UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

An integrated workflow for the multi-omic characterization of microorganisms

Permalink

https://escholarship.org/uc/item/9tf3c8vk

Author

Latif, Haythem

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

An integrated workflow for the multi-omic characterization of microorganisms

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Haythem Latif

Committee in charge:

Bernhard Ø. Palsson, Chair
Douglas Bartlett
Michael Heller
Milton H. Saier, Jr.
Karsten Zengler
Kun Zhang

2015

Copyright
Haythem Latif, 2015
All rights reserved.

The dissertation of Haythem Latif is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2015

DEDICATION

To my mother and father for all that you have sacrificed, all that you have provided, and all that you have taught me.

EPIGRAPH

"It is the characteristic of the magnanimous man to ask no favor but to be ready to do kindness to others."

-Aristotle

"What I see in Nature is a magnificent structure that we can comprehend only very imperfectly, and that must fill a thinking person with a feeling of humility. This is a genuinely religious feeling that has nothing to do with mysticism."

-Einstein

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures x

List of Tables xii

Acknowledgements xiii

Vita xvii

Abstract of the Dissertation xix

Chapter 1 Introduction 1
1.1 The genome that launched 1,000 genomes. 1
1.2 The complexity of microbial genomes revealed by multi-omic characterization. 2
1.3 Systems microbiology approach to omics data integration. 3
1.4 Previous implementations of the systems approach to multi-omic data integration. 4
1.5 Updating and expanding the previous workflow. 6
1.6 Bibliography 8

Chapter 2 Genome assembly improvements using next-generation sequencing technology. 13
2.1 Abstract 13
2.2 Introduction 14
2.3 Results and Discussion 15
2.3.1 *Thermotoga maritima* ATCC genomovar sequence and gene annotation for a missed ≈9 kb region . . 15
2.3.2 *Escherichia coli* serotype O157:H7 EDL933 assembly. 17
2.4 Conclusion 20
2.5 Acknowledgements 20
2.6 Bibliography 22

Chapter 3 The genome organization of *Thermotoga maritima* reflects its lifestyle . . . 27
3.1 Abstract . . . 27
3.2 Author Summary . . . 28
3.3 Introduction . . . 28
3.4 Results . . . 31
3.4.1 An integrative, multi-omic approach for the annotation of the genome organization . . . 31
3.4.2 Identification of promoters and RBSs followed by quantitative intra- and interspecies analysis of binding free energies . . . 35
3.4.3 *T. maritima* promoter-containing intergenic regions reveal a unique distribution of 5′UTRs and spatial limitations on regulation . . . 42
3.4.4 *T. maritima* has an actively transcribed genome that is tightly correlated to protein abundances . . . 45
3.5 Discussion . . . 47
3.6 Materials and Methods . . . 50
3.6.1 Culture conditions and physiology . . . 50
3.6.2 Genome resequencing and annotation updates . . . 50
3.6.3 Transcription start site determination . . . 51
3.6.4 Transcriptome characterization and gene expression 51
3.6.5 Proteomics, peptide mapping, and protein abundance quantitation . . . 52
3.6.6 Promoter element motif analysis and position weight matrix (PWM) generation . . . 53
3.6.7 Information content calculations . . . 54
3.6.8 Ribosome binding site energy calculations . . . 55
3.6.9 Rho-independent terminator site determination . 56
3.6.10 Prediction of small RNAs . . . 56
3.6.11 Transcription unit assembly . . . 56
3.6.12 Transcription factor binding site mapping . . . 57
3.6.13 Data deposition . . . 57
3.7 Acknowledgments . . . 57
3.8 Bibliography . . . 58

Chapter 4 Adaptive evolution of *Thermotoga maritima* reveals plasticity of the ABC transporter network . . . 67
4.1 Abstract . . . 67
4.2 Introduction . . . 68
4.3 Materials and Methods . . . 70
4.3.1 Culture conditions and physiology. . . . 70
4.3.2 Genomic DNA sequencing and variant analysis. . 71

4.3.3 RNA-seq and transcript abundance estimation. . . 71
4.3.4 Gene expression analysis. 72
4.3.5 Data Deposition. 72
4.4 Results 73
4.4.1 Glucose evolution and evolved phenotypic properties. 73
4.4.2 Genetic variants in evolved cultures on glucose. . 75
4.4.3 Gene expression analysis of eTMglc mutant cultures. 77
4.5 Discussion 80
4.6 Acknowledgments 84
4.7 Bibliography 84

