Nature*, vol. 500, no. 7464, pp. 608–11, Aug. 2013. DOI: 10.1038/nature12312.

[10] 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.

[11] N. J. Kuwada, B. Traxler, and P. A. Wiggins, "Genome-scale quantitative characterization of bacterial protein localization dynamics throughout the cell cycle," *Mol Microbiol*, vol. 95, no. 1, pp. 64–79, Jan. 2015. DOI: 10.1111/mmi.12841.

[12] D. W. Adams and J. Errington, "Bacterial cell division: Assembly, maintenance and disassembly of the Z ring," *Nat Rev Microbiol*, vol. 7, no. 9, pp. 642–53, Sep. 2009. DOI: 10.1038/nrmicro2198.

[13] T. Lo, D. Huang, H. Merrikh, and P. A. Wiggins, *In preparation*,

[14] E. J. Stewart, R. Madden, G. Paul, and F. Taddei, "Aging and death in an organism that reproduces by morphologically symmetric division," *PLoS Biol*, vol. 3, no. 2, e45, Feb. 2005. DOI: 10.1371/journal.pbio.0030045.

[15] S. Du and J. Lutkenhaus, "At the heart of bacterial cytokinesis: The Z ring," *Trends Microbiol*, vol. 27, no. 9, pp. 781–791, Sep. 2019. DOI: 10.1016/j.tim.2019.04.011.

[16] X. Ma, D. W. Ehrhardt, and W. Margolin, "Colocalization of cell division proteins FtsZ and FtsA to cytoskeletal structures in living *Escherichia coli* cells by using green fluorescent protein," *Proc Natl Acad Sci U S A*, vol. 93, no. 23, pp. 12 998–3003, Nov. 1996. DOI: 10.1073/pnas.93.23.12998.

[17] H. Niki, Y. Yamaichi, and S. Hiraga, "Dynamic organization of chromosomal DNA in *Escherichia coli*," *Genes Dev*, vol. 14, no. 2, pp. 212–23, Jan. 2000.

[18] I. F. Lau, S. R. Filipe, B. Søballe, O.-A. Økstad, F.-X. Barre, and D. J. Sherratt, "Spatial and temporal organization of replicating *Escherichia coli* chromosomes," *Mol Microbiol*, vol. 49, no. 3, pp. 731–43, Aug. 2003. DOI: 10.1046/j.1365-2958.2003.03640.x.

[19] R. E. Bird, J. Louarn, J. Martuscelli, and L. Caro, "Origin and sequence of chromosome replication in *Escherichia coli*," *J Mol Biol*, vol. 70, no. 3, pp. 549–66, Oct. 1972. DOI: 10.1016/0022-2836(72)90559-1.

[20] R. H. Pritchard, M. G. Chandler, and J. Collins, "Independence of F replication and chromosome replication in *Escherichia coli*," *Mol Gen Genet*, vol. 138, no. 2, pp. 143–55, 1975. DOI: 10.1007/BF02428118.

[21] K. P. Lemon and A. D. Grossman, "Movement of replicating DNA through a stationary replisome," *Mol Cell*, vol. 6, no. 6, pp. 1321–30, Dec. 2000. DOI: 10.1016/s1097-2765(00)00130-1.

[22] M. Wallden, D. Fange, E. G. Lundius, Ö. Baltekin, and J. Elf, "The synchronization of replication and division cycles in individual *E. coli* cells," *Cell*, vol. 166, no. 3, pp. 729–739, Jul. 2016. DOI: 10.1016/j.cell.2016.06.052.

[23] 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.

[24] 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.

[25] P. A. Wiggins, R. Phillips, and P. C. Nelson, "Exact theory of kinkable elastic polymers," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 71, no. 2 Pt 1, p. 021 909, Feb. 2005. DOI: 10.1103/PhysRevE.71.021909.

[26] E. Powell, "Growth rate and generation time of bacteria, with special reference to continuous culture," *Microbiology*, vol. 15, no. 492, 1956.

[27] F. Jafarpour *et al.*, "Bridging the timescales of single-cell and population dynamics," *Phys. Rev. X*, vol. 8, p. 021 007, 2 Apr. 2018. DOI: 10.1103/PhysRevX.8.021007.

[28] A. Kolmogorov, "Sur la notion de la moyenne," *Atti Accad. Naz. Lincei*, vol. 12, pp. 388–391, 1930.

[29] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes.," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.

[30] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. Chapman & Hall, 1974.

[31] H. Merrikh, Y. Zhang, A. D. Grossman, and J. D. Wang, "Replication-transcription conflicts in bacteria," *Nat Rev Microbiol*, vol. 10, no. 7, pp. 449–58, Jun. 2012. DOI: 10.1038/nrmicro2800.

[32] J. Lin and A. Amir, "The effects of stochasticity at the single-cell level and cell size control on the population growth," *Cell Systems*, vol. 5, no. 4, 2017. DOI: 10.1016/j.cels.2017.08.015.

[33] C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.*, vol. 78, no. 14, p. 2690, 1997.

[34] H. Merrikh, C. Machón, W. H. Grainger, A. D. Grossman, and P. Soultanas, "Co-directional replication-transcription conflicts lead to replication restart," *Nature*, vol. 470, no. 7335, pp. 554–7, Feb. 2011. DOI: 10.1038/nature09758.

Chapter 3

The *in vivo* measurement of replication fork velocity and pausing by lag-time analysis

Originally published as: [1] 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," *Nat Commun*, vol. 14, no. 1, p. 1762, Mar. 2023. DOI: 10.1038/s41467-023-37456-2.

Author contributions: D.H., A.E.J., T.W.L., H.M. and P.A.W. designed the experiments. D.H., A.E.J., and B.S.S. assembled input data and ran experiments. D.H. and P.A.W. developed the theory, wrote code, ran the model, and analyzed output data. D.H., A.E.J., H.M., and P.A.W. wrote the manuscript.

Abstract

An important step towards understanding the mechanistic basis of the central dogma is the quantitative characterization of the dynamics of nucleic-acid-bound molecular motors in the context of the living cell. To capture these dynamics, we develop lag-time analysis, a method for measuring *in vivo* dynamics. Using this approach, we provide quantitative locus-specific measurements of fork velocity, in units of kilobases per second, as well as replisome-pause durations, some with the precision of seconds. The measured fork velocity is observed to be both locus and time dependent, even in wild-type cells. In this work, we quantitatively characterize known phenomena, detect brief, locus-specific pauses at ribosomal DNA loci in wild-type cells, and observe temporal fork velocity oscillations in three highly-divergent bacterial species.

3.1 Introduction

At a single-molecule scale, all cellular processes are both highly stochastic as well as subject to a crowded cellular environment where they typically compete with a large number of potentially-antagonistic processes that share the same substrate [2, 3]. In spite of these challenges, essential processes must be robust at a cellular scale to facilitate efficient cellular proliferation. Understanding how these processes are regulated to achieve robustness remains an important and outstanding biological question [4–10]. However, a central challenge in investigating these questions is the quantitative characterization of the activity of enzymes in the context of the living cell. For instance, although single-molecule assays can resolve the pausing of molecular motors on nucleic-acid substrates in the context of *in vitro* measurements [11, 12], performing analogous measurements in the physiologically-relevant environment of the cell, where these processes are subject to antagonism, poses a severe challenge to the existing methodologies [13].

In this paper, we develop an approach, lag-time analysis, that facilitates the quantitative characterization of dynamics, with resolution of seconds, in the context of the living cell. The approach exploits exponential growth as the stopwatch to capture dynamics in exponentially-proliferating cellular cultures [14] and unlike competing approaches, it can circumvent the difficulties and potential artifacts introduced by cell synchronization [15] or fluorescent labeling. Lag-time analysis exploits the same data as marker frequency analysis, but it directly measures the locus-specific fork velocity, in units of kilobases per second, and the duration of replisome pauses in seconds. Lag-time analysis facilitates detailed comparisons to be made, not just between different loci in a single cell, but between wild-type and mutant cells as well as between bacterial species and unlike a recent competing analysis, no detailed stochastic models or simulations are employed [16]. We apply this approach to analyze three model bacterial systems: *Bacillus subtilis*, *Vibrio cholerae*, and *Escherichia coli*. In *B. subtilis*, we analyze transcription-induced replication antagonism which is the main determinant of replisome dynamics in a set of mutants with

retrograde (reverse-oriented) fork motion. An analysis of *V. cholerae* provides evidence that fork number is an important determinant of fork velocity, but also provides clear evidence that fork velocity is time dependent. To explore this time-dependence, we analyze the fork-velocity in *E. coli* which provides strong evidence for temporally-oscillating fork velocity, consistent with a recent report [16]. Finally, we demonstrate that these oscillations are observed in all three organisms. In summary, the observed phenomena demonstrate the central importance of characterizing central dogma processes in the context of the living cell, where their activity is regulated and modulated by the cellular environment.

3.2 Results

3.2.1 The bacterial cell cycle

The bacterial cell cycle is divided into three periods [17, 18]: The B period is analogous to the G_1 phase of the eukaryotic cell cycle, corresponding to the period between cell birth and replication initiation. The C period is analogous to the S phase (and early M phase) in which the genome is replicated and simultaneously and sequentially segregated [19]. The D period is analogous to a combination of phases G_2 and late-M, corresponding to a period of time between replication termination and cell division, including the process of septation (i.e., cytokinesis).

The demographics of cell-cycle periods of exponentially-growing bacterial cells were first quantitatively modeled by Cooper and Helmstetter in an influential paper [20] and then refined by multiple authors [21–23]. In the Methods Section, we generalize these models to demonstrate that the measured marker-frequency analysis quantitatively measures the cell-cycle replication dynamics. The key results are summarized below.

3.2.2 Lag-time analysis

Our strategy will be to use exponential growth as the stopwatch with which we resolve cell-cycle dynamics. In short, cells with greater cell-cycle progression (i.e., age) are depleted

in the population, equivalent to an independent, exponentially-proliferating species that lags newborn cells by a time equal to its age [14]. (See Fig. 3.1 for a schematic illustration of the approach.) Lag-time analysis is the measurement of this time lag. In principle, this approach can be applied to characterize the dynamics of any biological molecules or complexes; however, for concreteness, we will focus on replication dynamics since the replication process is of great biological interest and next-generation sequencing provides a powerful tool for digital, as well as genome-wide, quantitation of the number of genomic loci.

In marker-frequency experiments, the number of each sequence $N(\ell)$ in a steady-state, asynchronously growing population is determined by mapping next-generation-sequencing reads to the reference genome. This marker frequency can be reinterpreted as a measurement of the *lag time* $\tau(\ell)$:

$$\tau(\ell) = \tfrac{1}{k_G} \ln \tfrac{N_0}{N(\ell)}, \tag{3.1}$$

where $N(\ell)$ is the observed number of the locus at genomic position ℓ and N_0 is the observed number of the origin in the culture and k_G is the growth rate. This relation can be understood as a consequence of the exponential growth law [14].

In a deterministically-timed model, the measured lag time would be equal to the replication time relative to initiation. In reality, the timing of all processes in the cell cycle is stochastic. We previously showed that the measured lag time is related to the distribution of durations in single cells by the exponential mean [14]:

$$\tau_i \equiv -\tfrac{1}{k_G} \ln \mathbb{E}_t \exp(-k_G t), \tag{3.2}$$

where $\mathbb{E}_t$ is the expectation over stochastic time t with distribution $t \sim p_i(\cdot)$.

3.2.3 Determination of replisome pause durations

Replisome pause durations or the lag time difference between the replication of any two loci can be computed using the difference of lag times between the two loci:

$$\Delta\tau_{ij} \equiv \tau_j - \tau_i = \tfrac{1}{k_G} \ln \tfrac{N(\ell_i)}{N(\ell_j)}. \tag{3.3}$$

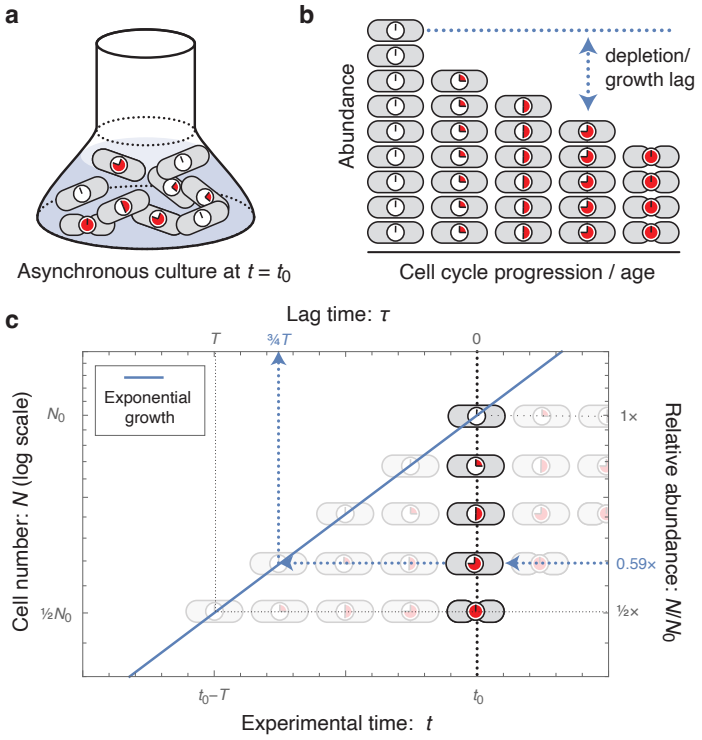


Figure 3.1: **Lag-time analysis. Panel a: Sample preparation.** An asynchronous culture in steady-state exponential growth is harvested at time $t = t_0$. **Panel b: Quantitation of demographics.** Cell abundance is quantified. For analyzing replication dynamics, cell quantitation is performed by next-generation sequencing. **Panel c: Measurement of lag time.** The dotted black line represents the culture at $t = t_0$. Cells with greater cell cycle progression (i.e., age) are depleted relative to newborn cells. For each cell age, the relative abundance determines the lag time. Their abundance is equivalent to an exponentially proliferating species that lags newborn cells by a time equal to its age. For instance, the nine-o'clock cell is at a relative abundance of 0.59 with a lag time of 3/4ths the mass-doubling time T. Schematically, start from the observed number of nine-o'clock cells, and follow that lineage horizontally (back in time) until you reach the newborn cell, born at $t = t_0 - \tau$ (blue dotted line). For a stochastic cell cycle, lag time measures the exponential mean of the stochastic time, Eq. 3.2.

We emphasize that the observed lag-time difference is the exponential mean of the stochastic time difference, which has important consequences for slow processes.

3.2.4 Determination of the fork velocity

For fast processes, like single nucleotide incorporation, the exponential mean leads to a negligible correction (see Methods); therefore, the fork velocity has a simple interpretation: it is the slope of the genomic position versus lag-time curve:

$$v(\ell) \equiv \frac{\mathrm{d}\ell}{\mathrm{d}\tau} = \frac{k_G}{\alpha(\ell)}, \tag{3.4}$$

or equivalently it is the ratio of the growth rate to the log-slope:

$$\alpha(\ell) \equiv -\frac{\mathrm{d}}{\mathrm{d}\ell} \ln N(\ell), \tag{3.5}$$

which can be directly determined from the marker frequency.

3.2.5 Lag-time analysis reveals *V. cholerae* replication dynamics

To explore the application of lag-time analysis to characterize replication dynamics, we begin our analysis in the bacterial model system *Vibrio cholerae*, which harbors two chromosomes: Chromosome 1 (Chr1) is 2.9 Mb and Chromosome 2 (Chr2) is 1.1 Mb. The origin of Chr1, *oriC1*, fires first and roughly the first half of replication is completed before the replication-initiator-RctB-binding-site *crtS* is replicated, triggering Chr2 initiation at *oriC2* [24–26]. Chr1 and Chr2 then replicate concurrently for the rest of the C period. (See Fig. 3.2a.)

To demonstrate the power of lag-time analysis, we compute the marker frequency, lag-time, and fork velocities. To measure pause times and replication velocities, we generate a piecewise linear model with a resolution set by the Akaike Information Criterion (AIC). The AIC-optimal model for fast growth (in LB) had 39 knots, spaced by 100 kb, generating 38 measurements of locus velocity across the two chromosomes. The replication dynamics for growth in LB is shown in Fig. 3.2. For tabulated velocities, see Supplementary Data 1 in Ref. [1].

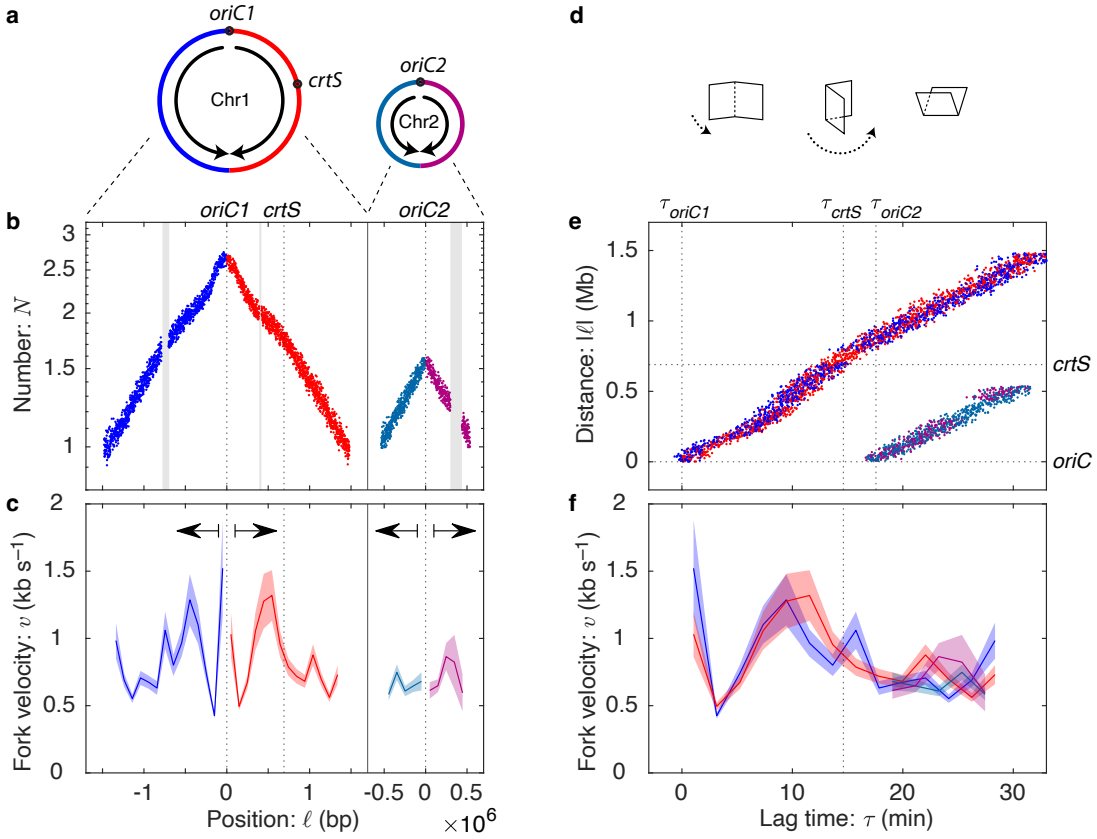


Figure 3.2: **Replication fork dynamics in *V. cholerae*. Panel a: Chromosome organization in *V. cholerae*.** *V. cholerae* harbors two chromosomes Chr1 and Chr2. *oriC2* initiates shortly after the *crtS* sequence is replicated on the right arm of Chr1. Data color represents chromosome identity (1 or 2) and arm (R or L) and is consistent throughout the panels. **Panel b: Marker frequency for *V. cholerae* grown on LB.** Repetitive sequences that cannot be mapped result in gaps. **Panel c: Fork velocity is locus-dependent.** The fork velocity is shown as a function of genomic position with an error region. Statistically significant differences in the fork velocity are observed between loci. There is significant bilateral (i.e., mirror) symmetry around the origin. Data are presented as mean values ± standard error of the mean (SEM). **Panel d:** A visual representation of the relation between the log-marker-frequency and lag-time plots: Fold at the origin and rotate. **Panel e: Lag-time analysis.** The replication forks start at the origin at lag time zero and then accelerate and decelerate synchronously, as the forks move away from the origin. The consistency in arm position is a manifestation of bilateral symmetry. **Panel f: Fork velocity as a function of lag time.** In addition to bilateral symmetry, after Chr2 initiates, all four forks show roughly consistent velocities.

3.2.6 The measurement of the duration of fast processes

We focus first on the duration of time between *crtS* replication and the initiation of *oriC2*. Fluorescence microscopy imaging reveals that this wait time is very short [27], but it is very difficult to quantify since the precise timing of the replication of the *crtS* sequence is difficult to determine by fluorescence imaging; however, this is a natural application for lag-time analysis. To measure the lag-time difference between *crtS* replication and *oriC2* replication, we use Eq. 3.3 to compute the replication time difference from the relative copy numbers. For this analysis, we generate a piecewise linear model with knots at the *crtS* and *oriC2* loci. The measured lag time is

$$\Delta\tau_{\text{pause}} = 3.5 \pm 0.1 \text{ min}, \tag{3.6}$$

a pause duration which is clearly resolved in the lag-time plot shown in Fig. 3.2e.

3.2.7 The fork velocity is locus dependent

It is qualitatively clear from the fork-distance-versus-time plot (Fig. 3.2e) that the fork velocity is locus dependent since the trajectory is not straight. To test this question statistically, we compare the 39-knot model to the null hypothesis (constant fork velocity), which is rejected with a p-value of $p \ll 10^{-30}$ and therefore the data cannot be described by a single fork velocity. (See Tab. 3.1.) The resulting velocity profiles are shown in Fig. 3.2cf.

3.2.8 Bilateral symmetry supports a time-dependent mechanism

Our understanding of the replication process motivated two general classes of mechanisms: (i) time-dependent and (ii) locus-dependent mechanisms. Time-dependent mechanisms, like a dNTP-limited replication rate, affect all forks uniformly and therefore loci equidistant from the origin should have identical fork velocities:

$$v(\ell) = v(-\ell), \tag{3.7}$$

Table 3.1: **Fork number and velocities under different growth conditions.** Increasing fork number by increasing cell metabolism does not consistently reduce fork velocity. Fork velocities in fast growth are higher in *E. coli* and lower in *V. cholerae*. The statistical significance column shows the p-value for the null hypothesis of constant fork velocity (likelihood ratio test). For more details on how these values are calculated, see Supplementary Methods Sec. 3.8.4-3.8.6.

Organism	Growth condition	Doubling time: T (min)	Fork statistics: C period: C (min)	Fork number: $\overline{N}_F$	Velocity mean: $\overline{v}$ (kb s^{-1})	std: σ_v (kb s^{-1})	Symmetry: f_S	Statistical significance p-value: p
E. coli	LB	19	30	3.8	1.3	0.19	84%	$\ll 10^{-30}$
	M9	69	46	1.2	0.85	0.12	59%	6×10^{-12}
V. cholerae	LB	22	31	4.3	0.82	0.27	76%	$\ll 10^{-30}$
	M9	50	32	1.5	0.84	0.28	70%	$\ll 10^{-30}$
B. subtilis	S7	64	42	0.82	1.1	0.68	50%	$\ll 10^{-30}$
rrnIHG	MOPS+CA	44	40	1.2	0.86	0.45	45%	$\ll 10^{-30}$
	MOPS	50	41	1.1	0.99	0.52	57%	$\ll 10^{-30}$

where ℓ is the genetic position relative to the origin. In contrast, in a locus-dependent mechanism, like replication-conflict-induced slowdowns, the slow regions are expected to be randomly distributed over the chromosome. In this scenario we expect to see no bilateral symmetry between arms. (See Methods.)

A bilateral symmetry between the arms is clearly evident in the data (the mirror symmetry about the origin in Figs. 3.2bc and is manifest in the lag-time analysis as the coincidence between the left and right arm trajectories and velocities in Figs. 3.2ef. To quantitate this symmetry, we divide the variance of the fork velocity into symmetric and antisymmetric contributions. (See Supplementary Methods Sec. 3.8.4.) A time-dependent mechanism would generate a $f_S = 100\%$ symmetric variance, whereas a locus-dependent mechanism would be expected to generate equal symmetric and antisymmetric variance contributions ($f_S = 50\%$). *V. cholerae* Chr1 and Chr2 have $f_S = 76\%$ symmetry, consistent with a time-dependent mechanism playing a dominant but not exclusive role in determining the fork velocity. (See Tab. 3.1.)

3.2.9 The replisome pauses briefly at rDNA in *B. subtilis*

To explore the possibility that locus-dependent mechanisms could play a dominant role in determining the fork velocity profile, we next characterized the fork dynamics in the context of replication conflicts, where the antagonism between active transcription and replication, have been reported to stall the replisome by a locus-specific mechanism [10, 28]. In *B. subtilis*, there are seven highly-transcribed rDNA loci on the right arm and only a single locus of the left arm. Consistent with the notion of rDNA-induced pausing, the *ter* locus is positioned asymmetrically on the genome, at 172° rather than 180°, leading the right arm of the chromosome to be shorter than the left arm. (See Fig. 3.3a.) In spite of the difference in length, both arms terminate roughly synchronously, implying that the average fork velocity is lower on the right arm, consistent with putative fork pausing at the rDNA loci. Are these conflict-induced pauses present in wild-type cells where the replication and transcription are

co-directional? We have previously reported evidence based on single-molecule imaging that they are [13], but there is as of yet no other unambiguous supporting evidence.

To detect putative short pauses at the rDNA loci in wild-type *B. subtilis*, a low-noise dataset was essential. We therefore examined a number of different datasets, including our own, to search for a dataset with the lowest statistical and systematic noise. A marker-frequency dataset for a nearly wild-type strain growing on minimal media was identified for which the noise level was extremely low. (See Supplementary Methods Sec. 3.8.3.2.) The lag-time analysis is shown in Fig. 3.3b. Replication pauses should result in discrete steps in the lag time; however, no clearly-defined steps are visible in the lag-time plot. The pauses are either absent or too small to be clearly visible without statistical analysis.

To achieve optimal statistical resolution, we used the AIC model-selection framework [29, 30] on four competing hypotheses: In Model 1, fork velocities are constant and equal on both arms with no pauses. In Model 2, fork velocities are constant but unequal on the left and right arms with no pauses. In Model 3, fork velocities are constant and equal on the left and right arms with equal-duration pauses at each rDNA locus. In Model 4, fork velocities are constant and unequal on the left and right arms with equal-duration pauses at each rDNA locus. AIC selected Model 3 (equal arm velocities with rDNA pauses) and a pause duration of:

$$\Delta\tau_{\text{pause}} = 17 \pm 8 \text{ s}, \tag{3.8}$$

is observed. The pause models were strongly supported over the non-pause models ($\Delta\text{AIC}_{23} = 4.3$ and $\Delta\text{AIC}_{43} = 9.4$). Therefore, statistical analysis supports the existence of short slowdowns (i.e., pauses) at the rDNA, even if these features are not directly observable without statistical analysis. In higher-noise datasets, the statistical inference was ambiguous.

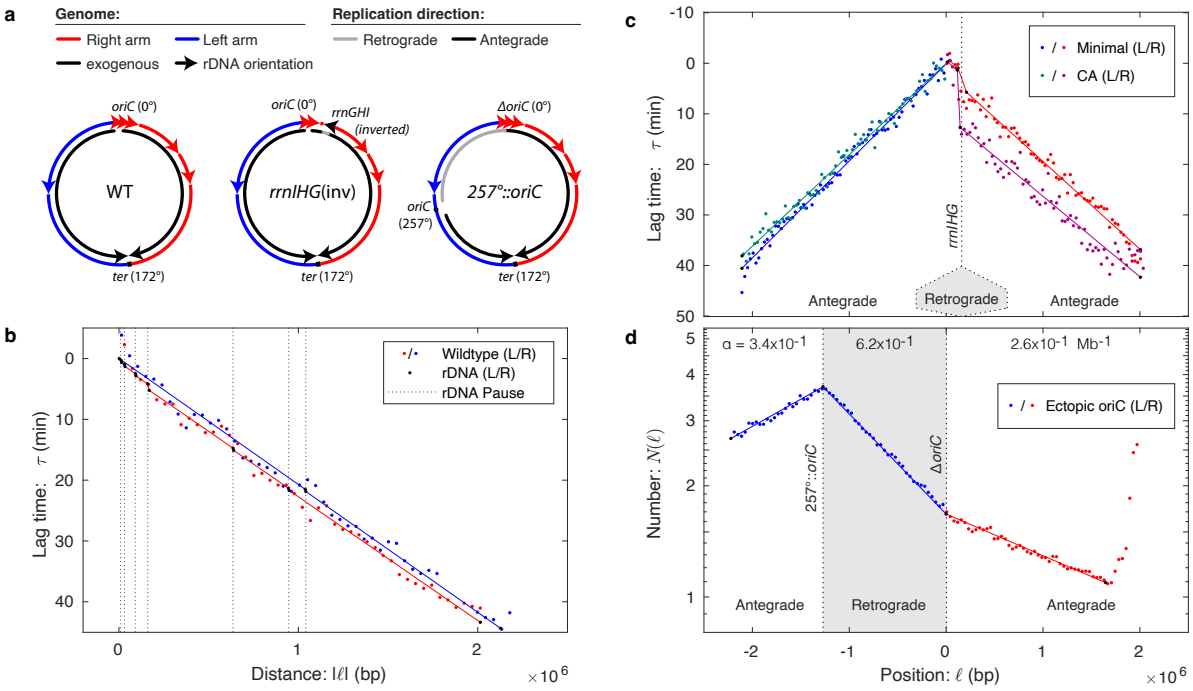


Figure 3.3: ***B. subtilis* fork dynamics and transcriptional conflicts. Panel a: Chromosomal structure for wild-type and mutant *B. subtilis* strains.** The *ter* region in wild-type *B. subtilis* is positioned at 172°, rather than 180°, making the right arm shorter than the left. In *rrnIHG*(inv), the *rrnIHG* locus is inverted so that it is transcribed in a head-on orientation with respect to replication. In 257°::*oriC*, the origin is moved to 257°, resulting in a short left arm that terminates at the terminus and a long right arm that replicates initially in the retrograde direction, before replicating the residuum of the right arm in the antegrade orientation. Data color (blue or red) represents the arm of the chromosome (R or L) and is consistent throughout the panels. Gray segments represent replication in the retrograde direction. **Panel b: Lag time in wild-type cells.** Replication on the right arm (red) is delayed relative to the left arm (blue) by multiple endogenous co-directional rDNA loci. **Panel c: Head-on conflicts lead to pausing.** The *rrnIHG* genes are inverted so that transcription of the rDNA locus is in the head-on direction. A longer lag-time pause is observed at intermediate growth rates (CA, purple) than slow growth (minimal, red). Fork velocities elsewhere are roughly consistent. **Panel d: Retrograde fork motion is slow.** The retrograde fork motion in R is slow compared to antegrade replication in A1. Late antegrade motion in A2 is faster than early antegrade motion in A1. Source data are provided as a Source Data file.

3.2.10 Strong, head-on conflicts lead to long pauses

Although we have just demonstrated that endogenous co-directional conflicts are detected statistically, they do not lead to a clear unambiguous signature. In contrast, strong, exogenous head-on conflicts in which the replisome and transcriptional machinery move in opposite directions can lead to particularly potent conflicts and even cell death [4–10, 31]. The ability to engineer conflicts at specific loci facilitates the use of lag-time analysis for measuring the duration of the replication pauses.

To measure the pause durations due to head-on conflicts, we analyze the marker frequency from a strain, *rrnIHG*(inv), generated by Srivatsan and coworkers with three rDNA genes (*rrnIHG*) inverted so that they are transcribed in the head-on orientation. Marker-frequency datasets were reported for this strain in two growth conditions: minimal supplemented with casamino acids, in which the strain grows at an intermediate growth rate, and unsupplemented minimal media [32]. (Mutant cells cannot proliferate in rich media, presumably because the transcription conflicts are so severe [32].) In both slow and intermediate growth conditions, a clearly resolved step at the head-on locus is observed in the marker-frequency and lag-time analysis (Fig. 3.3b), exactly analogous to the simulated pause. (See Methods.)

To determine the pause durations in the two growth conditions, we again consider a model with an unknown pause duration (at the inverted rDNA locus) and constant but unequal fork velocities on the left and right arms. The observed lag-time pauses are

$$\Delta\tau_{\text{pause}} = \begin{cases} 3.3 \pm 0.7 \text{ min (slow)} \\ 9.7 \pm 0.9 \text{ min (intermediate)} \end{cases}, \tag{3.9}$$

for the slow and intermediate growth rates, respectively.

Although lag-time analysis reports a precise pause duration, it is important to remember that the observed lag time corresponds to the exponential mean of the stochastic state lifetime, Eq. 3.2, including cells that arrest and therefore never complete the replication process. Eq. 3.18 accounts for the pause generated by this arrested cell fraction. Srivatsan

and coworkers report that 10% of the cells are arrested in intermediate growth, which accounts for 8.3 min of the lag time, leaving an estimated pause time of $\Delta\tau_{\text{pause}} = 1.4 \pm 0.9$ min for non-arrested cells, which is roughly consistent with the pause time observed in slow growth conditions.

3.2.11 Slow retrograde replication in *B. subtilis*

Are all conflict-induced slowdowns consistent with long pauses at a small number of rDNA loci? Wang et al. have previously engineered a head-on strain, 257°::*oriC*, with less severe conflicts by moving *oriC* down the left arm of the chromosome to 257° [33]. (See Fig. 3.3a.) The resulting strain has a very short left arm and a very long right arm, the first third of which is replicated in the retrograde (i.e., reverse to wild-type) orientation. This retrograde region contains only a single rDNA locus. All other regions are replicated in the antegrade (i.e., endogenous) orientation.

Consistent with the analysis of Wang et al., we position knots to divide the chromosome into three regions with three distinct slopes: an early antegrade region $A1$ (the short left arm) with log-slope $\alpha_{A1} = 0.34 \pm 0.01$ Mb^{-1}, a retrograde region R with log-slope $\alpha_R = 0.63 \pm 0.01$ Mb^{-1}, and a late antegrade region $A2$ with log-slope $\alpha_{A2} = 0.26 \pm 0.01$ Mb^{-1}, that replicates after the left arm terminates. (See Fig. 3.3c.) Due to the higher percentage of head-on genes in the R region compared with the $A1$ region, the conflict model predicts more rapid replication in region $A1$ versus R. Consistent with this prediction, the ratio of replication velocities is:

$$v_{A1}/v_R = 1.84 \pm 0.4, \tag{3.10}$$

revealing a strong replication-direction dependence. The slope appears relatively constant, consistent with a model of uniformly-distributed slow regions rather than a small number of long pauses as observed in the reversal of the rDNA locus *rrnIHG*. Our quantitative analysis is consistent with the interpretation of Wang et al. [33].

3.2.12 Rapid late replication due to genomic asymmetry

This dataset has a striking feature that is not emphasized in previous reports: Late antegrade fork velocity is faster than early antegrade velocity:

$$v_{A2}/v_{A1} = 1.29 \pm 0.05. \tag{3.11}$$

Although this effect is weaker than the replication-direction dependence discussed above, Eq. 3.10, its size is still comparable. An analogous late-time speedup is seen in two other ectopic origin strains. (See Supplementary Figs. 3.16 and 3.19.)

One potential hypothesis is that a locus-dependent mechanism slows the fork in the $A1$ region relative to the $A2$ region; however, no velocity difference is evident in these regions in the wild-type cells (Fig. 3.3b). Alternatively, one could hypothesize that there is some form of communication between forks that leads to a slowdown in region $A1$ due to the slowdown in region R; however, no coincident slowdown is observed in *rrnIHG*(inv) at a position opposite the *rrnIHG* locus, inconsistent with this hypothesis. Another possible hypothesis is that late-time replication is always rapid; however, no significant speedup is observed in either wild-type *B. subtilis* (Fig. 3.3a) or *V. cholerae* cells at the end of the replication process (Fig. 3.3b and Fig. 3.2e). However, there is one extremely important difference between 257°::*oriC* and the wild-type strains: Due to the asymmetric positioning of the origin and replication traps at the terminus (Fig. 3.3a), there is only a single active replication fork as the A2 region is replicated. We therefore hypothesize that the fork velocity is inversely related to active fork number.

3.2.13 Fork number determines velocities in *V. cholerae*

To explore the effects of changes in the fork number on fork velocity, it is convenient to return to *V. cholerae*. In slow growth conditions, the cells start the C period with a pair of replication forks, for which the fork-number model predicts faster fork velocity, and finish the replication cycle with two pairs of forks, predicting slower fork velocity.

Although the structure of the velocity profile is more complex than predicted by the fork-number model alone, the observed fork velocity is broadly consistent with its predictions. If a mean fork velocity is computed before and after *oriC2* initiates, the ratio is:

$$v_{\text{before}}/v_{\text{after}} = 1.46 \pm 0.02, \tag{3.12}$$

which is quantitatively consistent with the hypothesis that more forks lead to a slowdown in replication and the size of the effect is comparable to what is observed in *B. subtilis*, Eq. 3.11.

A mutant *V. cholerae* strain has been constructed that facilitates a non-trivial test of the fork-number model: In the monochromosomal strain MCH1, Chr2 is recombined into Chr1 at the terminus of Chr1, resulting in a single monochromosome (Chr 1-2). (See Fig. 3.4a.) Both the wild-type and MCH1 strains have essentially identical sequence content, implying the locus-dependent model would predict identical replication velocities; however, all replication in MCH1 occurs with only a single set of forks whereas the wild-type strain replicates the latter half of the C period with two pairs of forks, one pair on each chromosome.

The measured fork velocities are shown in Fig. 3.4b and support the fork-number model: MCH1 replicates the sequences after *crtS* at roughly 1.6 times the fork velocity of the wild-type cells, consistent with the fork-number model. Alternatively, we can consider the same quantitation of fork velocity we considered above: The ratio of fork velocities of loci replicated before *crtS* to those replicated afterwards:

$$v_{\text{before}}/v_{\text{after}} = 1.11 \pm 0.03, \tag{3.13}$$

therefore only a very small slowdown is observed after *crtS* is replicated in MCH1, even though exactly the same sequences are replicated, again consistent with the fork-number model.

3.2.14 The fork velocity oscillates in *E. coli*

Although experiments in *V. cholerae* clearly support the fork-number model, there is significant variability that cannot be explained by this model alone. Are time-dependent

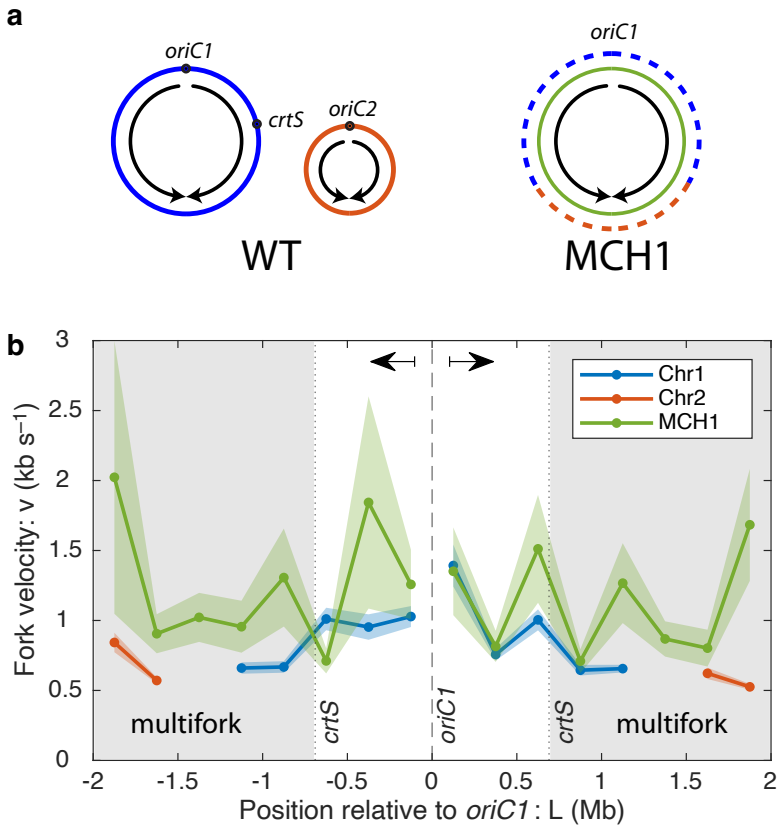


Figure 3.4: **Reducing fork number increases fork velocity. Panel a:** The monochromosomal strain MCH1 has a single chromosome (green) which was constructed by recombining Chr2 (orange) into the terminus of Chr1 (blue) [34]. Under slow growth conditions the first part of the chromosome in both strains is replicated by a single pair of forks. When the fork reaches the *crtS* sequence on the right arm, Chr2 is initiated at *oriC2* in the wild-type cells. All of Chr2 and the residuum of Chr1 replicate simultaneously, resulting in two pairs of active forks. In contrast, all sequences in MCH1 are replicated using a single pair of forks. Data color is consistent throughout the panels. **Panel b:** In MCH1, where all sequences are replicated by a single pair of forks, the fork velocity is faster than is observed in WT cells during the multifork region (gray shaded regions represent sequences replicated after *crtS*). Data are presented as mean values ± SEM.

variations in fork velocity also observed in organisms that replicate a single chromosome? To answer this question, we worked in the gram-negative model bacterium *Escherichia coli*, which harbors a single 4.6 Mb chromosome. A large collection of marker-frequency datasets have already been generated for both rapid and slow growth conditions by the Rudolph lab [35]. As with the *B. subtilis* marker-frequency datasets, we selected those that had the lowest statistical and systematic noise. (See the Supplementary Methods Sec. 3.8.3.2.)

The fork velocities are shown in Fig. 3.5. As before, statistically significant variation is observed in the fork velocity as a function of position. (See Tab. 3.1 and Supplementary Methods Sec. 3.8.6.) As discussed above in the context of *V. cholerae*, we had initially hypothesized that this variation might be a consequence of rDNA position or some other locus-dependent mechanism; however, there are three arguments against this hypothesis: (i) The slow-velocity regions are not coincident with rDNA locations (Fig. 3.5a) or relative GC content (Supplementary Fig. 3.9). (ii) Consistent with the time-dependent model, 84% (and 59%) of the observed variation in the fork velocity is symmetric for fast (and slow) growth. (iii) We would expect that a locus-dependent model would predict slow regions that are consistent between fast and slow growth, which is not observed. (See the purple arrows in Fig. 3.5a.) We therefore conclude that the dominant mechanism for determining the fork velocity is a time-dependent mechanism, consistent with our observations for *V. cholerae*.