Chapter 5 Integrated analysis of molecular and systems level function of Crp using ChIP-exo. 90
5.1 Abstract 90
5.2 Introduction 91
5.3 Results 93
5.3.1 ChIP-exo data provides genome-scale, *in vivo* mechanistic insights into bacterial transcription activation. 93
5.3.2 ChIP-exonuclease coupled with gene expression delineates the full Crp regulon 107
5.4 Discussion 117
5.5 Materials and Methods 121
5.5.1 Strains and Culturing Conditions 121
5.5.2 ChIP Experiments 122
5.5.3 Gene Expression 123
5.6 Author Contributions 124
5.7 Acknowledgements 124
5.8 Bibliography 124

Chapter 6 A streamlined ribosome profiling protocol for the characterization of microorganisms 133
6.1 Abstract 133
6.2 Introduction 134
6.3 Results and Discussion 135
6.4 Conclusion 137
6.5 Materials and Methods 139
6.5.1 Reagents 139
6.5.2 Procedure 140
6.5.3 Recipes 147
6.5.4 Troubleshooting 150
6.5.5 Equipment 151

6.6 Acknowledgements . . . 152
6.7 Author Contributions . . . 152
6.8 Bibliography . . . 152

Chapter 7 Trash to treasure: Production of biofuels and commodity chemicals via syngas fermenting microorganisms . . . 154
7.1 Abstract . . . 154
7.2 Graphical Abstract . . . 155
7.3 Introduction . . . 155
7.4 Syngas fermenting microorganisms . . . 157
7.5 The Wood-Ljungdahl pathway . . . 157
7.6 Energy conservation in acetogens . . . 162
7.7 Advances in genetic manipulation tools . . . 163
7.8 Strain engineering to obtain desired production phenotypes 165
7.9 Rational strain design and process optimization through a systems-level approach . . . 166
7.10 Summary and opportunities . . . 168
7.11 Acknowledgements . . . 169
7.12 Bibliography . . . 169

Chapter 8 Conclusion . . . 178
8.1 Model organisms and their knowledgebases. . . . 178
8.2 Closing the knowledge gap using an integrated, multi-omic characterization workflow. . . . 179
8.3 The benefits of a consolidated data generation platform. . 179
8.4 Applications of microbial knowledgebases. . . . 182
8.5 Perspective on the future of multi-omic data integration. 183
8.6 Bibliography . . . 186

LIST OF FIGURES

Figure 1.1: The four principle steps of integrating multi-omic datasets. . . . 5

Figure 2.1: *T. maritima* genome gap revealed. 18
Figure 2.2: Updated *E. coli* serotype O157:H7 EDL933 assembly. 19

Figure 3.1: Generation of multiple genome-scale datasets integrated with bioinformatics predictions reveals the genome organization 32
Figure 3.2: Identification and quantitative comparison of genetic elements for transcription and translation initiation 36
Figure 3.3: Arrangement of genomic features contained within promoter-containing intergenic regions (PIRs) 43
Figure 3.4: Global analysis of mRNA and protein expression levels 46

Figure 4.1: Glucose evolution time course 74
Figure 4.2: Mutations to the *gluEFK* and *bglEFGKL* ABC transporter operons. 76
Figure 4.3: Gene expression analysis of eTMglc cultures. 79
Figure 4.4: Gene expression analysis of the different functional categories of proteins found in the ABC2 importer families for carbohydrate uptake (3.A.1.1 & 3.A.1.2) and the peptide/opine/nickel uptake family (3.A.1.5) for cultures grown on glucose. 81

Figure 5.1: TSS aligned and oriented σ^{70} ChIP-exo data reveals DNA footprint patterns consistent with stable transcription initiation intermediates. 94
Figure 5.2: Crp promoter classes have unanticipated ChIP-exo footprint regions. 98
Figure 5.3: Contrasting ChIP-exo profiles of repressors and activators. . . . 103
Figure 5.4: The effect of genetic perturbation on Crp/RNAP interactions. . 105
Figure 5.5: The full Crp regulon defined through paired RNA-seq and ChIP-exonuclease data. 110
Figure 5.6: Distribution of expression levels for catabolic, anabolic, and chemiosmotic genes reflects physiological models of Crp regulation. . . . 113
Figure 5.7: Systems level circuitry of Crp regulation is in line with genetic perturbations. 116

Figure 6.1: Streamlined protocol for ribosome profiling of microorganisms. . 136
Figure 6.2: Benchmarking the streamlined protocol against publically available data. 138

Figure 7.1: Overview of syngas fermentation. 155

Figure 7.2: Syngas fermenting organisms known to produce multi-carbon organic compounds. 158
Figure 7.3: The Wood-Ljungdahl pathway and its connection to heterotrophic metabolism. 160
Figure 7.4: Energy conservation during syngas fermentation. 164

Figure 8.1: Systems level workflow for multi-omic data integration. 180
Figure 8.2: Applications of genome-scale knowledgebases. 184

LIST OF TABLES

Table 4.1: Physiological properties of glucose evolved cultures 74

Table 6.1: Recipe 1 Lysis Buffer (scale up as needed) 147
Table 6.2: Recipe 2 MNase Buffer (scale up as needed) 147
Table 6.3: Recipe 3 MNase Reaction Mix (200 μL) 148
Table 6.4: Recipe 4 Polysome Buffer (10 mL) 148
Table 6.5: Recipe 5 T4 PNK Reaction Mix (22.2 μL) 148
Table 6.6: Recipe 6 tRNA Depletion (30 μL) 149

ACKNOWLEDGEMENTS

First and foremost I would like to thank God for all he has blessed me with. Next, I want to thank my family for their unconditional support throughout this experience. I love you all so much. There is nothing I would not do for any of you and without hesitation. Despite all that we have faced we managed to make the best of situations together. In particular I want to recognize my parents for the amazing job they did raising us. I am not the man I am today without all that they have done for me, provided for me, and taught me. My mother is the strongest person I have ever known. She is the rock of our family and deserves the majority of the credit for anything I have, and will achieve, is attributed to her. Next, I want to thank my Dad for being so caring and providing us with what we needed to accomplish our goals. I also want to thank my brother Walead for always being there to advise me on just about everything going on in my life. He has always been a great role model from when we were children to this very day. To Moria, I love you so much and you can count on me and Walead forever. To Jylan, I want to say thank you so much for being there for me. I spent four years in San Diego but it wasn't until I met you that I began to feel like it was home.

My time in graduate school was a time of great personal and professional growth. This is attributed to the outstanding cadre of researchers and classmates I was fortunate enough to encounter in my time here. I would like to thank Dr. Karsten Zengler for all of his guidance, his mentorship, and above all, his friendship. I would also like to thank Dr. Bernhard Palsson and everyone in the Systems Biology Research Group here at UCSD. It is rare to find a group of individuals who are so creative, collaborative, and passionate about their work. In particular, I would like to thank a few individuals. First, I want to thank Mallory Embree. She has been a great friend seemingly from our first day on campus together. She is incredibly smart, organized, and obviously tolerant if she put up with me for all these years. Next, I want to thank Josh Lerman for teaching me so much on the *Thermotoga maritima* paper. You taught me how to code, how to analyze data, and even conjured up wild ideas about walkie talkies in a loud club to explain how this organism can withstand the heat. I also would like to thank Richard Szubin for without him my thesis would not

have been finished. Richard is an incredibly gifted scientist and an absolute joy to work with. I also must thank Kathy Andrews, our lab manager, for being able to move mountains when asked to. I have also had some great mentors along the way that I would like to recognize. These included Vasiliy Portnoy, who never hesitated to give me a wakeup call when I missed my bus in the morning; Harish Nagarajan, who is simply the Genius; the Tarasova sisters, who did much of the experimental work; Dr. Yu, who helped teach me the basics in molecular biology; and Ali Ebrahim, for his computations support and friendship. I would also like to thank all of the co-authors and collaborators I have had. There are too many to include here but I sincerely thank you all. Finally, thanks to Marc Abrams, Yana Campen, Helder Balelo, and Ying Hefner for their administrative support.

I would also like to thank my dissertation committee. Though not officially the chair of my committee, I once again would like to thank Dr. Karsten Zengler for guiding my research efforts. Next, I would like to thank Dr. Palsson for chairing my committee and for encouraging highly collaborative research. I would also like to thank Dr. Doug Bartlett and Dr. Priya Narasingarao for working with us on thermophilic projects that have result in really interesting findings. To Dr. Kun Zhang and his lab, I want to say thank you for being great neighbors and always being willing to share best practices, reagents, etc. Lastly, I want to thank Dr. Michael Heller and Dr. Milton H. Saier, Jr. for taking time out of their busy schedules to serve on my committee.