Lag-time analysis is particularly informative with respect to the mechanism of variation in the fork velocity: Although there is no alignment in the velocity with respect to locus position (Fig. 3.5a), there is clear alignment of the fork velocity variation with respect to lag time (Fig. 3.5b), not only between the left and right arms of the chromosome, but between slow and fast growth conditions. The oscillations do not align with respect to locus position (Fig. 3.5a) since the difference in average fork velocity leads the slow and fast temporal periods to correspond to different locus positions under slow and fast growth conditions.

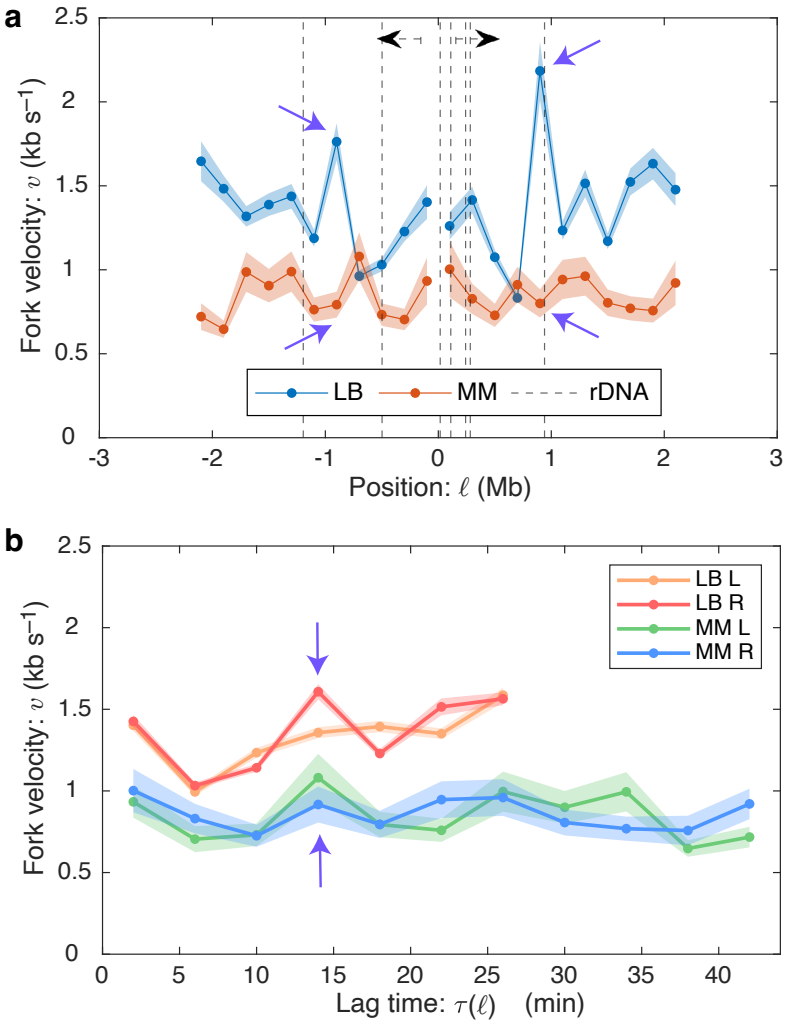


Figure 3.5: **Observed oscillations are consistent with a temporal mechanism. Panel a: Velocity oscillations with respect to position in *E. coli*.** We compare fork velocities as a function of genomic position (with respect to *oriC*) under rapid (LB) and slow (minimal media–MM) growth conditions. Motivated by conflict-induced pauses, we have annotated the rDNA positions; however, slow velocities are not consistently coincident with rDNA loci. Regions with high fork velocities are not consistent between rapid and slow growth (e.g., see the purple arrows). Data are presented as mean values ± SEM. **Panel b: Velocity oscillations with respect to lag time in *E. coli*.** The velocity profiles have significant bilateral symmetry: the right and left arm velocities oscillate up and down together. Furthermore, not only are the oscillations consistent between left and right arms, they are also consistent between rapid (LB) and slow growth (minimal media–MM). E.g., see the purple arrows.

3.2.15 Fork velocity oscillations are observed in three organisms

Temporal oscillations in the fork velocity are an unexpected phenomenon. Are these features a systematic error with a single dataset? First we note that these oscillations are present in two *E. coli* growth conditions (LB and minimal). This phenomenon would be on sounder footing if similar oscillations are observed in two evolutionarily distant species: the gram-negative *V. cholerae* and gram-positive *B. subtilis*. If this phenomenon is observed, to what extent are the oscillations of similar character (e.g., phase, amplitude, and period)?

We compared the lag-time-dependent fork velocity for all three species. In *B. subtilis*, we have already discussed a rDNA-induced pausing on the right arm, which could complicate the interpretation of the data. We therefore consider the fork velocity on just the left arm. For *E. coli* and *V. cholerae*, we compute the average velocity as a function of lag time between the two arms. Since the different organisms and growth conditions have different mean fork velocities, we compare the fork velocity relative to the overall mean. The results are shown in Fig. 3.6 and Tab. 3.2.

All three organisms show oscillations with the same qualitative features: Each fork velocity has roughly the same phase: The velocity begins high, before decaying. The relative amplitudes, roughly 0.5 peak-to-peak, are all comparable with the largest-amplitude oscillations observed in *V. cholerae* and the smallest in *E. coli*. When the relative velocities are compared, it is striking how much consistency there is between growth conditions in *E. coli* and *B. subtilis*. Finally, the period of oscillation is comparable but distinct in all three organisms, ranging from 10 to 15 minutes. The oscillation characteristics are summarized in a table in Fig. 3.6 and Tab. 3.2.

3.3 Discussion

The focus of this paper is on the development of lag-time analysis, which uses exponential growth as the timer to characterize replication dynamics.

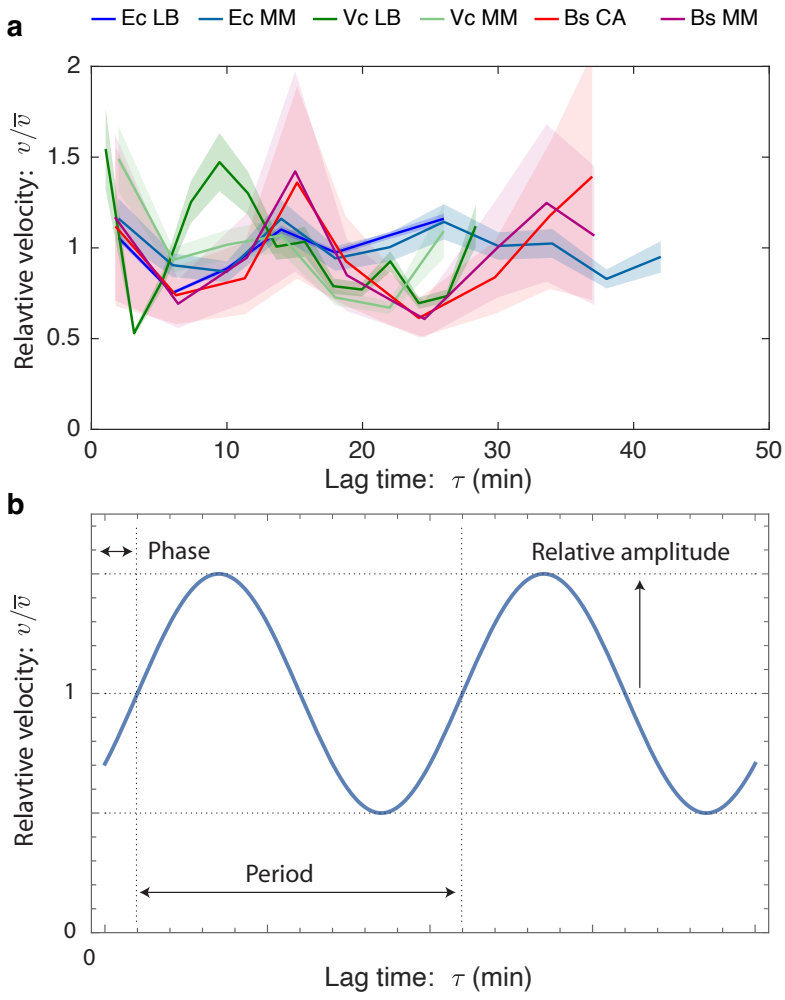


Figure 3.6: **Fork velocity oscillations. Panel a: Temporal velocity oscillations are observed in three bacterial species:** *E. coli* (Ec), *B. subtilis* (Bs), and *V. cholerae* (Vc). The fork velocity starts high before decaying rapidly and then recovering. Data are presented as mean values ± SEM. **Panel b: Oscillation characteristics.** The definition of the phase, amplitude, and period of the fork velocity oscillation.

Table 3.2: **Velocity oscillation characteristics for different bacterial species and growth conditions.** The oscillatory characteristics are broadly consistent both between conditions and species.

Organism	Growth condition	Period (min)	Phase (degrees)	Relative amplitude
E. coli	Fast (LB)	15	$-78°$	18%
	Slow (M9)	12	$-45°$	18%
V. cholerae	Fast (LB)	12	$-81°$	31%
	Slow (M9)	10	$-39°$	36%
B. subtilis	M9+CA	17	$-110°$	26%
	M9	15	$-150°$	30%

3.3.1 The significance of the fork velocity

Previous marker-frequency analyses have often reported a log-slope (e.g., [33, 36]), which is closely related to the fork velocity. What new insights does the measurement of the fork velocity offer over this closely related approach? The fork velocity approach has two important advantages: (i) The first advantage is a conceptual one. The underlying quantity of interest is velocity (or rate per base pair). This is the quantity that is measured *in vitro* and is relevant in a mechanistic model. In contrast, the log-slope is an emergent quantity that is only relevant in the context of exponential growth. (ii) The second advantage is concrete: Although log-slope measurements allow ratiometric comparisons between fork velocity at different loci in the same dataset, they cannot be used to make comparisons across datasets. Any comparison of the log-slope between cells with different growth rate (e.g., due to changes in growth conditions, mutations, species, etc.) are meaningless. For instance, the log-slopes of the wild-type and MCH1 *V. cholerae* strains are very different even though the changes in the fork velocity are small. Our wide-ranging comparisons between growth conditions, mutants, and organisms demonstrate the power of reporting fork velocity over the log-slope.

3.3.2 Applications to eukaryotic cells

Although our focus has been on replication in bacterial cells, an important question is to what extent our approach could be adapted to eukaryotic cells. First, we emphasize that the lag-time analysis is directly applicable without modification to the eukaryotic context. As such, the timing of the replication of loci can be analyzed; however, since the S phase is typically a smaller fraction of the cell cycle and the genomes of eukaryotic cells are larger, deeper sequencing will be required to achieve the same resolution we demonstrate in the context of bacterial cells. One significant potential refinement to this approach is the use of cell sorting (sort-seq) to enrich for replicating cells which can greatly increase the signal-to-noise ratio [37, 38]; however, this approach appears to lead to significant flattening near early-firing origins, as we have observed in other contexts (Supplementary

Methods Sec. 3.8.3.2), and therefore increasing sequencing depth is probably the most promising approach for eukaryotic systems when quantitative characterization is a priority. (See Methods Eqs. 3.22 and 3.23 for an estimate of resolution.)

Although lag-time analysis can easily be extended to the eukaryotic context, the measurement of the fork velocity will require some care. A critical assumption in our analysis is that replication forks move unidirectionally at any particular locus, i.e., it can be either rightward or leftward moving but not both. (See Supplementary Methods Sec. 3.8.3.9.) Fork traps prevent this bidirectionality in many bacterial cells. For loci in the terminus region, although the replication timing can be determined with high precision, the bidirectionality of the fork movement prevents the measurement of fork velocities in these regions. This is a more important limitation in eukaryotic cells where the number of origins is much greater; however, if regions of the chromosome can be found where fork movement is unidirectional, e.g., sufficiently close to early-firing origins, fork velocity measurements could be made in eukaryotic cells. For instance, these conditions appear to be met for a significant fraction of the *Saccharomyces cerevisiae* genome [38]. With significant increases in sequencing depth, we expect analogous replication phenomenology, including pausing and locus- and time-dependent fork velocities, will be observed in eukaryotic systems using lag-time analysis.

3.3.3 Importance of a model-independent approach

As we prepared this manuscript, we became aware of a competing group which also uses marker-frequency analysis to test a specific hypothesis: the fork velocity is oscillatory in *E. coli* [16], consistent with our own observations. Although our reports share some conclusions, this competing approach requires detailed models for the cell cycle and the fork velocity, along with explicit stochastic simulations. We demonstrate an approach to measure fork velocities independent of model assumptions or detailed hypotheses for the fork velocity, without the need for numerical simulation and complete with the ability to perform an explicit and tractable error analysis.

Although our initial investigations were dependent on explicit numerical simulations of stochastic models, the use of lag-time analysis not only circumvents the need to perform these numerical simulations, but demonstrates that stochastic models are equivalent to deterministic models as well as providing a framework to understand the effects of stochasticity on the growth of populations through the use of the exponential mean, Eq. 3.2 [14]. This significant simplification will make lag-time analysis both widely applicable as well as accessible to other investigators who lack specialized analytical skills and modeling expertise.

3.3.4 Slow growth increases noise

One unintuitive feature of fork velocity measurements is that slow growth tends to increase noise. This is typically not the result of poor sequencing depth, but rather the consequence of changes in the marker frequency. For a fixed fork velocity, a decrease in growth rate implies a decrease in the log-slope α (Eq. 3.15); however even as the signal in α decreases, the magnitude of the noise remains roughly constant (Eq. 3.22). As a result, slow growth tends to require deeper sequencing to achieve the same velocity sensitivity.

3.3.5 Systematic error in datasets

It may seem perplexing that we have not pooled many existing datasets from multiple independent experiments. This would naively increase the statistical resolution and sensitivity from an analytical perspective. However, it is important to emphasize that not all marker-frequency experiments are of equal quality and that many datasets we have analyzed have clear signatures of systematic error. (See the Supplemental Methods Sec. 3.8.3.2.) In our analysis, we have prioritized the selection of artifact-free datasets over the indiscriminate pooling of data. We emphasize that to date, datasets have not been generated with quantitative replication dynamics analysis as a goal and we are confident that experimental protocols can be optimized to improve the data. Ref. [39] describes a

promising approach, including harvesting populations earlier in exponential phase. We too are developing new protocols to increase data quality.

3.3.6 Multiple factors determine replisome dynamics

Our measurements of the replication velocity reveal that there are multiple important determinants that results in complex velocity profiles.

3.3.7 dNTP pools regulate the fork velocity

Previous work had already demonstrated that increases (or decreases) in dNTP pool levels lead to concomitant decreases (or increases) in the C period duration, consistent with a dNTP-limited model of the replication velocity [40–43]. Our data is broadly consistent with these previous results, but in a subcellular context: (i) The fork-number model, in which fork velocities decrease as the number of active forks increase, is clearly consistent with a mechanism in which the nucleotide pool levels, although highly-regulated [44], cannot completely compensate for the increased incorporation rate associated with multiple forks. (ii) The observation of the fork velocity oscillations is also consistent with an analogous failure of the regulatory response to compensate, this time temporally. The initial fall in the fork velocity is consistent with a model in which dNTP levels initially fall as replication initiates and nucleotides begin to be incorporated into the genome. Reduction in the dNTP levels causes a regulatory response to increase dNTP synthesis by ribonucleotide reductase [44], but the finite response time of the network could lead to dynamic overshoot in the regulatory feedback, leading to oscillations [45]. Ref. [16] has also argued that this oscillating-dNTP-level model would lead to time-dependent oscillations in the mutation rate which are consistent with the origin-mirror-symmetric distribution of the mutation observed in *E. coli*. However, this interesting phenomenon and this hypothesized mechanism will require further investigation.

3.3.8 Fork-velocity oscillations are observed in divergent species

A key clue to the potential significance of the fork-velocity oscillations comes from their observation, not only in *E. coli*, but in *B. subtilis* and *V. cholerae*, three highly-divergent species, as well as their observation under multiple growth conditions. Although it has long been assumed that homeostatic regulation keeps key cellular metabolites in a relatively narrow range, our observations, as well as the recent reports of oscillations in other key nucleotides in bacteria (e.g., ATP in *E. coli* [46]), suggest that key metabolites are in fact subject to significant temporal oscillations even in the context of steady-state log-phase growth. These observations, if their ubiquity is supported by future work, may require a significant revision of our understanding of the metabolic environment of the cell.

3.3.9 Retrograde fork motion leads to slow replication velocities

Retrograde fork motion, where the fork moves in the opposite direction from wild-type cells, lead to the largest changes in fork velocity observed. To what extent is the observed slowdown a consequence of a few long-duration pauses versus a region-wide slowdown? In regions which exclude the rDNA, the effect appears well distributed. However, it is important to note that the genomic resolution of lag-time analysis is still much too low to resolve individual transcriptional units. We anticipate that with increased sequencing depth as well as improvements in sample preparation, this approach could detect genomic structure in the fork velocity at the resolution of individual transcriptional units. Although we did analyze a number of mutants with retrograde fork movement in *V. cholerae* and *E. coli* (analysis not shown), the competing effect of increased fork number as well as the genomic instability of these strains made these experiments difficult to interpret quantitatively, since fork number and direction were both affected in these strains [33, 47]. We concluded qualitatively that retrograde replication direction appears not to play as large a role in these gram-negative bacteria as it does in gram-positive *B. subtilits*, consistent with previous evidence [32, 33, 48–50]. However, we expect lag-time analysis could be used to characterize even small effects

of the retrograde fork orientation in more-carefully engineered strains, analogous to those that we analyzed in the context of *B. subtilis* [32, 33].

3.3.10 Replisome pausing

Previous reports [32, 33], including our own [13, 51–54], had reported long-duration replication-conflict induced pauses, especially in mutant strains where the orientation of rDNA [32] or other highly transcribed genes [51] are inverted to give rise to a head-on conflict between replication and transcription. The contribution of lag-time analysis in this context is multifold: First, we provide a quantitative number in the context of the very-short-duration pauses for co-directional transcription in wild-type cells. This analysis supports a long-standing hypothesis that the right arm of the *B. subtilis* chromosome is shorter than the left arm to compensate for pausing at the rDNA loci that arm predominately located on this arm.

We also report quantitative measurements for the longer pauses that results from head-on conflicts in mutants where highly transcribed genes are inverted. Our analysis gives us the ability to quantitatively differentiate the contributions of fork pausing and arrest in the analysis of the marker frequency, which was previously impossible. Our measurement of a timescale of minutes is consistent with our previous *in vivo* single-molecule measurements in which we report transcription-dependent disassembly of the core replisome [13]. Could the observed fork-velocity oscillations be misinterpreted as pauses? The observed lag-time offset between the two arms (e.g., Fig. 3.3b) is not predicted by fork-velocity oscillations.

3.3.11 Conclusion

In this paper, we introduce a method for quantitively characterizing cellular dynamics by lag-time analysis. Although more broadly applicable, we focus our analysis on the characterization of replication dynamics using next-generation sequencing to quantitate DNA locus copy number genome-wide. The approach has the ability to make precise, even at the resolution of seconds, measurements of time differences and pause durations, as well

as the ability to quantitatively measure fork velocities *in vivo* in physiological units of $\mathrm{kb\,s^{-1}}$, at genomic resolutions of roughly 100 kb. Importantly, unlike marker-frequency analysis, our approach allows direct quantitative comparisons to be made between growth conditions, mutant strains, and even different organisms. The resulting measurements of replication dynamics reveal complex phenomenology, including temporal oscillations in the fork velocity as well as evidence for multiple mechanisms that determine the fork velocity. The lag-time analysis has great potential for application beyond bacterial systems as well as the potential to significantly increase in resolution and sensitivity as sequencing depth and sample preparation improve.

3.4 Methods

3.4.1 Strains used in this study

Detailed information about the strains used in this study are included in Supplementary Table 3.3.

3.4.2 Introduction to marker-frequency analysis

Our focus will be on marker-frequency analysis, which measures the total number of a genetic locus in an asynchronous population. The model was generalized to predict the marker frequency $N(\ell)$ of a locus a genomic distance ℓ away from the origin [21–23]:

$$N(\ell) \;=\; N_0\, \mathrm{e}^{-\alpha|\ell|}, \tag{3.14}$$

where N_0 is the number of origins, which grows exponentially in time with the rate of mass doubling of the culture, k_G. Since the origin is replicated first, the number of origins is always largest compared to the numbers of other loci. Quantitatively, the copy number is predicted to decay exponentially with log-slope:

$$\alpha \;=\; -\tfrac{\mathrm{d}}{\mathrm{d}\ell} \ln N(\ell) = k_G/v, \tag{3.15}$$

where k_G is the population growth rate and v is the fork velocity, typically expressed in units of kilobases per second. To derive this result, two critical assumptions were made: (i) the timing of the cell cycle is deterministic and (ii) the fork velocity is constant [20, 21].

Initially, our naive expectation was that the interplay between the significant stochasticity of the cell-cycle timing with the asynchronicity of the culture would prevent marker-frequency analysis from being used as a quantitative tool for characterizing cell-cycle dynamics. For instance, significant stochasticity is observed in the duration of the B period [55] (i.e., the duration of time between cell birth and the initiation of replication). Does this stochasticity lead to a failure of the log-slope relation, Eq. 3.15?

3.4.3 Stochastic simulations support the log-slope relation

To explore the role of stochasticity and a locus-dependent fork velocity in shaping the marker frequency, we simulated the cell cycle using a stochastic simulation. Our aim was not to perform a simulation whose mechanistic details were correct, but rather to study how strong violations of the Cooper-Helmstetter assumptions, in particular how stochasticity, as strong or stronger than that observed, influenced the observed marker frequency and the log-slope relation, Eq. 3.15. In short, we used a Gillespie simulation [56] where the B period duration and the lifetime of replisome nucleotide incorporation steps are exponentially distributed, and we added regions of the genome where the incorporation rate was fast as well as a single slow step on one arm. See Fig. 3.7a and Supplementary Notes Sec. 3.9 for a detailed description of the model, as well as movies of the marker frequency approaching steady-state growth, starting from a single-cell progenitor (Supplementary Movies 1 and 2 from Ref. [1]).

To our initial surprise, the stochasticity of the model had no effect on the predicted log-slope of the locus copy number. (See Fig. 3.7b.) In spite of the stochastic duration of the B period and the locus-dependence, the marker frequency still decays exponentially with the same decay length locally:

$$\alpha(\ell) \equiv -\tfrac{\mathrm{d}}{\mathrm{d}\ell} \ln N(\ell) = k_G/v(\ell), \tag{3.16}$$

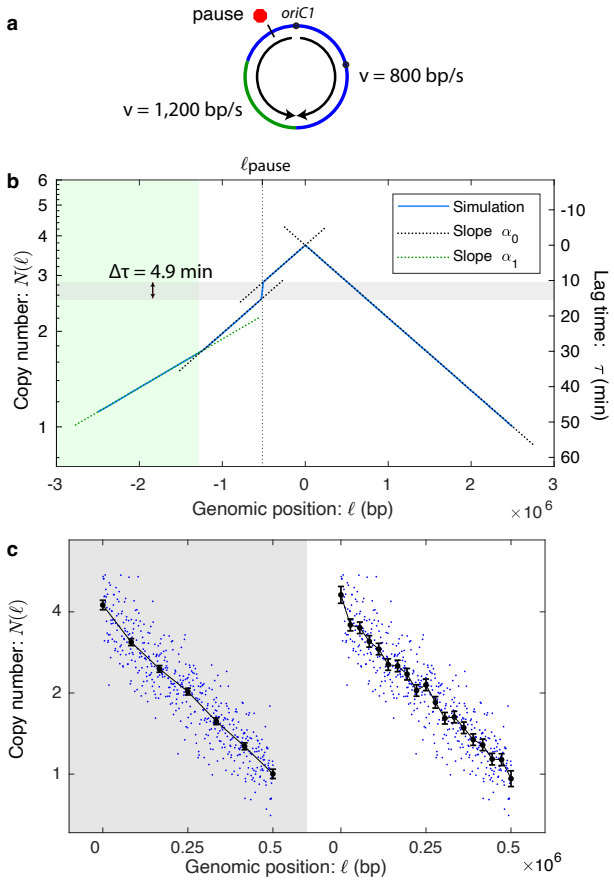


Figure 3.7: **Analysis of simulated data. Panel a: A schematic of the simulated chromosome.** Replication initiates at the origin, pauses at a locus (red octagon) on the left arm and the velocity is increased on the lower left arm (green). **Panel b: Simulated marker frequency obeys the log-slope law.** The stochastic simulation generates a marker-frequency curve (blue). The model is stochastic in the timing of replication initiation as well as the fork dynamics and it includes two regions (blue and green) with different fork velocities as well as a pause with a stochastic lifetime. (See the Terminus 4 model in the Supplementary Notes Sec. 3.9.) In spite of the stochasticity, it obeys the log-slope law locally, Eq. 3.16. Furthermore, the inferred lag-time pause (4.9 min) is predicted by the exponential mean, Eq. 3.2. **Panel c: Tradeoff between genomic resolution and velocity precision.** As the spacing between knots decreases, increasing the genomic resolution, the error in the velocity measurement increases. These plots are generated with $n = 500$ simulated data points that are independently Gaussian-distributed about their means. The mean values correspond to a model with 16 genomic segments that each have different fork velocities. Data are presented as mean values ± SEM.

where k_G was the empirically determined growth rate in the simulation and $v(\ell)$ was the local fork velocity at the locus with position ℓ.

We hypothesized that this result might be a special case of choosing an exponential lifetime distribution, since this is consistent with a stochastic realization of chemical kinetics. To test this hypothesis, we simulated several different distributions, including a uniform distribution, for the duration of the B period and the stepping lifetime for the replisome (as well as simulating multifork replication). In each case, the local log-slope relation held, Eq. 3.16, even as the growth rate and fork velocities changed with the changes in the underlying simulated growth dynamics. We therefore hypothesized that Eq. 3.16 was a universal law of cell-cycle dynamics and independent of Cooper and Helmstetter's original assumptions.

3.4.4 The exponential-mean duration

Motivated by this empirical evidence, we exactly computed the population demography in a class of stochastically-timed cell models [14]. In short, we showed that there is an exact correspondence between these stochastically-timed models and deterministically-timed models in exponential growth. The relationship between the corresponding deterministic lifetime τ_i of a state i and the underlying distribution p_i in the stochastic model is the exponential mean, Eq. 3.2 [14]. The exponential mean biases the mean towards short times, the growth rate k_G determines the strength of this bias, and the biological mechanism for this bias is due to the enrichment of young cells relative to old cells in an exponentially growing culture [14].

To understand the consequences of this result, we consider two special cases of this exponential mean. For processes with lifetimes short compared to the doubling time, Eq. 3.2, can be Taylor expanded to show that the exponential mean is:

$$\tau \approx \mu_t - \tfrac{1}{2} k_G \sigma_t^2 + ..., \tag{3.17}$$

the regular arithmetic mean μ_t with a leading-order correction proportional to the product of the growth rate and variance σ_t^2. In the context of single-nucleotide incorporation, this correction is on order one-part-in-a-million and therefore can be ignored. As a consequence, Eq. 3.16, corresponding to the transitions between states with short-lifetimes, is unaffected by the stochasticity, exactly as we observed in our simulations.

Another important case to consider is the strong disorder limit, in which a small fraction of the population ϵ stochastically arrests, i.e., with lifetime ∞, while the other individuals have exponential-mean lifetime τ_0. Using the definition in Eq. 3.2, it is straightforward to show that the deterministic lifetime is:

$$\tau = \tau_0 - T \log_2(1-\epsilon) \approx \tau_0 + \tfrac{\epsilon}{\ln 2} T, \tag{3.18}$$

where T is the population doubling time and the second equality is an approximation for small ϵ. The exponential mean duration is extended by the arrest, but remains finite. Therefore, an arrest of a subpopulation is indistinguishable from a longer duration pause in an exponentially proliferating population. (See Ref. [14].)

3.4.5 Marker-frequency demography

For a stochastic model with locus-dependent fork velocity, we showed that Eqs. 3.14 and 3.15 generalize to:

$$N(\ell) = N_0 \, \mathrm{e}^{-k_G \tau(\ell)}, \tag{3.19}$$

where we will call $\tau(\ell)$ the lag time of a locus at position ℓ, which is equal to the sum of the differential lag times for each sequential step:

$$\tau_j = \sum_{i=0}^{j-1} \delta\tau_i, \tag{3.20}$$

where $\delta\tau_i$ is the differential lag time for state i or the exponential mean of the state lifetime [14]. In the continuum limit, it is more convenient to represent this sum as an integral:

$$\tau(\ell_i) = \int_0^{\ell_i} \mathrm{d}\ell \tfrac{1}{v(\ell)}, \tag{3.21}$$

where the fork velocity is defined: $v(\ell_i) \equiv 1$ bp$/\delta\tau_i$. To demonstrate that the generalized stochastic model predicts the log-slope relation, Eq. 3.16, the log-slope can be derived by substituting Eq. 3.21 into Eq. 3.19, as was observed in the stochastic simulations, demonstrating the universality of Eq. 3.4. We note that Wang and coworkers had previously derived an equivalent expression using the deterministic framework of the Cooper-Helmstetter model in the Material and Methods Section of Ref. [32].

3.4.6 Stochasticity has a minimal effect on the marker frequency

We initially had hypothesized that stochasticity should affect the marker frequency. As explained above, it is the rapidity of base incorporation that explains why stochasticity is dispensable in this context. The same argument does not apply to the B period which is comparable to the duration of the cell cycle. However, for the marker frequency, it is lag-time differences between the replication times of loci that is determinative, and therefore the lag time of the B period cancels from these lag-time differences. Although it is mostly irrelevant for understanding wild-type cell dynamics, stochasticity and an arrested subpopulation will play an important role in one phenomenon we analyze: replication-conflict induced pauses.

3.4.7 Time resolution

Due to the large number of reads achievable in next generation sequencing, the time resolution will be high in carefully designed analyses. The number of reads is subject to counting or Poisson noise. It is therefore straightforward to estimate the experimental uncertainty in the lag time due to finite read number:

$$\sigma_{\tau_j} \quad = \quad k_G^{-1} \frac{1}{\sqrt{N_j}} = 1 \text{ s} \cdot \left(\frac{6\times10^6}{N_j}\right)^{1/2}, \tag{3.22}$$

where we have used a read number inspired by the replication-conflict pausing example. This estimate suggests that under standard conditions, time measurements with an uncertainty of seconds are possible using this approach.

3.4.8 Fork-velocity resolution

To compute the slope in Eq. 3.4, the log-marker-frequency is fit to a piecewise linear function with equal spacing between knots. See Fig. 3.7b. There is an important tradeoff between genomic resolution (i.e., the genomic distance between knots) and fork velocity precision (i.e., the uncertainty in velocity measurement): Increasing the genomic distance between knots reduces the genomic resolution but also reduces the uncertainty in the velocity measurement. We therefore consider a series of models with increasing genomic resolution and use the Akaike Information Criterion (AIC) to select the optimal model [29, 30]. See Supplementary Methods Sec. 3.8.3.10. This approach balances the desire to resolve features by increasing the genomic resolution with the loss of velocity precision.

Given a knot spacing, it is straightforward to estimate the relative error:

$$\frac{\sigma_v}{v} = \sqrt{\frac{2}{n\,(\Delta\ell)^3}}\frac{v}{k_G} \approx 0.1 \cdot \left(\frac{1.5}{n}\right)^{1/2} \left(\frac{100\ \text{kb}}{\Delta\ell}\right)^{3/2}, \tag{3.23}$$

where n is the read depth in reads per base and $\Delta\ell$ is the spacing between knots in base pairs. Therefore, for a canonical next-generation-sequencing experiment, we can expect to achieve roughly 10% error in the fork velocity for 100 kb genomic resolution. Note that in our error analysis, we have included only the error from cell number N, not the error from the uncertainty in the cell-cycle duration, which covaries between loci in a particular experiment.

3.5 Data and code availability

The sequencing datasets generated during the current study are available from the NCBI Sequence Read Archive with the BioProject accession code PRJNA919081. The data from Galli et al. [36] and Midgley-Smith et al. [57] are both available from the European Nucleotide Archive (ENA), with the accession codes PRJEB28538 and PRJEB25595, respectively. The digitized data from Wang et al. [33] and Srivatsan et al. [32] are available in the Source Data file. More detailed information about data availability is provided in Supplementary

Table 3.4. MATLAB scripts written for this study are available on the GitHub repository and on reasonable request.

3.6 Acknowledgments

The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, and J. Mittler for many useful conversations. We would like to thank P. Levin, J. Wang, L. Simmons, and S. Pigolotti for advice on our manuscript. We thank S. B. Peterson and A. Schaefer for help with *V. cholerae*. We would like to thank J. Wang, C. Possoz, F.-X. Barre, and C. Rudolph for detailed conversations about their data. This work was supported by NIH grant R01-GM128191, which was awarded to P.A.W and H.M.

3.7 Supplementary tables

3.7.1 Bacterial strains used in this study

Table 3.3: Strains used in the study.

Short name:	Strain name:	Description:	Source:
V. cholerae WT	O1 biovar El Tor N16961	*ChapR* Δ*lacZ* gm^R	[36]
V. cholerae MCH1	MCH1	Integration of N16961 Chr2, without *oriC2* and partition machinery region, in place of the *dif1* site in N16961 Chr1.	[36]
V. cholerae oriR4	EGV111	N16961 *ChapR* Δ*lacZ oriC1* @ R4 (1,898Mb) gm^R	[36]
B. subtilis WT	JH642	*trpC2 pheA1*	[33]
B. subtilis 257°::*oriC*	MMB703	*trpC2 pheA1 argG*(257°)::(*oriC*/*dnaAN kan*) Δ(*oriC-L*)::*spc*	[33]
B. subtilis 94°::*oriC*	JDW258	*trpC2 pheA1 aprE*(94°)::(*oriC*/*dnaAN kan*) Δ(*oriC-L*)::*spc dnaB134ts-zhb83*::Tn*917*(*cat*)	[33]
B. subtilis oriN	MMB208	*pheA1* (*ypjG-hepT*)*122 spoIIIJ*(359°)::(*oriN kan tet*) Δ*oriC-S*	[33]
B. subtilis 257°::*oriN*	MMB700	*pheA1* (*ypjG-hepT*)*122 argG*(257°)::(*oriN kan*) Δ*oriC-S*	[33]
B. subtilis YB886	YB886	*trpC2 metB5 sigB amyE sp*β*-ICEBs*° *xin-*	[32]
B. subtilis rrnIHG(pre-inv)	JDW858	YB886 *kbaA'::neo' cat::'ybaN rrnG-5S'::erm 'neo::'ybaR*	[32]
B. subtilis rrnIHG(inv)	JDW860	YB886 *kbaA'::neo erm* inv(*ybaN::rrnG-5S*) *cat::'ybaR*	[32]
E. coli WT	K-12 MG1655	F- lambda- *ilvG- rfb*-50 *rph*-1	[57]

3.7.2 Datasets used in this study

All datasets were obtained from cells grown at 37 °C. Next-generation-sequencing FASTQ files and marker frequencies that were generated for this study are available from the NCBI Sequence Read Archive with the BioProject accession code PRJNA919081. The data from Galli et al. [36] and Midgley-Smith et al. [57] are both obtained from the European Nucleotide Archive (ENA), with the accession codes PRJEB28538 and PRJEB25595, respectively. The individual run accessions are also included. The digitized data from Wang et al. [33] and Srivatsan et al. [32] are available in the Source Data file, with the corresponding sheets listed in the table below. In the following table, Exp is shorthand for exponential phase growth and Stat is shorthand for stationary phase. Growth media are described in more detail in Supplementary Methods Sec. 3.8.1.

Table 3.4: Datasets used in the study.

Short name:	Growth media:	Doubling time:	Source:	Project accession:	Run accession:	Sample size (read count):
V. cholerae WT	LB	22 ± 1 min	This study	SRA: PRJNA919081	Exp: SRR23003324 Stat: SRR23003328	5.4×10^7 6.0×10^7
	M9 fructose	50 ± 4 min	From study [36]	ENA: PRJEB28538	Exp: ERX2796386 Stat: ERX2796387	3.2×10^7 2.0×10^7
V. cholerae MCH1	LB	90 ± 6 min	This study	SRA: PRJNA919081	Exp: SRR23003321 Stat: SRR23003311	5.4×10^7 6.1×10^7
	M9 fructose	50 ± 4 min	From study [36]	ENA: PRJEB28538	Exp: ERX2796384 Stat: ERX2796385	1.1×10^7 1.5×10^7
V. cholerae oriR4	M9 fructose	55 ± 5 min	From study [36]	ENA: PRJEB28538	Exp: ERX2796379 Stat: ERX2796380	1.5×10^7 1.3×10^7
B. subtilis WT	S7 fumarate	61.0 ± 1.2 min	From study [33]	N/A	N/A	N/A (Microarray)
B. subtilis 257°::*oriC*	S7 fumarate	Not reported	From study [33]	Digitized	Source Data: Bs_257	N/A (Microarray)
B. subtilis 94°::*oriC*	S7 fumarate	Not reported	From study [33]	Digitized	Source Data: Bs_94	N/A (Microarray)
B. subtilis oriN	S7 fumarate	64.3 ± 2.9 min	From study [33]	Digitized	Source Data: Bs_oriN	N/A (Microarray)
B. subtilis 257°::*oriN*	S7 fumarate	Not reported	From study [33]	Digitized	Source Data: Bs_oriN257	N/A (Microarray)
B. subtilis rrnIHG(pre-inv)	LB	20 ± 1 min	From study [32]	N/A	N/A	N/A (Microarray)
	MOPS glucose CA	28 ± 1 min	From study [32]	N/A	N/A	N/A (Microarray)
	MOPS glucose MM	42 ± 1 min	From study [32]	N/A	N/A	N/A (Microarray)
B. subtilis rrnIHG(inv)	LB	> 160 min	From study [32]	N/A	N/A	N/A (Microarray)
	MOPS glucose CA	44 ± 5 min	From study [32]	Digitized	Source Data: Bs_IHG_MM	N/A (Microarray)
	MOPS glucose MM	44 ± 1 min	From study [32]	Digitized	Source Data: Bs_IHG_CA	N/A (Microarray)
E. coli WT	LB	19.3 ± 1.7 min	From study [57]	ENA: PRJEB25595	Exp: ERS2298483 Stat: ERS2298484	1.4×10^7 1.3×10^7
	M9 glucose	68.8 ± 6.2 min	From study [57]	ENA: PRJEB25595	Exp: ERS2298504 Stat: ERS2298505	1.7×10^7 1.4×10^7

3.8 Supplementary methods and derivations

3.8.1 Growth media and determination of growth phase

As we have used data from multiple sources, the minimal media and the determination of population growth phase are not consistent across all studies. All studies use the standard recipe for Luria-Bertani (LB) rich medium (1% tryptone, 0.5% yeast extract, and 1% NaCl in H2O), with the exception of Midgley-Smith et al. (2018), where 0.05% NaCl is used instead. All studies using M9 minimal media have the same base recipe (1X M9 salts, 2 mM MgSO4, and 0.1 mM CaCl2), with different supplements and carbon sources. In our study and in Galli et al. (2019) [36], M9 was supplemented with 10 μg mL^{-1} thiamine HCl for the *V. cholerae* strains. The carbon source for M9 is 0.4% glucose for our study, 0.4% fructose for Galli et al. [36], and 0.2% glucose for Midgley-Smith et al. [57]. In Wang et al. (2007) [33], the relevant datasets exclusively use S7 minimal medium (50 mM MOPS, 10 mM (NH4)2SO4, 5 mM potassium phosphate, 2 mM MgCl2, 0.9 mM CaCl2, 50 μM MnCl2, 5 μM FeCl3, 10 μM ZnCl2, and 2 μM thiamine hydrochloride), supplemented with 1% sodium fumarate as the carbon source, 0.1% glutamate, 40 μg mL^{-1} tryptophan, and 40 μg mL^{-1} phenylalanine. For Srivatsan et al. (2010) [32], the minimal medium consists of 50 mM MOPS with 1% glucose, along with different supplements for the CA (0.5% casamino acids) and MM (40 μg mL^{-1} tryptophan, 40 μg mL^{-1} methionine, 40 μg mL^{-1} phenylalanine, and 100 μg mL^{-1} arginine) growth media.

All populations were grown at 37 °C. In our study, exponential phase cultures were grown to OD_{600} of 0.60-0.80. Stationary phase samples were grown to OD_{600} of 1.50 or greater. In Galli et al. [36], exponential phase cultures were grown to an OD_{650} of 0.05 and 0.2 in M9 and LB, respectively. In Midgley-Smith et al. [57], exponential phase cultures were grown to an OD_{600} of 0.48. In Srivatsan et al. [32], exponential phase cultures were grown to an OD_{600} of 0.2-0.6.

3.8.2 Generation of marker frequency data for this study

For detailed protocols of marker-frequency generation for each dataset, see the corresponding references in Sec. 3.7.2. In this study, strains were struck on solid agar and grown overnight at 37 °C. Three individual colonies were selected from each strain and used to inoculate 10 mL of LB media, which was grown overnight at 37 °C with 260 rpm shaking. The following morning, these overnight pre-cultures were back-diluted to OD_{600} of 0.05 in 5 mL of either LB or M9 media. The same biological replicate was used to inoculate both growth conditions. Exponential phase cultures were grown to OD_{600} of 0.60-0.80. Stationary phase samples were grown to OD_{600} of 1.50 or greater. To harvest, cultures were centrifuged at $8,000 \times g$ for 5 mins at 25 °C, and the supernatant was removed. gDNA was prepared immediately following harvest using GeneJet Genomic DNA Purification Kit (Thermo Scientific, K0721). Purified gDNA was quantified using Qubit dsDNA HS Assay Kit (Invitrogen, Q32851) and quality was assayed using a Nanodrop 2000 Spectrophotometer (Thermo Scientific, ND-2000). Whole-genome libraries were prepared from purified gDNA using Twist 96-Plex Library Prep Kit (Twist Bioscience, 104950) with dual adapters. Libraries were subsequently pooled and sequenced on Illumina NovaSeq6000 (S4) to a depth of 6×10^7 reads per sample.