I would lastly like to acknowledge funding sources that made this work possible. First and foremost I must recognize the National Science Foundation for awarding me their Graduate Research Fellowship under grant DGE1144086. I also want to thank The Office of Science of the U.S. Department of Energy (DOE) under grants DE-FG02-08ER64686 and DE-FG02-09ER25917. Proteomics capabilities were developed under support from the DOE Office of Biological and Environmental Research (BER) Pan-omics Project and the NIH National Center for Research Resources (RR018522), and a significant portion of this work was performed in the Environmental Molecular Sciences Laboratory (EMSL), a DOE-BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multi-

program national laboratory operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RLO 1830. The Novo Nordisk Foundation also provided support for works presented herein as did the National Institutes of Health under U01 grant GM102098-01 and under grant GM098105.

Chapter 2 is in part adapted from two published manuscripts: (1) Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K. (2013) The Genome Organization of *Thermotoga maritima* Reflects Its Lifestyle. PLoS Genet 9(4): e1003485. doi:10.1371/journal.pgen.1003485. * indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Joshua A. Lerman (equal contributor), Vasiliy A. Portnoy, Yekaterina Tarasova, Harish Nagarajan, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Dae-Hee Lee, Yu Qiu, and Karsten Zengler. (2) Latif H, Li HJ, Charusanti P, Palsson BØ., Aziz RK. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 Strain EDL933. Genome Announc. 2014 Aug 14;2(4). pii: e00821-14. doi: 10.1128/genomeA.00821-14. The dissertation author was the primary author of this paper responsible for the research. The other authors were Howard J Li, Pep Charusanti, Bernhard Ø. Palsson, and Ramy K Aziz.

Chapter 3 in full is a reprint of a published manuscript: Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, et al. (2013) The Genome Organization of *Thermotoga maritima* Reflects Its Lifestyle. PLoS Genet 9(4): e1003485. doi:10.1371/journal.pgen.1003485. * indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Joshua A. Lerman (equal contributor), Vasiliy A. Portnoy, Yekaterina Tarasova, Harish Nagarajan, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Dae-Hee Lee, Yu Qiu, and Karsten Zengler.

Chapter 4 has been submitted for publication of the material as it may appear in Latif H, Sahin M, Tarasova J, Tarasova Y, Portnoy VA, Zengler K. Adaptive

evolution of *Thermotoga maritima* reveals plasticity of the ABC transporter network. Submitted to Appl. Environ. Microbiol. 2014. The dissertation author was the primary author of this paper responsible for the research. The other authors are Merve Sahin, Janna Tarasova, Yekaterina Tarasova, Vasiliy A. Portnoy, and Karsten Zengler.

Chapter 5 has been submitted for publication of the material as it may appear in Latif H*, Federowicz S*, Szubin R, Tarasova J, Utrilla J, Ebrahim A, Zengler K, Palsson BØ. Integrated analysis of molecular and systems level function of Crp using ChIP-exo. Submitted to Cell. 2014. *indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Stephen Federowicz (equal contributor), Richard Szubin, Janna Tarasova, Jose Utrilla, Ali Ebrahim, Karsten Zengler, and Bernhard Ø. Palsson.

Chapter 6 has been submitted for publication of the material as it may appear in Latif H, Szubin R, Tan J, Brunk E, Lechner A, Zengler K, Palsson BØ. A streamlined ribosome profiling protocol for the characterization of microorganisms. Submitted to Biotechniques. 2014. The dissertation author was the primary author of this paper responsible for the research. The other authors are Richard Szubin, Justin Tan, Elizabeth Brunk, Anna Lechner, Karsten Zengler, and Bernhard Ø. Palsson.

Chapter 7 in full is a reprint of a published manuscript: Latif H, Zeidan AA, Nielsen AT, Zengler K. Trash to treasure: production of biofuels and commodity chemicals via syngas fermenting microorganisms. Curr Opin Biotechnol. 2014 Jun;27:79-87. doi: 10.1016/j.copbio.2013.12.001. The dissertation author was the primary author of this paper responsible for the research. The other authors are Ahmad A. Zeidan, Alex T. Nielsen, and Karsten Zengler.