3.8.3 Marker frequency analysis

3.8.3.1 Sequence alignment

To align deep sequencing read outputs to the bacterial reference genomes for MFA, we used the read alignment tool Bowtie 2 (v2.4.5) via the MATLAB (R2022b) function `bowtie2` [58], which requires the Bowtie 2 Support Package from the MATLAB Bioinformatics Toolbox (v4.16.1). We then extracted the position information from the resulting SAM file using an in-lab written MATLAB script and the command-line tool SAMtools [59] (v1.16.1).

3.8.3.2 Data selection

Not all marker-frequency experiments are of the same quality and many datasets we have analyzed have clear signatures of systematic error. We use two key criteria to determine the quality of the data. The first is a qualitative measure involving the shape of the marker-frequency profile and the second is a quantitative measure based on resolution analysis.

Marker profile near origin. Our analysis uses exponential growth as the stop-watch for resolving dynamics, so it is essential that the population is harvested at the right time for sequencing. We have found that the marker-frequency profile has a signature flattening near the origin of replication if the population is harvested too late in exponential phase, as shown in Supplementary Fig. 3.24. The population is entering stationary phase, where nutrient depletion begins to cause slower growth. Entering stationary phase also causes a decrease in the population-wide rate of replication initiation, which leads to the flattening of the profile near the origin. For the rest of the chromosome, replication continues unchanged for the majority of the population. Since exponential growth is crucial to our analysis, we have chosen to only use datasets with cusp-like behavior near the origin(s), as opposed to a rounded concave-down shape.

Resolution analysis. In the case where there is a single origin of replication for each chromosome, we expect only a single maximum in the marker frequency profile, corresponding to the origin. In an ideal noiseless situation, any local maxima distinct from the origin would correspond to other points of replication initiation and the fit would predict a retrograde fork velocity. However, due to the stochastic nature of the sequencing data, there can be spurious local maxima if the resolution is chosen to be too high. Fluctuations due to noise can cause the knots in the piecewise-linear fit along each arm of the chromosome to be non-monotonic, as shown in Panel c of Fig. 1 in the main paper. Thus, given two datasets for the same organism and growth condition, data quality was determined based on the best resolution that can be achieved without introducing spurious maxima.

3.8.3.3 To divide or not to divide by stationary phase data

To further remove any sequencing induced bias, we divide the copy number data obtained during exponential growth by the data obtained during stationary phase. This normalization step is done with the 1 kb binned values. Since the division of two noisy uncorrelated data sets typically results in a decreased signal-to-noise ratio (SNR), we used a comparison of the variance divided by the mean before and after division by stationary phase.

We expect the relative variance to decrease by normalization if the reduction in sequencing bias resulting from division has a larger effect than the increase in noise from dividing two noisy data sets. We first divide the noisy stationary phase data by its expected mean value. This ensures that it is of unit order and will not introduce a global scale factor during division, which would affect only the normalized data and not the pre-normalization data. We find that after division, the variance is roughly halved, suggesting that the normalization by stationary phase is successful in reducing sequencing bias without resulting in excessive noise. We thus chose to divide all exponential phase data with the corresponding stationary phase data. For some datasets, the stationary phase results were not provided for some of the ectopic origin mutants. The only difference in these mutants from their original strains is the addition of a short segment of DNA for the ectopic origin, which does not significantly change the locations of sequencing bias. Therefore, we chose to divide the exponential phase data for the ectopic origin strains by the stationary phase data from the original strains.

3.8.3.4 Filtering the data

To remove outliers from the marker-frequency data, we used a centered 200 kb median filter, such that at each position ℓ along the chromosome, we set the median marker-frequency from $\ell - 100$ kb to $\ell + 100$ kb as a baseline. The data was given periodic boundary conditions such that the centered medians still had 200 kb regions at the boundaries of the data set. We then discarded the 10% of points that most deviated from this centered median baseline, sorted by the absolute value of the difference. We also removed data points from regions

known to be associated with mobile elements of low GC content, such as the Super Integron on Chr2 of *V. cholerae*, which is subject to anomalous sequencing bias.

3.8.3.5 Nearest-neighbor variance estimator

We used a nearest-neighbor estimate of variance:

$$\hat{\sigma}^2 \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(x_{i+1}-x_i)^2, \tag{3.24}$$

where $x_{N+1} = x_1$, in agreement with periodic boundary conditions. This approach allows us to obtain a variance measurement without assuming the underlying distribution of the data and the corresponding expected value of each bin.

To see how this relates to the usual definition of variance, we use the notation $\langle x\rangle \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$ to denote the average:

$$\begin{aligned}
\hat{\sigma}^2 &\approx \frac{1}{2}\left\langle (x_{i+1}-x_i)^2\right\rangle, && (3.25)\\
&= \frac{1}{2}\left\langle x_{i+1}^2 + x_i^2 - 2x_{i+1}x_i\right\rangle, && (3.26)\\
&= \frac{1}{2}\big(\left\langle x_{i+1}^2\right\rangle + \left\langle x_i^2\right\rangle - 2\left\langle x_{i+1}\right\rangle\left\langle x_i\right\rangle\big), && (3.27)\\
&= \left\langle x^2\right\rangle - \left\langle x\right\rangle^2. && (3.28)
\end{aligned}$$

where $\langle x_{i+1}x_i\rangle = \langle x_{i+1}\rangle\langle x_i\rangle = \langle x\rangle^2$ because the values of an independent random variable are uncorrelated and $\left\langle x_{i+1}^2\right\rangle = \left\langle x_i^2\right\rangle = \left\langle x^2\right\rangle$ because the expectation value is taken over the full data set. Eq. 3.28 is the usual definition of the variance.

3.8.3.6 Fitting slopes to the log of the copy number

After taking the natural logarithm of the copy number, we expect the data to follow a roughly linear trend along each arm of the chromosome. Since our goal is to obtain higher resolution measurements of fork velocity along the chromosome, it is necessary to subdivide each arm into smaller segments with variable slopes. The fit to the data should be continuous, since discontinuities would create copy number ambiguity at segment junctions. Therefore, we

use a continuous piecewise linear least squares fit to obtain slope measurements from the data. We call the MATLAB function `lsqnonlin` as part of an in-lab written fitter function. The `lsqnonlin` function is part of the MATLAB Optimization Toolbox (v4.16.1). The fitter obtains vertical control point values at the junctions between segments. Each control point is used in the measurement of the slope both to the left and to the right of it.

3.8.3.7 Estimating errors for the control points

To obtain error estimates for the control points of the least squares fit, we use the Jacobian that lsqnonlin returns, which is the Jacobian of the difference between the fit data and the experimental data. If we have k fit parameters θ_α, where $\alpha = 1, ..., k$, and N data points y_i, where $i = 1, ..., N$, then the Jacobian takes the form:

$$J_{i\alpha} = \frac{\partial}{\partial\theta^\alpha}(y_i - \mu_i(\theta)), \tag{3.29}$$

where $\mu_i(\theta)$ is the fitter estimate of y_i, based on the model with the set of parameters θ. We use the matrix form of the Fisher information:

$$[I(\theta)]_{\alpha,\beta} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta^\alpha}\log f(Y;\theta)\right)\left(\frac{\partial}{\partial\theta^\beta}\log f(Y;\theta)\right)\middle|\theta\right], \tag{3.30}$$

where $f(Y;\theta)$ is the probability density function (PDF) of the random variable Y, given parameters θ. In our case, we expect the noise to be Gaussian distributed around the mean, so we have the PDF:

$$f(Y;\theta) = \frac{1}{(\sigma\sqrt{2\pi})^N}\prod_{i=1}^{N}\exp\left(-\frac{(y_i - \mu_i(\theta))^2}{2\sigma^2}\right), \tag{3.31}$$

where σ is the standard deviation. Now we take the derivative of the logarithm:

$$\frac{\partial}{\partial\theta^\alpha}\log f(Y;\theta) = \frac{\partial}{\partial\theta^\alpha}\left(-\frac{\sum_i(y_i - \mu_i(\theta))^2}{2\sigma^2} - N\log\left(\sigma\sqrt{2\pi}\right)\right), \tag{3.32}$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \mu_i(\theta))\frac{\partial}{\partial\theta^\alpha}(y_i - \mu_i(\theta)), \tag{3.33}$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \mu_i(\theta))J_{i\alpha}. \tag{3.34}$$

Plugging this into Eq. 3.30 and noting that the matrix multiplication contracts over the i index, we are left with:

$$I(\theta) = \frac{J^T J}{\sigma^2}. \tag{3.35}$$

The Cramér-Rao bound states that the inverse of the Fisher information provides a lower bound for the covariance matrix of unbiased estimators. More explicitly, the relation is as follows:

$$\mathrm{cov}_\theta\left(\hat{\theta}\right) \geq I(\theta)^{-1}. \tag{3.36}$$

We take the equality for our error estimates. The variances of the parameter estimates are given by the diagonal elements of the covariance matrix.

3.8.3.8 Transformation matrices for obtaining the fork velocity

To convert the control points of the piecewise linear fit into slopes and fork velocities, we use coordinate transformation matrices, which have the added benefit of also transforming the variance estimates from the Fisher information. The coordinate transformation matrices can either be obtained through direct reasoning or as Jacobians of the final coordinates relative to the initial coordinates. The following examples will be square matrices of dimension 3, acting on sets of 3 parameters. These examples illustrate the procedure of obtaining the transformation matrices and are easily generalizable. To convert from a vector of control point values to a vector where the first element is the leftmost control point value and the other elements are the slopes, we use the following matrix:

$$\texttt{CtrlPts2Slopes} = \begin{pmatrix} 1 & 0 & 0 \\ -\Delta_{12}^{-1} & \Delta_{12}^{-1} & 0 \\ 0 & -\Delta_{23}^{-1} & \Delta_{23}^{-1} \end{pmatrix}, \tag{3.37}$$

where Δ_{ij} is the chromosomal position (horizontal value) of the j-th control point minus the position of the i-th control point. We can thus relate the vector of slopes $\vec{\alpha}$ and the vector

of (vertical) control point values $\vec{\theta}$:

$$\vec{\alpha} = \texttt{CtrlPts2Slopes} \cdot \vec{\theta} = \begin{pmatrix} \theta_1 \\ (\theta_2 - \theta_1)/\Delta_{12} \\ (\theta_3 - \theta_2)/\Delta_{23} \end{pmatrix} \tag{3.38}$$

From the slopes, we can obtain the fork velocities with the following matrix:

$$\texttt{Slopes2Vels} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -k_G/\alpha_2^2 & 0 \\ 0 & 0 & -k_G/\alpha_3^2 \end{pmatrix}. \tag{3.39}$$

The velocities $\vec{v}$ are related to $\vec{\alpha}$ like so:

$$\vec{v} = \texttt{Slopes2Vels} \cdot \vec{\alpha} = \begin{pmatrix} \alpha_1 \\ -k_G/\alpha_2 \\ -k_G/\alpha_3 \end{pmatrix}. \tag{3.40}$$

To properly transform the covariance matrix obtained in Sec. 3.8.3.7, we must use the transformation matrix on both sides, like so:

$$\text{cov}_\alpha = \texttt{CtrlPts2Slopes}^T \cdot \text{cov}_\theta \cdot \texttt{CtrlPts2Slopes}, \tag{3.41}$$

$$\text{cov}_v = \texttt{Slopes2Vels}^T \cdot \text{cov}_\alpha \cdot \texttt{Slopes2Vels}. \tag{3.42}$$

The error estimates for each parameter are the square roots of the corresponding diagonal elements.

3.8.3.9 Behavior near the terminus

Due to the stochastic nature of replisome progression, replication forks do not always meet at the terminus, which corresponds to the *dif* site [36]. If one fork arrives first, it can continue past the *dif* site, converting from antegrade to retrograde motion along the other arm. While there are Tus-*Ter* traps in *E. coli* and *B. subtilis* to limit the amount of retrograde motion, they are not 100% efficient and they are not present in *V. cholerae* [36]. Thus, in a population,

the true termination points along the chromosome would form a distribution around the *dif site*. This means that in regions near the terminus, there can be antegrade and retrograde replication forks both contributing to the marker frequency, which contradicts one of the assumptions required for our analysis: replication proceeding only in one direction. Such retrograde motion would cause a flattening out around any fork convergence points observed in the marker-frequency data. Therefore, when doing a multi-segment analysis, we remove the two fork velocity measurements on either side of the marker-frequency minimum.

3.8.3.10 Selection of the number of fitted knots

First we bin the data into 1 kb regions, following the convention of previous marker frequency analysis studies. To determine the number of segments that would best fit the data, we use the Akaike Information Criterion (AIC). We compare a series of nested models with 2^k continuous piecewise-linear least squares fits, where $k = 1, ..., 10$. The AIC-optimal model for fast growth (in LB) had 39 knots, spaced by 100 kb, generating 38 measurements of locus velocity across the two chromosomes of WT *V. cholerae*. Other strains either have similar or smaller region sizes that minimized the AIC value, so for ease of comparison with other mutant strains, we take 100 kb to be the step size for determining fork velocity.

3.8.4 Bilateral symmetry analysis

To analyze whether an observed velocity profile was consistent with the predictions of the time-dependent mechanism, we test for bilateral symmetry between the left and right arms:

$$v(\ell) = v(-\ell), \tag{3.43}$$

where ℓ is the locus position relative to the origin.

Assume the fork velocity has been computed over $2m$ equal-length genome segments, arranged symmetrically about the origin ($\ell = 0$). The segments $i = -1... - m$ correspond to sequentially labeled segments along the left arm and the segments $i = 1...m$ correspond to sequentially labeled segments along the right arm such that $\ell = \Delta\ell i$ corresponds to the

end point of each segment where $\Delta\ell$ is the segment length. Let the mean fork velocity be defined:

$$\overline{v} \equiv \frac{1}{2m}\sum_{i=1}^{m} v_i + v_{-i}, \tag{3.44}$$

where v_i is the velocity over segment i.

To divide the variance into symmetric and antisymmetric contributions, we define symmetrized, (i), and antisymmetrized velocities, $[i]$, for index pairs $\pm i$:

$$\delta v_{(i)} \equiv \tfrac{1}{2}(v_i + v_{-i} - 2\overline{v}), \tag{3.45}$$

$$\delta v_{[i]} \equiv \tfrac{1}{2}(v_i - v_{-i}). \tag{3.46}$$

We define the symmetric and antisymmetric variances

$$\sigma_S^2 \equiv \frac{1}{m}\sum_{i=1}^{m} \delta v_{(i)}^2 \tag{3.47}$$

$$\sigma_A^2 \equiv \frac{1}{m}\sum_{i=1}^{m} \delta v_{[i]}^2, \tag{3.48}$$

and the total variance is the sum of the symmetric and antisymmetric variances:

$$\sigma^2 \equiv \frac{1}{2m}\sum_{i=1}^{m} (v_i - \overline{v})^2 + (v_{-i} - \overline{v})^2 \tag{3.49}$$

$$= \sigma_S^2 + \sigma_A^2. \tag{3.50}$$

Finally, we define the symmetric fraction of the variance:

$$f_S \equiv \sigma_S^2/\sigma^2. \tag{3.51}$$

If the variation in the velocity obeys the bilateral symmetry (Eq. 3.43), the variance is all symmetric (i.e., $f_S = 1$). If, on the other hand, the variation is randomly distributed along the genome, we expect equal symmetric and antisymmetric combinations (i.e., $f_S = 1/2$).

3.8.5 Estimation of average fork number per cell cycle

The following analysis is only an estimation. There are a few key assumptions that are inconsistent with cell phenomenology, but are kept to make the analysis tractable. In

particular, we take the assumption that neither arm of the chromosome has fork arrest, which is inconsistent with the results for *B. subtilis*. This assumption allows us to obtain an estimate without prior knowledge of where fork arrest may occur. We also assume that the D period (time between end of chromosomal replication and cell division) is negligible, which is inconsistent when growth is extremely slow. This assumption allows us to use the copy number of the terminus as an approximation of the number of cells, which cannot be determined from MFA.

The population number density of forks $n_f(\ell)$ at any position ℓ is given by the change in copy number between two consecutive positions along the chromosome:

$$n_f(\ell) = -\frac{\mathrm{d}}{\mathrm{d}\ell} N(\ell). \tag{3.52}$$

This equation is a result of the following reasoning: If there are two copies at $\ell = x$ and one copy at $\ell = x + 1$, then there must be a fork that has just replicated $\ell = x$. We need to integrate the number density over the entire chromosome to get the total number of forks for the population:

$$N_{fork,population} = 2\int_{ori}^{ter} n_f(\ell)\,\mathrm{d}\ell = 2\int_{ori}^{ter} -\frac{\mathrm{d}}{\mathrm{d}\ell} N(\ell)\,\mathrm{d}\ell = 2(N_{ori} - N_{ter}), \tag{3.53}$$

where the factor of 2 came from having two arms. To get the average number of forks per cell $\bar{N}_f$, we divide by the total number of cells, which is roughly equal to the number of termini N_{ter}:

$$\bar{N}_f = 2\frac{N_{ori} - N_{ter}}{N_{ter}}. \tag{3.54}$$

This can be related to lag time:

$$\tau_\ell = -k_G^{-1} \ln \frac{N_\ell}{N_{ori}} \implies \frac{N_{ori}}{N_{ter}} = \exp(k_G \tau_{ter}) = e^{\tau_{ter} \ln 2/T} = \left(e^{\ln 2}\right)^{\tau_{ter}/T} = 2^{\tau_{ter}/T}, \tag{3.55}$$

which can be substituted into the equation above to get:

$$\bar{N}_f = 2 * (2^{\tau_{ter}/T} - 1). \tag{3.56}$$

This is closely related to the result derived in [21]. For two chromosomes, we add the number of forks for both, but divide everything just by the minimum $N_{ter,i}$, which corresponds to the terminus of Chr1 in this case:

$$\bar{N}_{f,2chr} = 2\frac{(N_{ori,1} - N_{ter,1})}{N_{ter,1}} + 2\frac{(N_{ori,2} - N_{ter,2})}{N_{ter,1}}. \tag{3.57}$$

In general, when there are multiple chromosomes:

$$\bar{N}_{fork,multichr} = \frac{2}{\min(N_{ter,i})}\sum_i (N_{ori,i} - N_{ter,i}), \tag{3.58}$$

where i denotes which chromosome. To get a relative lag time, we use:

$$\tau_{\ell\to m} = \tau_m - \tau_\ell = -k_G^{-1}\left(\ln\frac{N_m}{N_{ori}} - \ln\frac{N_\ell}{N_{ori}}\right) = k_G^{-1}\ln\frac{N_\ell}{N_m}. \tag{3.59}$$

Letting the subscript *end* denote the lag time corresponding to $\min(N_{ter,i})$. We thus have:

$$\bar{N}_{fork,multichr} = 2\sum_i \left(2^{(\tau_{end}-\tau_{ori,i})/T} - 2^{(\tau_{end}-\tau_{ter,i})/T}\right), \tag{3.60}$$

where τ_{end} is the maximum lag time measured. This can be simplified to:

$$\bar{N}_{fork,multichr} = 2\times 2^{\tau_{end}/T}\times\sum_i \left(2^{-\tau_{ori,i}/T} - 2^{-\tau_{ter,i}/T}\right), \tag{3.61}$$

which enables a calculation of average fork number per cell cycle using lag times.

3.8.6 Statistically significant deviations of local fork velocity from the global mean

To test the statistical significance of the measured velocities, we use a null hypothesis test. The null hypothesis corresponds to a model with uniform fork velocity on the right and left arms. The alternative hypothesis corresponds to a model with a model-selected number of knots (m), corresponding to $m-1$ genomic region-specific velocities.

To calculate the p-value, we begin by obtaining the χ^2-test statistic:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{\sigma_i^2} \tag{3.62}$$

where O_i denotes the observed values, E_i denotes the expected values, and σ_i^2 is the variance of the fork velocities measured using the fitter function (as described in Supplementary Methods Sec. 3.8.3.8). In this case, E_i is the global mean fork velocity for all i, since that is the null hypothesis. Since there is only one degree of freedom for the model, we use the one-dimensional χ^2 cumulative distribution function to determine the probability of measuring a χ^2 statistic as extreme as the one we've measured, assuming the null hypothesis is true. This p-value is found to be $\ll 10^{-30}$ for all datasets except one with a p-value of 6×10^{-12}. All of the p-values are much smaller than the standard threshold of 0.05, so we reject the null hypothesis. We thus conclude that there are statistically significant variations of the fork velocity relative to the mean.

3.9 Supplementary notes on the stochastic simulation

3.9.1 Stochastic simulations method

The purpose of the stochastic simulation was to investigate the role of stochasticity in determining the log-phase growth demography of the population. It was *not* to generate a mechanistically realistic and detailed model of the cell cycle.

To simulate the cell cycle, we performed an exact stochastic simulation using the Gillespie Algorithm [56]. We idealized the cell cycle as follows: **Genome representation:** The genome was divided into two equal length arms each consisting to 100 course-grained bases. An explicit realization of all 5 Mb would be too slow to rapidly explore different cell models. Finer coarse-grained basepairs were explored but made no difference to the results. **Replication initiation**: We experimented with a number of different models for initiation since we initially believed that the stochasticity of initiation would limit the ability to quantitate the fork velocity. (i) We initially investigate a *Terminus Model* in which there was a constant rate of initiation k_{init} after the termination of the on-going replication. (ii) In order to simulate a more realistic model, we then considered an *Origin Model* where there was constant rate k_{init} of initiation at each origin. (iii) We explored a *Precise Model*

with precise timed B period length T_B. (iv) We explored a *Uniform Model* with a uniform distribution of B periods. **Fork velocity:** We used the replication of the course-grained bases had exponentially distributed wait times $k_{\text{rep}} = v/\Delta\ell$ (i.e., constant rate per unit time) where v is the replication velocity and $\Delta\ell$ is the length of the course-grained bases. We note that this assumption makes replication *more stochastic* than a more realistic model in which wait times are exponentially distributed at the single-base level. We experimented with a number of different types of locus-dependent fork velocities. In addition, we added exponentially-distributed pauses of various durations at a specific locus. **Termination:** We only allowed a single direction of fork propagation. When the fork reached the end of the arm, replication terminated. **Cell division:** We assumed cells divided immediately after termination of the slowest arm of the chromosome. **Initiation of the simulation:** All simulations initiated from a single new-born cell with an un-replicated chromosome. **Stop conditions:** Simulation were run until there were 10^5 cells. **Determination of growth rate.** The number of cells was fit to an exponential to determine the growth rate k_G between $N_{\text{cell}}(t) = 10^4$ and $N_{\text{cell}}(t) = 10^5$ cells. **Generation of simulated marker frequency.** The marker frequency was defined as the number of each genetic locus in the population at the termination of the simulation. In summary, the model was described by the parameters k_{init} and the fork velocity $v(\ell)$ (or pause time distribution) at each locus. The model output was the marker frequency at each locus $N(\ell)$ and the growth rate k_G. See Supplementary Fig. 3.8.

3.9.2 Stochastic simulations match the predictions of the log-slope

To test whether the log-slope law applies locally, irrespective of stochasticity in cell-cycle timing, we used a stochastic simulation to generate simulated marker-frequency data. The basic strategy was to simulate a stochastically timed cell cycle, including stochastically timed B periods as well as stochastic fork dynamics, and compare the observed population demographics to the prediction of non-stochastic models.

3.9.2.1 Log slope

We define the log slope:

$$\alpha(\ell) \equiv \frac{\mathrm{d}}{\mathrm{d}\ell} \log N(\ell), \tag{3.63}$$

where $N(\ell)$ is the population copy number of the locus at position ℓ. Theoretically, the slope is predicted to be:

$$\alpha(\ell) = \frac{k_G}{v(\ell)}, \tag{3.64}$$

where k_G is the growth rate and $v(\ell)$ is the locus-dependent fork velocity.

3.9.2.2 Log difference

For long pauses, the dynamics are characterized by a step rather than a slope. The step is defined:

$$\Delta(\ell) \equiv \lim_{\delta \to 0^+} \log N(\ell + \delta) - \log N(\ell), \tag{3.65}$$

where $N(\ell)$ is the population copy number of the locus at position ℓ. Theoretically, the log difference is predicted to be:

$$\Delta(\ell) \equiv k_G \delta\tau(\ell) = \log\bigl(1 + \tfrac{k_G}{k}\bigr), \tag{3.66}$$

where k_G is the growth rate, $\delta\tau(\ell)$ is the exponential mean of the step wait time and the last equality applies in the special case of an exponentially-distributed wait time with rate k.

3.9.3 Simulation models

The model descriptions and motivations are summarized below (See Fig. 3.8 for a schematic representation):

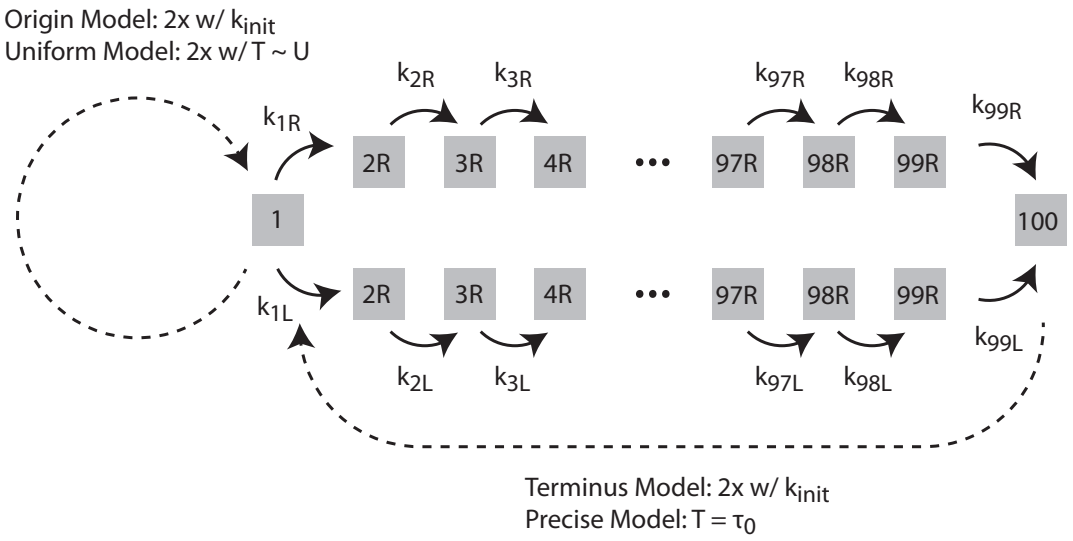


Figure 3.8: **Schematic of stochastic simulation model.** Replication initiates at the origin (state 1) and proceeds bi-directionally along the left (L) and right arms (R). Each transition between simulation bases (i.e., state number) represents the replication of roughly 25 kb. States have exponentially-distributed lifetimes with a rate equal to the fork velocity at that position: $k_X = v(\ell_X)$. A number of different initiation schemes were simulated. In the *Origin Model*, the initiation rate was proportional to the number origins (per cell). In the *Terminus Model*, the initiation rate was proportional to the number termini (per cell). In the *Uniform Model*, there is a uniformly distributed wait time after initiation before the next initiation occurs. In the *Precise Model*, there is a fixed time interval after termination before the next initiation occurs.

Terminus Model 1: In the terminus model, the replication initiation rate is proportional to the terminus number (per cell). In this model, there is a locus-independent fork velocity.

Terminus Model 2: Same as above, but with two fork velocities. On the last half of the right arm (region 1), the fork velocity is higher than the rest of the chromosome (region 0).

Terminus Model 3: Same as above, but with an exponentially-distributed pause close to the origin on the right arm.

Terminus Model 4: Same as Terminus 1, but with a lower initiation rate.

Origin Model: In the origin model, the replication initiation rate is proportional to the origin number.

Uniform Model: The timing between initiations is uniformly distributed. This model was included to explore whether the observations were a special case of exponentially-distributed wait times.

Precise Model: The timing between initiations is uniformly distributed. This model was included to explore whether the observations were a special case of exponentially-distributed wait times.

3.9.4 Simulation results

The results from the simulations are summarized in the table below. All numbers have at least precision of 1% and are in simulation units: simulation bases (sb) and simulation time (st).

Table 3.5: Stochastic simulation results.

Model	Parameter Initiation	Parameter Fork dynamics	Observed Growth rate: k_G	Observed Log slope/difference	Predicted Log slope/difference
Terminus 1	$k_{\text{init}} = 1\ \text{st}^{-1}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$2.6 \times 10^{-1}\ \text{st}^{-1}$	$\alpha = 9.0 \times 10^{-3}\ \text{sb}^{-1}$	$\alpha = 9.0 \times 10^{-3}\ \text{sb}^{-1}$
Terminus 2	$k_{\text{init}} = 1\ \text{st}^{-1}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$2.7 \times 10^{-1}\ \text{st}^{-1}$	$\alpha_0 = 1.3 \times 10^{-2}\ \text{sb}^{-1}$	$\alpha_0 = 1.3 \times 10^{-2}\ \text{sb}^{-1}$
		$v_1 = 30\ \text{sb}\,\text{st}^{-1}$		$\alpha_1 = 8.8 \times 10^{-3}\ \text{sb}^{-1}$	$\alpha_1 = 8.8 \times 10^{-3}\ \text{sb}^{-1}$
Terminus 3	$k_{\text{init}} = 1\ \text{st}^{-1}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$2.6 \times 10^{-1}\ \text{st}^{-1}$	$\alpha_0 = 1.3 \times 10^{-2}\ \text{sb}^{-1}$	$\alpha_0 = 1.3 \times 10^{-2}\ \text{sb}^{-1}$
		$v_1 = 30\ \text{sb}\,\text{st}^{-1}$		$\alpha_1 = 8.8 \times 10^{-3}\ \text{sb}^{-1}$	$\alpha_1 = 8.8 \times 10^{-3}\ \text{sb}^{-1}$
		$k_3 = 2\ \text{st}^{-1}$		$\Delta = 1.3 \times 10^{-1}$	$\Delta = 1.3 \times 10^{-1}$
Terminus 4	$k_{\text{init}} = 0.3\ \text{st}^{-1}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$1.4 \times 10^{-1}\ \text{st}^{-1}$	$\alpha_0 = 7.3 \times 10^{-3}\ \text{sb}^{-1}$	$\alpha_0 = 7.3 \times 10^{-3}\ \text{sb}^{-1}$
Origin	$k_{\text{init}} = 0.3\ \text{st}^{-1}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$3.0 \times 10^{-1}\ \text{st}^{-1}$	$\alpha = 1.5 \times 10^{-2}\ \text{sb}^{-1}$	$\alpha = 1.5 \times 10^{-2}\ \text{sb}^{-1}$
Uniform	$T \sim U([0, 6])\ \text{st}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$2.7 \times 10^{-1}\ \text{st}^{-1}$	$\alpha = 1.3 \times 10^{-2}\ \text{sb}^{-1}$	$\alpha = 1.3 \times 10^{-2}\ \text{sb}^{-1}$
Precise	$T = 0.2\ \text{st}$	$v_0 = 20\ \text{sb}\,\text{st}^{-1}$	$1.4 \times 10^{-1}\ \text{st}^{-1}$	$\alpha = 6.8 \times 10^{-3}\ \text{sb}^{-1}$	$\alpha = 6.8 \times 10^{-3}\ \text{sb}^{-1}$

3.9.5 Re-scaling simulation units to compare to measured data

The simulations were performed in simulation units. We call the coarse-grained bases simulation bases (sb) and time units simulation time (st). To compare these simulations to experimental data, the prediction need to be rescaled to place the observations in biological units. To model a 5 Mb bacterial genome, with typical fork velocity of 1 kb s^{-1}, the conversion

factors are:

$$\frac{5 \times 10^6 \text{ bp}}{200 \text{ sb}} = 2.5 \times 10^4 \text{ bp sb}^{-1}, \tag{3.67}$$

$$\left(\frac{1 \text{ s}}{10^3 \text{ bp}}\right)\left(\frac{20 \text{ sb}}{1 \text{ st}}\right)\left(\frac{5 \times 10^6 \text{ bp}}{200 \text{ sb}}\right) = 5 \times 10^2 \text{ s st}^{-1}, \tag{3.68}$$

since we simulated typical fork velocities of 20 sb st^{-1} which we rescale the units to be equivalent to 1 kb s^{-1}.

Note that due to the smaller number of sb relative to bp, the stochasticity of fork dynamics should be exaggerated in the simulations, strengthening our argument that the stochasticity does not effect the model predictions.

3.9.6 Movies of marker frequency dynamics approaching steady state growth

In the paper, we include two examples of movies tracking the marker frequency dynamics from a single cell to steady state growth. To see the movies, see the Supplementary Material of Ref. [1]. Supplementary Movies 1 and 2 show the dynamics of the Terminus 1 and Terminus 3 models respectively. Movie time is shown in simulation units. The dynamics emphasizes the importance of performing the calculations of the marker frequency in the steady-state limit. Not taking this limit correctly leads to anomalous results (e.g., [60]).

3.10 Supplementary figures and data tables

3.10.1 Comparison with GC content

One potential rate-limiting factor for replication is the distribution of GC content across the genome. However, at the current genomic resolution of 100 kb, we find no significant evidence of GC content determining the rate of replication. See Supplementary Fig. 3.9 for the results of wild-type *E. coli* strain MG1655 in LB and M9.

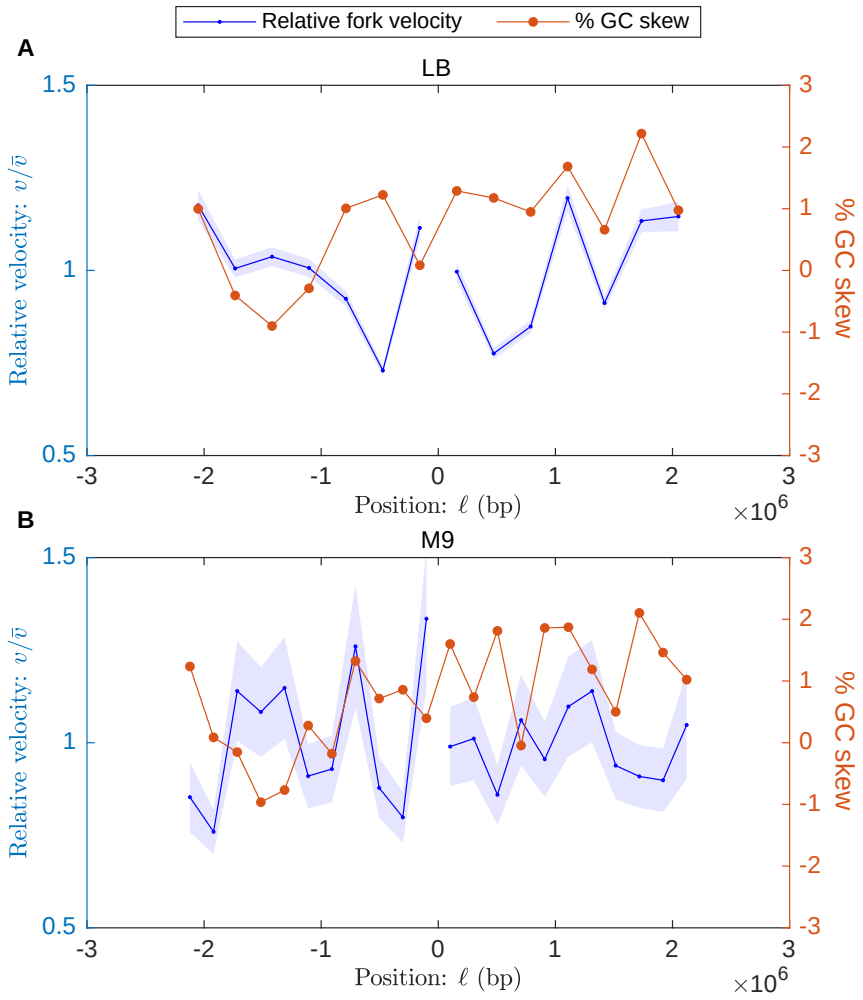


Figure 3.9: **Relative fork velocity and % GC skew as a function of position.** Results for wild-type *E. coli* strain MG1655. The blue curve represents the fork velocity divided by the mean fork velocity, with shaded error bars. The orange curve represents the % GC skew. Data are presented as mean values ± standard error of the mean (SEM). **Panel A:** Growth in LB medium. **Panel B:** Growth in M9 glucose minimal medium.

3.10.2 *V. cholerae* WT on LB

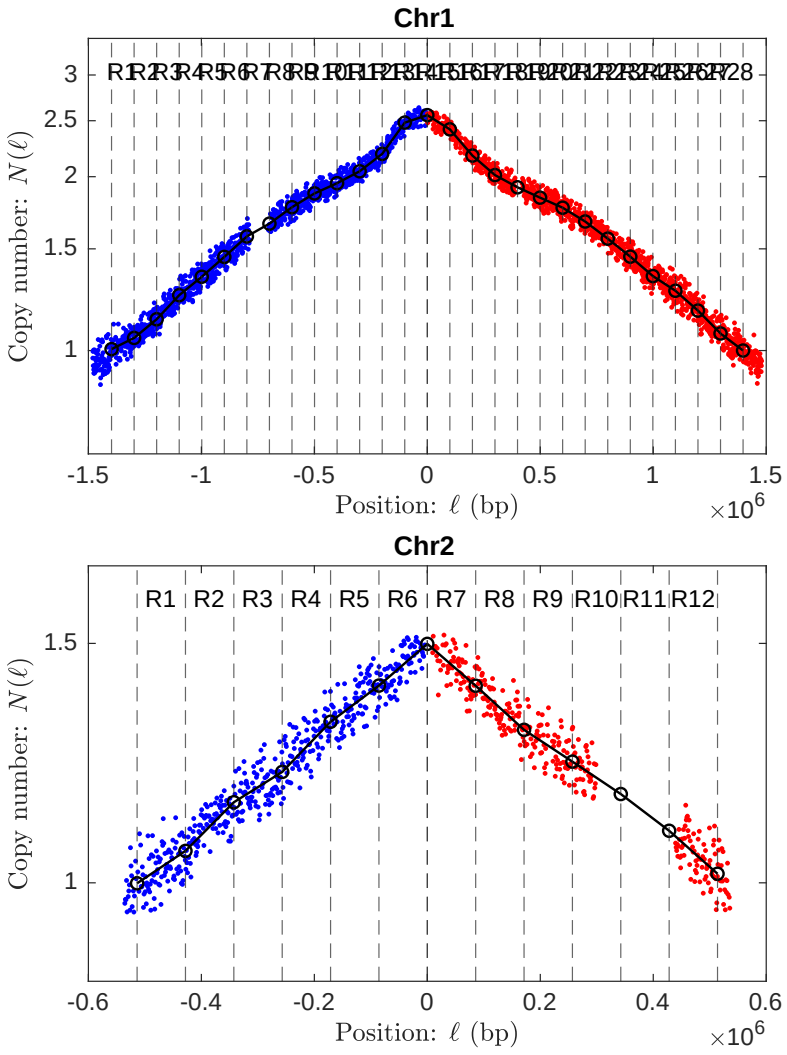


Figure 3.10: **Marker frequency data for *V. cholerae* WT on LB.** Regions chosen by AIC model selection are denoted by vertical dashed lines. Piecewise-linear fits are denoted by black lines, connected by black circles representing the control-point parameters. 1 kb-binned marker frequency data are denoted by blue (left arm) and red (right arm) dots. Detailed results tabulated in the `Supplementary Data.xlsx` file of Ref. [1].

3.10.3 *V. cholerae* WT on M9 fructose

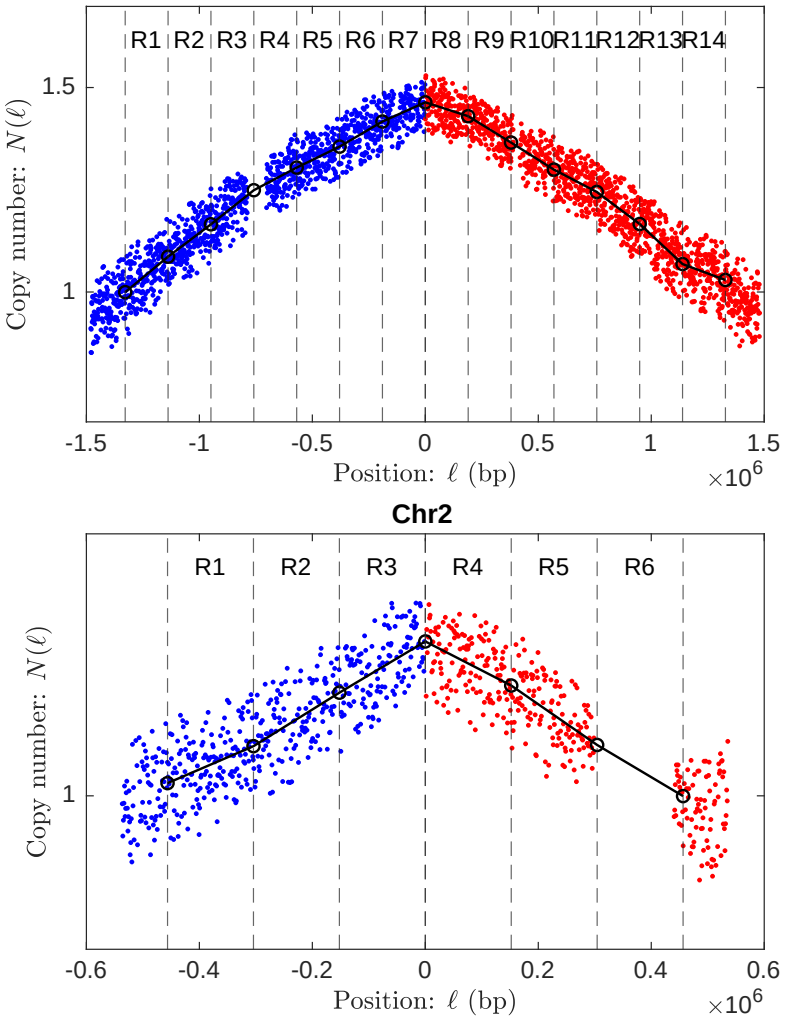


Figure 3.11: **Marker frequency data for *V. cholerae* WT on M9 fructose.** Regions chosen by AIC model selection are denoted by vertical dashed lines. Piecewise-linear fits are denoted by black lines, connected by black circles representing the control-point parameters. 1 kb-binned marker frequency data are denoted by blue (left arm) and red (right arm) dots. Detailed results tabulated in the `Supplementary Data.xlsx` file of Ref. [1].

3.10.4 *V. cholerae* MCH1 on LB

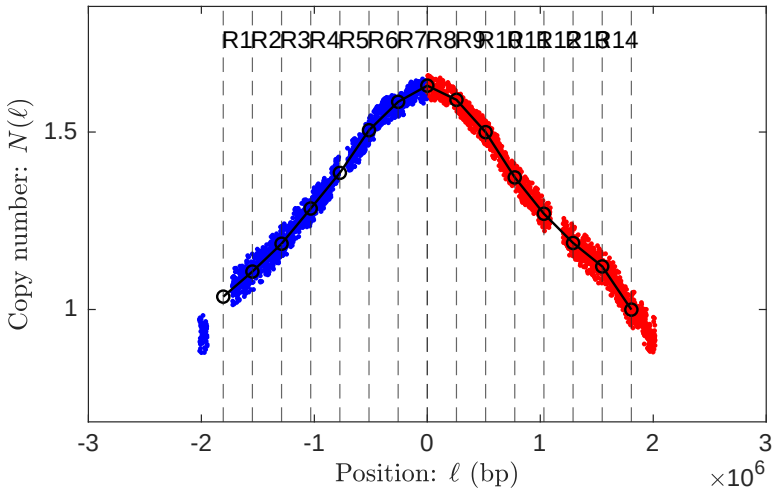


Figure 3.12: **Marker frequency data for *V. cholerae* MCH1 on LB.** Regions chosen by AIC model selection are denoted by vertical dashed lines. Piecewise-linear fits are denoted by black lines, connected by black circles representing the control-point parameters. 1 kb-binned marker frequency data are denoted by blue (left arm) and red (right arm) dots. Detailed results tabulated in the `Supplementary Data.xlsx` file of Ref. [1].

3.10.5 *V. cholerae* MCH1 on M9 fructose

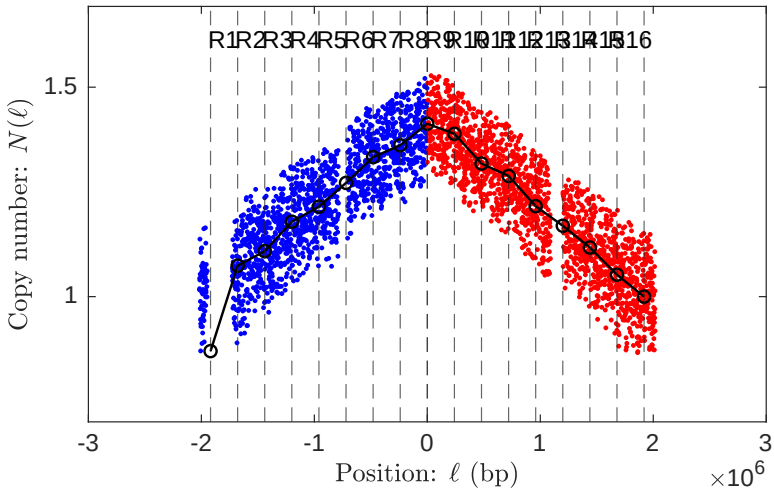
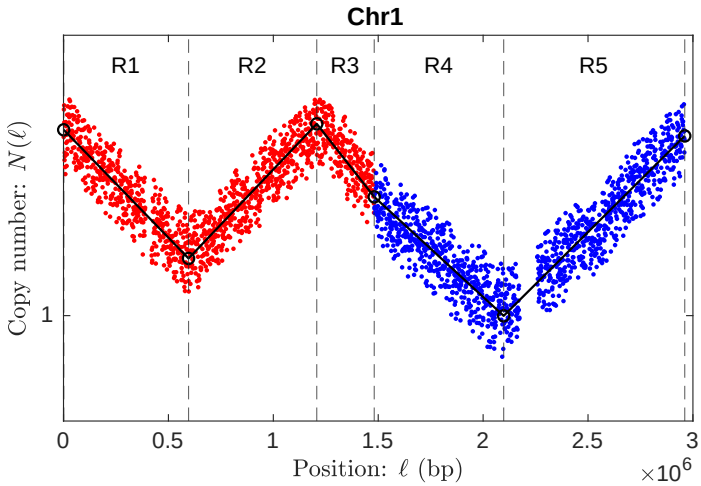


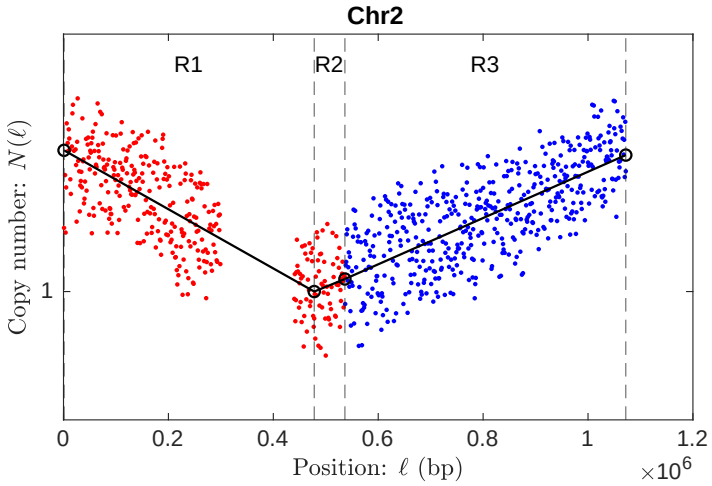
Figure 3.13: **Marker frequency data for *V. cholerae* MCH1 on M9 fructose.** Regions chosen by AIC model selection are denoted by vertical dashed lines. Piecewise-linear fits are denoted by black lines, connected by black circles representing the control-point parameters. 1 kb-binned marker frequency data are denoted by blue (left arm) and red (right arm) dots. Detailed results tabulated in the `Supplementary Data.xlsx` file of Ref. [1].

3.10.6 *V. cholerae oriR4* on M9 fructose

Figure 3.14: **Marker frequency data and tabulated results for *V. cholerae oriR4* on M9 fructose.**



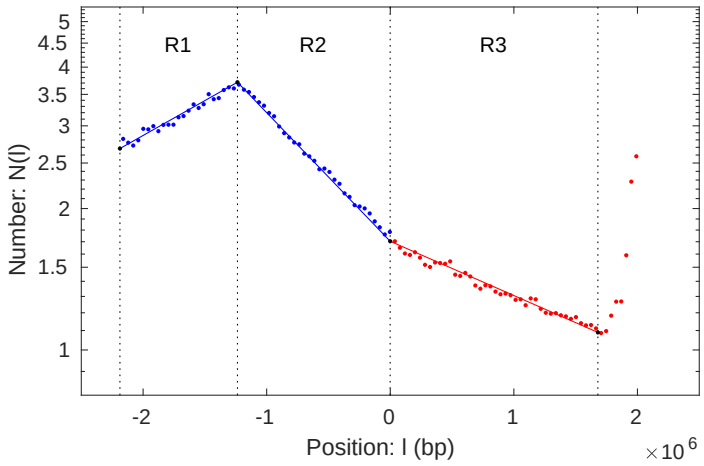
Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v (kb s^{-1})	Log difference: Δ	Duration: $\Delta\tau$ (min)	Percent error
R1	0	0.596	+	A	0.308	0.750	0.183	13.2	±2.0%
R2	0.596	1.21	-	R	0.314	0.736	0.192	13.8	±1.7%
R3	1.21	1.48	+	A	0.380	0.608	0.104	7.51	±3.3%
R4	1.48	2.10	+	R	0.275	0.840	0.170	12.2	±1.9%
R5	2.10	2.96	-	A	0.297	0.777	0.256	18.5	±1.3%



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v (kb s^{-1})	Log difference: Δ	Duration: $\Delta\tau$ (min)	Percent error
R1	0	0.478	+	A	0.257	0.898	0.123	8.85	±5.0%
R2	0.478	0.537	-	R	0.186	1.24	0.0109	0.787	±52%
R3	0.537	1.07	-	A	0.201	1.15	0.108	7.76	±4.4%

3.10.7 *B. subtilis oriC*-257° on S7 fumarate

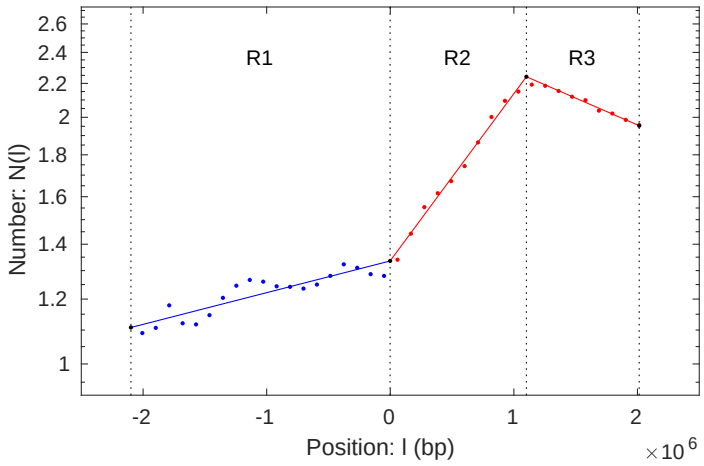
Figure 3.15: **Digitized marker frequency data and tabulated results for *B. subtilis oriC*-257° on S7 fumarate.** Regions chosen by ectopic and WT origin position are denoted by vertical dashed lines. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on WT origin location.



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-2.19	-1.24	-	A	$0.342 \pm 4.3\%$	NA	$0.325 \pm 4.3\%$	NA
R2	-1.24	-8.72×10^{-4}	+	R	$0.631 \pm 1.3\%$	NA	$0.779 \pm 1.3\%$	NA
R3	-8.72×10^{-4}	1.68	+	R	$0.266 \pm 2.6\%$	NA	$0.447 \pm 2.6\%$	NA

3.10.8 *B. subtilis oriC*-94° on S7 fumarate

Figure 3.16: **Digitized marker frequency data and tabulated results for *B. subtilis oriC*-94° on S7 fumarate.** Regions chosen by ectopic and WT origin position are denoted by vertical dashed lines. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on WT origin location.

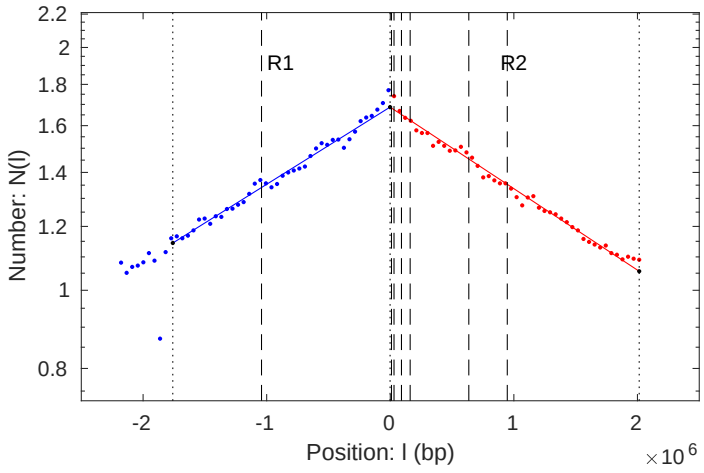


Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-2.10	-1.44×10^{-3}	-	A	$8.91 \times 10^{-2} \pm 10\%$	NA	$0.187 \pm 10\%$	NA
R2	-1.44×10^{-3}	1.10	-	R	$0.469 \pm 3.5\%$	NA	$0.518 \pm 3.5\%$	NA
R3	1.10	2.01	+	A	$0.151 \pm 18\%$	NA	$0.138 \pm 18\%$	NA

3.10.9 *B. subtilis oriN* on S7 fumarate

3.10.9.1 Two-slopes model

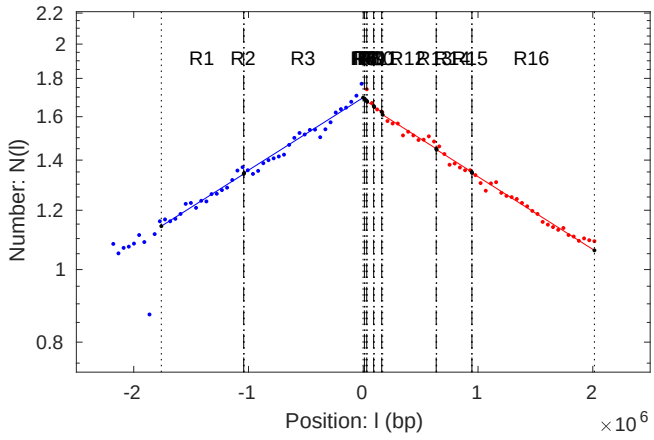
Figure 3.17: **Digitized marker frequency data and tabulated results for *B. subtilis oriN* on S7 fumarate, fit with two-slope model.** Just two regions, based on position relative to the origin. Only fit with two slopes. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on origin location.



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-1.76	-1.59×10^{-3}	-	A	$0.221 \pm 2.3\%$	$0.948 \pm 2.3\%$	$0.388 \pm 2.3\%$	$30.9 \pm 2.3\%$
R2	-1.59×10^{-3}	2.01	+	A	$0.232 \pm 1.8\%$	$0.900 \pm 1.8\%$	$0.469 \pm 1.8\%$	$37.3 \pm 1.8\%$

3.10.9.2 Pause model

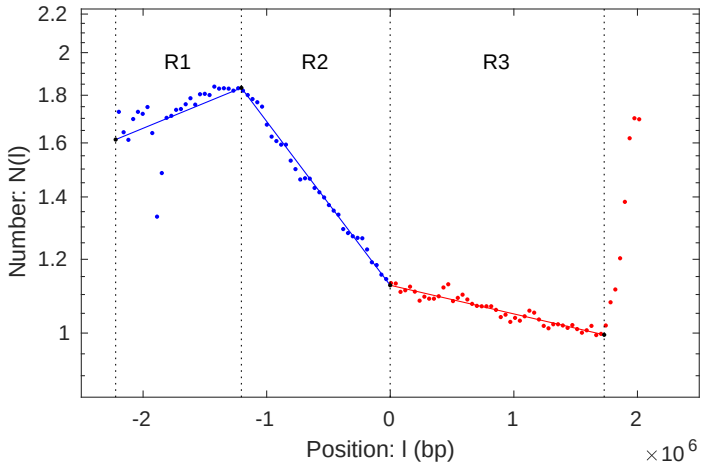
Figure 3.18: **Digitized marker frequency data and tabulated results for *B. subtilis oriN* on S7 fumarate, fit with pause model.** Regions based on rDNA position. Slopes fit with the pause model. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on origin location.



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-1.76	-1.04	-	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$0.160 \pm 1.9\%$	$12.8 \pm 1.9\%$
R2	-1.04	-1.04	-	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$3.13 \times 10^{-3} \pm 21\%$	$0.250 \pm 21\%$
R3	-1.04	-1.59×10^{-3}	-	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$0.232 \pm 1.9\%$	$18.4 \pm 1.9\%$
R4	-1.59×10^{-3}	1.13×10^{-2}	+	R	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$2.88 \times 10^{-3} \pm 1.9\%$	$0.229 \pm 1.9\%$
R5	1.13×10^{-2}	1.45×10^{-2}	+	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$3.13 \times 10^{-3} \pm 21\%$	$0.250 \pm 21\%$
R6	1.45×10^{-2}	3.06×10^{-2}	+	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$3.59 \times 10^{-3} \pm 1.9\%$	$0.286 \pm 1.9\%$
R7	3.06×10^{-2}	3.38×10^{-2}	+	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$3.13 \times 10^{-3} \pm 21\%$	$0.250 \pm 21\%$
R8	3.38×10^{-2}	9.18×10^{-2}	+	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$1.29 \times 10^{-2} \pm 1.9\%$	$1.03 \pm 1.9\%$
R9	9.18×10^{-2}	9.50×10^{-2}	+	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$3.13 \times 10^{-3} \pm 21\%$	$0.250 \pm 21\%$
R10	9.50×10^{-2}	0.159	+	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$1.44 \times 10^{-2} \pm 1.9\%$	$1.15 \pm 1.9\%$
R11	0.159	0.169	+	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$9.40 \times 10^{-3} \pm 21\%$	$0.749 \pm 21\%$
R12	0.169	0.636	+	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$0.104 \pm 1.9\%$	$8.30 \pm 1.9\%$
R13	0.636	0.639	+	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$3.13 \times 10^{-3} \pm 21\%$	$0.250 \pm 21\%$
R14	0.639	0.945	+	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$6.83 \times 10^{-2} \pm 1.9\%$	$5.44 \pm 1.9\%$
R15	0.945	0.948	+	A	$0.973 \pm 21\%$	$0.215 \pm 21\%$	$3.13 \times 10^{-3} \pm 21\%$	$0.250 \pm 21\%$
R16	0.948	2.01	+	A	$0.223 \pm 1.9\%$	$0.937 \pm 1.9\%$	$0.238 \pm 1.9\%$	$19.0 \pm 1.9\%$

3.10.10 *B. subtilis oriN*-257° on S7 fumarate

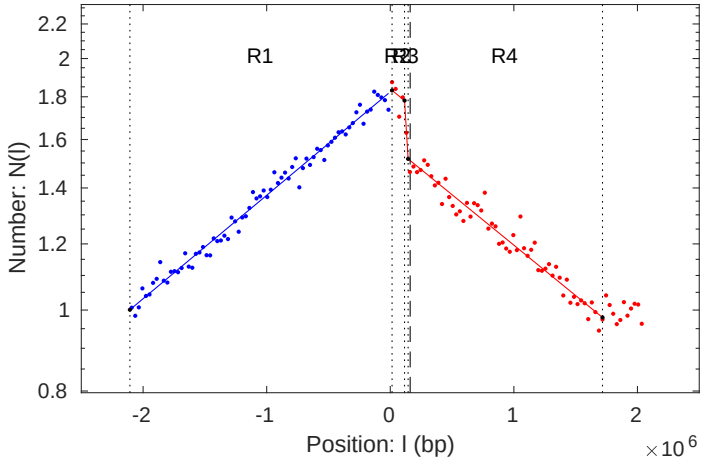
Figure 3.19: **Digitized marker frequency data and tabulated results for *B. subtilis oriN*-257° on S7 fumarate.** Regions chosen by ectopic and WT origin position are denoted by vertical dashed lines. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on WT origin location.



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-2.22	-1.20	-	A	$0.126 \pm 18\%$	NA	$0.128 \pm 18\%$	NA
R2	-1.20	-9.18×10^{-4}	+	R	$0.405 \pm 3.5\%$	NA	$0.488 \pm 3.5\%$	NA
R3	-9.18×10^{-4}	1.73	+	A	$7.08 \times 10^{-2} \pm 16\%$	NA	$0.123 \pm 16\%$	NA

3.10.11 *B. subtilis rrnIHG* inversion on MOPS glucose w/ Casamino Acids

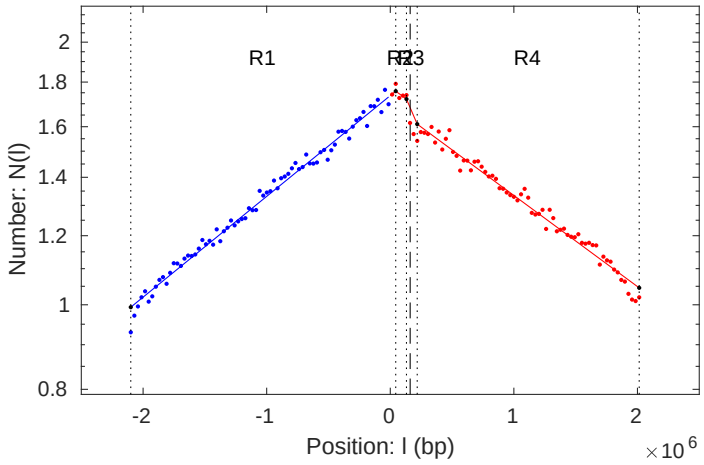
Figure 3.20: **Digitized marker frequency data and tabulated results for *B. subtilis rrnIHG* inversion on MOPS glucose w/ Casamino Acids.** Regions chosen by ectopic and WT origin position are denoted by vertical dashed lines. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on WT origin location.



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-2.13	-6.09×10^{-3}	-	A	$0.285 \pm 2.3\%$	$0.921 \pm 2.3\%$	$0.605 \pm 2.3\%$	$38.4 \pm 2.3\%$
R2	-6.09×10^{-3}	9.49×10^{-2}	+	A	$0.278 \pm 4.0\%$	$0.945 \pm 4.0\%$	$2.81 \times 10^{-2} \pm 4.0\%$	$1.78 \pm 4.0\%$
R3	9.49×10^{-2}	0.124	+	A	$5.57 \pm 8.4\%$	$4.71 \times 10^{-2} \pm 8.4\%$	$0.161 \pm 8.4\%$	$10.2 \pm 8.4\%$
R4	0.124	1.70	+	A	$0.278 \pm 4.0\%$	$0.945 \pm 4.0\%$	$0.437 \pm 4.0\%$	$27.7 \pm 4.0\%$

3.10.12 *B. subtilis rrnIHG* inversion on MOPS glucose – Minimal

Figure 3.21: **Digitized marker frequency data and tabulated results for *B. subtilis rrnIHG* inversion on MOPS glucose – Minimal.** Regions chosen by ectopic and WT origin position are denoted by vertical dashed lines. Digitized marker frequency data are denoted by blue (left arm) and red (right arm) dots, colored based on WT origin location.



Region name	Region start position: ℓ_- (Mb)	Region end position: ℓ_+ (Mb)	Fork direction	Replication orientation	Slope: α (Mb^{-1})	Velocity: v ($\mathrm{kb\,s}^{-1}$)	Log difference: Δ	Duration: $\Delta\tau$ (min)
R1	-2.10	4.41×10^{-2}	-	A	$0.266 \pm 1.7\%$	$0.869 \pm 1.7\%$	$0.570 \pm 1.7\%$	$41.1 \pm 1.7\%$
R2	4.41×10^{-2}	0.131	+	A	$0.241 \pm 2.5\%$	$0.961 \pm 2.5\%$	$2.09 \times 10^{-2} \pm 2.5\%$	$1.51 \pm 2.5\%$
R3	0.131	0.218	+	A	$0.761 \pm 13\%$	$0.304 \pm 13\%$	$6.61 \times 10^{-2} \pm 13\%$	$4.77 \pm 13\%$
R4	0.218	2.01	+	A	$0.241 \pm 2.5\%$	$0.961 \pm 2.5\%$	$0.432 \pm 2.5\%$	$31.2 \pm 2.5\%$

3.10.13 *E. coli* WT on LB

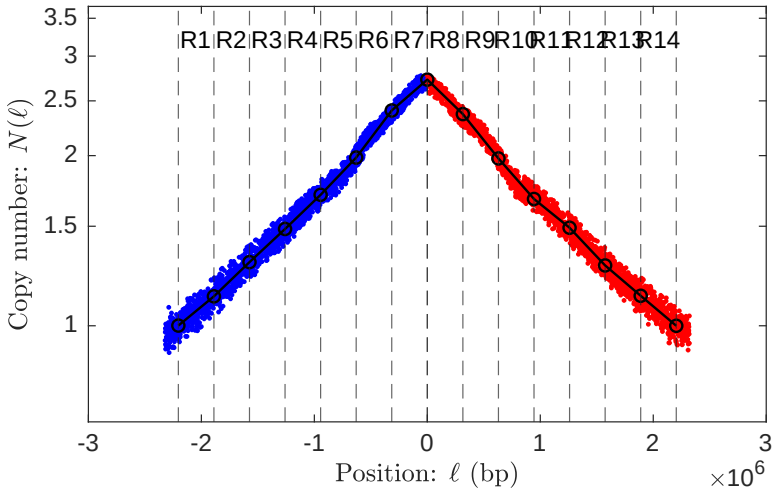


Figure 3.22: **Marker frequency data for *E. coli* WT on LB.** Regions chosen by AIC model selection are denoted by vertical dashed lines. Piecewise-linear fits are denoted by black lines, connected by black circles representing the control-point parameters. 1 kb-binned marker frequency data are denoted by blue (left arm) and red (right arm) dots. Detailed results tabulated in the `Supplementary Data.xlsx` file of Ref. [1].

3.10.14 *E. coli* WT on M9 glucose

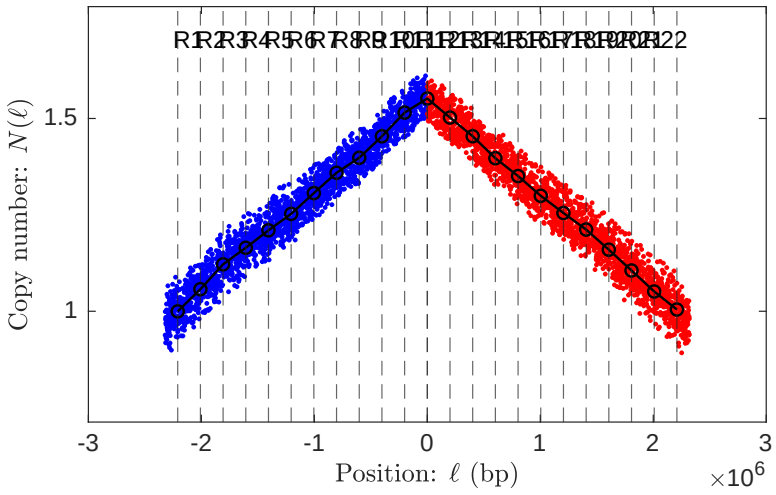


Figure 3.23: **Marker frequency data for *V. cholerae* WT on M9 glucose.** Regions chosen by AIC model selection are denoted by vertical dashed lines. Piecewise-linear fits are denoted by black lines, connected by black circles representing the control-point parameters. 1 kb-binned marker frequency data are denoted by blue (left arm) and red (right arm) dots. Detailed results tabulated in the `Supplementary Data.xlsx` file of Ref. [1].

3.10.15 Flattening of marker frequency profile of *V. cholerae* MCH1 in LB

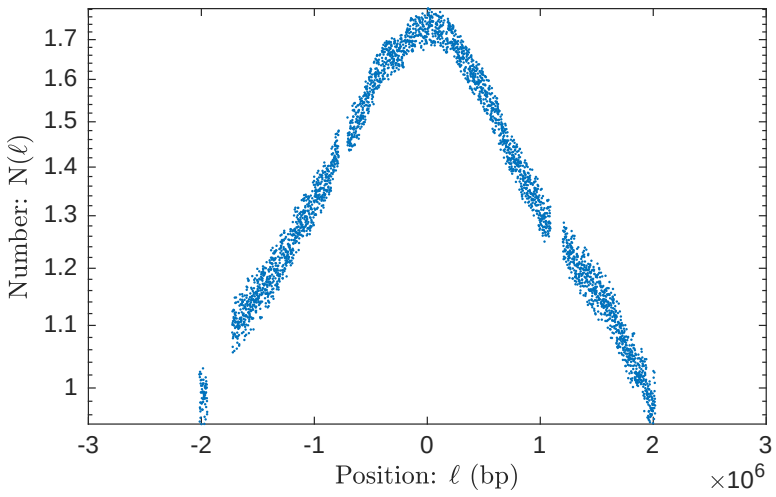


Figure 3.24: **Flattening of marker frequency profile of *V. cholerae* MCH1 in LB.** MCH1 in LB rich media. Flattening of the profile near the origin of replication suggests a slowdown in the rate of replication initiation, consistent with the population entering stationary phase.

3.11 Supplementary discussion

In this work, we have formalized the lag-time analysis approach. Although the approach has been understood at a conceptual level since the pioneering work of Cooper and Helmstetter [20], our recent exploration of stochastic models and the introduction of the exponential mean have clarified its interpretation [14] and, from an experimental perspective, the introduction of next-generation sequencing greatly expanded the potential of the approach for characterizing the dynamics of nucleic acids and in particular replication dynamics, where it has the potential to make precise measurements of replication timing. In multiple applications, we have used this approach to quantitatively measure time durations as short as seconds, a time resolution that is challenging, if not impossible, to achieve using other methods.

To appreciate the power of our approach, it is useful to compare our results to recent results of Nieduszynski and coworkers who have used experimental methods, cell synchronization (sync-seq), to directly resolve replication dynamics [37, 38]. In this case, the time resolution is limited by a combination of the precision of cell synchronization, which is an imperfect tool [61], and the frequency of fraction collection (every 5 minutes). Although it would be interesting to compare the relative precision of our approach to this competing method, the authors do not report fork velocities, pause durations, or provide an error analysis of their reported replication times, questioning to what extent the approach is truly quantitative. Since the fractions are collected on five-minute intervals, this time resolution is the floor of the direct time resolution achieved by this approach. In contrast, we report on a range of pause durations that are shorter than 5 minutes.

A significant experimental shortcoming of the sync-seq approach is the necessity of cell synchronization. In most systems, synchronization requires cell-cycle arrest, which introduces a significant potential for artifactual results [15]; whereas lag-time analysis probes dynamics in steady-state growth. Our own preliminary analysis suggests that the timing of initiation at a subset of loci in *Saccharomyces cerevisiae* is changed by the cell synchronization procedure relative to steady-state growth.

In addition to these quantitative and high-time-resolution applications, we have also demonstrated the approach in a more conceptual context: using lag-time analysis to argue that the observed oscillations in fork velocity were temporal rather than locus dependent.

It may seem perplexing that we have not pooled many existing datasets from multiple independent experiments. This would naively increase the statistical resolution and sensitivity from an analytical perspective. However, it is important to emphasize that not all marker-frequency experiments are of equal quality and that many datasets we have analyzed have clear signatures of systematic error. (See the Supplementary Methods Sec. 3.8.3.2.) A signature of systematic error that appears in many datasets is significant flattening of the marker frequency in the vicinity of the origin which appears intermittently. This feature is consistent with a culture that has begun to reduce the rate of initiation and

suggests that early harvesting and rapid cell processing may be essential for the generation of optimal datasets. In our analysis, we have prioritized the selection of artifact-free datasets over the indiscriminate pooling of data.

We emphasize that to date, datasets have not been generated with quantitative replication dynamics analysis as a goal and we are confident that experimental protocols can be optimized to improve the data. Ref. [39] describes a promising approach, including harvesting populations earlier in exponential phase. We too are developing new protocols to increase data quality.

3.12 Bibliography

[1] 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," *Nat Commun*, vol. 14, no. 1, p. 1762, Mar. 2023. DOI: 10.1038/s41467-023-37456-2.

[2] R. Phillips, J. Kondev, J. Theriot, and N. Orme, *Physical Biology of the Cell*. Garland Science, 2013, ISBN: 9780815344506.

[3] I. Tinoco Jr and R. L. Gonzalez Jr, "Biological mechanisms, one molecule at a time," *Genes Dev*, vol. 25, no. 12, pp. 1205–31, Jun. 2011. DOI: 10.1101/gad.2050011.

[4] M. Elías-Arnanz and M. Salas, "Resolution of head-on collisions between the transcription machinery and bacteriophage phi29 DNA polymerase is dependent on RNA polymerase translocation," *EMBO J*, vol. 18, no. 20, pp. 5675–82, Oct. 1999. DOI: 10.1093/emboj/18.20.5675.

[5] A. M. Deshpande and C. S. Newlon, "DNA replication fork pause sites dependent on transcription," *Science*, vol. 272, no. 5264, pp. 1030–3, May 1996. DOI: 10.1126/science.272.5264.1030.

[6] B. Liu and B. M. Alberts, "Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex," *Science*, vol. 267, no. 5201, pp. 1131–7, Feb. 1995. DOI: 10.1126/science.7855590.

[7] E. P. C. Rocha, "The organization of the bacterial genome," *Annu Rev Genet*, vol. 42, pp. 211–33, 2008. DOI: 10.1146/annurev.genet.42.110807.091653.

[8] E. V. Mirkin and S. M. Mirkin, "Replication fork stalling at natural impediments," *Microbiol Mol Biol Rev*, vol. 71, no. 1, pp. 13–35, Mar. 2007. DOI: 10.1128/MMBR.00030-06.

[9] N. Y. Yao and M. O'Donnell, "Replisome structure and conformational dynamics underlie fork progression past obstacles," *Curr Opin Cell Biol*, vol. 21, no. 3, pp. 336–43, Jun. 2009. DOI: 10.1016/j.ceb.2009.02.008.

[10] R. T. Pomerantz and M. O'Donnell, "The replisome uses mRNA as a primer after colliding with RNA polymerase," *Nature*, vol. 456, no. 7223, pp. 762–6, Dec. 2008. DOI: 10.1038/nature07527.

[11] K. Adelman and J. T. Lis, "Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans," *Nature Reviews Genetics*, vol. 13, no. 10, pp. 720–731, 2012. DOI: 10.1038/nrg3293.

[12] R. J. Davenport, G. J. Wuite, R. Landick, and C. Bustamante, "Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase," *Science*, vol. 287, no. 5462, pp. 2497–500, Mar. 2000. DOI: 10.1126/science.287.5462.2497.

[13] 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.

[14] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014 420, 1 Jan. 2022. DOI: 10.1103/PhysRevE.105.014420.

[15] 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.

[16] 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.

[17] J. D. Wang and P. A. Levin, "Metabolism, cell growth and the bacterial cell cycle," *Nat Rev Microbiol*, vol. 7, no. 11, pp. 822–7, Nov. 2009. DOI: 10.1038/nrmicro2202.

[18] L. Willis and K. C. Huang, "Sizing up the bacterial cell cycle," *Nat Rev Microbiol*, vol. 15, no. 10, pp. 606–620, Oct. 2017. DOI: 10.1038/nrmicro.2017.79.

[19] R. Reyes-Lamothe and D. J. Sherratt, "The bacterial cell cycle, chromosome inheritance and cell growth," *Nat Rev Microbiol*, vol. 17, no. 8, pp. 467–478, Aug. 2019. DOI: 10.1038/s41579-019-0212-7.

[20] 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.

[21] 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.

[22] R. E. Bird, J. Louarn, J. Martuscelli, and L. Caro, "Origin and sequence of chromosome replication in *Escherichia coli*," *J Mol Biol*, vol. 70, no. 3, pp. 549–66, Oct. 1972. DOI: 10.1016/0022-2836(72)90559-1.

[23] R. H. Pritchard, M. G. Chandler, and J. Collins, "Independence of F replication and chromosome replication in *Escherichia coli*," *Mol Gen Genet*, vol. 138, no. 2, pp. 143–55, 1975. DOI: 10.1007/BF02428118.

[24] M.-E. Val, A. Soler-Bistué, M. J. Bland, and D. Mazel, "Management of multipartite genomes: The *Vibrio cholerae* model," *Curr Opin Microbiol*, vol. 22, pp. 120–6, Dec. 2014. DOI: 10.1016/j.mib.2014.10.003.

[25] P. Srivastava, R. A. Fekete, and D. K. Chattoraj, "Segregation of the replication terminus of the two *Vibrio cholerae* chromosomes," *J Bacteriol*, vol. 188, no. 3, pp. 1060–70, Feb. 2006. DOI: 10.1128/JB.188.3.1060-1070.2006.

[26] T. Rasmussen, R. B. Jensen, and O. Skovgaard, "The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle," *EMBO J*, vol. 26, no. 13, pp. 3124–31, Jul. 2007. DOI: 10.1038/sj.emboj.7601747.

[27] M.-E. Val *et al.*, "A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*," *Sci Adv*, vol. 2, no. 4, e1501914, Apr. 2016. DOI: 10.1126/sciadv.1501914.

[28] S. French, "Consequences of replication fork movement through transcription units *in vivo*," *Science*, vol. 258, no. 5086, pp. 1362–5, Nov. 1992. DOI: 10.1126/science.1455232.

[29] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference.* 2nd. Springer-Verlag New York, Inc., 1998.

[30] H. Akaike, "Information theory and an extension of the maximum likelihood principle.," in *2nd International Symposium of Information Theory.*, P. B. N. and E. Csaki, Eds., Akademiai Kiado, Budapest., 1973, pp. 267–281.

[31] D. Dutta, K. Shatalin, V. Epshtein, M. E. Gottesman, and E. Nudler, "Linking RNA polymerase backtracking to genome instability in *E. coli*," *Cell*, vol. 146, no. 4, pp. 533–43, Aug. 2011. DOI: 10.1016/j.cell.2011.07.034.

[32] A. Srivatsan, A. Tehranchi, D. M. MacAlpine, and J. D. Wang, "Co-orientation of replication and transcription preserves genome integrity," *PLoS Genet*, vol. 6, no. 1, e1000810, Jan. 2010. DOI: 10.1371/journal.pgen.1000810.

[33] J. D. Wang, M. B. Berkmen, and A. D. Grossman, "Genome-wide coorientation of replication and transcription reduces adverse effects on replication in *Bacillus subtilis*," *Proc Natl Acad Sci U S A*, vol. 104, no. 13, pp. 5608–13, Mar. 2007. DOI: 10.1073/pnas.0608999104.

[34] M.-E. Val, O. Skovgaard, M. Ducos-Galand, M. J. Bland, and D. Mazel, "Genome engineering in *Vibrio cholerae*: A feasible approach to address biological issues," *PLoS Genetics*, vol. 8, no. 1, 2012. DOI: 10.1371/journal.pgen.1002472.

[35] C. J. Rudolph, A. L. Upton, A. Stockum, C. A. Nieduszynski, and R. G. Lloyd, "Avoiding chromosome pathology when replication forks collide," *Nature*, vol. 500, no. 7464, pp. 608–11, Aug. 2013. DOI: 10.1038/nature12312.

[36] E. Galli *et al.*, "Replication termination without a replication fork trap," *Scientific Reports*, vol. 9, no. 1, 2019. DOI: 10.1038/s41598-019-43795-2.

[37] D. G. Batrakou, C. A. Müller, R. H. C. Wilson, and C. A. Nieduszynski, "DNA copy-number measurement of genome replication dynamics by high-throughput sequencing: The sort-seq, sync-seq and MFA-seq family," *Nat Protoc*, vol. 15, no. 3, pp. 1255–1284, Mar. 2020. DOI: 10.1038/s41596-019-0287-7.

[38] C. A. Müller *et al.*, "The dynamics of genome replication using deep sequencing," *Nucleic Acids Res*, vol. 42, no. 1, e3, Jan. 2014. DOI: 10.1093/nar/gkt878.

[39] A. Knöppel, O. Broström, K. Gras, D. Fange, and J. Elf, "The coordination of replication initiation with growth rate in *Escherichia coli*," *Preprint at bioRxiv*, 2022. DOI: 10.1101/2021.10.11.463968.

[40] 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.

[41] A. Zaritsky and R. H. Pritchard, "Changes in cell size and shape associated with changes in the replication time of the chromosome of *Escherichia coli*," *J Bacteriol*, vol. 114, no. 2, pp. 824–37, May 1973. DOI: 10.1128/jb.114.2.824-837.1973.

[42] I. Odsbu, Morigen, and K. Skarstad, "A reduction in ribonucleotide reductase activity slows down the chromosome replication fork but does not change its localization," *PLoS One*, vol. 4, no. 10, e7617, Oct. 2009. DOI: 10.1371/journal.pone.0007617.

[43] 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.

[44] 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.

[45] G. B. Arfken and H. J. Weber, *Mathematical methods for physicists; 4th ed.* San Diego, CA: Academic Press, 1995.

[46] 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.

[47] J. U. Dimude *et al.*, "Origins left, right, and centre: Increasing the number of initiation sites in the *Escherichia coli* chromosome," *Genes (Basel)*, vol. 9, no. 8, Jul. 2018. DOI: 10.3390/genes9080376.

[48] E. V. Mirkin and S. M. Mirkin, "Mechanisms of transcription-replication collisions in bacteria," *Mol Cell Biol*, vol. 25, no. 3, pp. 888–95, Feb. 2005. DOI: 10.1128/MCB.25.3.888-895.2005.

[49] E. V. Mirkin, D. Castro Roa, E. Nudler, and S. M. Mirkin, "Transcription regulatory elements are punctuation marks for DNA replication," *Proc Natl Acad Sci U S A*, vol. 103, no. 19, pp. 7276–81, May 2006. DOI: 10.1073/pnas.0601127103.

[50] K. S. Lang and H. Merrikh, "The clash of macromolecular titans: Replication-transcription conflicts in bacteria," *Annu Rev Microbiol*, vol. 72, pp. 71–88, Sep. 2018. DOI: 10.1146/annurev-micro-090817-062514.

[51] K. S. Lang *et al.*, "Replication-transcription conflicts generate r-loops that orchestrate bacterial stress survival and pathogenesis," *Cell*, vol. 170, no. 4, 787–799.e18, Aug. 2017. DOI: 10.1016/j.cell.2017.07.044.

[52] H. Merrikh, C. Machón, W. H. Grainger, A. D. Grossman, and P. Soultanas, "Co-directional replication-transcription conflicts lead to replication restart," *Nature*, vol. 470, no. 7335, pp. 554–7, Feb. 2011. DOI: 10.1038/nature09758.

[53] C. N. Merrikh, B. J. Brewer, and H. Merrikh, "The *B. subtilis* accessory helicase PcrA facilitates DNA replication through transcription units," *PLoS Genet*, vol. 11, no. 6, e1005289, Jun. 2015. DOI: 10.1371/journal.pgen.1005289.

[54] S. Million-Weaver *et al.*, "An underlying mechanism for the increased mutagenesis of lagging-strand genes in *Bacillus subtilis*," *Proc Natl Acad Sci U S A*, vol. 112, no. 10, E1096–105, Mar. 2015. DOI: 10.1073/pnas.1416651112.