VITA

2000-2004	B. S. in Chemical Engineering, Rutgers University
2004-2009	Staff Biochemical Engineer, Merck Research Laboratories, Merck & Co. Inc.
2009-2015	Ph. D. in Bioengineering, University of California, San Diego

PUBLICATIONS

Latif H*, Federowicz S*, Szubin R, Tarasova J, Utrilla J, Ebrahim A, Zengler K, Palsson BØ. Integrated analysis of molecular and systems level function of Crp using ChIP-exo. Submitted to Cell. 2014.

Latif H, Szubin R, Tan J, Brunk E, Lechner A, Zengler K, Palsson BØ. A streamlined ribosome profiling protocol for the characterization of microorganisms. Submitted to Biotechniques. 2014.

Latif H, Sahin M, Tarasova J, Tarasova Y, Portnoy VA, Zengler K. Adaptive evolution of *Thermotoga maritima* reveals plasticity of the ABC transporter network. Submitted to Appl. Environ. Microbiol. 2014.

Seo SW, Kim D, **Latif H**, O'Brien EJ, Szubin R, Palsson BØ. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. Nat Commun. 2014 Sep 15;5:4910. doi:10.1038/ncomms5910.

Latif H, Li HJ, Charusanti P, Palsson BØ, Aziz RK. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 Strain EDL933. Genome Announc. 2014 Aug 14;2(4). pii: e00821-14. doi: 10.1128/genomeA.00821-14.

Bordbar A, Nagarajan H, Lewis NE, **Latif H**, Ebrahim A, Federowicz S, Schellenberger J, Palsson BØ. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. Mol Syst Biol. 2014 Jul 1;10(7):737.doi: 10.15252/msb.20145243.

Latif H, Zeidan AA, Nielsen AT, Zengler K. Trash to treasure: production of biofuels and commodity chemicals via syngas fermenting microorganisms. Curr Opin Biotechnol. 2014 Jun;27:79-87. doi: 10.1016/j.copbio.2013.12.001.

Nagarajan H, Sahin M, Nogales J, **Latif H**, Lovley DR, Ebrahim A, Zengler K. Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of *Clostridium ljungdahlii*. Microb Cell Fact. 2013 Nov 25;12:118. doi: 10.1186/1475-2859-12-118.

Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'Brien E, Bordbar A, Roth AM, Rosenbloom J, Bian C, Xie M, Chen W, Li N, Baycin-Hizal D, **Latif H**, Forster J, Betenbaugh MJ, Famili I, Xu X, Wang J, Palsson BØ. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. Nat Biotechnol. 2013 Aug;31(8):759-65. doi: 10.1038/nbt.2624.

Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K. The genome organization of *Thermotoga maritima* reflects its lifestyle. PLoS Genet. 2013 Apr;9(4):e1003485. doi: 10.1371/journal.pgen.1003485.

Lerman JA*, Hyduke DR*, **Latif H**, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, Palsson BØ. *In silico* method for modeling metabolism and gene product expression at genome scale. Nat Commun. 2012 Jul 3;3:929. doi: 10.1038/ncomms1928.

Ravcheev DA, Li X, **Latif H**, Zengler K, Leyn SA, Korostelev YD, Kazakov AE, Novichkov PS, Osterman AL, Rodionov DA. Transcriptional regulation of central carbon and energy metabolism in bacteria by redox-responsive repressor Rex. J Bacteriol. 2012 Mar;194(5):1145-57. doi: 10.1128/JB.06412-11.

Ye J, Alvin K, **Latif H**, Hsu A, Parikh V, Whitmer T, Tellers M, de la Cruz Edmonds MC, Ly J, Salmon P, Markusen JF. Rapid protein production using CHO stable transfection pools. Biotechnol Prog. 2010 Sep-Oct;26(5):1431-7. doi: 10.1002/btpr.469.

* Authors contributed equally.

ABSTRACT OF THE DISSERTATION

An integrated workflow for the multi-omic characterization of microorganisms

by

Haythem Latif

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2015

Professor Bernhard Ø. Palsson, Chair

In this dissertation, I provide a generalized framework for the in depth molecular characterization of microorganisms using multi-omic data integration. Next-generation sequencing is rapidly becoming a staple in biological research. However, as data generation becomes more routine, greater attention must be placed on the analytics needed to extract useful information from these datasets. Recent work using multi-omic characterization approaches have revealed that microbial genomes and their organizational features are far more complex than previously thought. Here, I present a multi-omic data integration strategy that updates and expands upon previously implemented workflows that follow four principles to study microbial transcription, translation, and regulation: (1) data generation, (2) data processing, (3) data in-

tegration, (4) data analysis. At the core of this integrative workflow is the a complete and accurate reference genome assembly. Chapter 2 discusses updated sequences and gene annotation for *Thermotoga maritima* and *Escherichia coli* O157:H7 EDL933 using next-generation sequencing. With the genome sequence revealed, expanded annotation of cellular features is achieved quantitatively using a blend of genome-scale experimental methods and complimentary bioinformatics approaches where the genome serves as a normalizing factor. In Chapter 3 this multifaceted approach is used to elucidate the genome organization of *T. maritima* and revealed novel insights into its hyperthermophilic lifestyle. The detailed characterization of the *T. maritima* genome organization was applied in Chapter 4 where the genotype-to-phenotype relationship associated with laboratory evolved cultures were revealed. A genomic region excluded from the original genome sequence was found to be modulated in response to the applied selective pressure, underscoring the importance of the accuracy of the genome sequence. Furthermore, recently developed next-generation approaches were implemented in the multi-omic workflow to provided *in vivo*, genome-scale data on transcriptional regulation (ChIP-exo, Chapter 5) and protein translation (ribosome profiling, Chapter 6). These assays provide stark improvements compared to previously implemented methodologies with respect to their resolution, signal-to-noise, and cost. They also reveal detailed molecular interactions that previously could not be discerned at genome-scale. Collectively, the workflow utilized here will enable researchers to rapidly and cost-effectively characterize microbial systems beyond the one-dimensional genome annotation and towards complete elucidation of the multi-dimensional genome organization.

Chapter 1

Introduction

1.1 The genome that launched 1,000 genomes.

Sequencing of the 5.4 kb viral ΦX174 genome in 1977 by Frederick Sanger [1] marks a seminal moment in the life sciences. This genome was the first for which the entire nucleotide sequence was revealed. The following two and half decades would see substantial progress in the size of whole genomes assembled beginning with larger viruses [2] and organelles [3, 4], then to free-living bacteria [5, 6], unicellular eukaryotes [7], and culminated with the completion of the human genome project in 2003. However, as the genomes grew so did the consortia needed to piece them together. For example, assembly of the 4.2 Mb *Bacillus subtilis* genome included 25 European laboratories, 7 in Japan, 1 in Korea, and 2 biotechnology companies and took approximately 10 years to complete [5].

The once monumental task of genome sequencing is now considered routine thanks in large part to advances in sequencing technologies and the accompanying bioinformatics tools. In the mid 1990's technologies were actively being developed that leveraged massively parallel sequencing to rapidly and accurately reveal the composition of short DNA fragments. These technologies included pyrosequencing [8], cyclic reversible termination [9], oligonucleotide ligation [10], and real-time sequencing [11]. Collectively, these 'next-generation' sequencing approaches have, for the most part, replaced Sanger sequencing by reducing the time and cost associated with producing a genome sequence while simultaneously improving accuracy

[12, 13, 14, 15, 16]. Now, the *Bacillus subtilis* genome can be sequenced and assembled by a single researcher in a matter of days for under $7 (based on 2014 estimates of the cost per Mb of DNA with a 30X coverage depth [17]). In 2008, an international effort was launched with the goal of sequencing 1,000 human genomes. Thanks to next-generation sequencing, this project was completed in 2012 and detailed genetic variations among 1,092 individuals [18].

1.2 The complexity of microbial genomes revealed by multi-omic characterization.

Microbial whole-genome sequencing projects have largely out paced those for eukaryotes [19]. Though sequencing of these organisms is comparatively easier, multi-omic characterization studies have made it evident that microbial genome sequences are not unidimensional, but rather intricately structured and multifaceted. Several studies utilizing omics approaches have demonstrated that the genome and the control of phenotypic states found in microorganisms is far more complex than previously thought. For example, many studies led to the alteration of the traditional operon model. Operons, thought to rigidly transcribe all genes contained within, were shown to be broken up into operonic and intraoperonic transcription units (TUs) [20, 21, 22, 23, 24, 25]. TUs afford microorganisms increased versatility in coordinating the expression of gene cassettes. TUs are also often associated with multiple transcription start sites (TSSs) [20, 21, 23, 24, 22]. Multiple TSSs can be indicative of RNA polymerase recruitment by different sigma factors and transcription factors [26, 27]. They have also been implicated in regulatory mechanisms involving the $5'$ untranslated region [24]. Microbial transcriptomes have also been shown to contain numerous small RNAs, non-coding RNAs, and antisense RNAs that were undetectable using bioinformatics alone [23, 24]. Studies characterizing proteomes have revealed that annotated start codons are often incorrect [20, 23]. These studies also showed the existence of open reading frames that were missed by bioinformatics annotation methods. Lastly, studies of regulation have shown highly interconnected networks that govern the flow of information in microorganisms both at the transcriptional