[55] F. Si, G. Le Treut, J. T. Sauls, S. Vadia, P. A. Levin, and S. Jun, "Mechanistic origin of cell-size control and homeostasis in bacteria," *Curr Biol*, vol. 29, no. 11, 1760–1770.e7, Jun. 2019. DOI: 10.1016/j.cub.2019.04.062.

[56] 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. eprint: https://doi.org/10.1021/j100540a008.

[57] S. L. Midgley-Smith *et al.*, "Chromosomal over-replication in *Escherichia coli recG* cells is triggered by replication fork fusion and amplified if replichore symmetry is disturbed," *Nucleic Acids Research*, vol. 46, no. 15, pp. 7701–7715, 2018. DOI: 10.1093/nar/gky566.

[58] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012. DOI: 10.1038/nmeth.1923.

[59] P. Danecek *et al.*, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, no. 2, Feb. 2021, giab008, ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf.

[60] R. Retkute, M. Hawkins, C. J. Rudolph, and C. A. Nieduszynski, "Modeling of DNA replication in rapidly growing bacteria with one and two replication origins," *Preprint at bioRxiv*, 2018. DOI: 10.1101/354654. eprint: https://www.biorxiv.org/content/early/2018/06/29/354654.full.pdf.

[61] 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.

Chapter 4

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

Content for this chapter first appeared in: [1] T. W. Lo, H. K. J. Choi, D. Huang, and P. A. Wiggins, "Noise robustness and metabolic load determine the principles of central dogma regulation," *preprint*, Oct. 2023. DOI: 10.1101/2023.10.20.563172.

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

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 predicts novel principles of central dogma regulation: Optimal protein expression levels for many gene 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 show that each of these novel regulatory principles is observed. These results reveal that robustness and metabolic load determine the global regulatory principles that govern central dogma processes, and these principles have broad implications for cellular function.

4.1 Introduction

What rationale determines the transcription and translation level of a gene in the cell? Both experiment and theory support the idea that gene expression levels maximize cell fitness and that cells can rapidly adapt genetically to new environments [2, 3]. Although it is clear that fitness optimization is the rationale for protein expression levels, the consequences of this optimization on protein expression are still poorly understood. For instance, multiple approaches suggest that protein expression is much higher than one would predict based on protein activity [4–11]. How can these elevated protein levels be reconciled with fitness optimization? One possibility is that growth robustness may explain not only this putative protein overabundance but the relative levels of both transcription and translation. We explore this hypothesis in this chapter.

Achieving growth robustness is nontrivial since all processes at the cellular scale are stochastic, including gene expression [12]. This biological noise leads to significant cell-to-cell variation in protein numbers, even for essential proteins that are required for growth [13, 14]. How does the cell ensure robust expression of hundreds of distinct essential gene products required for cellular function? There are three qualitative strategies for achieving robustness: (i) increased protein expression levels, (ii) reducing the noise amplitude, and (iii) metabolic regulatory feedback control. The analysis in this chapter will demonstrate that the central dogma is regulated to implement both of the first two approaches, with strategy (i) leading to protein overabundance and strategy (ii) shaping the relation between transcription and translation. Approach (iii) is explored in Chapter 5.

To study the consequences of growth robustness on the central dogma quantitatively, we propose and analyze a minimal model: the Robustness-Load Trade-Off (RLTO) Model. The 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 (i.e., high multiplicity). Implementing this model required a key theoretical innovation: the analysis

of the consequences of noise on a highly-asymmetric fitness landscape. Our novel approach predicts new phenomenology absent from previous models (e.g., [15]).

In the RLTO model, growth rate maximization predicts protein overabundance generically, and vast overabundance for low-expression essential genes. Protein overabundance explains the paradox of protein expression levels being simultaneously optimal and in excess of what is required for function [10, 16–18]. The model predicts that there is a transcriptional floor of roughly one message transcribed per cell cycle for essential genes. We demonstrate that just such a lower threshold is observed in *Escherichia coli*, yeast, and human. The RLTO model also predicts a central dogma regulatory program that balances transcription and translation, which we call *load balancing*. Load balancing predicts the dependence of gene proteome fraction on message number, as well as the global regulatory response of cells to changes in the metabolic cost of transcription. Furthermore, load balancing predicts that noise should have a non-canonical scaling with protein abundance. We demonstrate that the predicted scaling is observed in yeast. Taken together, these results reveal that noise robustness and metabolic load fundamentally shape the function of the central dogma processes and that global function is quantitatively described by a few simple emergent principles.

4.2 Results

4.2.1 Defining the RLTO Model

We start by developing a minimal quantitative model for cell fitness (i.e., growth rate), which includes the stochasticity of gene expression, the dependence of growth on essential proteins, as well as the metabolic load of gene expression. We will assume the numbers of proteins for gene i, N_{pi}, are gamma-distributed independent random variables: $N_{pi} \sim \Gamma(a_i, \theta_i)$, where the distribution is described by two gene-specific statistical parameters: the scale parameter θ_i and the shape parameter a_i [14, 19, 20]. In what follows, we will drop the explicit gene subscript i for readability. These two statistical parameters have clear biological

interpretations in terms of the *Telegraph Model*, the kinetic model for the central dogma [20]. The scale parameter is proportional to the *translation efficiency* (ε), the mean number of proteins translated from each message transcribed for gene i: $\theta = \varepsilon \ln 2$, and the shape parameter (a) is proportional to the message number: $a = \mu_m / \ln 2$, where the *message number* μ_m is defined as the mean number of messages transcribed per cell cycle for gene i. See the Supplementary Material Sec. 4.6.1.1 for a detailed description of the model and the relationships between the model parameters.

In terms of these gene-specific central dogma parameters, the *protein number* (μ_p), defined as the mean protein number per cell at cell birth, and the coefficient of variation (CV_p^2) are:

$$\mu_p = \mu_m \varepsilon, \tag{4.1}$$

$$\mathrm{CV}_p^2 = \frac{\ln 2}{\mu_m}, \tag{4.2}$$

for gene i. The protein number can be understood as the product of the gains of a two-stage amplification process: transcription followed by translation. (See Fig. 4.1A.) However, under typical biological conditions, the noise is dominated by transcription and therefore is inversely proportional to the message number μ_m [20].

Next, we consider the metabolic load on the cell as a consequence of protein expression. In the absence of cell cycle arrest, we will assume that the cell cycle duration τ is proportional to the overall metabolic load of the messages, proteins, and other cellular components [21]. Focusing on the metabolic load of a particular gene, the inverse relative growth rate is:

$$\frac{k_0}{k} = 1 + \frac{\lambda + \varepsilon}{N_0} \mu_m, \tag{4.3}$$

where k_0 is the growth rate in the absence of the metabolic load of gene i, N_0 is the total metabolic load of all genes (in protein equivalents) and λ is the metabolic message cost. (See Supplementary Material Sec. 4.6.2 for a detailed derivation.) The λ-term represents the metabolic cost of transcription and the ε-term represents the metabolic cost of translation of gene i. Although the global parameters N_0 and λ provide an intuitive representation of the model, the relative growth rate depends on fewer parameters. We define the relative load

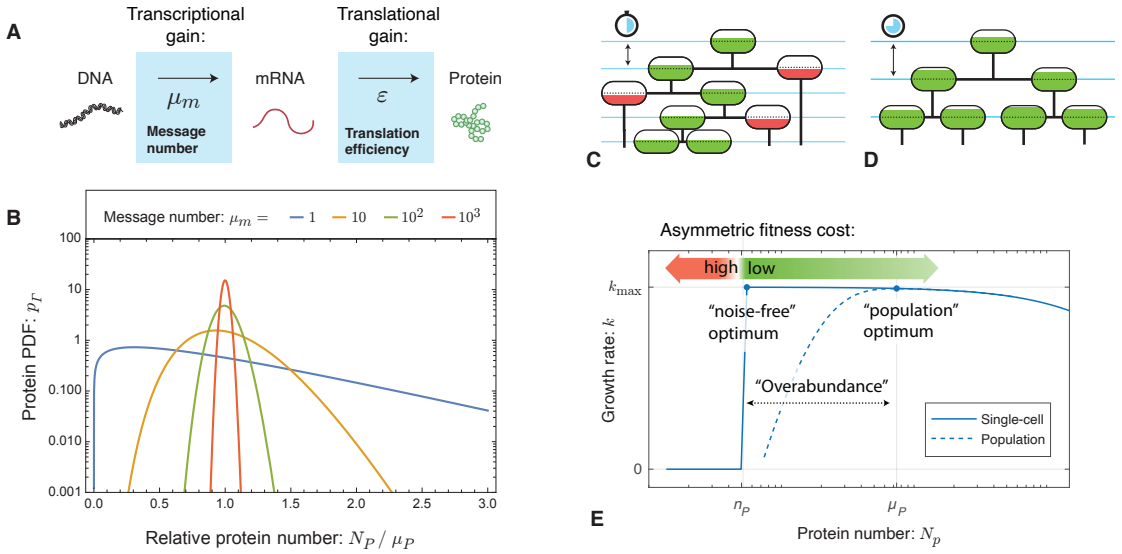


Figure 4.1: **Features of the RLTO model. Panel A: Two-stage amplifier model of expression.** The central dogma describes a two-step stochastic process by which genes are first transcribed and then translated. The transcription process generates an average of μ_m, the message number, of messages per cell cycle. Translation generates an average of ε, the translation efficiency, proteins per message transcribed. **Panel B: Gene expression noise.** The protein number is observed to be Gamma-distributed due to the stochasticity of the central dogma processes. The noise is determined by the message number μ_m. For highly-transcribed genes (e.g., $\mu_m = 10^3$), the distribution is tightly distributed about its mean; however, for low-expression genes (e.g., $\mu_m = 1$), the distribution is extremely wide, resulting in many cells with extremely low protein abundance. **Panel C & D: The RLTO Model.** A schematic cell lineage tree is shown during exponential growth. The cell fill represents an essential protein expression level. The dotted line represents the threshold number n_p required for cell growth. **Panel C:** Reducing the expression level reduces the metabolic load (the spacing between blue lines); however, below-threshold cells arrest (red). **Panel D:** Increasing protein expression increases the metabolic load (the spacing between blue lines); however, all cells are above threshold. **Panel E: Protein overabundance optimizes growth rate.** The solid blue line represents the growth rate as a function of protein number for a single cell. Below the threshold level n_p, there is no growth. Above the threshold, the growth rate decreases slowly as a consequence of the metabolic load. The dashed blue line represents the population growth rate as a function of mean protein level. The growth rate is optimized at μ_p, far above the threshold due to the cell-to-cell variation in protein number. The asymmetric fitness landscape causes the optimal protein expression level to be overabundant.

as $\Lambda \equiv \lambda/N_0$ as the ratio of the metabolic load of a single message to the total load and the relative translation efficiency $E \equiv \varepsilon/N_0$ as the ratio of the number of proteins translated per message to the total metabolic load N_0. In *E. coli*, we estimate that both Λ and E are roughly 10^{-5} and they are smaller still for eukaryotic cells. (See Supplementary Material Sec. 4.6.10.)

The final task is to link protein expression with cellular function. Motivated by the concept of rate-limiting reactants [22] as well as single-cell growth rate measurements [23], we will propose that the essential processes of the cell have a threshold-like dependence on each essential protein: We will assume that each essential protein i has a critical gene-specific threshold number n_p such that growth arrests below this critical number. (See Fig. 4.1CDE.) In our analysis, we will treat the thresholds n_p as gene-specific unknown parameters that could be determined experimentally. It will usually be more convenient to work in terms of the threshold fraction, $\phi_p \equiv n_p/N_0$, which can be interpreted as the threshold protein fraction required for growth.

We have previously analyzed a model in which the duration of cell-cycle phases (or periods) are of stochastic duration [24]. In particular, we have already considered the case of stochastic cell-cycle arrest, which can be computed analytically. Assuming the expression of each protein is independent, the protein number in subsequent cell cycles are uncorrelated [14] and the failure probability is small, the growth rate of the RLTO model can also be computed analytically. (See Supplementary Material Sec. 4.6.2.2.) The relative growth rate is:

$$\ln \tfrac{k}{k_0} = -(\Lambda + E)\mu_m - \tfrac{1}{\ln 2}\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{\phi_p}{E \ln 2}), \tag{4.4}$$

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. Eq. 4.4 represents an explicit analytic model for cell fitness that accounts for growth robustness to noise, metabolic load, and high multiplicity. In the RLTO model, the relative growth rate depends only on a single global parameter: the relative metabolic load Λ, and three gene-specific parameters: the threshold fraction ϕ_p, the message number μ_m, and the relative translation efficiency

E. We propose that the cell is regulated to optimize the message number and translation efficiency to maximize the growth rate. The model parameters are summarized in Tab. 4.2.1.

4.2.2 The fitness landscape of a trade-off

The fitness landscape predicted by the RLTO model for representative parameters is shown in Fig. 4.2. The figure displays a number of important model phenomena: There is no growth for a protein fraction Φ_p, defined as $\Phi_p \equiv \mu_p/N_0$, below the threshold value ϕ_p, and for high noise, Φ_p must be in significant excess of ϕ_p. Rapid growth can be achieved by the two mechanisms described in Fig. 4.2: (i) high expression levels (Φ_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 represented by the dotted white line. The optimal fitness corresponds to a balance between increasing the mean protein expression (Φ_p) and decreasing noise (CV_p^2). This optimal central dogma program strategy leads to significant overabundance. (See Fig. 4.2B.)

4.2.3 RLTO predicts protein overabundance

How does this optimal strategy depend on the threshold fraction (ϕ_p) and the relative load (Λ)? To understand the phenomenology of the model, we optimize the growth analytically. Although the threshold fraction ϕ_p is intuitive from a mechanistic perspective, it is less convenient from an experimental perspective. We therefore define the overabundance o, the protein number relative to the threshold:

$$o \equiv \Phi_p/\phi_p, \tag{4.5}$$

where Φ_p is the protein fraction and ϕ_p is the threshold fraction required for growth (e.g., [17]).

To maximize the growth rate, we set the partial derivatives of Eq. 4.4 with respect to message number and translation efficiency to zero and then solve the coupled transcendental

Table 4.1: **Summary of RLTO parameters.** A summary of model parameters is given for the RLTO model. The top three parameters are global. The middle and bottom sets of parameters are all gene specific (Gene i). The top and middle parameter sets are assumed to be fixed in the model; whereas the bottom set are optimized to maximize growth rate. Per gene i, there are two independent parameters (μ_m and ε) and one fixed parameter genes-specific parameter (n_p). Model parameters appearing with a hat are optimal values (e.g., optimal message number: $\hat{\mu}_m$).

Model parameter name:	**Symbol:**	**Units:**	**Global/ Specific:**	**Description:**
Total metabolic load	N_0	Protein molecules	global	Total metabolic load in protein molecule equivalents
Message cost	λ	Protein molecules	global	Metabolic cost per message transcribed per cell cycle
Relative load	$\Lambda \equiv \lambda/N_0$	Number	global	Message cost relative to total load
Threshold number	n_p	Protein molecules	gene i	Protein number required for function
Threshold fraction	$\phi_p \equiv n_p/N_0$	Number	gene i	Protein fraction required for function
Message number	μ_m	Number	gene i	Total number of messages transcribed per cell cycle
Translation efficiency	ε	Protein molecules	gene i	Number of protein made per message
Relative translation efficiency	$E \equiv \varepsilon/N_0$	Number	gene i	Number of protein made per message relative to total load
Protein number	$\mu_p \equiv \mu_m\varepsilon$	Protein molecules	gene i	Protein number
Protein fraction	$\Phi_p \equiv \mu_p/N_0$	Number	gene i	Protein number relative to total load
Overabundance	$o \equiv \mu_p/n_p$	Number	gene i	Protein number relative to threshold number

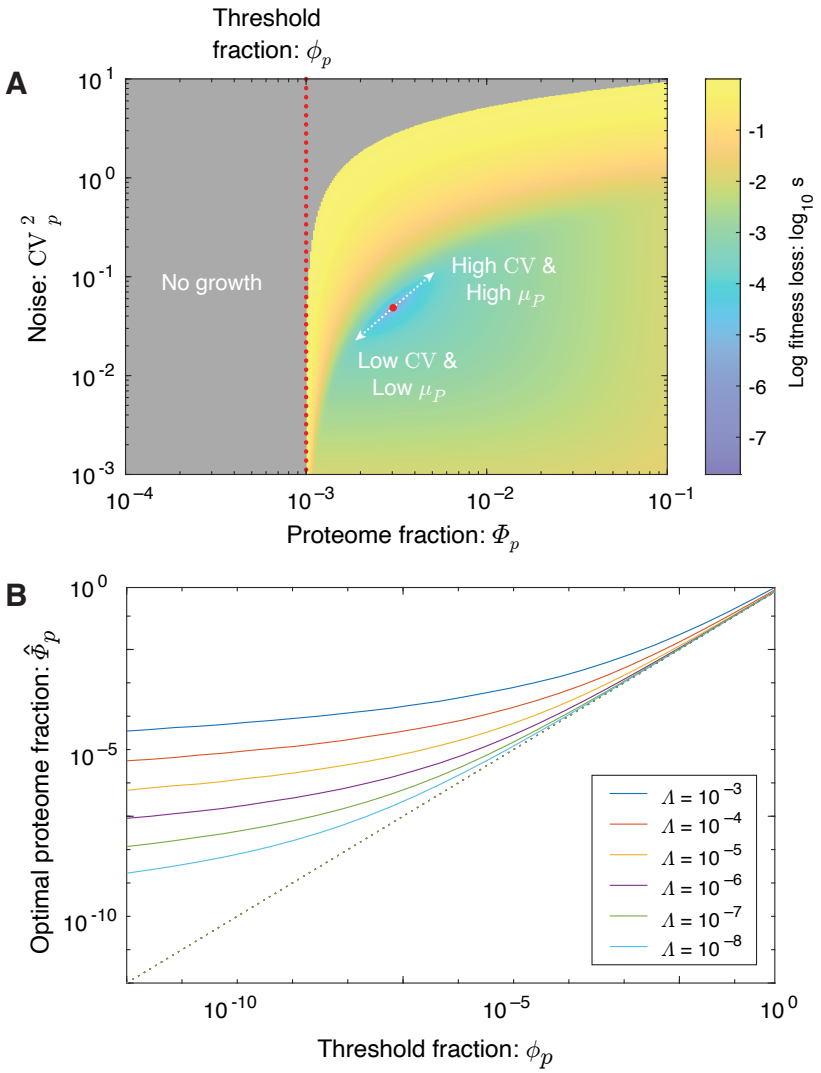


Figure 4.2: **RLTO model prediction of the fitness landscape. Panel A: Fitness landscape.** The fitness loss ($s \equiv \ln k_{\max}/k$) is shown as a function of noise (CV_p^2) and protein fraction. The red dotted curve shows the threshold ϕ_p, the red dot represents the optimum fitness, and the distance between these levels represents the overabundance ($o = \Phi_p/\phi_p$). Robustness to noise drives the optimum expression level significantly above the threshold ϕ_p. The low and higher noise strategies discussed above are shown with the white-dotted line. **Panel B: Optimal protein fraction.** The optimal protein fraction is shown as a function of the threshold fraction. To prevent growth arrest, the protein fraction is always larger than the threshold fraction, leading to overabundance. However, the overabundance grows rapidly as the threshold is reduced. The optimal protein fraction also depends on the relative load Λ. Higher relative load leads to a smaller protein fraction at fixed threshold.

equations for the optimal message number $\hat{\mu}_m$ and relative translation efficiency $\hat{E}$. As shown in Supplementary Material Sec. 4.6.3, the equation for the optimal message number is:

$$\Lambda \ln 2 = -\frac{\partial}{\partial \hat{\mu}_m} \gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o} \ln 2}\right). \tag{4.6}$$

The resulting equation depends on only a single parameter: the relative load Λ, and is independent of the optimal relative translation efficiency.

The predicted relation between the optimal message number ($\hat{\mu}_m$) and overabundance ($\hat{o}$) is shown in Fig. 4.3A. The RLTO model generically predicts that the optimal protein fraction is overabundant ($o > 1$). 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 does depend on the relative load (Λ), but its phenomenology is qualitatively unchanged over orders of magnitude variation in Λ.

4.2.4 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 [25] and bacterial gene expression is subject to a large-magnitude noise floor that increases the noise for high-expression genes [14]. The optimization of message number at fixed translation efficiency does result in a slightly modified optimization condition for the message number (Supplementary Material Sec. 4.6.4.1); however, the predicted overabundance is only subtly perturbed (Fig. 4.3A). 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. 4.3A.)

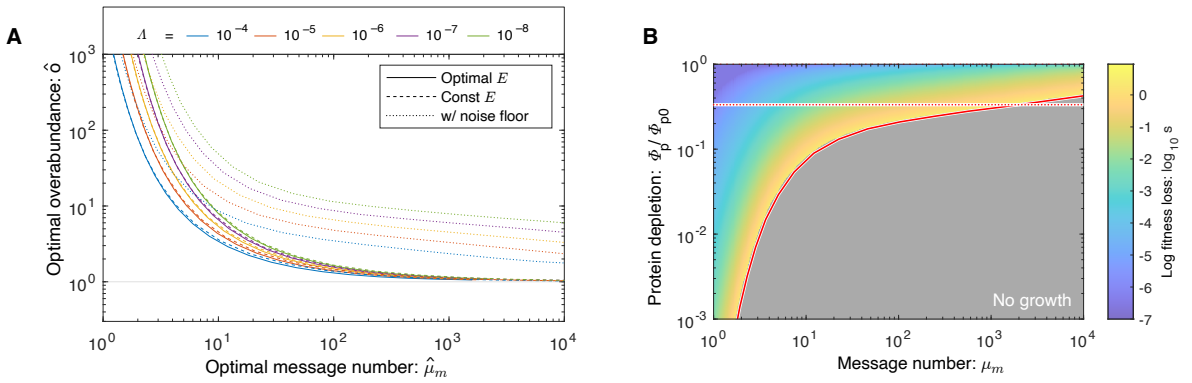


Figure 4.3: **RLTO prediction of overabundance. Panel A: Overabundance as a function of message number.** In the limit of high message number (μ_m), the noise is small, and the optimal expression level is close to the threshold ϕ_p, leading to minimal overabundance ($o \to 1$); however, as the message number approaches unity, the noise is comparable to the mean, driving vast overabundance ($o \gg 1$). Overabundance increases as the relative metabolic load decreases, and is essentially identical for constant (dashed) versus optimized translation efficiency (solid). To model overabundance in bacterial cells, we included the contribution of the noise floor with fixed translation efficiency (dotted curves), which increases overabundance, especially for high expression genes. **Panel B: 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.

4.2.5 Overabundance is a robust prediction

Is the predicted overabundance an artifact of the RLTO model? We hypothesize that overabundance might be a consequence of growth arrests for protein numbers below the threshold. To test whether growth arrest is required for the observed RLTO model phenomenology, we consider a *slow-growth model* where the growth rate smoothly decreases to zero at low protein number, but does not arrest. (See the Supplementary Material Sec. 4.7 for a detailed description of the models.) This numerical analysis reveals that the slow-growth and RLTO models had qualitatively identical phenomenology (i.e., $o \gg 1$). We therefore concluded that explicit arrest is not required to make overabundance optimal.

Next, we hypothesized that overabundance is the consequence of the highly-asymmetric fitness landscape (Fig. 4.1E). In both the slow-growth and RLTO models, the growth rate rapidly decreases for underabundance but decays gradually for overabundance. To explore the consequences of asymmetry, we analyzed a *symmetric model* with a symmetric fitness landscape. (See the Supplementary Material Sec. 4.7 for a detailed description of the models.) Consistent with the hypothesis that growth rate asymmetry is the key mathematical mechanism to drive overabundance, we observe that the symmetric model was optimized very close to the noise-free optimum protein number (i.e., the model did not predict overabundance and $o = 1$). We conclude that fitness asymmetry, but not growth arrest, is the characteristic that drives overabundance in the models.

4.2.6 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. 4.3B for the RLTO model with parameters representative of a bacterial cell. (See Supplementary Material Sec. 4.6.4.2.)

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.

4.2.7 Overabundance is observed in a range of experiments

The RLTO Model predicts that all essential proteins are overabundant. 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 [10]. 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 [16, 18, 26]. 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 [16]; 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 [16]. As shown in Fig. 4.3B, 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.

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. We will present an explicit experimental test of this prediction elsewhere [23].

4.2.8 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. 4.4A. 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. The model therefore predicts a lower floor on transcription for essential genes of one message per cell cycle.

4.2.9 A lower threshold is observed for message number

To identify a putative transcriptional floor, we consider the central dogma in three model organisms: the bacterium *E. coli*, *Saccharomyces cerevisiae* (yeast) and *Homo sapiens* (human). For *E. coli*, we consider both rapid and slow growth conditions. We analyze three different transcriptional statistics for each gene: transcription rate (β_m), cellular message number ($\mu_{m/c}$), defined as the average number of messages instantaneously, and message number (μ_m), defined as the number of messages transcribed in a cell cycle. Analysis of these organisms explores orders-of-magnitude differences in characteristics of the central dogma, including total message number, protein number, doubling time, message lifetime, and number of essential genes. (See Supplementary Tab. 4.3.)

In the previous section, we hypothesized that cells must express essential genes above some threshold message number for robust growth; however, we expect to see that non-essential genes can be expressed at much lower levels since growth is not strictly

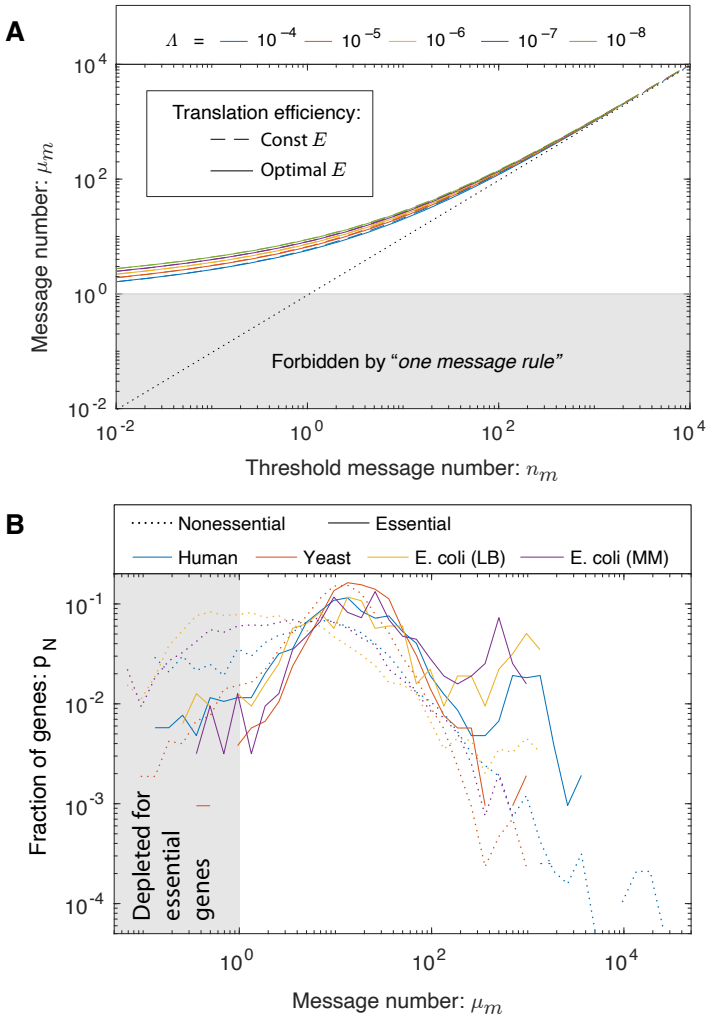


Figure 4.4: **The one-message-rule. Panel A: One-message-rule for essential genes.** For highly transcription genes (high μ_m), little compensation for noise is required and the optimal message number tracks with the threshold message number n_m. However, as the threshold message number approaches one ($n_m \to 1$), 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.

dependent on their expression. A detailed description of the estimation of transcription rates, cellular message number, and message number, and a detailed description of the sources of the data are provided in the Supplementary Material Sec. 4.8. To identify the putative transcriptional threshold, we generated histograms of each statistic in each organism and growth condition.

As expected, there does not appear to be any consistent lower threshold between *E. coli*, yeast, or human transcription, either as characterized by the transcription rate (β_m) or the cellular message number ($\mu_{m/c}$). (See Supplementary Fig. 4.13.) However, as predicted by the RLTO model, there is a consistent lower limit on message number (μ_m) of roughly one message per cell cycle for essential genes. (See Fig. 4.4B.) This floor is consistent not only between *E. coli*, growing under two different conditions, but also between the three highly-divergent organisms: *E. coli*, yeast and human. We will conservatively define the minimum message number as:

$$\mu_m \geq 1, \tag{4.7}$$

and summarize this observation as the *one-message-rule* for essential gene expression.

In addition to the common floor for essential genes, there is a commonality in message number distribution between organisms. To appreciate the significance of this observation, we note that no such similarity is seen in the distribution of transcription rates or cellular message numbers. (See Supplementary Fig. 4.13.) This similarity emphasizes both the conservation of the central dogma regulatory program and the importance of message number as the key transcriptional statistic.

In contrast to essential genes, non-essential genes can be expressed with message numbers below the threshold. (See Fig. 4.4B.) However, it is important to emphasize that the difference between essential and non-essential genes is less significant than it initially appears. These non-essential genes include those that are inducibly expressed but are not induced under the growth conditions studied (e.g., the *lac* operon in *E. coli*). The similarity between essential and non-essential gene expression can be seen in the yeast message number distributions: The essential and non-essential curves are very similar. We believe that the

principal analytical significance of the analysis of essential genes is to confine our analysis to genes that are transcriptionally active under the conditions studied.

4.2.10 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:

$$R \equiv \frac{\mu_p}{\lambda \mu_m} = \frac{\varepsilon}{\lambda}. \tag{4.8}$$

In the Supplementary Material Sec. 4.6.3, we show that the optimal load ratio is:

$$\hat{R} \equiv \frac{1}{\Lambda \ln 2}\ p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{1}{\delta \ln 2}), \tag{4.9}$$

where p_Γ is the PDF of the gamma distribution. The optimal load ratio is shown in Fig. 4.5A.

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.

4.2.11 Translation efficiency is predicted to increase with transcription

Now that we have defined the optimal load ratio, the equation for optimal translation efficiency can be written concisely:

$$\hat{\varepsilon} = \lambda \hat{R} \qquad \text{or} \qquad \hat{E} = \Lambda \hat{R}, \tag{4.10}$$

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

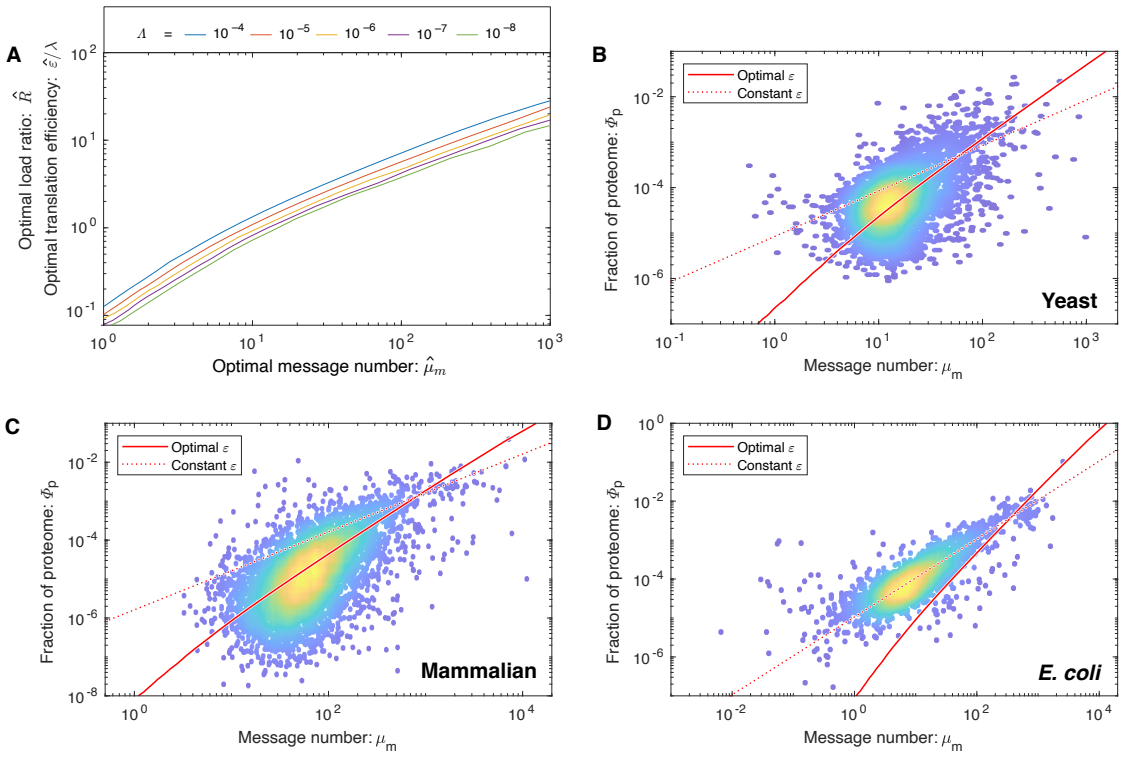


Figure 4.5: **Load balancing for three model species. Panel A: Load balancing: the optimal translation efficiency increases with message number and cost.** The ratio of the optimal translation efficiency ($\hat{\varepsilon}$) to the message cost (λ) is roughly independent of the relative load (Λ); therefore, the optimal translation efficiency is expected to be proportional to the message cost. The translation efficiency is also roughly proportional to the optimal message number $\hat{\mu}_m$. 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. **Gene proteome fraction versus message number for three model systems.** Two models for relation between gene proteome fraction and message number are compared with observations: The RLTO model makes a parameter-free prediction of the optimal relation (*optimal*, solid red line). A second competing model is constant translation efficiency (*constant*, dotted red line). **Panel B: Yeast proteome fraction.** The RLTO prediction (solid) is superior to the constant-translation-efficiency prediction (dashed). **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).

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. 4.5A.) 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*.

4.2.12 Message number also responds to message cost

We will first focus on analyzing the implications of the message cost dependence in Eq. 4.10. At a fixed load ratio, Eq. 4.10 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. 4.6A. The resulting relation between the optimal message numbers is roughly linear on a log-log plot, predicting the approximate power-law relation:

$$\hat{\mu}_m(\Lambda) \propto \hat{\mu}_m(\Lambda_0)^{\alpha}, \tag{4.11}$$

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

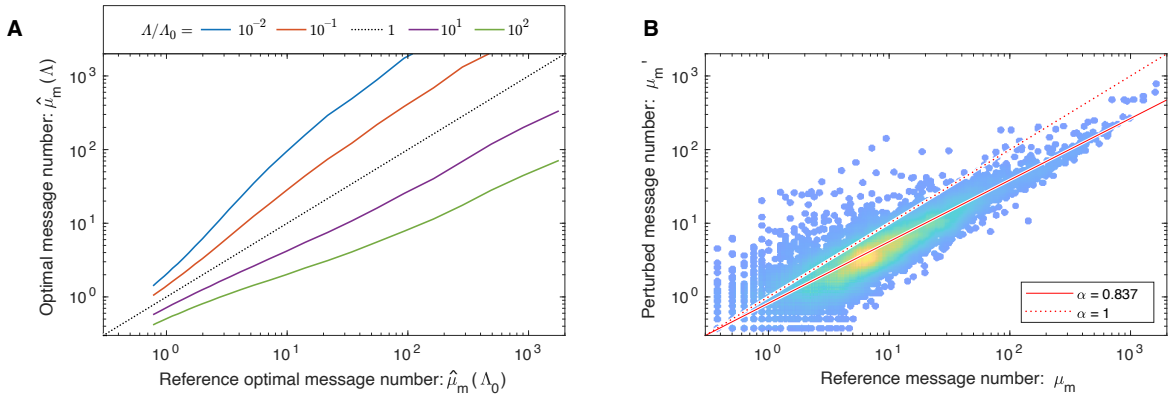


Figure 4.6: **RLTO prediction of message number. Panel A: Message number decreases with increased relative load Λ.** The optimal message number responds to changes in the message cost. The RLTO model predicts an approximate power-law relation (linear on a log-log plot) between message numbers. **Panel 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 [27]. (Data from Ref. [28].) 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.

4.2.13 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 [27], to a reference condition [28]. As predicted, the relative transcriptome data was well described by a power law (Eq. 4.11) and the observed slope was smaller than one: $\hat{\alpha} = 0.837 \pm 0.001$, as predicted by the increased message cost. See Fig. 4.6B.

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.

4.2.14 Parameter-free prediction of proteome fraction

We now turn our focus to an analysis of the implications of the message number dependence of the translation efficiency (Eq. 4.10). The most direct test of this prediction is measuring the relation between proteome fraction and message number. The RLTO model predicts protein fraction:

$$\hat{\Phi}_p = \hat{E}\hat{\mu}_m \propto \hat{\mu}_m^2, \tag{4.12}$$

where μ_m is the observed message number and the optimal relative translation efficiency is predicted by Eq. 4.10. 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.

4.2.15 RLTO predicts proteome fractions in eukaryotic cells.

To test the RLTO predictions, we compare observed proteome measurements in three evolutionarily divergent species, *E. coli* [25], yeast [29] and mammalian cells [30], to two models: the RLTO and the constant-translation-efficiency models. The results of the parameter-free predictions are shown in Fig. 4.5BCD. The RLTO model clearly captures the global trend in the proteome-fraction message-number relation in eukaryotic cells, but not in *E. coli*, where the constant-translation-efficiency models better describes the data.

What gives rise to the spread on the data around the optimal protein fraction and why does the RLTO model fail to describe *E. coli*? In the Supplementary Material, we analyze a number of refinements to the RLTO model. Motivated by the data spread, we investigate models in which there are gene-specific supplemental loads associated with proteins (e.g., toxicity) and models in which gene and protein length are treated explicitly (Supplementary Material Sec. 4.6.8). We conclude that the observed data could be explained by supplemental load, but not gene and protein length. Motivated by the failure of the optimal translation efficiency to describe the proteome fraction in *E. coli*, we consider two different mechanisms which limit translation (Supplementary Material Sec. 4.6.9). The *E. coli* data is consistent with a ribosome-per-message limit, as proposed by Hausser et al. [15].

4.2.16 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 [14, 31]:

$$\mathrm{CV}_p^2 \propto \mu_p^{-1}, \tag{4.13}$$

for low-expression proteins, as observed in *E. coli* [14]; however, the more fundamental prediction is that the noise is inversely proportional to the message number (Eq. 4.2). In

E. coli, the translation efficiency is roughly constant (i.e., $\hat{\mu}_p \propto \hat{\mu}_m$) and therefore Eq. 4.2 is consistent with the canonical noise model (Eq. 4.13). However, in eukaryotes the translation efficiency grows with message number (i.e., $\hat{\mu}_p \propto \hat{\mu}_m^2$). If we substitute this proportionality into Eq. 4.2, we predict the non-canonical noise scaling:

$$\mathrm{CV}_p^2 \propto \mu_p^{-1/2}, \tag{4.14}$$

for eukaryotic cells.

4.2.17 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 [13]. Since the competing models (Eqs. 4.13 and 4.14) make different scaling predictions, we first apply a statistical test to determine whether the observed scaling is consistent with the canonical model (Eq. 4.13). We consider the null hypothesis of canonical model (Eq. 4.13) and the alternative hypothesis with an unknown scaling exponent. To test the models, we performe a null hypothesis test. (A detailed description of the statistical analysis, which include the contribution of the noise floor, is given in the Supplementary Material Sec. 4.9.) 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 ($-\frac{1}{2}$).

4.2.18 Prediction of noise from protein-message relation

By combining the noise model (Eq. 4.2) with a protein-message abundance relation, the relation between protein abundance and noise can be predicted without additional fitting parameters. 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. (See the Supplementary Material Sec. 4.9 for a detailed description of the analysis.) The fit of the competing protein-message abundance

models are shown in Fig. 4.7A. 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. 4.7B.

In both its ability to capture the protein abundance and predict the noise, the RLTO model vastly out performs the constant-translation-efficiency model. However, the purely empirical model which, as a consequence of directly fitting both the y-offset and slope, best capture of the protein abundance data also performs best at 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. 4.2) quantitatively predicts the observed noise from the message number and that eukaryotic noise has non-canonical scaling due to load balancing.

4.3 Discussion

4.3.1 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 (Eq. 4.6). 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. Should this strategy surprise us? No. We routinely use this strategy in our own lives. For instance, when using a pipette in lab, low-cost items that are used stochastically (e.g., pipette tips) are purchased in great excess (overabundance), while the higher cost items that are less stochastic (e.g., pipette) are purchased as needed. (See Fig. 4.8A.)

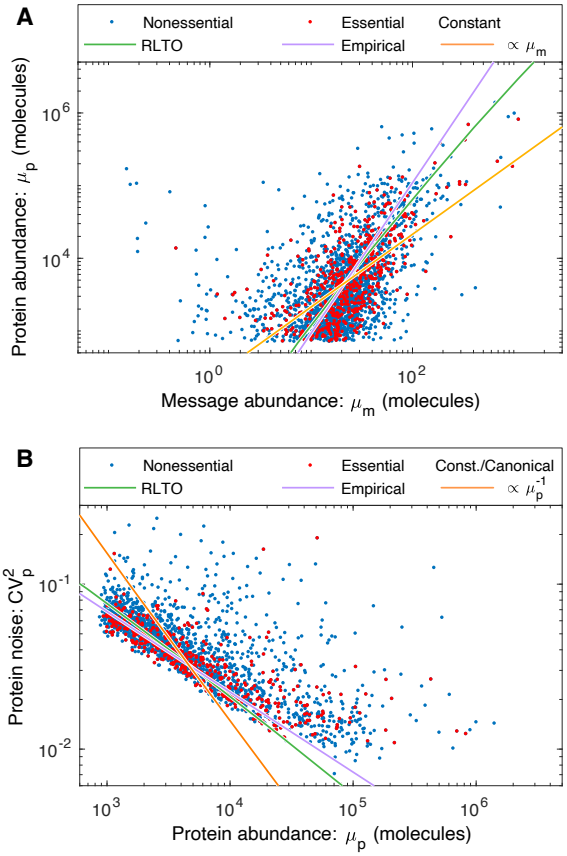


Figure 4.7: **Comparison of noise models in yeast. Panel A: Three competing models for protein abundance in yeast.** The empirical model (purple) fits the slope and the y offset. The RLTO (green) and constant-translation-efficiency (orange) models fit a parameter corresponding to the y offset only. As 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 [13]. **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. 4.2. The canonical noise model (Eq. 4.13) 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 Supplementary Material Fig. 4.15.

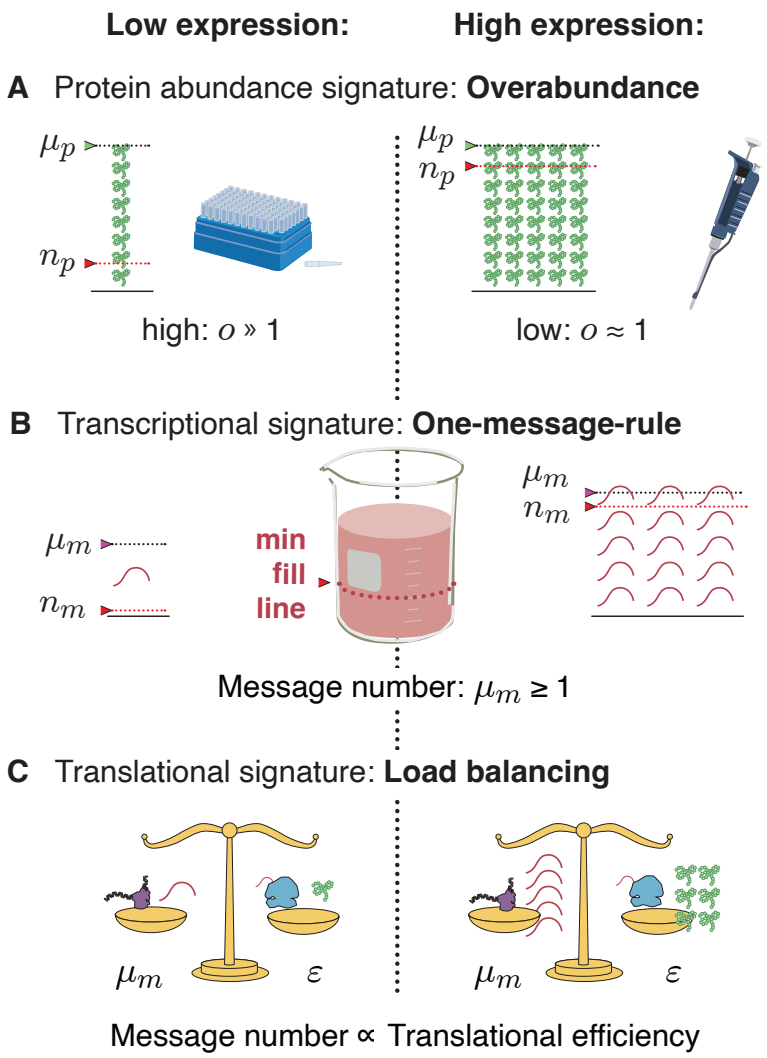


Figure 4.8: **Central dogma regulatory principles. Panel A: Overabundance.** Low-expression essential genes are expressed with high overabundance; whereas, high-expression essential genes are expressed with low overabundance. Lab supply analogy: Low-cost items that are used stochastically (e.g., pipette 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.

4.3.2 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 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 [16]. This may help explain why inhibitors targeting translation make such effective antimicrobial drugs. (See Fig. 4.3B.) 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 [32].

4.3.3 Implications for non-essential 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 non-essential genes share the same optimization? In support of the proposal that RLTO optima describe non-essential genes is the success of the model in predicting the translation efficiency for all genes, not just essential genes. (See Fig. 4.5.) 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 non-essential on other carbon sources [33]. Under growth conditions where the *lacZ* gene is essential, we predict that LacZ should be overabundant. The interesting aspect to *lac* operon expression is that it is not constitutive, but induced in the presence of lactose. It is therefore natural to predict that the *lac* operon is either *off* or *on-and-overabundant*. This biphasic behavior has long been known: Novick and Weiner reported that *enzyme induction [is] an all-or-none phenomenon* [34]. Furthermore, the *lac* operon expression does appear to be overabundant when expressed, since a *metabolic memory effect* is observed, in which multiple generations of cell-growth-induced protein dilution are required before cells lose their adaptation for growth on lactose [11]. In analogy to the *lac* operon, we expect all gene

products, most especially those with low expression, to be overabundant, under conditions where their activity is essential.

4.3.4 Load balancing

A second 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 message number which matches the observations in eukaryotic cells (Fig. 4.5BC). However, in *E. coli*, the translation efficiency and message number are not strongly correlated (Fig. 4.5D). 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 limited, based on the ribosome footprint on mRNA molecules [15]. (See Supplementary Material Sec. 4.6.9.) 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.

4.3.5 Comparisons to previous work

Our analysis is not the first to consider the trade-off between noise and metabolic load. Notably, Hausser et al. performed a more limited analysis [15]; however, their model does not share the rich phenomenology we report. What is the difference between these two analyses?

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 [15].

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. (See Supplementary Material Sec. 4.10.)

4.3.6 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. [12]. 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. 4.8B.) 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. 4.8A.) Robustness to noise also gives rise to load balancing, the proportionality of the optimal transcription and translation rates. (See Fig. 4.8C.) 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. 4.7B.)

4.3.7 The principles that govern central dogma regulation

In summary, the Robustness-Load Trade-Off (RLTO) model describes a number of key regulatory principles that predict the function of the central dogma and are summarized in Fig. 4.8. 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 ensure 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 significant impact, not only on our understanding of central dogma regulation specifically, but in understanding the rationale for protein expression level and function in many biological processes.

4.4 Data availability

A source data file which includes the estimated message numbers as well as essential/nonessential classifications for each organism can be found in the Supplemental Material of Ref. [1].

4.5 Acknowledgments

The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, K. Cutler, M. Cosentino-Lagomarsino, S. van Teeffelen, and S. Murray. This work was supported by NIH grant R01-GM128191.

4.6 Supplemental analysis of the RLTO model

4.6.1 Detailed description of the noise model

4.6.1.1 Stochastic kinetic model for central dogma

The telegraph model for the central dogma describes multiple steps in the gene expression process: Transcription generates mRNA messages [35]. These messages are then translated to synthesize the protein gene products [35]. Both mRNA and protein are subject to degradation and dilution [36]. At the single cell level, each of these processes are stochastic.

We will model these processes with the stochastic kinetic scheme [35]:

$$\begin{array}{ccccc} \text{DNA} & \xrightarrow{\beta_m} & \text{mRNA} & \xrightarrow{\beta_p} & \text{Protein} \\ & & \gamma_m \downarrow & & \gamma_p \downarrow \\ & & \varnothing & & \varnothing \end{array} \tag{4.15}$$

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 [14, 37, 38]: $\gamma_p = T_{cc}^{-1} \ln 2$, where T_{cc} is the doubling time.

4.6.1.2 Statistical model for protein abundance

To study the stochastic dynamics of gene expression, we used a stochastic Gillespie simulation [39, 40]. (See Supplemental Material Sec. 4.6.1.3.) 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 [19, 20], 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 Supplementary Material Sec. 4.6.1.4.) The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [20]: The scale parameter:

$$\theta = \varepsilon \ln 2, \tag{4.16}$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \tag{4.17}$$

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 [20]:

$$a = \frac{\beta_m}{\gamma_p}; \tag{4.18}$$

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

$$\mu_m \equiv \beta_m T_{cc} = a \ln 2, \tag{4.19}$$

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.

4.6.1.3 Gillespie simulation of the telegraph model

Protein distributions of the telegraph model for *E. coli* were simulated with a Gillespie algorithm. Assuming the lifetime of the cell cycle ($T_{cc} = 30$ min) [41], mRNA lifetime ($T_m = 2.5$ min) [42], 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.

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 statistical noise model (i.e., gamma function) as shown in Fig. 4.9.

4.6.1.4 Gamma distribution

The gamma distributed random variable X will be written:

$$X \sim \Gamma(a, \theta), \tag{4.20}$$

where a is the shape parameter and θ is the scale parameter [43]. The PDF of the distribution is:

$$p_\Gamma(x|a, \theta) \equiv \frac{x^{a-1}}{\theta^a \Gamma(a)} e^{-x/\theta}, \tag{4.21}$$

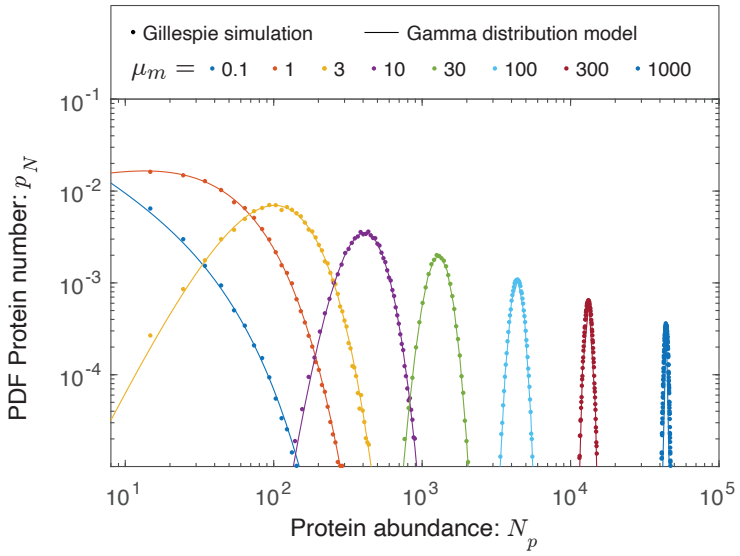


Figure 4.9: **The protein abundance is approximately gamma distributed.** Protein abundance was modeled for eight different transcription rates using a Gillespie simulation, including the stochastic partitioning of the proteins between daughter cells at cell division. The range in abundance matches the observed range of expression levels in the cell. We observed that the simulated protein abundances were well fit by gamma distributions.

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

$$\begin{aligned} P_\Gamma(x|a,\theta) &= \int_0^x \mathrm{d}x'\ p_\Gamma(x'|a,\theta), && (4.22)\\ &= P_\Gamma(\tfrac{x}{\theta}|a,1), && (4.23)\\ &= \int_0^{x/\theta} \mathrm{d}x'\ \tfrac{x^{a-1}}{\Gamma(a)}e^{-x}, && (4.24)\\ &= \gamma(a,x/\theta), && (4.25) \end{aligned}$$

where γ is the regularized incomplete gamma function.

4.6.2 The derivation of the RLTO growth rate

4.6.2.1 Metabolic load

In the absence of cell cycle arrest, we will assume that the cell cycle duration τ is proportional to the overall metabolic load of the messages, proteins, and other cellular components [21]. Focusing on a particular gene:

$$\tfrac{\tau}{\delta\tau} = N_0 + \lambda\mu_m + \mu_p, \tag{4.26}$$

where $\delta\tau$ is a constant with units of time, N_0 is the metabolic load of everything but gene i (in units of protein equivalents), λ is the message cost, the metabolic load associated with an mRNA molecule relative to a single protein molecule of the gene product. We will assume the parameters N_0, $\delta\tau$ and λ are global (i.e., gene independent), while the message number μ_m and translation efficiency ε, and therefore the mean protein number:

$$\mu_p = \mu_m\varepsilon, \tag{4.27}$$

are gene specific. The total metabolic load in protein equivalents is:

$$N_0 \equiv L_0 + \sum_i \{(\lambda+\varepsilon)\mu_m\}_i, \tag{4.28}$$

where L_0 is the load of non-protein and message cellular components and the sum runs over all genes. The metabolic load of transcription for gene i is $\lambda\mu_{m,i}$ and for translation $\mu_{p,i} = \varepsilon_i\mu_{m,i}$.

Although the global parameters N_0, $\delta\tau$ 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:

$$\frac{k_0}{k} = 1 + (\Lambda + E)\mu_m, \tag{4.29}$$

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 $E \equiv \varepsilon/N_0$, which is the ratio of the number of proteins translated per message to the total metabolic load N_0. If we neglect the difference between the total metabolic load and the number of proteins, the proteome fraction for gene i is $\Phi_p = E\mu_m$. Both reduced parameters, Λ and E are extremely small. In *E. coli*, we estimate that both Λ and E are roughly 10^{-5} and they are smaller still for eukaryotic cells. (See Supplementary Material Sec. 4.6.10.)

4.6.2.2 Growth rate with arrest

For completeness, we provide a derivation of the growth rate with stochastic cell-cycle arrest that we have previously described [24]. Starting from the exponential mean expression for the population growth rate [24]:

$$k = \frac{\ln 2}{\overline{T}}, \tag{4.30}$$

where k is the population growth rate and

$$\overline{T} \equiv -\frac{1}{k} \ln \mathbb{E}_T \exp(-kT), \tag{4.31}$$

is the exponential mean where T is the stochastic cell cycle duration [24, 44].

Let P_+ be the probability of growth. When the cells are growing, the cell cycle duration is given by the metabolic load cell-cycle duration τ (Eq. 4.26). Evaluating the expectation gives:

$$\ln 2 = -\ln P_+ + k\tau, \tag{4.32}$$

which can be rearranged to give an expression for the population growth rate k:

$$k = \tau^{-1} \ln 2P_+. \tag{4.33}$$

As expected, the growth rate goes down as the probability of growth P_+ decreases, stopping completely at $P_+ = \frac{1}{2}$.

4.6.2.3 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:

$$P_+ = \prod_{i \in \mathcal{E}} \Pr\{N_p > n_p\}_i, \tag{4.34}$$

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:

$$\begin{aligned} \ln P_+ &= \sum_{i \in \mathcal{E}} \ln\left(1 - \Pr\{N_p < n_p\}_i\right), && (4.35) \\ &\approx -\sum_{i \in \mathcal{E}} \Pr\{N_p < n_p\}_i\}, && (4.36) \\ &= -\sum_{i \in \mathcal{E}} \left[\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2})\right]_i, && (4.37) \end{aligned}$$

where γ is the regularized incomplete gamma function and the CDF of the gamma distribution. Finally, we will be interested in the quantity:

$$\ln \ln 2P_+ \approx \ln \ln 2 + -\sum_{i \in \mathcal{E}} \tfrac{1}{\ln 2} \left[\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2})\right]_i, \tag{4.38}$$

where we have again expanded the natural logarithm in the limit that the CDF is small.

4.6.2.4 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 figure (Figs. 4.1F), we compute the fitness not just at the optimal values but far from them. Here we cannot approximate the natural log and we use the full expression in Eq. 4.35.

4.6.2.5 Single-gene equation

By combining Eqs. 4.33, 4.26 and 4.38, we can write an expression for the growth rate:

$$\begin{aligned} \ln k &= -\ln\left(L_0 + \sum_i \left[(\lambda+\varepsilon)\mu_m\right]_i\right) + ... \\ &\quad + \ln\ln 2 - \sum_{i\in\mathcal{E}} \tfrac{1}{\ln 2}\left[\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2})\right]_i . \end{aligned} \tag{4.39}$$

We will work in the large multiplicity limit where each of the arrest probabilities is very small. To optimize the expression of gene i we can drop all constant terms and write the simplified expression:

$$\begin{aligned} \ln k &= -\ln\left(N_0' + (\lambda+\varepsilon)\mu_m\right) + ... \\ &\quad + \ln\ln 2 - \tfrac{1}{\ln 2}\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{n_p}{\varepsilon \ln 2}) + ..., \end{aligned} \tag{4.40}$$

where we have made the essential gene i subscript implicit and we consider a total load minus gene i:

$$N_0' \equiv L_0 + \sum_{j\neq i} \left[(\lambda+\varepsilon)\mu_m\right]_j , \tag{4.41}$$

interpreted as the metabolic load of all genes but i. However, we will work in the large multiplicity limit where we ignore the distinction between N_0 and N_0'.

4.6.3 Message number and translation efficiency optimization

Starting with Eq. 4.4 for the growth rate, we set the partial derivative with respect to message number equal to zero:

$$0 = -\frac{\lambda+\hat{\varepsilon}}{N_0} - \frac{1}{(\ln 2)^2}\gamma_{,1}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon}\ln 2}), \tag{4.42}$$

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:

$$0 = -\frac{\hat{\mu}_m}{N_0} + \frac{n_p}{\hat{\varepsilon}^2(\ln 2)^2}\gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon}\ln 2}). \tag{4.43}$$

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:

$$\hat{o} \equiv \frac{\hat{\mu}_p}{n_p} = \frac{\hat{\varepsilon}\hat{\mu}_m}{n_p}, \tag{4.44}$$

in both Eqs. 4.42 and 4.43. Eq.4.43 can now be solved for the optimal translation efficiency:

$$\hat{\varepsilon} = \frac{N_0}{\hat{o}(\ln 2)^2}\gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}). \tag{4.45}$$

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

$$\hat{\varepsilon} = \frac{N_0}{\ln 2}p_\Gamma(\mu_m|\frac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2), \tag{4.46}$$

which will be the optimization equation for the translation efficiency.

To derive the optimization condition for the message number μ_m, we substitute Eq. 4.45 into Eq. 4.42:

$$\frac{\lambda\ln 2}{N_0} = -\frac{1}{\hat{o}\ln 2}\gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}) - \frac{1}{\ln 2}\gamma_{,1}(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}). \tag{4.47}$$

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

$$\Lambda\ln 2 = -\partial_{\hat{\mu}_m}\gamma(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}), \tag{4.48}$$

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

The two optimization equations are summarized below:

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m} \gamma(\tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{\hat{\mu}_m}{\hat{o} \ln 2}), \tag{4.49}$$

$$\tfrac{\hat{E}}{\Lambda} = \tfrac{\hat{\varepsilon}}{\lambda} = \tfrac{1}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \tag{4.50}$$

where despite the factor of Λ in the denominator of the RHS of Eq. 4.50, the RHS depends very weakly on Λ. See Fig. 4.5A.

4.6.4 Modifications of the RLTO model for bacterial cells

There are two features of the bacterial cell that are distinct relative to eukaryotes: (i) constant translation efficiency and (ii) a high noise floor. We will consider both.

4.6.4.1 Optimization of message number only

Consider the special case of optimizing the message number only at fixed translation efficiency. Eq. 4.42 is essentially the condition; however, in this case it makes sense to adsorb both the message and protein metabolic load into a single metabolic load. The optimum message number satisfies the equation:

$$\tfrac{(\lambda+\varepsilon) \ln 2}{N_0} = -[\partial_{\hat{\mu}_m} \gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\hat{o}}}. \tag{4.51}$$

We define a modified relative load:

$$\Lambda' \equiv \tfrac{(\lambda+\varepsilon)}{N_0}, \tag{4.52}$$

and substitute this into the optimum message number equation:

$$\Lambda' \ln 2 = -[\partial_{\hat{\mu}_m} \gamma(\hat{\mu}_m, \hat{n}_m)]_{\hat{n}_m = \frac{\hat{\mu}_m}{\hat{o}}}, \tag{4.53}$$

which is closely related to Eq. 4.48.

We compare this modified expression to the original for optimum overabundance as a function of message number in Fig. 4.3A and demonstrate that the two make nearly identical predictions.

4.6.4.2 Adaptation of the RLTO model to a noise floor

In bacterial cells, the noise is dominated by the noise floor for higher expression levels. Including the noise floor, the coefficient of variation squared is [14]:

$$\mathrm{CV}_p^2 = \tilde{a}(\mu_m)^{-1} = \frac{\ln 2}{\mu_m} + C_0, \tag{4.54}$$

where $C_0 = 0.1$ for bacterial cells [14]. 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 [14]; however, we need to modify the statistical parameters to account for the noise floor. The modified gamma parameters are:

$$a = \tilde{a}, \tag{4.55}$$

$$\theta = \varepsilon\frac{\mu_m}{\tilde{a}}, \tag{4.56}$$

chosen such that the noise is determined by Eq. 4.54 but the protein number remains:

$$\mu_p = \varepsilon\mu_m, \tag{4.57}$$

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 expression 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. 4.3A and demonstrate that bacterial cells are predicted to have much higher overabundance at high expression levels.

4.6.5 Arrest probability

The gene i arrest probability at optimal expression is:

$$P_- = \gamma(\tfrac{\hat{\mu}_m}{\ln 2}, \tfrac{\hat{\mu}_m}{\hat{o}\ln 2}), \tag{4.58}$$

which is independent of the message cost λ. As expected, the arrest probability is extremely low. (See Fig. 4.10.)

A more interesting experimental quantity is the total arrest probability. We can estimate the total arrest probability in *E. coli* by summing up the arrest probability for all essential genes. The estimated total arrest probability is $P_{-}^{\text{tot}} = 8.5 \times 10^{-4}$. This result is consistent with single-cell analysis of cell growth where the filamentation rate of MG1655 is roughly 1% per generation, but smaller in the B/r strain [45].

4.6.6 Discussion of *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 [46]. By this classification, 10 essential genes were below threshold. (See Supplementary Material Tab. 4.2.) Our first step was to determine what fraction of these genes were also classified as essential using transposon-based mutagenesis [47, 48]. 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 [49]. 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.

4.6.7 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. [27] who analyzed the differences in fitness between transcription and transcriptional-and-translation of a fluorescent protein driven by the

Table 4.2: **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.

Gene name	Message number: μ_m	Annotated function from Ecocyc	Essential (E)/ Nonessential (N) Ref. [46], [47], [48]
alsK	0.3	The alsK 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 *E. coli* K-12.	E, N, N
bcsB	0.4	BcsB is encoded in a predicted operon together with *bcsA*, *bcsZ* and *bcsC*. 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 *E. coli* does not show a rdar morphotype and does not produce cellulose.	E, N, N
entD	0.4	AcpS is the founding member of a 4'-phosphopantetheinyl (P-pant) transferase protein family that includes *E. coli* EntD, *E. coli* o195 protein, and *Bacillus subtilis* Sfp; family members share two conserved motifs but relatively low sequence identity overall.	E, N, N
yafF	0.4	No information about this protein was found by a literature search conducted on April 19, 2017.	E,-, N
yagG	0.6	*yagGH* 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 *yagGH*.	E,-, N
yceQ	0.2	No information about this protein was found by a literature search conducted on July 12, 2017.	E, E, N
ydiL	0.2	No information about this protein was found by a literature search conducted on April 7, 2017.	E, N, N
yhhQ	0.4	YhhQ is an inner membrane protein implicated in the uptake of queuosine (Q) precursors - 7-cyano-7-deazaguanine (preQ0) and 7-aminomethyl-7-deazaguanine (*preQ1*) - for Q salvage. Q-modified tRNA is absent in Δ*queD* and Δ*queD* Δ*yhhQ* strains grown in minimal media with glycerol; Q-modified tRNA is detected when a Δ*queD* strain is grown in minimal media plus 10 nM *preQ0* or *preQ1* but is absent when a Δ*queD* Δ*yhhQ* strain is grown under these conditions. *yhhQ* expressed from a plasmid restores the presence of Q-modified tRNA in a Δ*queD* Δ*yhhQ* strain.	E,-, N
yibJ	0.3	No information about this protein was found by a literature search conducted on July 9, 2018.	E, N, N
ymfK	0.4	YmfK is a component of the relic lambdoid prophage e14 and is likely the SOS-sensitive repressor. It is similar to the P34 gene of the *Shigella flexneri* bacteriophage SfV and belongs to the LexA group of SOS-response transcriptional repressors.	E, E, E

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.

4.6.8 Increased protein load analysis

Although the overall trend in the relation between the translation efficiency and message number in eukaryotic cells is captured by the RLTO model, a significant amount of scatter is observed around this optimal relation. One important consideration in a more realistic model are proteins whose fitness cost is greater than their metabolic load. (For instance, consider an ATPase with non-specific activity.) A second potentially interesting question is how message and protein length affect the optimal parameters. For instance, are a 50- and 500-amino-acid proteins expressed using a different regulatory program?

To explore the consequences of these added complexities, we can modify the metabolic load term in the growth rate equation (Eq. 4.48):

$$E\mu_m \rightarrow \lambda_p E\mu_m, \tag{4.59}$$

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.

4.6.8.1 Derivation

To analyze the affect of increased protein load, we modify Eq. 4.4:

$$\ln \tfrac{k}{k_0} = -(\Lambda + \lambda_p E)\mu_m - \tfrac{1}{\ln 2}\gamma(\tfrac{\mu_m}{\ln 2}, \tfrac{\phi_p}{E \ln 2}), \tag{4.60}$$

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 additions load (e.g., toxicity). The optimization conditions (Eqs. 4.42 and 4.43) become:

$$0 = -\frac{\lambda+\lambda_p\hat{\varepsilon}}{N_0} - \frac{1}{(\ln 2)^2}\gamma_{,1}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon}\ln 2}), \tag{4.61}$$

$$0 = -\frac{\lambda_p\hat{\mu}_m}{N_0} + \frac{n_p}{\hat{\varepsilon}^2(\ln 2)^2}\gamma_{,2}(\frac{\hat{\mu}_m}{\ln 2}, \frac{n_p}{\hat{\varepsilon}\ln 2}). \tag{4.62}$$

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

$$\Lambda \ln 2 = -\partial_{\hat{\mu}_m}\gamma(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}), \tag{4.63}$$

$$\hat{R} = \frac{1}{\Lambda \ln 2}p_\Gamma(\hat{\mu}_m|\frac{\hat{\mu}_m}{\ln 2}, \hat{o}\ln 2); \tag{4.64}$$

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

$$R = \frac{\lambda_p\varepsilon\mu_m}{\lambda\mu_m} = \frac{\lambda_p\varepsilon}{\lambda}, \tag{4.65}$$

representing the modified total load ratio.

4.6.8.2 Increased protein cost reduces the optimal translation efficiency.

The relation between the overabundance and message number is unchanged (Eq. 4.6). 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:

$$\hat{\varepsilon} = \frac{\lambda}{\lambda_p}\hat{R}, \tag{4.66}$$

where $\hat{R}$ is the optimal load ratio, defined by Eq. 4.10. The optimal curves are shown in Fig. 4.11.

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,

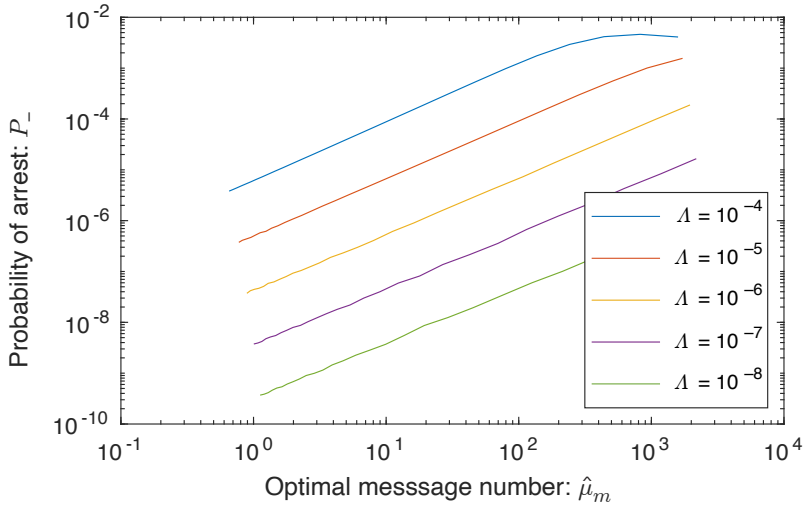


Figure 4.10: **Cell cycle arrest probability.** Low protein number slows growth by arresting the cell cycle. The per-essential-gene arrest probabilities are shown as a function of message number μ_m. The arrest probability has an approximately linear dependence of the message number μ_m. The arrest probability is roughly proportion to the relative load Λ.

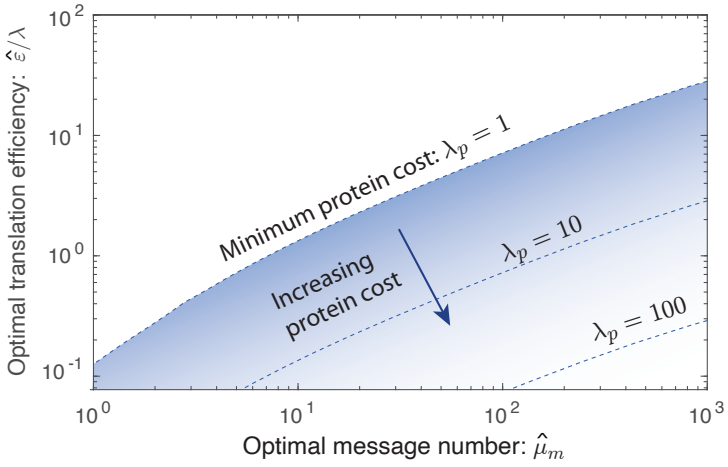


Figure 4.11: **Increased protein cost decreases optimal translation efficiency.** A protein cost of $\lambda_p = 1$ corresponds to the metabolic cost of protein synthesis only, and is the minimum protein cost. For larger protein costs, the optimal translation efficiency is lower. As a result, the $\lambda_p = 1$ curve represents an upper bound of the optimal translation efficiency.

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:

$$\hat{\mu}_m \rightarrow 10 \cdot \hat{\mu}_m, \tag{4.67}$$

$$\hat{\varepsilon}_m \rightarrow \tfrac{1}{10} \cdot \hat{\varepsilon}_m, \tag{4.68}$$

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. 4.5. 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.

4.6.9 Analysis of translational limits.

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. 4.5D) 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 ribosome 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., [50]). 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 [15].

4.6.10 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 in not determinative of the translation efficiency.)

The optimal translation efficiency is (Eq. 4.46):

$$\hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \tag{4.69}$$

and therefore the optimal protein number is:

$$\hat{\mu}_p = \hat{\mu}_m \hat{\varepsilon} = \frac{\lambda}{\Lambda \ln 2} \hat{\mu}_m p_\Gamma(\hat{\mu}_m | \frac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2). \tag{4.70}$$

We define the normalization constant A:

$$A = \sum_i [\hat{\mu}_m \cdot \tfrac{1}{\Lambda \ln 2} p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2)]_i, \tag{4.71}$$

where the index i runs over all genes. Now, by summing Eq. 4.70, 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:

$$N_p^{\text{tot}} = \lambda A. \tag{4.72}$$

Solving for the protein cost results in the estimate:

$$\hat{\lambda} = \tfrac{N_p^{\text{tot}}}{A}. \tag{4.73}$$

This message cost estimate $\hat{\lambda}$ can then be plugged into the metabolic load definition (Eq. 4.41) to estimate the multiplicity:

$$\hat{N}_0 \equiv L_0 + \hat{\lambda} N_m^{\text{tot}} + N_p^{\text{tot}} \tag{4.74}$$

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

4.6.10.1 Detailed protocol

We first estimate the message numbers, as described in Sec. 4.8.2.3, 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. 4.49. (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. 4.71. In Eqs. 4.73 and 4.74, we use the N_p^{tot} from Tab. 4.3. N_m^{tot} is computed by summing the estimated message numbers.

4.6.10.2 Estimate the message cost and metabolic load in yeast

In yeast, the estimates are:

$$A = 4.8 \times 10^5, \tag{4.75}$$

$$\hat{\lambda} = 1.0 \times 10^2, \tag{4.76}$$

$$\hat{N}_0 = 6.2 \times 10^7, \tag{4.77}$$

$$\hat{\Lambda} = 1.6 \times 10^{-6}, \tag{4.78}$$

where the data sources are described in detail in Sec. 4.8.1.2.

4.6.10.3 Estimate the message cost and metabolic load in human cells

In human cells, the estimates are:

$$A = 4.3 \times 10^6, \tag{4.79}$$

$$\hat{\lambda} = 7.1 \times 10^2, \tag{4.80}$$

$$\hat{N}_0 = 2.4 \times 10^9, \tag{4.81}$$

$$\hat{\Lambda} = 2.9 \times 10^{-7}. \tag{4.82}$$

where the data sources are described in detail in Sec. 4.8.1.3.

4.6.11 Prediction of the proteome fraction

4.6.11.1 RLTO: proteome fraction

Starting from Eq. 4.70, clearly:

$$\hat{\mu}_p \propto \hat{\mu}_m \; p_\Gamma(\hat{\mu}_m | \tfrac{\hat{\mu}_m}{\ln 2}, \hat{o} \ln 2), \tag{4.83}$$

which can be used to predict the proteome fraction:

$$\hat{\Phi}_{pi} \equiv \tfrac{\hat{\mu}_{pi}}{\sum_j \hat{\mu}_{pj}}, \tag{4.84}$$

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

$$C \equiv \sum_i \left[\mu_m p_\Gamma(\mu_m|\tfrac{\mu_m}{\ln 2}, o \ln 2)|_{o=\hat{o}(\mu_m)}\right]_i , \tag{4.85}$$

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. 4.48) for message number μ_{mi} and the sum index i runs over all genes. The predicted optimal proteome fraction is:

$$\hat{\Phi}_{pi} = C^{-1}\left[\mu_m p_\Gamma(\mu_m|\frac{\mu_m}{\ln 2}, o \ln 2)|_{o=\hat{o}(\mu_m)}\right]_i , \tag{4.86}$$

which generates the predicted solid curves shown in Fig. 4.5BCD.

4.6.11.2 Constant-translation-efficiency model: proteome fraction

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

$$C' \equiv \sum_i \mu_{mi}, \tag{4.87}$$

and the predicted proteome fraction is:

$$\Phi'_{pi} = C'^{-1}\mu_{mi}, \tag{4.88}$$

which generates the predicted dotted curves shown in Fig. 4.5BCD.

4.6.11.3 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 measurements [25]. For the yeast data, the protein abundance data is measured by mass spec and message abundances are determined by [29]. For the mammalian data, we used mouse data. The protein abundance data is measured by mass-spec and message abundances are determined by [30].

We estimated the message number μ_m as described in Sec. 4.8.2.3. For the mouse data, the study provided message lifetimes, the cell cycle duration and abundances in molecules

per cell [30]. For the *E. coli* and yeast data, the total number of proteins, messages, cell cycle duration, and message lifetimes for each organism and their sources are described in Tab. 4.3.

4.7 Analysis of alternative models

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 a *threshold-like model* (the RLTO model), a *slow-growth model*, and a *symmetric model*. In each case, we will assume that the protein number is described by a gamma distribution:

$$N_p \sim \Gamma(\tfrac{\mu_m}{\ln 2}, \varepsilon \ln 2). \tag{4.89}$$

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. 4.31 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 $\mu_p = \mu_m \varepsilon$.

4.7.1 Threshold (RLTO) model

For the threshold-like (RLTO) model has cell cycle duration:

$$T = \tau_0 \begin{cases} \infty, & N_p < n_p \\ N_0 + N_p, & N_p > n_p \end{cases}, \tag{4.90}$$

where protein expression below threshold n_p results in growth arrest.

4.7.2 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:

$$T = \tau_0 \max\{N_0 + N_p, \tfrac{2n_p N_0}{N_p}\} \tag{4.91}$$

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

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

4.7.3 Model 3: Symmetric model

For the symmmetric 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

$$T = \tau_0 N_0 \exp\left(\tfrac{(N_p - n_0)^2}{2\sigma_n^2}\right) \tag{4.92}$$

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

4.7.4 Conclusions from fitness-landscape analysis

The growth rates as a function of the mean expression level μ_p are shown in Fig. 4.12. The symmetric model has a population optimum in close proximity to its single-cell optimum, as we intuitively expect. However, both the threshold-like (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 the threshold-like growth rate 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. [24], we showed that the population doubling time $\overline{T}$ can be understood as the exponential mean of the stochastic cell-cycle duration:

$$\overline{T} \equiv f^{-1}[\mathbb{E}_T f(T)], \tag{4.93}$$

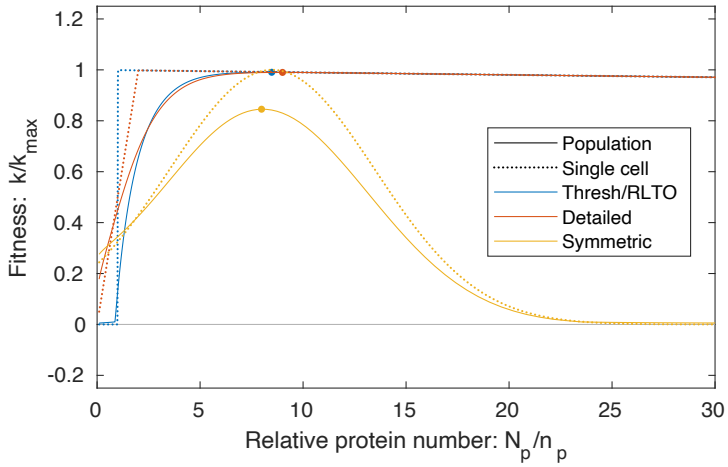


Figure 4.12: **Exploring the mathematical mechanism of overabundance.** Single-cell and population growth rate are compared for three different models: threshold-like, slow-growth, and symmetric models. In the threshold-like model (RLTO), the growth rate goes to zero below threshold protein level n_p. In the slow-growth model, the growth rate transitions continuously to zero as the N_p is depleted below n_p. In both the threshold-like 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.

where $\mathbb{E}_T$ is the expectation over the stochastic duration T and $f(t) \equiv \exp(-kt)$, where $k = \overline{T}^{-1} \ln 2$ 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 faction of growing cells.

4.8 Quantitation of central dogma parameters for the one-message-rule

4.8.1 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 [51]. 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. 4.3, 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. [15].

4.8.1.1 E. coli data

Message lifetimes: The message lifetimes (and median lifetime) were taken from a recent transcriptome-wide study by Chen et al. [42]. 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 [14]. 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 Bartholomaus et al. study [52], 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 [51]).

Doubling time: The source of the doubling times for rapid (LB) and slow (M9) growth of *E. coli* comes from Bernstein [41].

Essential gene classification: The classification of essential genes in yeast comes from the construction of the Keio knockout collection from Baba et al. [46].

Protein number: The total protein number in *E. coli* came from Milo's recent review of this subject [53].

4.8.1.2 Yeast data

Message lifetimes: The message lifetimes (and median lifetime) were taken from Chia et al. [54].

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 [13].

mRNA abundance: The transcriptome data comes from the very recent Blevins et al. study [55].

Total cellular message number: There are a wide range of estimates for the total cellular message number in yeast: 1.5×10^4 [56] (BNID 104312 [51]), 1.2×10^4 [57] (BNID 102988 [51]), 6.0×10^4 [58] (BNID 103023 [51]), 2.6×10^4 [59] (BNID 106763 [51]) and 3.0×10^4 [60]. We used the compromise value of 2.9×10^4.

Doubling time: The doubling time was taken from [61].

Protein number: The total protein number in yeast comes from Futcher et al. [62].

Essential gene classification: The classification of essential genes in yeast comes from van Leeuwen et al. [63].

Proteome abundance data: The proteome abundance data came from two sources: flow cytometry of fluorescent fusions from Newman et al. [13] as well as mass-spec data from de Godoy et al. [64].

4.8.1.3 Human data

Message lifetimes: The message lifetimes (and median lifetime) were taken from Yang et al. [65] 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 [66], which we averaged over tissue types.

Total cellular message number: The total cellular message number in human comes from Velculescu et al. [67] (BNID 104330 [51]).

Doubling time: The doubling time was taken from [68].

Protein number: The total protein number in human came from Milo's recent review of this subject [53].

Essential gene classification: The classification of essential genes in human comes from Wang et al. [69].

4.8.2 Quantitative estimates of central dogma parameters

4.8.2.1 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 $N_{m/c}^{\rm tot}$. This number and its source are provided in Tab. 4.3. To estimate the number of messages corresponding to gene i, we re-scaled the un-normalized abundance level r_i:

$$N_{m/c,i} = N_{m/c}^{\rm tot} \frac{r_i}{\sum_j r_j}, \tag{4.94}$$

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

4.8.2.2 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 telegraph model prediction for the cellular message number:

$$N_{m/c,i} = \beta_{m,i}/\gamma_{m,i}, \tag{4.95}$$

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* [42]), we estimate the decay rate as the inverse gene-median message life time:

$$\gamma_{m,i} = \tau_m^{-1}, \tag{4.96}$$

for which a consensus value was found from the literature. This number and its source are provided in Tab. 4.3. We then estimate the gene-specific transcription rate:

$$\beta_{m,i} = N_{m/c,i}/\tau_m. \tag{4.97}$$

4.8.2.3 Estimating the message number: μ_m

To estimate the message number of gene i, we use the predicted value from the telegraph model:

$$N_{m,i} = T\beta_{m,i} = \frac{T}{\tau_m} N_{m/c,i}, \tag{4.98}$$

where T is the doubling time and $N_{m/c,i}$ is the cellular message number (Eq. 4.94).

4.8.2.4 Histograms

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. 4.13; however, the histogram for message number does show a consistent lower bound for the three model organisms and is shown in Fig. 4.4B.

Table 4.3: **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.