and post-transcriptional levels [22, 28, 24, 23].

1.3 Systems microbiology approach to omics data integration.

The growing availability of high-throughput omics technologies is providing researchers with unprecedented access to genome-scale data for a continually expanding list of biological subsystems. As a field, systems biology not only wants to understand living systems but to also predict their behavior. These high-throughput approaches provide a basis upon which further analyses can be conducted such as functional characterization, cross-species comparison, evolutionary studies, and *in silico* predictive modeling. Systems biology is inherently integrative as it offers researchers the ability to investigate biological phenomena in the context of global networks. Thus, systems approaches offer the best opportunity to not only standardize and centralize the abundance of omics data currently being generated but also provides a framework from which contextual information found therein can be extracted.

At its core, systems biology is comprised of a comprehensive catalog of components. The interactions among these components is subsequently defined to construct motifs, pathways, and ultimately networks [29]. In this thesis, the genome organization refers to the collection of annotated cellular features including (but not limited to) protein coding genes, structural RNAs, functional RNAs, promoters, transcription start sites, regulatory non-coding regions, untranslated regions, transcription units (TUs), transcriptional pause sites, translational pause sites, and ribosome binding sites (RBS) [20, 23]. This level of annotation is multi-dimensional and captures the elements responsible for the flow of information from genotype-to-phenotype. These cellular components can be projected onto defined genomic coordinates. Thus, the genome is the ideal molecular vantage point for normalizing multi-omic datasets and this representation of cellular features provides a basis upon which the elucidation of complex, interconnected cellular processes such as transcription, translation, and regulation can occur.

A general systems biology workflow for integrating multiple genome-scale data-

sets and data types across various levels of cellular organization (e.g., transcriptome, proteome, and regulation) has been defined and is comprised of four principle stages: (1) data generation, (2) data processing, (3) data integration, and (4) data analysis (Figure 1.1) [30]. With this methodology, relationships between data types are established that lead to the systematic elucidation of the genome organization. First, predetermined genome-scale experimental datasets are generated. Individual data types are then processed using appropriate quantitative metrics. For example, whole-genome tiling arrays are quantified by aligning the raw image file, then determining the fluorescence intensity, and ultimately normalizing the fluorescence intensity relative to those observed in other experimental conditions using an algorithm such as RMA (robust multi-array analysis). Features identified within the datasets are then oriented relative to genomic coordinates and strand assignments are made when appropriate. This process provides a foundation for concurrent and integrated data analysis of diverse data types. Furthermore, orthogonal sources of evidence are generated to support annotation of any given feature giving greater confidence that the annotation is accurate.

1.4 Previous implementations of the systems approach to multi-omic data integration.

The first iterations of this multi-omic workflow were conducted on two microorganisms: *Escherichia coli* [20] and *Geobacter sulfurreducens* [23]. These studies used experimental datasets to improve the gene annotation and determine the TU architecture. Gene annotation was improved by using whole-genome, shotgun proteomics. Individual peptide fragments determined by LC-FTICR-MS (Fourier transform ion cyclotron resonance mass spectrometry) were mapped to the genome using a stop-to-stop peptide database. This enabled the identification of previously unannotated protein coding regions and the adjustment of the start codon assignment for numerous coding sequences. For *G. sulfurreducens* this yielded 55 new protein coding genes and 77 corrected open reading frames [23].