Model organism	Growth condition	Doubling time: T	Message lifetime: $\tau_m = \gamma_m^{-1}$	Message recycling ratio: $m = T/\tau_m$	Total number of messages /cell: $N_{m/c}^{\text{tot}}$	Total number of messages /cell-cycle: N_m^{tot}	Total number of proteins: N_p^{tot}	Average translation efficiency: ε	Average translation rate: β_p (h^{-1})
Escherichia coli (*E. coli*)	LB	30 min [41]	2.5 min [42]	12	7.8×10^3 [52]	9.4×10^4	3×10^6 [53]	22	530
	M9	90 min [41]	2.5 min [42]	36	2.4×10^3 [52]	8.6×10^4	3×10^6 [53]	24	580
Saccharomyces cerevisiae (Yeast–haploid)	YEPD	90 min [61]	22 min [54]	4	2.9×10^4 [59]	1×10^5	5×10^7 [62]	4×10^2	410
Mus musculus (Mammalian mouse)	Tissue	27.5 h [30]	15 h [30]	1.8	1.7×10^5 [30]	3×10^5 [30]	3×10^9 [30]	1×10^4	660
Homo sapiens (Human)	Tissue	24 h [68]	14 h [65]	1.7	3.6×10^5 [61]	5×10^5	2×10^9 [53]	4×10^3	120

4.9 Supplemental analysis of Noise-Protein-Abundance Relation in Yeast

4.9.1 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 [13] were given in arbitrary units (AU). In order to convert from AU to protein number, the fluorescence values were rescaled by comparing with

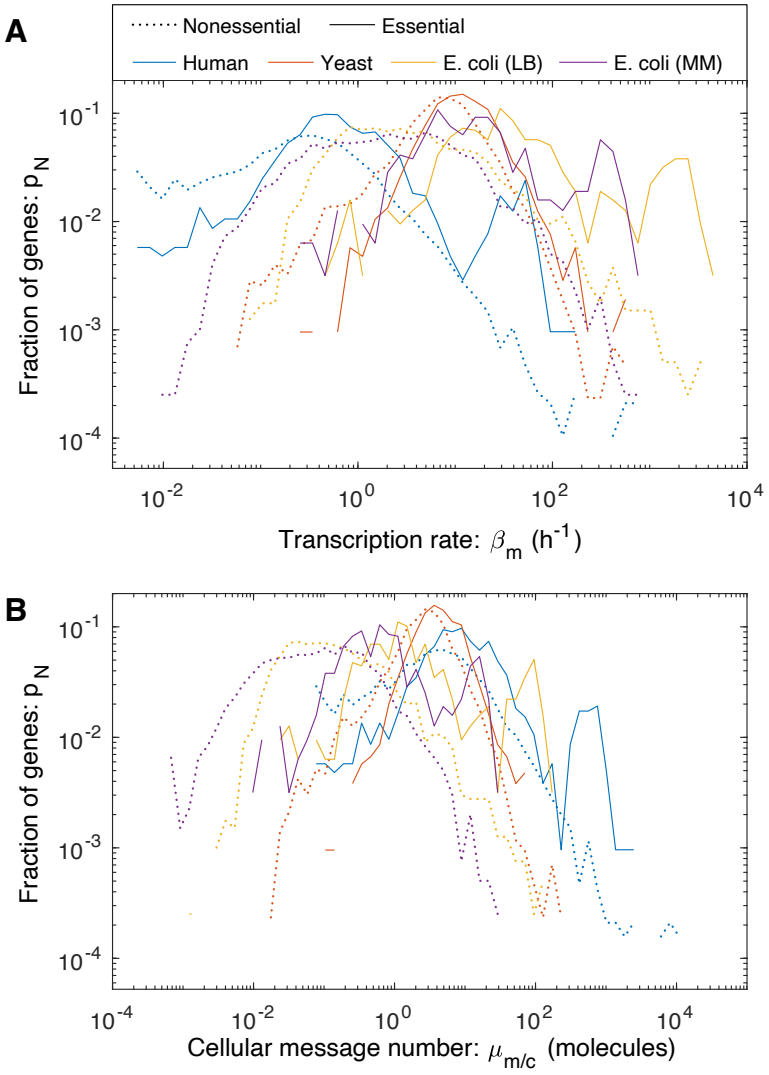


Figure 4.13: **Transcription in three model organisms.** We characterized different gene transcriptional statistics in three model organisms. In *E. coli*, two growth conditions were analyzed: Rapid growth in LB and Slow growth in minimal media. **Panel A: The distribution of gene transcription rate.** The typical transcription rate varies by two orders-of-magnitude between organisms. **Panel B: 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.

mass-spectrometry protein abundance data for yeast grown in YNB media [64]. 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. 4.8.1.2.) 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 (see Fig. 4.14), which corresponds to a multiplicative scaling factor. We then used that offset value to rescale the fluorescence data from AU to protein number. We also compared to yeast grown in SD media [13] and found a similar offset result.

4.9.2 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. [13], re-scaled to estimate protein number (Sec. 4.9.1) and transcriptome data from Lahtvee et al. [70], re-scaled to estimate message number (Sec. 4.8.2.3).

4.9.2.1 The meaning of the error estimates

Before providing a detailed error analysis, it is important to place our error estimates in perspective. The error that we will be estimating is the statistical error associated with the finite sample size; however, *this is not the dominant source of error*. A far more important consideration are systematic problems with our analysis and the underlying experiments. For instance, since we do not have a detailed model for the error of the experiments analyzed, there are multiple distinct analyses (i.e., assumptions about the error model) that could be implemented for the data fitting, each giving slightly different model parameters. These model to model differences still give rise to predictions consistent with our qualitative

conclusions; however, they are likely larger than the statistical uncertainty we compute (while assuming a particular model).

4.9.2.2 Empirical model for protein number

We initially fit the empirical model for protein number,

$$\mu_p = C_0 \mu_m^{\alpha_0}, \tag{4.99}$$

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 [71], which assumes comparable error in both variables. The model parameters are:

$$\alpha_0 = 2.1 \pm 0.04, \tag{4.100}$$

$$C_0 = 8.0 \pm 1.0, \tag{4.101}$$

where the uncertainties are the estimated standard errors.

4.9.2.3 Empirical model for message number

For the prediction of the coefficient of variation, it is useful to invert Eq. 4.99 to generate a model for message number as a function of protein number:

$$\mu_m = C_0^{-1/\alpha_0} \mu_p^{1/\alpha 0}, \tag{4.102}$$

$$= C_1 \mu_p^{\alpha_1}, \tag{4.103}$$

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

$$\alpha_1 \equiv 1/\alpha_0, \tag{4.104}$$

$$= 0.48 \pm 0.01, \tag{4.105}$$

$$C_1 \equiv C_0^{-1/\alpha_0}, \tag{4.106}$$

$$= 0.37 \pm 0.02, \tag{4.107}$$

where the uncertainties are the estimated standard errors.

4.9.2.4 Empirical model for translation efficiency

To generate an empirical model for translation efficiency, we started from the empirical model for protein number (Eq. 4.99), and then use Eq. 4.1 to relate protein number, message number, and translation efficiency:

$$\begin{aligned} \varepsilon &= \tfrac{\mu_p}{\mu_m}, && (4.108)\\ &= C_0 \mu_m^{\alpha_0 - 1}, && (4.109)\\ &= C_2 \mu_m^{\alpha_2}, && (4.110) \end{aligned}$$

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

$$\begin{aligned} \alpha_2 &= \alpha_0 - 1, && (4.111)\\ &= 1.07 \pm 0.04, && (4.112)\\ C_2 &= C_0, && (4.113)\\ &= 8.0 \pm 1.0, && (4.114) \end{aligned}$$

where the uncertainties are the estimated standard errors.

4.9.2.5 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. 4.103), and then substitute this into the statistical model prediction for CV_p^2 (Eq. 4.2):

$$\begin{aligned} \mathrm{CV}_p^2 &= \tfrac{\log 2}{\mu_m}, && (4.115)\\ &= C_0^{1/\alpha_0} \log 2 \cdot \mu_p^{-1/\alpha_0}, && (4.116)\\ &= C_3 \mu_p^{\alpha_3}, && (4.117) \end{aligned}$$

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

$$\begin{aligned} \alpha_3 &\equiv -1/\alpha_0, & (4.118)\\ &= -0.48 \pm 0.01, & (4.119)\\ C_3 &\equiv C_0^{1/\alpha_0} \log 2, & (4.120)\\ &= 1.9 \pm 0.1, & (4.121) \end{aligned}$$

where the uncertainties are the estimated standard errors.

4.9.3 Supplemental analysis of gene expression noise

The quantitative model for gene expression noise includes multiple contributions:

$$\mathrm{CV}_p^2 \approx \tfrac{1}{\mu_p} + \tfrac{\log 2}{\mu_m} + c_0, \quad (4.122)$$

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. [72, 73].

4.9.3.1 Inclusion of noise floor in the yeast analysis

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 ($\mathrm{CV}_p^2 = 0.1$) and is determinative of the noise associated with almost all essential genes [14, 72, 73], in yeast the noise floor is much lower ($\mathrm{CV}_p^2 = 0.01$) 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. 4.15).

In the null hypothesis, we will consider a model:

$$\eta_0(\mu_p; b, c) = \frac{b}{\mu_p} + c, \tag{4.123}$$

and an alternative hypothesis with an extra exponent parameter a:

$$\eta_1(\mu_p; a, b, c) = \frac{b}{\mu_p^a} + c. \tag{4.124}$$

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:

$$S_I(\boldsymbol{\theta}) \equiv \sum_i [\mathrm{CV}_{p,i}^2 - \eta_I(\mu_{p,i}; \boldsymbol{\theta})]^2 \tag{4.125}$$

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

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} S_I(\boldsymbol{\theta}), \tag{4.126}$$

with residual norm:

$$\hat{S}_I = S_I(\hat{\boldsymbol{\theta}}). \tag{4.127}$$

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

$$\Lambda \equiv 2 \log \frac{q_1}{q_0}, \tag{4.128}$$

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

4.9.3.2 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

$$\hat{\sigma}_\eta^2 = \tfrac{1}{2} \left\langle (\mathrm{CV}_{p,i}^2 - \mathrm{CV}_{p,i+1}^2)^2 \right\rangle_i = 6.3 \times 10^{-4}, \tag{4.129}$$

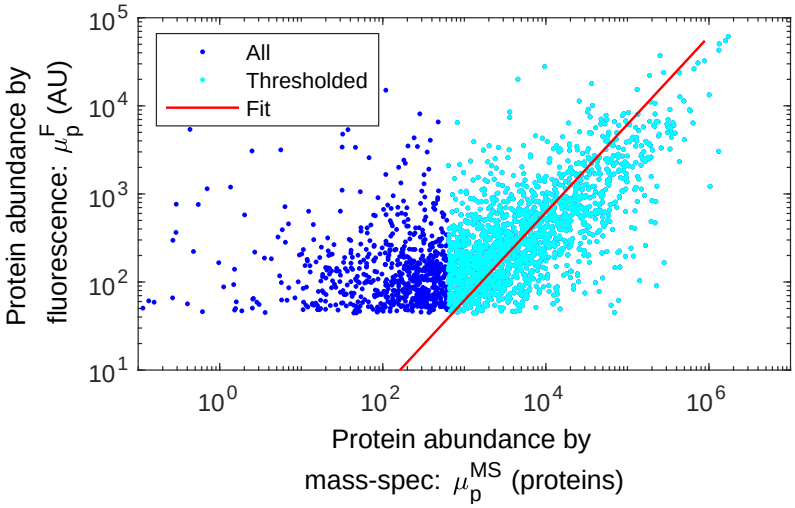


Figure 4.14: **Fit to rescale fluorescence intensity to protein number.** Protein abundance from flow cytometry fluorescence [13] as a function of mass-spectrometry scaled abundance [64]. The mass-spectrometry data was thresholded at 10 proteins, and then a linear fit was performed to find the multiplicative offset of 3.9, which was used to convert protein fluorescence AU to number.

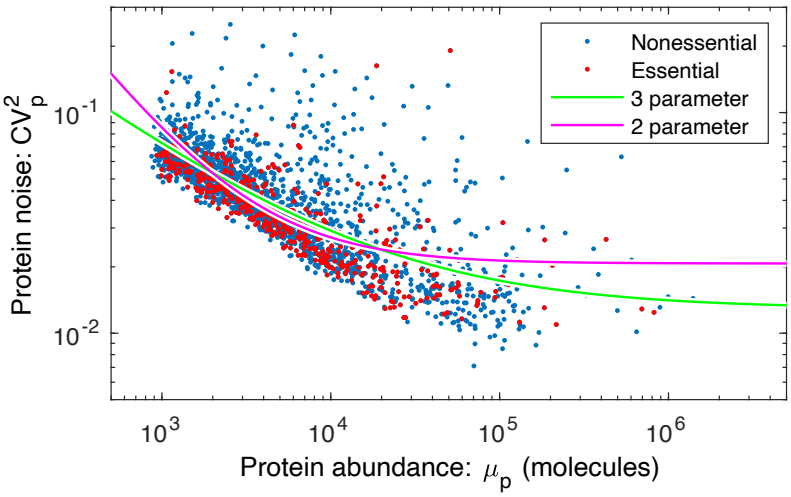


Figure 4.15: **Yeast noise fit against canonical noise model, with a noise floor.** Yeast noise data fit with the 2- (null hypothesis with μ_p^{-1} dependence) and 3- parameter (μ_p^a) models. The two-parameter model corresponds to the canonical noise model (Eq. 4.13) and fails to quantitatively fit the data.

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:

$$\Lambda = (\hat{S}_1 - \hat{S}_2)/\hat{\sigma}_\eta^2, \tag{4.130}$$

$$= 3.3 \times 10^4, \tag{4.131}$$

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

4.9.3.3 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:

$$\Lambda = N \log \tfrac{\hat{S}_1}{\hat{S}_2}, \tag{4.132}$$

$$= 1.6 \times 10^2, \tag{4.133}$$

where N is the number of data points. (Details of derivation are in Sec. 4.9.3.5.) In this case, the p-value can be computed assuming the Wilks' theorem (i.e., the chi-squared test):

$$p = 6 \times 10^{-36}, \tag{4.134}$$

again, strongly rejecting the null hypothesis.

4.9.3.4 Maximum likelihood estimates of the parameters

In the alternative hypothesis, the maximum likelihood estimate (MLE) of the empirical noise model (Eq. 4.124) parameters are (Fig. 4.15):

$$a = 0.57 \pm 0.02, \tag{4.135}$$

$$b = 3.0 \pm 0.5, \tag{4.136}$$

$$c = 0.013 \pm 0.001, \tag{4.137}$$

where the parameter uncertainty has been estimated using the Fisher Information in the usual way using the MLE estimate of the variance.

4.9.3.5 Details: Statistical details MLE estimate of the variance

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

$$h_I(\hat{\boldsymbol{\theta}}, \sigma^2) = \tfrac{N}{2} \log 2\pi\sigma^2 + \tfrac{1}{2\sigma^2} \hat{S}_I, \tag{4.138}$$

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

$$\partial_{\sigma^2} h|_{\hat{\sigma}^2} = 0, \tag{4.139}$$

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

$$\hat{\sigma}^2 = \tfrac{1}{N} \hat{S}_I. \tag{4.140}$$

Next we evaluate h at the variance estimator:

$$h_I(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \tfrac{N}{2} \left[\log 2\pi \tfrac{\hat{S}_I}{N} + 1 \right]. \tag{4.141}$$

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

$$\begin{aligned} \Lambda &= 2h_0(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - 2h_1(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2), & (4.142) \\ &= N \log \tfrac{\hat{S}_0}{\hat{S}_1}, & (4.143) \end{aligned}$$

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

4.10 Comments on the Hausser et al. analysis

Hausser et al. have previously performed a more limited analysis of the trade-off between metabolic load and gene-expression noise [15]. In this section, we will provide some more context into the differences between the two approaches.

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

$$k = k_0 - \tfrac{1}{2}|k''|(N_p - \mu_p)^2 - k_0\Lambda\mu_m, \tag{4.144}$$

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:

$$k = k_0 - \tfrac{1}{2}|k''|\sigma_p^2 - k_0\Lambda\mu_m, \tag{4.145}$$

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

$$k = k_0 - \tfrac{1}{2}|k''|\mu_p^2(\tfrac{\ln 2}{\mu_m} + C_0) - k_0\Lambda\mu_m, \tag{4.146}$$

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:

$$\hat{\mu}_m^2 = \tfrac{1}{2}\tfrac{|k''|}{k_0}\mu_p^2\tfrac{\ln 2}{\Lambda}, \tag{4.147}$$

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:

$$\mu_p \propto \hat{\mu}_m\sqrt{\Lambda}, \tag{4.148}$$

$$\hat{\varepsilon} \propto \sqrt{\Lambda}, \tag{4.149}$$

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,

$$\hat{\mu}_m \propto 1/\sqrt{\Lambda}, \tag{4.150}$$

$$\hat{\varepsilon} \propto \mu_p\sqrt{\Lambda}, \tag{4.151}$$

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:

$$\mu_p \propto \hat{\mu}_m^2\Lambda, \tag{4.152}$$

$$\hat{\varepsilon} \propto \hat{\mu}_m\Lambda, \tag{4.153}$$

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 [15]. At best, this is somewhat fuzzy logic since, as we demonstrate in the paper, Eq. 4.153 implies that the protein variance does not scale $\sigma_p^2 \propto \mu_p$!

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.

4.11 Bibliography

[1] T. W. Lo, H. K. J. Choi, D. Huang, and P. A. Wiggins, "Noise robustness and metabolic load determine the principles of central dogma regulation," *preprint*, Oct. 2023. DOI: 10.1101/2023.10.20.563172.

[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.

[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.

[4] 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.

[5] S. Pedersen, "*Escherichia coli* ribosomes translate *in vivo* with variable rate," *EMBO J*, vol. 3, no. 12, pp. 2895–8, Dec. 1984.

[6] R. Young and H. Bremer, "Polypeptide-chain-elongation rate in *Escherichia coli* B/r as a function of growth rate," *Biochem J*, vol. 160, no. 2, pp. 185–94, Nov. 1976. DOI: 10.1042/bj1600185.

[7] P. P. Dennis and H. Bremer, "Macromolecular composition during steady-state growth of *Escherichia coli* B/r," *J Bacteriol*, vol. 119, no. 1, pp. 270–81, Jul. 1974. DOI: 10.1128/jb.119.1.270-281.1974.

[8] A. L. Koch, "Overall controls on the biosynthesis of ribosomes in growing bacteria," *J Theor Biol*, vol. 28, no. 2, pp. 201–31, Aug. 1970. DOI: 10.1016/0022-5193(70)90053-6.

[9] S. Klumpp, M. Scott, S. Pedersen, and T. Hwa, "Molecular crowding limits translation and cell growth," *Proc Natl Acad Sci U S A*, vol. 110, no. 42, pp. 16 754–9, Oct. 2013. DOI: 10.1073/pnas.1310377110.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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. 18 010–18 017, Jul. 2020. DOI: 10.1073/pnas.2001507117.

[18] 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.

[19] 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.

[20] 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. 168 302, Oct. 2006. DOI: 10.1103/PhysRevLett.97.168302.

[21] H. G. S. Joseph W. Lengeler Gerhart Drews, Ed., *Biology of the Prokaryotes*. Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998.

[22] J. I. Steinfeld, J. S. Francisco, and W. L. Hase, *Chemical Kinetics and Dynamics*, 2nd. Prentice-Hall, 1999.

[23] H. K. J. Choi, K. J. Cutler, D. Huang, T. W. Lo, W. R. Will, and P. A. Wiggins, "In preparation," 2024.

[24] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014 420, 1 Jan. 2022. DOI: 10.1103/PhysRevE.105.014420.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] S. Ghaemmaghami *et al.*, "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–41, Oct. 2003. DOI: 10.1038/nature02046.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–3, Aug. 1970. DOI: 10.1038/227561a0.

[36] 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.

[37] 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.

[38] 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.

[39] D. Gillespie, "A rigorous derivation of the chemical master equation.," *Physica A*, vol. 188, no. 1, pp. 404–425, 1992.

[40] 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. eprint: https://doi.org/10.1021/j100540a008.

[41] 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.

[42] 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.

[43] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to mathematical statistics*. Pearson, 2020.

[44] 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," *Nat Commun*, vol. 14, no. 1, p. 1762, Mar. 2023. DOI: 10.1038/s41467-023-37456-2.

[45] P. Wang *et al.*, "Robust growth of *Escherichia coli*," *Curr Biol*, vol. 20, no. 12, pp. 1099–103, Jun. 2010. DOI: 10.1016/j.cub.2010.04.045.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] G. B. Arfken and H. J. Weber, *Mathematical methods for physicists; 4th ed.* San Diego, CA: Academic Press, 1995.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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. 11 005, Jul. 2019. DOI: 10.1038/s41598-019-47424-w.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] B. Alberts *et al.*, *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, 2015.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] M. Uhlén *et al.*, "Proteomics. Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1 260 419, Jan. 2015. DOI: 10.1126/science.1260419.

[67] V. E. Velculescu *et al.*, "Analysis of human transcriptomes," *Nat Genet*, vol. 23, no. 4, pp. 387–8, Dec. 1999. DOI: 10.1038/70487.

[68] G. M. Cooper, *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates 2000, 2000, ISBN: 0-87893-106-6.

[69] 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.

[70] 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, 495–504.e5, May 2017. DOI: 10.1016/j.cels.2017.03.003.

[71] 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.

[72] 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.

[73] 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. 12 795–800, Oct. 2002. DOI: 10.1073/pnas.162041399.

Chapter 5

Metabolic homeostasis, oscillations, and instability

Content for this chapter first appeared in: P. A. Wiggins, *Metabolic homeostasis, oscillations, and instability*, Proposal for National Science Foundation Physics of Living Systems Grant: NSF-Phys-2412326, May 2024.

Author contributions: D.H. and P.A.W. conceived the research, conducted analyses, and wrote the proposal.

A note on the structure: This chapter primarily functions as a road map for future research on the topic of metabolic oscillations. It is therefore structured differently, with sections composed of research Aims and Sub-aims instead of Results and Discussions.

Abstract

Cellular proliferation is dependent on metabolic homeostasis: the robust maintenance of the concentrations of a host of critical metabolites, including nucleotides. Mechanistically, homeostasis is the result of a complex network of regulatory feedback control. Although many specific interactions have now been characterized, important examples exhibit an unexpected phenomenon: temporal oscillations. The focus of this chapter is on understanding these metabolic oscillations: characterizing their dynamics and determining their mechanism and biological rationale.

5.1 Introduction

The function of cellular enzymes is dependent on the maintenance of critical metabolite levels (metabolic homeostasis) [2]. The traditional view is that metabolite levels (e.g., ATP

abundance) are tightly maintained since they are critical to nearly every biological process; however, this hypothesis has never been systematically tested and recent evidence demonstrates significant cell-to-cell variation [3], as well as dynamic fluctuations in metabolite levels [4, 5]. The existence of these fluctuations raises questions about whether there are fundamental limits on the precision of cellular homeostasis. It also prompts consideration of whether these fluctuations have a biological rationale or functional motivation.

5.1.1 Feedback-based regulatory control

Mechanistically, homeostasis is the result of a complex network of regulatory feedback control in which the levels of essential metabolites are sensed, and this state feeds back upon the metabolic process itself, affecting the rates at which individual metabolites are created and metabolized to robustly maintain metabolic homeostasis [2, 6–10]. Feedback-based control is found in many human-engineered systems (Fig. 5.1AB) and has been extensively studied in this context [11, 12]. In these engineered systems, desirable attributes of feedback control include stability, accuracy, rapidity, and the minimization of overshoot (Fig. 5.1B) [12]. Different applications call for different relative weighting of these attributes. For instance, the shock absorbers of a conventional car are tuned to be critically damped, eliminating oscillations (i.e., bouncing) of the car in response to perturbations (potholes) and improving passenger comfort; however, high-performance race cars have stiffer suspension that reduces the amplitude of spring deformation (i.e., increasing accuracy) at the expense of oscillations (i.e., increasing overshoot) and therefore reducing passenger comfort [13]. For what attributes has evolution optimized physiological homeostasis? Our naive intuition is that cellular regulatory control should be characterized by high-stability and overdamped dynamics which minimize overshoot and oscillations [14]; however, relatively few systematic investigations have yet been made to test this hypothesis in the context of cellular regulation [14]. In fact, contrary to our expectations, there are a number of important examples of metabolic

oscillations [11, 15–18] and we have recently reported a potentially striking example: dNTP pool oscillations [4]. (See Fig. 5.1C.)

5.1.2 Background on dNTP metabolism

All four aims describe a specific metabolic pathway, dNTP synthesis, which is summarized below: Nucleotides are synthesized by a complicated network of enzymes in the nucleotide metabolic pathway, shown schematically in Fig. 5.2A. These complicated synthesis pathways are regulated to achieve metabolic homeostasis by multiple layers of regulatory control (Fig. 5.2B). Under aerobic growth conditions, the protein gene products of *E. coli* genes *nrdAB* form the enzyme RiboNucleotide Reductase (RNR), which reduces nucleoside diphosphates (NDPs) to form deoxynucleoside diphosphates (dNDPs), the key precursor for DNA synthesis in DNA replication [19]. Because this step is the rate-limiting step in the production of deoxynucleoside triphosphates (dNTPs) [20] and the dNTP pool levels are rate-limiting in DNA synthesis [20, 21], the regulation of RNR is believed to play a key role in determining replication and cell-cycle dynamics [22]. RNR is subject to multiple levels of regulatory control: (i) The enzyme itself is subject to allosteric regulation [23–27]. (ii) Transcriptional regulation plays a key role in regulation as well [22]. Although a transcriptional burst of RNR is produced roughly coincident with replication initiation, the mechanism for this regulation is still poorly understood [22, 28–30]. The *nrd* promoter is regulated by DnaA, as well as being negatively regulated by the dNTP levels, although the mechanism is not yet understood [22].

5.1.3 Significance

The observed oscillatory phenomena motivate a more detailed investigation into the characteristics of homeostasis regulatory control and question our assumptions about what control attributes are important to the cell. Although regulatory networks have been modeled using chemical kinetics for 60 years [31], important questions relating to the stability of these networks are essentially unexamined. Learning what attributes are

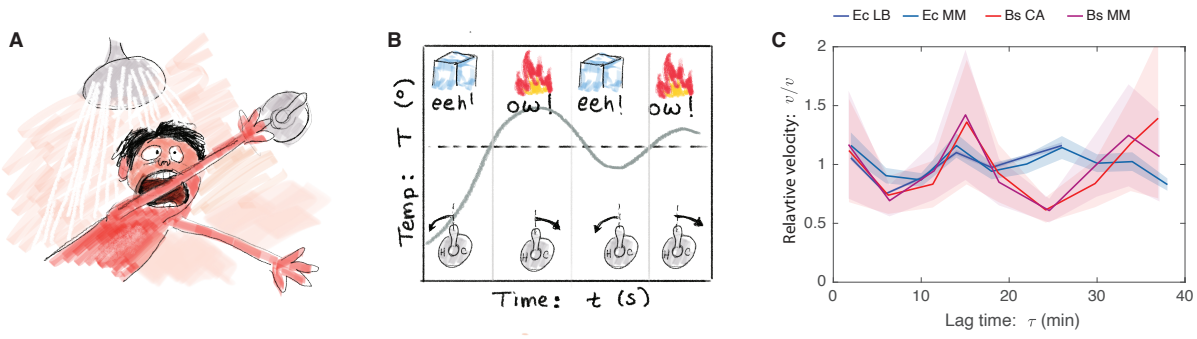


Figure 5.1: **Characteristics of regulatory feedback control. Panel A: Some feedback is required!** Most of us apply feedback-based regulatory control every day. Although I know the approximate position of the shower tap, I always need to adjust the tap to achieve the ideal temperature. **Panel B: Overshoot in regulation.** We apply negative feedback to control the water temperature. If the temperature is too high (low), we adjust the controls to decrease (increase) the temperature; however, we typically overcorrect due to the finite response time, leading to the phenomena of *overshoot* and oscillations. In a well-designed control scheme, these oscillations are quickly damped and the optimal conditions (e.g., temperature) are achieved. **Panel C: Evidence for the overshoot phenomena in metabolic regulation.** Temporal replication-fork-velocity oscillations are observed in dNTP-limited reaction conditions, consistent with metabolic oscillations in key metabolites.

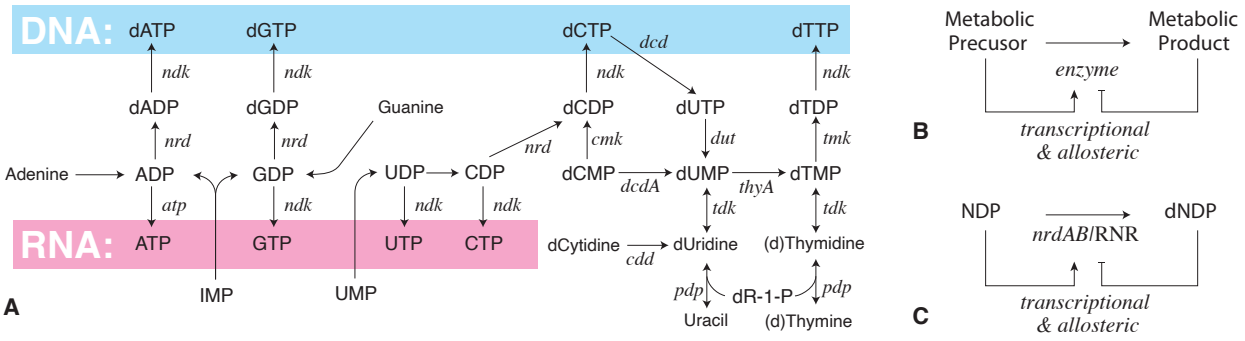


Figure 5.2: **Schematic and model of nucleotide metabolism in *E. coli*. Panel A: Nucleotide metabolism in *E. coli*.** The last few steps of nucleotide metabolism are shown schematically, including metabolites (sans serif) as well as enzyme genes (italic serif). *Homeostasis requires the active regulation of enzyme activity (not shown).* **Panel B: Homeostatic regulation.** Homeostatic regulation is believed to be achieved by negative feedback: The increase (decrease) in metabolic product (precursor) downregulates the catalytic enzyme, while the decrease (increase) in product (precursor) upregulates the enzyme. **Panel C: dNTP homeostasis.** The genes *nrdAB* and their gene product RNR, are regulated by both transcriptional and allosteric mechanisms.

optimized in homeostatic regulation is a challenge of central importance to understanding not only systems biology, but to understanding the regulatory imperatives that necessitate the observed transcriptional and allosteric regulation. The generic observation of oscillations in regulatory systems would fundamentally change our understanding of how the cell functions. Confirmation of these generic oscillations would have far-reaching implications to many important biological questions, not only in bacterial cells, which will be the focus of our experimental investigations, but in eukaryotic cells as well [32]. Some examples of metabolic oscillations are already known in this context [11, 15–18]; however, their biological rationale is still hotly debated [11]. The identification, characterization, and understanding of such seemingly exotic phenomena in *E. coli* could illustrate the unexceptional nature and, potentially, the necessity of oscillations in cellular regulatory control. The fundamental biophysical questions that animate this proposal are: (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?

5.2 *Aim 1: Modeling—Identify the determinants of regulatory oscillations and instability*

Motivation: Motivated by the observation of fork-velocity oscillations (Aim 2), we hypothesized that the finite response-time associated with dNTP homeostasis is responsible. This mechanism is familiar to anyone who has ever stepped into the shower and struggled with adjusting the water temperature. (See Fig. 5.1A.) Due to the delay in the response between control manipulation (turning the tap) and its response (a change in water temperature), one often overcompensates, leading to decaying (hopefully) temporal oscillations in temperature. (See Fig. 5.1B.) In this first aim, our initial focus will be on building a quantitative understanding of the phenomenology of homeostatic regulation to determine whether the observed oscillations are predicted by kinetic models based on our

current understanding of the regulatory interactions. We will then try to generalize these results to analyze homeostatic feedback more generally.

Significance: Although the canonical understanding of homeostatic regulation is that it ensures that the levels of essential metabolites are maintained in a narrow physiological range [33], this assumption does not appear to be wholly consistent with emerging experimental evidence. Instead, these studies reveal that the cellular environment is subject to significant fluctuations (or oscillations) of critical metabolites [5, 15]. In light of these experimental observations, a thorough theoretical analysis is necessary to explore the behavior of these regulatory networks. Understanding the fundamental limitations of robust regulatory control networks may have fundamental implications for cellular function and provide new insights into cellular biophysics [11].

Preliminary work—Approach: We modeled the regulatory network using chemical-kinetics rate equations. We began with the analysis of minimal models where protein expression is a single-step process (Fig. 5.3Aa) and then incorporated a number of realistic refinements, including transcription, translation, the protein maturation processes, and allosteric regulation (i.e., transition of enzymes between active and inactive states, Fig. 5.3Ab). For *E. coli*, representative rate constants for most of these processes are known. In each case, we determined steady-state solutions in metabolite space. The nonlinear rate equations were then linearized by analyzing small perturbations in the reactant vector $\delta\boldsymbol{N}$, leading to the linear-coupled rate equation:

$$\delta\dot{\boldsymbol{N}} = \boldsymbol{K}\delta\boldsymbol{N}, \tag{5.1}$$

where $\boldsymbol{K}$ is an effective rate matrix. This linear ODE was solved using an eigenvalue approach. The characteristics of the homeostatic response can then be inferred from the spectrum of eigenvalues: Modes are stable if the rate eigenvalue has a negative real part and unstable if the real part is positive. Rates with imaginary parts result in oscillating solutions. (See Fig. 5.3BC.)

Preliminary results: Our preliminary investigations provide some important insights into the rationale for oscillations: (i) To what extent do the observed fork-velocity oscillations

match the predictions of the models? The models make robust predictions about the phases of various network components, including for fork-velocity oscillation, and predicts the phase observed in experiment. (See Aim 2.) Regarding oscillation period, our preliminary calculations suggest that the observed dynamics are consistent with a transcriptional response, since it requires roughly 10 minutes to express protein [31], consistent with the observed period of 12–15 minutes [4]. (See Aim 2.) (ii) Why isn't the regulatory response overdamped (i.e., oscillation free)? There is a trade-off between high precision and rapidity versus minimizing oscillations and overshoot: Networks with precise and rapid regulatory response are subject to high-frequency oscillations. Overdamped networks, free from oscillations, are subject to large-amplitude and long-lived deviations from homeostasis. (See Fig. 5.3B and the race car analogy from the introduction.) (iii) To what extent do we predict that nucleotide metabolism is affected as a whole, rather than just dNTP levels? Our preliminary models predict a nucleotide-metabolism-wide response to replication initiation, going far beyond transcription of the *nrdAB* genes.

5.2.1 Sub-aim 1.1: What are the determinants of regulatory oscillations and instability?

Motivation: An important but putative takeaway from our analysis of nucleotide homeostasis is that achieving stability is non-trivial and oscillations are generic. (See Fig. 5.3B.) In this sub-aim, we will attempt to identify the determinants of the regulatory network structure that give rise to oscillations and instability [12], using nucleotide metabolism as a motivating example. Given physiological parameter ranges, are there limits on the speed and precision of homeostatic response? To what extent do biological regulatory systems saturate these constraints? Are oscillations endemic for strongly-regulated metabolites? We hope to identify the *emergent principles of regulatory network phenomenology and dynamics.*

Approach: We will start by making a more systematic analysis of dNTP homeostatic regulation. We will establish what functions of rates are determinative of stability and use

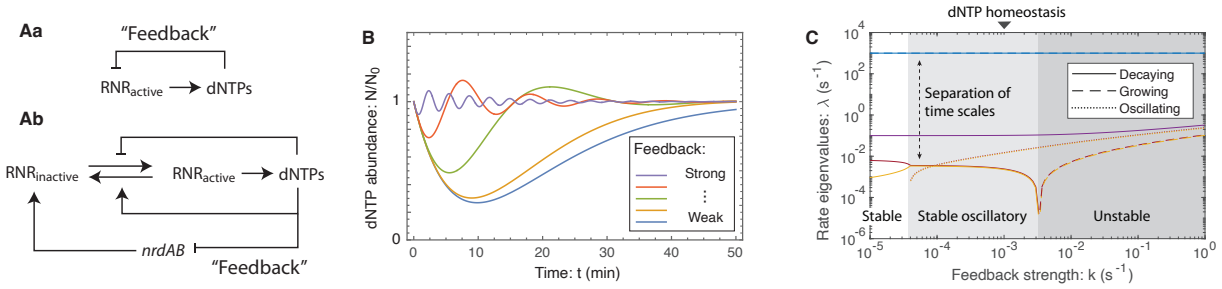


Figure 5.3: **Homeostatic model results. Panel A: Successive levels of complexity in the homeostatic model.** Panel Aa shows a minimal model with only two components (RNR and dNTPs) which is always stable, but can oscillate. Panel Ab shows a higher dimensional model which includes a more detailed and realistic mechanism of the homeostatic network, including transcription and translation, etc. **Panel B: Trade-off between overshoot, precision, and rapidity.** The oscillation period is controlled by the strength of the regulatory response. The homeostatic control can be overdamped (blue) to eliminate overshoot (i.e., oscillations); however, the size of dNTP pool depletion is larger and response time is slower. **Panel C: The spectrum of eigenvalues for a higher dimensional model.** Achieving rapid but stable regulation in higher dimensional models is subtle. For the weakest (i.e., slowest) feedback, the response is overdamped and is both stable and non-oscillating. As the strength of the feedback increases, the network first becomes oscillatory (light gray) and then unstable (medium gray). For dNTP homeostasis, we hypothesize that the regulatory network is positioned relatively close to the stability threshold.

these expressions to place limits on the feedback strength, response time, etc. After dNTP homeostasis has been analyzed in detail, we will attempt to generalize this analysis to other networks and consider whether there are generic principles of homeostatic control.

Preliminary and anticipated results: Although these networks are high-dimensional dynamical systems and therefore not generically amenable to an analytic approach, using biological relevant parameters, the separation in time scales appears to give rise to significant simplifications. Our preliminary investigations have led to a consistent qualitative picture for the phenomenology of networks: Eigenvalues are generically complex, implying the oscillatory responses are generic. Perhaps more surprisingly, in networks containing the negative-feedback control necessitated by homeostasis (i.e., to maintain a target metabolite level), eigenvalues with a positive real part are also generic, implying that these networks are unstable. Increasing the strength (i.e., the rapidity) of the network response appears to make the instability more severe, as does adding more realistic details (e.g., explicitly modeling both transcription and translation). (See Fig. 5.3C.) There are a number of biologically relevant network topologies for which we can determine limits on feedback strength in order to maintain network stability. We hypothesize that homeostatic regulation may be close to the stability threshold as a consequence of the desirability of rapid and precise control. (See the arrow at the top of Fig. 5.3C.) This proximity to the stability threshold implies that oscillations in metabolite levels should be generic.

5.2.2 Sub-aim 1.2: Explore the putative role of small RNA and proteins in stabilizing regulation

Motivation: A potentially important but poorly understood feature of many biological regulatory circuits is the presence of small RNA or proteins that are transcribed as part of or in conjunction with a larger ORF (e.g., [34]). For a variety of technical reasons, these factors have been difficult to identify and study. Although these factor have been implicated in a host of different processes, it is clear that many play an important regulatory role

[34]. One potentially intriguing possibility is that they play a critical role is stabilizing the regulatory feedback circuitry.

Approach: To explore the potential impact of small RNAs and proteins, we will add these factors to the regulatory model. We will explore their effect on regulatory stability. Our goal will be to determine when such factors could stabilize otherwise unstable regulatory feedback and use these conditions as a method for identifying candidate networks to identify novel regulators.

Preliminary and anticipated results: We have made some preliminary models to explore the plausibility of this mechanism. These models suggest that the mechanism can significantly reduce the overshoot phenomenon that leads to instability by reducing the delay time between generating a transcriptional response and detecting its presence.

5.3 Aim 2: Test the dNTP-oscillation model by characterizing fork-velocity oscillations

Motivation: How can high-temporal-resolution measurements of dNTP levels be made? The replication fork velocity (i.e., the elongation rate measured in bp/s) is a direct reporter of the relative abundance of dNTPs in a nucleotide-limited reaction. 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 [21]. Are fork velocities constant during the cell cycle? We have recently measured the fork velocity in three evolutionarily-divergent bacterial species and report fork-velocity oscillations of 50% with a form exactly analogous to those predicted by modeling [4]. (See Fig. 5.4C.) As demonstrated in our recent paper, the quantitative characterization of these oscillations by next-generation sequencing is highly tractable. This approach circumvents the necessity of synchronizing the cells, making this measurement an attractive approach to test the predictions of dNTP homeostasis models.

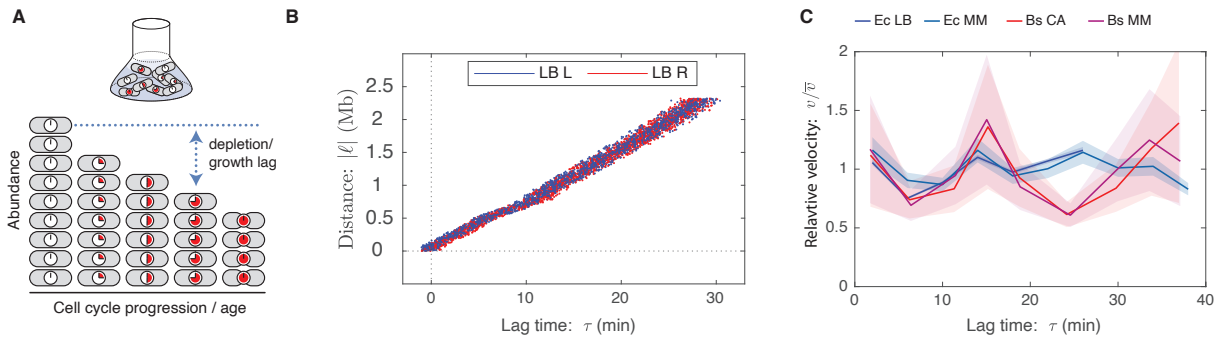


Figure 5.4: **Measuring replication fork velocity by lag-time analysis. Panel A: Measuring marker frequency.** Cells are harvested from a log-phase culture. Genomic marker frequency is determined by deep sequencing. **Panel B: Lag-time analysis** is performed to convert marker frequency into locus replication time. Average timing can be determined to a precision of seconds. Lag-time analysis shown for *E. coli* growing on LB. **Panel C: Measured fork-velocity oscillations in two organisms.** The fork velocity is then computed, in units of bp/s, from the locus replication timing. Fork-velocity oscillations of the same qualitative character are observed in at least three bacterial species: *E. coli*, *B. subtilis* and *V. cholerae*. The period of oscillation is observed to be 10-15 min.

Significance: dNTP synthesis is a highly-regulated pathway of central importance to cell fitness. The observation and characterization of regulatory oscillations in this model regulatory system is extremely important, since their discovery would suggest that the regulatory oscillation phenomenology is generic in regulatory networks, even if it may be much more difficult to characterize in other pathways. Demonstration of oscillations in multiple divergent species under various growth conditions suggests that regulatory oscillations may be a generic and fundamental feature for all biological systems.

Preliminary work. Two lines of evidence already strongly support the existence of dNTP oscillations: (i) Measurements of mutation rate, as function of locus position relative to the origin of replication, show a symmetric wave-like dependence on the distance from the origin [35, 36]. Since elevated dNTP pool levels result in increased mutation rates, it has been proposed that oscillations in the dNTP levels could account for these mutation rate waves [35, 36]. (ii) Furthermore, our own measurement of temporal dependence of

the fork velocity in living cells [4], as well as other reports [37], show oscillations in the fork velocity in time, not genomic position [4, 38]. We observed temporal oscillations not only in *E. coli*, under multiple growth conditions, but in two other evolutionarily-distant species: *Vibrio cholerae* and *Bacillus subtilis*. Why do these velocity oscillations implicate dNTP pool oscillations? Previous measurements demonstrate that the rate of replication elongation is nucleotide limited, a model strongly supported by our own measurements [4]. Nucleotide-limited dynamics implies that these oscillations are caused by changes in the nucleotide levels. Furthermore, the shape of the oscillation (e.g., phase) of the fork velocity matches model predictions.

Approach: We have recently developed and described the *Lag-Time Analysis* approach for measuring replication fork dynamics *in vivo* [4]. In short: Cells are grown in asynchronous culture and harvested in early log-phase growth. (See Fig. 5.4A.) The genomic DNA is extracted and prepared for next generation sequencing. The marker frequency is then measured by mapping sequencing reads to the reference genome. (See Fig. 5.4B.) The locus replication timing is then determined from the relative locus abundance:

$$\tau(\ell) = \gamma^{-1} \ln \tfrac{N(0)}{N(\ell)}, \tag{5.2}$$

where γ is the culture growth rate and $N(\ell)$ is the abundance of a locus at position ℓ and $N(0)$ is the abundance of the origin of replication [4, 38–41].

Does the stochasticity of cell-cycle timing affect lag-time analysis? Eq. 5.2 is predicted by a model with deterministic cell-cycle timing; however, single-cell imaging clearly reveals the replication dynamics are stochastically timed [42]. How does this stochasticity affect the measurements of the fork velocity? Although this approach was originally formulated in the context of deterministic cell-cycle dynamics, we have generalized this approach for use in the context of stochastic cell-cycle dynamics [4, 38]. We have shown that the exponential mean of stochastic lifetimes correspond exactly to an effective deterministic lifetime [38]. This approach allows us to measure average durations relative to the time of replication initiation [4]. After establishing the approach, we have used

next-generation sequencing to measure replication timing with a precision of seconds and the fork velocity in absolute units (bp/s) [4]. (See Fig. 5.4C.) This approach has the ability to make precise replication timing measurements *independent of our ability to synchronize cell growth* and *independent of the stochasticity of cell-cycle timing or growth rate* [38].

5.3.1 Sub-aim 2.1: Increase the temporal resolution and precision of fork velocity measurement

Motivation: Our original measurements focused on a wide range of determinants of fork velocity and were *not* optimized to capture fork-velocity oscillations with high sensitivity and genomic resolution. In this sub-aim, we will optimize these measurements by increasing their resolution and sensitivity.

Approach: Two mechanisms limit the resolution and sensitivity of the lag-time analysis approach: Poisson error due to finite sequencing depth and systematic error [4]. (i) Increasing sequencing depth reduces the Poisson error; however, reducing the systematic error is more subtle. (ii) Two key changes in the protocol for measuring marker frequency are expected to reduce systematic error: using non-amplifying library preparation and harvesting the cell in very early log-phase growth (OD_{600} <0.05) [43]. We will investigate both proposed methods to reduce noise (i-ii). (We describe a detailed protocol for characterizing the resolution in Ref. [4].)

Preliminary data and anticipated results: Our current analysis suggests that mechanism (ii) currently limits our resolution. *We emphasize that our existing protocol is precise enough to capture fork-velocity oscillations and the phenotypes described in the other sub-aims.*

5.3.2 Sub-aim 2.2: Test role of RNR in limiting fork velocity by inhibition

Motivation: A key determinant of the period and amplitude of the oscillations is the rapidity with which RNR can respond to dNTP depletion. A reduction in the activity of

RNR is predicted to increase both the period and amplitude of the dNTP oscillations, as well as decreasing the average fork velocity due to nucleotide-limiting [4, 21]. (See Fig. 5.3B.)

Approach: Hydroxyurea (HU) is a highly-specific inhibitor of RNR activity [44]. Treatment of cells with sub-Minimum-Inhibitory-Concentration (MIC) concentrations of HU reduces the RNR activity while leaving the cells viable and capable of log-phase growth. We will then measure the fork-velocity oscillations using lag-time analysis for a range of HU concentrations.

Anticipated results: We predict that this reduction in RNR activity will still result in oscillating fork velocity, but it will be more consistent with the *weak response* (yellow, green, and blue curves) shown in Fig. 5.3B. HU treatment should also reduce the average fork velocity. If neither effect is observed, it suggests that RNR activity is not limiting, in contrast to previous reports [22].

5.3.3 Sub-aim 2.3: Test feedback-regulation model for oscillations

Motivation: Our models predict that transcriptional regulatory feedback plays a key role in generating fork-velocity oscillation; therefore, we can directly test this hypothesis by circumventing the endogenous regulation.

Approach 1: First, we will revisit the question of whether fork velocity is dNTP limited. We will express *nrdAB* from an ectopic IPTG-inducible promoter using an existing construct: the plasmid pSMG7 [22]. We will then knock out the endogenous *nrdAB* locus. Using this mutant strain, we will characterize the changes in fork velocity using lag-time analysis, quantifying the expression level of *nrdAB* by both RT-qPCR and western and measuring average dNTP pool levels by LC-MS. (See Sub-aim 4.)

Anticipated results: Overexpression of RNR should increase pool levels and fork velocity, whereas underexpression of RNR should decrease pool levels and fork velocity, consistent with previous finding describing their effects (e.g., [21]) and our own recent measurements [38].

Approach 2: Our model is consistent with oscillations being driven by transcriptional-regulatory feedback. To test this model, we will use our plasmid-based *nrdAB* mutant strain to circumvent the endogenous regulation, and measure the temporal dependence of the fork velocity.

Preliminary data and anticipated results: The wild type temporal dependence of the oscillation is shown in Fig. 5.4C. If transcriptional regulation is responsible for the observed oscillations (as hypothesized), the exogenous promoter driving *nrdAB* should eliminate fork-velocity oscillations.

Approach 3: We have also hypothesized that the proteases ClpXP or Lon may play a critical role in RNR abundance by targeted degradation (e.g., [45]). To test this degradation hypothesis, we will measure the fork velocity in *clpXP*, *lon* and *clpXP lon* backgrounds (knockout strains were supplied from Ref. [46]).

Anticipated results: If ClpXP or Lon target RNR to increase the strength of the feedback, we would predict that the *clpXP lon* mutant should have a slower response time and therefore a longer period of oscillation as well as increasing the fork velocity due to higher RNR protein levels and accompanying elevated dNTP pool levels [22].

Approach 4: RNR is also regulated by an allosteric mechanism, in addition to a transcriptional mechanism. To test the role of allosteric regulation, we will exploit mutants in the allosteric regulatory pathway. Detailed activity-based characterization has been performed on a wide range of RNR mutants [47]. For the purpose of this experiment, we will select a number of different mutants which target the activation domain in order to disrupt allosteric regulation.

Anticipated results: If activation/inactivation is required to generate oscillations, the loss of function of the activation domain should eliminate (or at least suppress) fork-velocity oscillations.

5.4 Aim 3: Test transcriptional-regulation model by characterizing cell-cycle-dependent transcription

Motivation: Our models predict that the transcriptional regulation of *nrdAB* plays a key role both in the maintenance of the dNTP pool levels, as well as the generation of the regulatory oscillations. A key test of the model is therefore characterizing the cell-cycle-dependent transcription of *nrdAB*, as well as other key players in the network. Our model predicts that two bursts of *nrdAB* should be observed. The first will be synchronized with initiation. The second should be observed 15 minutes later as an oscillatory correction to regulatory overshoot. In addition to making an unambiguous prediction for cell-cycle-dependent *nrdAB* transcription, our systems-scale analysis suggests that the abundances of many more important metabolites may respond as dNTPs are synthesized. For instance, since NDPs are reduced to synthesize dNDPs, and the levels of NDP must be rapidly replenished as they are depleted, it is natural to hypothesize that nearly every enzyme in the nucleotide synthesis pathway is transcribed.

Significance: The current canonical view is that cell-cycle-dependent transcription is the exception rather than the rule in *E. coli*, since it can undergo overlapping rounds of replication in rapid growth. The observation that many genes undergo cell-cycle-dependent transcriptional regulation would therefore be highly significant. In addition, the classification of genes by high-temporal-resolution expression profiles could provide new insights into the mechanism of gene regulation by identifying groups of genes with similarities in their temporal expression profiles.

Existing support for cell-cycle dependent expression: In his influential book on growth and divisions, S. Cooper forcefully argued *against* the existence of any cell-cycle-dependent regulation in *E. coli* and this view appears to widely accepted in the research community [48]; however, subsequent experimental analyses of a number of *E. coli* messages and proteins strongly support the existence of cell-cycle-dependent regulation, including two genes of particular interest in our model: *nrdAB* [49, 50]. In addition to this

evidence from *E. coli*, it is also helpful to consider evidence from *Caulobacter crescentus*, a well-studied bacterial-cell-cycle system. M. Laub and coworkers have characterized the cell-cycle dependence transcriptome-wide [51]. The observed cell-cycle dependence is exactly analogous to what our model predicts should be observed in *E. coli*: A burst of transcription of nucleotide metabolism enzymes occurs roughly coincident with replication. (See Fig. 5.5B.) More recently, the analysis of *Mycobacterium tuberculosis* expression has shown analogous cell-cycle temporal patterning [52]. Our hypothesis is therefore that our RNA-Seq experiments in synchronized *E. coli* cells will reveal that a comparable number of genes show cell-cycle dependence in *E. coli*, especially in slow growth conditions which minimize the overlapping nature of the C period (i.e., replication).

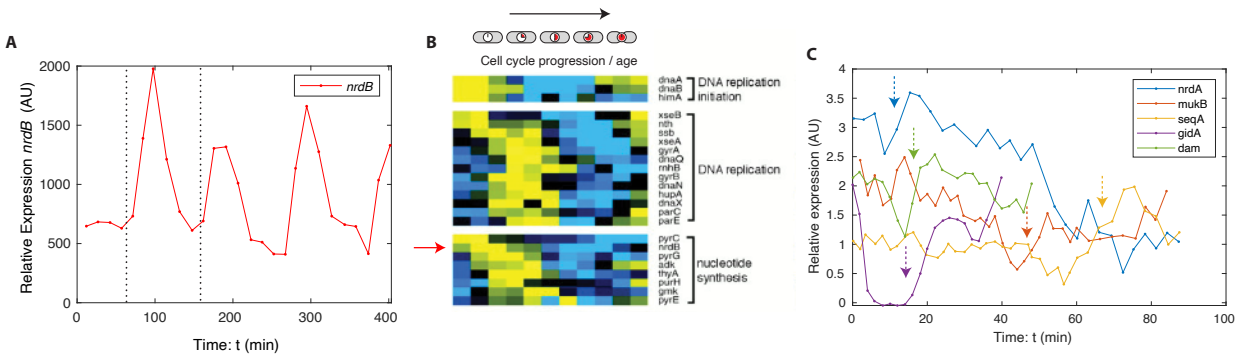


Figure 5.5: **Evidence of cell-cycle-dependent expression. Panel A: *nrdAB* expression is cell-cycle dependent.** Low-time-resolution measurements confirm a pulse of *nrdB* transcription roughly coincident with replication, as predicted. The dashed lines show the position of a complete cell cycle. (Data from Ref. [49].) **Panel B: Cell-cycle dependent expression profile.** In *Caulobacter*, other enzymes in the nucleotide-synthesis pathway show a similar temporal profile to *nrdB* (red arrow), consistent with the predictions of a pathway-wide induction. (Panel adapted from Ref. [51].) **Panel C: Cell-cycle dependent transcription of 5 genes in *E. coli* [53].** Cells are synchronized using the *dnaC2* allele. After release, the transcription of many genes, including *nrdA*, show distinct temporal signatures. The dotted arrows show putative regulatory overshoot dynamics.

5.4.1 Sub-aim 3.1: Test RNR-expression oscillation hypothesis by fluorescence imaging

Motivation: One of the most direct methods to resolve cellular dynamics is by fluorescence imaging [54]. The challenge in this context is that the dynamics of the oscillation are quite fast (period $\approx$ 15 min), which is a fraction of the maturation time of wild-type GFP (60 min) under our imaging conditions [55]. The Nyquist criterion demands a time resolution of roughly 8 minutes. Fortunately, the fastest-folding alleles barely satisfy this condition, making fluorescent imaging a promising approach to resolving oscillations [55].

Approach: We will quantitatively measure protein expression dynamics in single cells by imaging fluorescent fusions to NrdA and NrdB. The cell-cycle dynamics of single-cell protein expression will be characterized by full-cell-cycle imaging [54]. *Strain construction:* S. Xie and coworkers previously constructed a chromosomal functional fusion to NrdA at the endogenous locus [56]; however, the Venus variant that the Xie lab used in their collection takes 25 minutes to maturate, five-fold the maturation time of alternative fast-maturating alleles [55] and two times the length of the period of the predicted oscillations. We will therefore reconstruct this fusion using the mVenus NB allele, with a reported maturation time of 4.7 minutes at 32° and 4.1 minutes at 37° [55]. We will use an identical linker to that successfully used by Taniguchi et al. [56].

Preliminary data and anticipated results: Before learning that Taniguchi et al. [56] had used a slow-maturation allele of Venus, we imaged their strain. The endogenous expression levels were high enough that we could quantitate the protein expression level in time-lapse imaging; however, no unambiguous temporal oscillations were observed. In retrospect, this result should have been expected due to the extremely slow maturation time. We expect the mVenus NB allele to mature fast enough to resolve oscillations if they are observable.

5.4.2 Sub-aim 3.2: Optimize cell synchronization protocol to test hypotheses using bulk assays

Motivation: A wide range of bulk biochemical assays can measure the cell-cycle dynamics if the cell culture can be synchronized to sufficient precision. Two classes of synchronization approaches exist: *arrest-based* and *baby-cell-enrichment-based* methods. In the **cell-cycle-arrest-based approaches**, the cell cycle is arrested at initiation and then released. Ideally, cells begin the C period (replication) in synchrony. These approaches include the treatment of bacterial cells with serine hydroxamate (SHX) [57] and the use of a number of temperature sensitive alleles (*dnaC2* and *dnaC28* [58], as well as *dnaA46* [59]). The downside of these strategies is that cell-cycle arrest may significantly perturb cell physiology. Alternatively, **baby-cell-enrichment-based** approaches or *baby machines* use a number of different approaches to purify the newborn cells [60–63]. These approaches have the advantage that they do 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) [64], they may have poorer synchronization of replication.

Approach: To measure the efficiency of cell synchronization, we will exploit the visualization of a fluorescent fusion to a replisome protein: the beta-clamp (DnaN) [65]. This fusion is expressed from the endogenous promoter and has minimal effect on function. During ongoing replication, the DnaN forms a punctate focus in close proximity to the replication fork, whereas the protein has a diffuse localization in non-replicating cells [42, 64, 66]. Cells will be synchronized, then released. The proportion of replicating cells will be determined as a function of time since the release by fluorescence microscopy.

Preliminary data and anticipated results: We have preliminary synchronization data from both the *dnaC2*- and SHX-based approaches. Our expectation is that these approaches will lead to more precise replication initiation synchronization than the baby-machine-based approach.

5.4.3 Sub-aim 3.3: Test *nrdAB*-transcription oscillation hypothesis

Motivation: A burst of *nrdAB* transcription coincident with replication initiation is predicted by our model as well as a secondary pulse of transcription, due to the oscillation of dNTP levels, roughly 15 minutes after the first. Previous studies strongly support *nrdAB* transcription roughly coincident with replication [49, 53]; however, increased time resolution is required to test this multiple-pulse hypothesis.

Approach 1: To characterize the dynamics of *nrdAB* transcription, we will first assay mRNA levels of *nrdAB* using RT-qPCR in synchronized cells. To avoid the confounding effects of multi-fork replication, cells will be grown in minimal media (M9 glycerol w/o casamino acids, 30°C), in which replication cycles are non-overlapping [42, 64, 66]. The cells will be synchronized using two approaches (*dnaC2* and baby-machine approaches). The cell fractions will then be collected on time intervals of 2 minutes (the expected period is $T = 15$ min). The mRNA from each fraction will then be analyzed by RT-qPCR.

Preliminary data and anticipated results: Published results are already broadly consistent with our model: Refs. [49, 53] report that *nrdAB* is expressed coincident with replication using two different baby-cell-enrichment-based approaches. With the higher temporal precision achievable with *dnaC2* synchronization, Ref. [53] has also published data *consistent with the existence of transient oscillations with a period of roughly 15 min* (Fig. 5.5C), as predicted by our model; however, this observation was not the focus of these measurements and was not carefully characterized [53].

Approach 2: Our system-scale model predicts that replication initiation and the concomitant depletion of free dNTPs will also deplete the levels of NTPs and other closely related metabolites. We predict that this depletion will lead to homeostatic regulatory feedback by the transcription of many enzymes in the nucleotide synthesis pathway, and potentially transcriptome-wide changes in the regulatory program. In other words, *we predict that transcription in E. coli has significant cell-cycle dependence.* To test this

transcriptome-scale prediction, we can use RNA-Seq rather than RT-qPCR to quantify message abundance for all messages, not just *nrdAB*.

Anticipated results: Consistent with this model, cell-cycle-dependent transcription has already been reported in the model bacterial systems *Caulobacter* and *Mycobacterium* [51, 52]; however, analogous measurements have not been successfully made in *E. coli* and other bacteria capable of rapid growth with overlapping replication cycles. (See Fig. 5.5C.) The canonical expectation is that *minimal* cell-cycle dependence would be observed in *E. coli* transcription and this model is consistent with one transcriptome study [67]; however, there are repeated and consistent reports of significant cell-cycle-dependent transcription in studies focusing on single genes: *nrdAB* [22, 49, 53], *ftsZ* [50, 68] as well as others [53]. These results, in conjunction with our predictions and a consistent picture in *Caulobacter* and *Mycobacterium* [51, 52], convince us that analogous transcriptional dependence *is* present in *E. coli*, although it may be more difficult to detect due to the added complication of synchronizing the cells.

Approach 3: If synchronization is too imprecise, even large amplitude oscillations can be hidden due to incoherence in the population (e.g., [67]). An alternative strategy is the direct and simultaneous visualization of *nrdAB* transcription and replication *in vivo* in single cells. We will simultaneously visualize the transcription of *nrdAB* using the MS2 system (inserting an array of MBS binding sequences in the 3'UTR and expressing an MCP-GFP fusion) and using an mCherry-DnaN fusion to visualize replication [64, 66]. An updated version of the MS2 system fixes a shortcoming of the initial system: prolonged message lifetime [69].

Preliminary data and anticipated results: We have used both systems independently in the past [42, 64, 66, 70]. We anticipate that we will see multiple transcription events at replication initiation, followed by a second burst of transcription roughly 15 minutes after the first.

5.5 *Aim 4: Test metabolite oscillation model by direct cell-cycle dependent measurement of levels*

Motivation: Despite strong indirect evidence from measurements of the fork velocity and mutation rate, no direct temporal measurement of dNTP levels as a function of cell-cycle time has yet been made. In addition to the oscillation of dNTPs, our models also predict that NTP levels oscillate in response to replication elongation. In this aim, we will attempt to capture these predicted nucleotide oscillations using two different approaches: Liquid Chromatography Mass Spectrometry (LC-MS) in a synchronized culture and using a biosensor to measure levels in single cells.

Significance: The direct detection of dNTP and NTP oscillations would have great significance to our fundamental understanding of cell metabolism. Although many types of metabolic oscillations have previously been reported, our analysis would demonstrate that this phenomenon has a common and unremarkable origin: it is the natural consequence of the underlying mechanism of homeostatic regulatory control of metabolites.

5.5.1 Sub-aim 4.1: Test NTP-oscillation model by LC-MS

Motivation: NTP levels are critical for a host of essential biological processes: ATP serves as the canonical medium of cellular energy, four NTP monomers (GTP, ATP, CTP, and UTP) are polymerized to form RNA during transcription, and NTPs are also a key precursor to dNTPs [71]. (See Fig. 5.2.) Under aerobic conditions, RNR (gene product of *nrdAB*) reduces NDPs to form dNDPs, a key focus of this proposal. Our model predicts that NTP levels are affected by dNTP synthesis and that their levels oscillate coincident with the dNTP levels. A critical and tractable test of the model is the measurements of dynamics of NTP levels by LC-MS, which are present at 100-fold the concentration of dNTPs under physiological conditions. This makes their detection more tractable than dNTPs [20].

Approach: Cells growing on minimal media will be synchronized and fractions collected as described in Sub-aim 3. Metabolites will then be quantified as described in Ref. [72] in

collaboration with J. Wang (University of Wisconsin). (See the letter of collaboration.) In short: Harvested cells will be resuspended in ice-cold extraction solvent and centrifuged to remove cell debris. Samples will then be analyzed on a high-performance LC-MS system. Metabolite levels will be quantified from raw LC-MS data using MAVEN software and then normalized to their respective optical densities and sample volumes [72].

Preliminary data and anticipated results: The Wang lab developed the proposed approach to nucleotide quantitation and has extensive experience quantifying a wide range of nucleotides in the context of studying bacterial metabolism and routinely quantify NTP levels. The greater than 30% variation in ATP during the cell cycle (Fig. 5.6C) should be easily resolvable by LC-MS. We expect to observe both an overall depletion of NTP levels with replication initiation (Fig. 5.6C), as well as oscillations analogous to what is predicted for dNTP levels. The resolution of the predicted 15-minute-period oscillation will only be possible if sufficiently precise cell synchronization is achieved.

Alternative approach: Although LC-MS approaches quantitative measure of NTP and dNTP levels, our proposed experiments are dependent on efficient cell synchronization; however, an alternative cell biology approach allows quantitation of ATP levels in live cells grown asynchronously using a fluorescent biosensor, QUEEN-2m [3]. (See Fig. 5.6B.) Oscillations of ATP levels have already been reported using this approach [5], broadly consistent with our predictions; however, the coincidence of these oscillations with replication have not yet been analyzed. Our model predicts transient ATP depletion on the initiation of replication followed by smaller amplitude oscillations, followed by a return of ATP levels to pre-initiation levels on the completion of replication. We will test for the oscillation of dNTP levels using time-lapse fluorescence microscopy. As before, we will use an mCherry-DnaN fusion to monitor the replication state of the cell. The relative ATP level will be assayed by QUEEN-2m fluorescence, as previously described [3, 5].

Preliminary data and anticipated results: Based on published data, both rapid (period $T < 30$ min) and slow (period $T \geq 30$ min) oscillations have been observed, qualitatively consistent with the two predicted phenomenologies (homeostatic regulatory oscillations and

upregulation during the C period). Using the published data, we have computed a mean ATP level in the cells as a function of cell cycle progression. ATP levels are depressed during a period coincident with replication as predicted (Fig. 5.6C); however, the observation of putative synchronized oscillations will require the proposed simultaneous visualization of replication initiation and QUEEN-2m fluorescence.

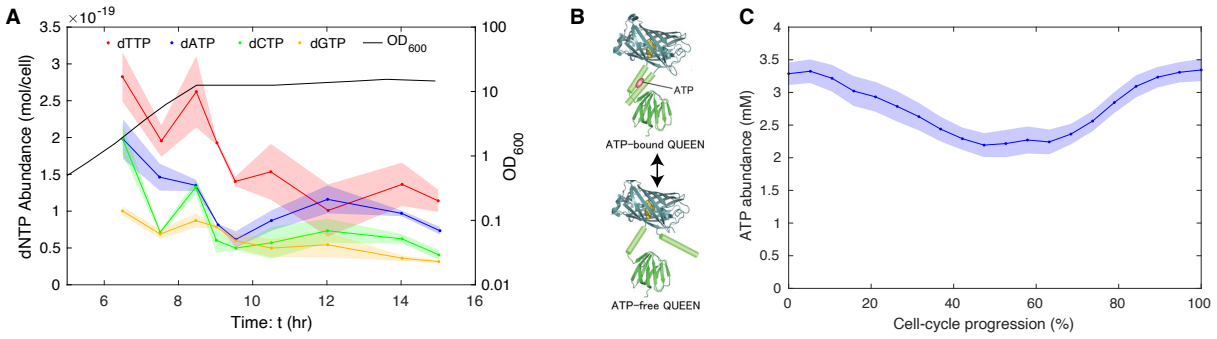


Figure 5.6: **dNTP fluctuations. Panel A: dNTP pool fluctuate as cells transition from log to stationary phase.** Oscillatory behavior is observed in the dNTP pools as the culture transitions from log to stationary phase in an asynchronous culture. Data reproduced from Ref. [73]. **Panel B: Measuring ATP levels using a biosensor.** The biosensor QUEEN-2m reports on ATP levels *in vivo* [3]. **Panel C: Mean ATP levels during the cell cycle.** Preliminary measurements are consistent with a depletion of ATP during the C period (mid-cell-cycle).

5.5.2 Sub-aim 4.2: Test dNTP-oscillation model by LC-MS

Motivation: dNTP oscillations are a critical prediction of the regulatory model. Once the less technically challenging detection of NTP levels has been achieved, we will then attempt the quantitation of dNTP levels.

Approach: Identical to the previous sub-aim. The protocol can be slightly modified to match previous successful protocols for dNTP quantitation, as described in Ref. [73].

Preliminary data and anticipated results: Although more challenging to detect than NTPs, dNTPs are routinely characterized: In fact there is already interesting precedent for

dNTP oscillations as cells transition from log to stationary phase [73]. (See Fig. 5.6A.) The oscillations we predict are of comparable magnitude to those previously detected and therefore the sensitivity of the LC-MS approach should be more than adequate to detect these levels. As before, the detection of 15-minute-period oscillations is dependent on achieving cell synchronization.

5.6 Outlook

The long term goal of our laboratory is the identification of new emerging laws for complex biological and biophysical systems. The analysis of analytically tractable quantitative models, often inspired by biological as well as non-biological physics, plays a central role in this work (Aim 1). For instance, the basis of our measurements of fork velocity (Aim 2) are dependent on the development of a detailed understanding of the role of temporal disorder in an exponentially growing system [38]. Our preliminary measurements of dNTP oscillations suggest that the conception of what constitutes homeostasis may be fundamentally flawed, and may neglect fundamental limitations on what can be achieved at a cellular scale. Measuring the properties of homeostasis and understanding the rationale for these properties (Aims 3 and 4) constitutes both a biological and biophysical problem of fundamental importance and significance. It is important to emphasize that although regulatory networks have been modeled using chemical kinetics for 60 years [31], no one has yet explored the biological implications of the limits placed by considerations of network stability, despite these phenomena being a generic prediction of the biophysical models. Oscillations may also be a generic, but underappreciated, feature of regulatory feedback in biological circuits.

Transcriptome impulse response? In addition to testing an important hypothesis, this work has the potential to develop a novel approach to studying regulatory network topology and function. Sub-aim 3.2 describes an experiment which captures the temporal dependence of the transcriptome in response to a temperature shift. Data from the literature already

supports the hypothesis that genes have distinct temporal signatures. (See Fig. 5.5C.) The dynamic transcriptome data has the potential to be interpreted as a high-dimensional impulse-response measurement–like an oscilloscope measuring the time-dependent voltage at > 4000 nodes simultaneously. We hypothesize that by comparing the temporal response between genes, we could discover novel regulatory mechanisms and differentiate between direct and indirect regulatory responses. Should this approach prove fruitful, it could easily be generalized to a broad range of different impulse stimuli beyond temperature shifts. We believe this approach (measurement and analysis) has great potential, including implications for both basic science, as well as medical and biotech applications.

Other rationales for oscillations? Our current hypothesis is that regulatory oscillations are an inescapable consequence of tight regulation; however, the characterization of our mutants may demonstrate that oscillations can be effectively eliminated with no reduction in the fork velocity, etc. These results could support an intriguing hypothesis: the oscillations themselves may be mechanistically advantageous. For instance, single-molecule measurements demonstrate that the inherent stochasticity of the molecular scale is not overcome but rather harnessed as a fundamental mechanism in the function of molecular motors (i.e., the thermal ratchet). Could metabolic oscillations be an essential functional mechanism of the cell? This study could provide evidence in support of this intriguing hypothesis.

5.7 Bibliography

[1] P. A. Wiggins, *Metabolic homeostasis, oscillations, and instability*, Proposal for National Science Foundation Physics of Living Systems Grant: NSF-Phys-2412326, May 2024.

[2] R. P. Carlson, "Decomposition of complex microbial behaviors into resource-based stress responses," *Bioinformatics*, vol. 25, no. 1, pp. 90–7, Jan. 2009. DOI: 10.1093/bioinformatics/btn589.

[3] H. Yaginuma *et al.*, "Diversity in ATP concentrations in a single bacterial cell population revealed by quantitative single-cell imaging," *Sci Rep*, vol. 4, p. 6522, Oct. 2014. DOI: 10.1038/srep06522.

[4] 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," *Nat Commun*, vol. 14, no. 1, p. 1762, Mar. 2023. DOI: 10.1038/s41467-023-37456-2.

[5] 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.

[6] V. M. Boer, J. H. de Winde, J. T. Pronk, and M. D. W. Piper, "The genome-wide transcriptional responses of *Saccharomyces cerevisiae* grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur," *J Biol Chem*, vol. 278, no. 5, pp. 3265–74, Jan. 2003. DOI: 10.1074/jbc.M209759200.

[7] M. Csete and J. Doyle, "Bow ties, metabolism and disease," *Trends Biotechnol*, vol. 22, no. 9, pp. 446–50, Sep. 2004. DOI: 10.1016/j.tibtech.2004.07.007.

[8] H. Holms, "Flux analysis and control of the central metabolic pathways in *Escherichia coli*," *FEMS Microbiol Rev*, vol. 19, no. 2, pp. 85–116, Dec. 1996. DOI: 10.1111/j.1574-6976.1996.tb00255.x.

[9] M. Schaechter, O. Maaloe, and N. O. Kjeldgaard, "Dependency on medium and temperature of cell size and chemical composition during balanced grown of *Salmonella typhimurium*," *J Gen Microbiol*, vol. 19, no. 3, pp. 592–606, Dec. 1958. DOI: 10.1099/00221287-19-3-592.

[10] 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.

[11] F. A. Chandra, G. Buzi, and J. C. Doyle, "Glycolytic oscillations and limits on robust efficiency," *Science*, vol. 333, no. 6039, pp. 187–92, Jul. 2011. DOI: 10.1126/science.1200705.

[12] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback Control Theory*. Prentice Hall Professional Technical Reference, 1991, ISBN: 0023300116.

[13] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics* (Halliday & Resnick Fundamentals of Physics). John Wiley & Sons Canada, Limited, 2010, ISBN: 9780470547939.

[14] H. El-Samad, "Biological feedback control-respect the loops," *Cell Syst*, vol. 12, no. 6, pp. 477–487, Jun. 2021. DOI: 10.1016/j.cels.2021.05.004.

[15] A. Papagiannakis, B. Niebel, E. C. Wit, and M. Heinemann, "Autonomous metabolic oscillations robustly gate the early and late cell cycle," *Mol Cell*, vol. 65, no. 2, pp. 285–295, Jan. 2017. DOI: 10.1016/j.molcel.2016.11.018.

[16] J. L. Robinson and M. P. Brynildsen, "Discovery and dissection of metabolic oscillations in the microaerobic nitric oxide response network of *Escherichia coli*," *Proc Natl Acad Sci U S A*, vol. 113, no. 12, E1757–66, Mar. 2016. DOI: 10.1073/pnas.1521354113.

[17] D. C. Andersen, J. Swartz, T. Ryll, N. Lin, and B. Snedecor, "Metabolic oscillations in an *E. coli* fermentation," *Biotechnol Bioeng*, vol. 75, no. 2, pp. 212–8, Oct. 2001. DOI: 10.1002/bit.10018.

[18] J. Liu *et al.*, "Metabolic co-dependence gives rise to collective oscillations within biofilms," *Nature*, vol. 523, no. 7562, pp. 550–4, Jul. 2015. DOI: 10.1038/nature14660.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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. 14 496–501, Apr. 2004. DOI: 10.1074/jbc.M310142200.

[24] 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. 33 021–6, Oct. 2000. DOI: 10.1074/jbc.M005337200.

[25] 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.

[26] 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.

[27] 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.

[28] L. B. Augustin, B. A. Jacobson, and J. A. Fuchs, "*Escherichia coli* Fis and DnaA proteins bind specifically to the *nrd* promoter region and affect expression of an *nrd-lac* fusion," *J Bacteriol*, vol. 176, no. 2, pp. 378–87, Jan. 1994. DOI: 10.1128/jb.176.2.378-387.1994.

[29] S. J. Elledge, Z. Zhou, J. B. Allen, and T. A. Navas, "DNA damage and cell cycle regulation of ribonucleotide reductase," *Bioessays*, vol. 15, no. 5, pp. 333–9, May 1993. DOI: 10.1002/bies.950150507.

[30] E. Torrents *et al.*, "NrdR controls differential expression of the *Escherichia coli* ribonucleotide reductase genes," *J Bacteriol*, vol. 189, no. 14, pp. 5012–21, Jul. 2007. DOI: 10.1128/JB.00440-07.

[31] R. Phillips, J. Kondev, J. Theriot, and N. Orme, *Physical Biology of the Cell*. Garland Science, 2013, ISBN: 9780815344506.

[32] P. L. Freddolino and S. Tavazoie, "Beyond homeostasis: A predictive-dynamic framework for understanding cellular behavior," *Annu Rev Cell Dev Biol*, vol. 28, pp. 363–84, 2012. DOI: 10.1146/annurev-cellbio-092910-154129.

[33] H. Modell, W. Cliff, J. Michael, J. McFarland, M. P. Wenderoth, and A. Wright, "A physiologist's view of homeostasis," *Adv Physiol Educ*, vol. 39, no. 4, pp. 259–66, Dec. 2015. DOI: 10.1152/advan.00107.2015.

[34] G. Storz, Y. I. Wolf, and K. S. Ramamurthi, "Small proteins can no longer be ignored," *Annu Rev Biochem*, vol. 83, pp. 753–77, 2014. DOI: 10.1146/annurev-biochem-070611-102400.

[35] 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.

[36] 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.

[37] 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.

[38] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, "Characterizing stochastic cell-cycle dynamics in exponential growth," *Phys. Rev. E*, vol. 105, p. 014 420, 1 Jan. 2022. DOI: 10.1103/PhysRevE.105.014420.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] A. Knöppel, O. Broström, K. Gras, J. Elf, and D. Fange, "Regulatory elements coordinating initiation of chromosome replication to the *Escherichia coli* cell cycle," *Proceedings of the National Academy of Sciences*, vol. 120, no. 22, e2213795120, 2023. DOI: 10.1073/pnas.2213795120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2213795120.

[44] B. W. Davies, M. A. Kohanski, L. A. Simmons, J. A. Winkler, J. J. Collins, and G. C. Walker, "Hydroxyurea induces hydroxyl radical-mediated cell death in *Escherichia coli*," *Mol Cell*, vol. 36, no. 5, pp. 845–60, Dec. 2009. DOI: 10.1016/j.molcel.2009.11.024.

[45] R. D. Zeinert, H. Baniasadi, B. P. Tu, and P. Chien, "The Lon protease links nucleotide metabolism with proteotoxic stress," *Mol Cell*, vol. 79, no. 5, 758–767.e6, Sep. 2020. DOI: 10.1016/j.molcel.2020.07.011.

[46] A. Kuroda *et al.*, "Role of inorganic polyphosphate in promoting ribosomal protein degradation by the Lon protease in *E. coli*," *Science*, vol. 293, no. 5530, pp. 705–8, Jul. 2001. DOI: 10.1126/science.1061315.

[47] D. Ahluwalia, R. J. Bienstock, and R. M. Schaaper, "Novel mutator mutants of *E. coli nrdAB* ribonucleotide reductase: Insight into allosteric regulation and control of mutation rates," *DNA Repair (Amst)*, vol. 11, no. 5, pp. 480–7, May 2012. DOI: 10.1016/j.dnarep.2012.02.001.

[48] S. Cooper, *Bacterial Growth and Division–Biochemistry and Regulation of Prokaryotic and Eukaryotic Division Cycles*, 1st Edition. Academic Press, Feb. 1991.

[49] L. Sun and J. A. Fuchs, "*Escherichia coli* ribonucleotide reductase expression is cell cycle regulated," *Mol Biol Cell*, vol. 3, no. 10, pp. 1095–105, Oct. 1992. DOI: 10.1091/mbc.3.10.1095.

[50] J. Männik, B. E. Walker, and J. Männik, "Cell cycle-dependent regulation of FtsZ in *Escherichia coli* in slow growth conditions," *Mol Microbiol*, vol. 110, no. 6, pp. 1030–1044, Dec. 2018. DOI: 10.1111/mmi.14135.

[51] M. T. Laub, H. H. McAdams, T. Feldblyum, C. M. Fraser, and L. Shapiro, "Global analysis of the genetic network controlling a bacterial cell cycle," *Science*, vol. 290, no. 5499, pp. 2144–8, Dec. 2000. DOI: 10.1126/science.290.5499.2144.

[52] A. C. Bandekar, S. Subedi, T. R. Ioerger, and C. M. Sassetti, "Cell-cycle-associated expression patterns predict gene function in *Mycobacteria*," *Curr Biol*, vol. 30, no. 20, 3961–3971.e6, Oct. 2020. DOI: 10.1016/j.cub.2020.07.070.

[53] P. Zhou *et al.*, "Gene transcription and chromosome replication in *Escherichia coli*," *J Bacteriol*, vol. 179, no. 1, pp. 163–9, Jan. 1997. DOI: 10.1128/jb.179.1.163-169.1997.

[54] N. J. Kuwada, B. Traxler, and P. A. Wiggins, "High-throughput cell-cycle imaging opens new doors for discovery," *Curr Genet*, vol. 61, no. 4, pp. 513–6, Nov. 2015. DOI: 10.1007/s00294-015-0493-y.

[55] E. Balleza, J. M. Kim, and P. Cluzel, "Systematic characterization of maturation time of fluorescent proteins in living cells," *Nat Methods*, vol. 15, no. 1, pp. 47–51, Jan. 2018. DOI: 10.1038/nmeth.4509.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] S. M. Mangiameli, J. A. Cass, H. Merrikh, and P. A. Wiggins, "The bacterial replisome has factory-like localization," *Curr Genet*, vol. 64, no. 5, pp. 1029–1036, Oct. 2018. DOI: 10.1007/s00294-018-0830-z.

[67] A. Løbner-Olesen, M. Slominska-Wojewodzka, F. G. Hansen, and M. G. Marinus, "DnaC inactivation in *Escherichia coli* K-12 induces the SOS response and expression of nucleotide biosynthesis genes," *PLoS One*, vol. 3, no. 8, e2984, Aug. 2008. DOI: 10.1371/journal.pone.0002984.

[68] P. Zhou and C. E. Helmstetter, "Relationship between *ftsZ* gene expression and chromosome replication in *Escherichia coli*," *J Bacteriol*, vol. 176, no. 19, pp. 6100–6, Oct. 1994. DOI: 10.1128/jb.176.19.6100-6106.1994.

[69] E. Tutucci, M. Vera, J. Biswas, J. Garcia, R. Parker, and R. H. Singer, "An improved MS2 system for accurate reporting of the mRNA life cycle," *Nat Methods*, vol. 15, no. 1, pp. 81–89, Jan. 2018. DOI: 10.1038/nmeth.4502.

[70] S. Stylianidou, N. J. Kuwada, and P. A. Wiggins, "Cytoplasmic dynamics reveals two modes of nucleoid-dependent mobility," *Biophys J*, vol. 107, no. 11, pp. 2684–92, Dec. 2014. DOI: 10.1016/j.bpj.2014.10.030.

[71] B. Alberts *et al.*, *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, 2015.

[72] D. K. Fung, J. Yang, D. M. Stevenson, D. Amador-Noguez, and J. D. Wang, "Small alarmone synthetase SasA expression leads to concomitant accumulation of pGpp, ppApp, and AppppA in *Bacillus subtilis*," *Front Microbiol*, vol. 11, p. 2083, 2020. DOI: 10.3389/fmicb.2020.02083.

[73] M. H. Buckstein, J. He, and H. Rubin, "Characterization of nucleotide pools as a function of physiological state in *Escherichia coli*," *J Bacteriol*, vol. 190, no. 2, pp. 718–26, Jan. 2008. DOI: 10.1128/JB.01020-07.

Appendix A

An interbacterial DNA deaminase toxin directly mutagenizes surviving target populations

I have published one other paper during my time in the Wiggins lab. It does not fit within the rest of the narrative of this dissertation, so I have not included it as one of the chapters. It does include a lot of microscopy and data analysis that I conducted, and is deeply related to the recent advances in CRISPR-free mitochondrial base editing. For those who are interested, this is the reference:

M. H. de Moraes, F. Hsu, D. Huang, D. E. Bosch, J. Zeng, M. C. Radey, *et al.*, "An interbacterial DNA deaminase toxin directly mutagenizes surviving target populations," *eLife*, vol. 10, e62967, Jan. 2021, ISSN: 2050-084X. DOI: 10.7554/eLife.62967.

Abstract: When bacterial cells come in contact, antagonism mediated by the delivery of toxins frequently ensues. The potential for such encounters to have long-term beneficial consequences in recipient cells has not been investigated. Here, we examined the effects of intoxication by DddA, a cytosine deaminase delivered via the type VI secretion system (T6SS) of *Burkholderia cenocepacia*. Despite its killing potential, we observed that several bacterial species resist DddA and instead accumulate mutations. These mutations can lead to the acquisition of antibiotic resistance, indicating that even in the absence of killing, interbacterial antagonism can have profound consequences on target populations. Investigation of additional toxins from the deaminase superfamily revealed that mutagenic activity is a common feature of these proteins, including a representative we show targets single-stranded DNA and displays a markedly divergent structure. Our findings suggest

that a surprising consequence of antagonistic interactions between bacteria could be the promotion of adaptation via the action of directly mutagenic toxins.