In both studies, the transcriptome was initially characterized using high-

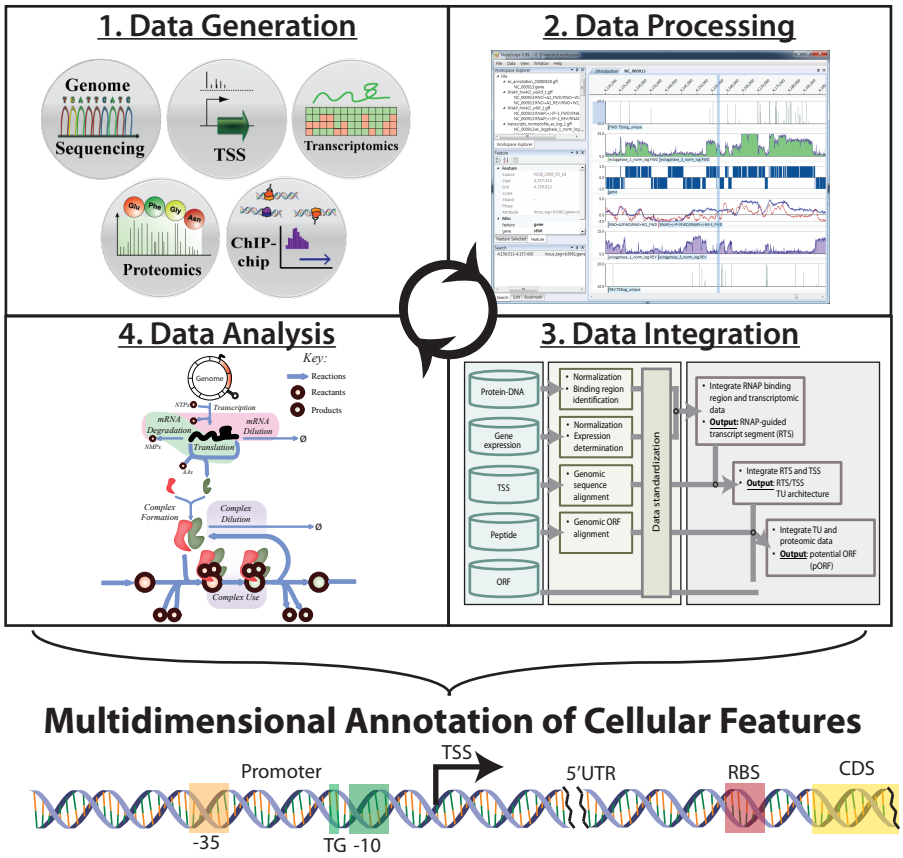


Figure 1.1: **The four principle steps of integrating multi-omic datasets.** This figure illustrates the systematic approach used to produces a multidimensional annotation of cellular features. These features are analyzed quantitatively for both intra- as well as interspecies comparison. The annotation is expanded to include features critical for transcription and translation initiation including promoter elements, RBSs, TSSs, and 5′ UTRs.

density tiling arrays to estimate transcript bounds. The use of high-density tiling arrays allowed for strand-specific detection of transcripts across the entire genome sequence with a resolution of 20-25 bp. The 5′ and 3′ bounds of transcripts were defined by merging consecutive probes with a "present" assignment. Gene expression libraries were generated from total RNA isolated under numerous growth conditions to ensure comprehensive elucidation of the transcriptome. The 5′ end of transcripts were then refined by locally searching for TSSs detected using a sequencing based methodology. Lastly, promoter regions were identified through ChIP-chip experiments performed on RNA polymerase and a variety of sigma factors. These experiments provide or-

thogonal confirmation of TSSs but the resolution is on the order of kilobases thus it cannot be used for precise determination of the actively transcribed TSS.

1.5 Updating and expanding the previous workflow.

Recent advances in experimental and bioinformatics approaches present an opportunity to iteratively improve upon previous multi-omic data integration efforts while remaining true to the principles outlined in Figure 1.1. Here, these advances are implemented to produce a data integration strategy that has distinct advantages over its predecessors. First, a suite of bioinformatics tools were incorporated to complement experimental approaches. This was applied to the characterization of the *Thermotoga maritima* genome organization ([31], Chapter 3). In doing so, the cellular features revealed included not only improvements to the gene annotation and definition of the TU architecture but also expanded the annotation to include elements critical for transcription and translation initiation and termination. The elements include promoter elements (-10, extended -10, and -35 motifs), RBSs, intrinsic termination sites, and $5'$ and $3'$ UTRs. This expanded set of cellular features enabled detailed, cross-species comparative analysis which indicated a possible relationship between the sequence conservation of genetic elements and the environment. The data generated in this study in combination with parallel efforts to detail the regulatory network of *T. maritima* proved to be of great importance to elucidating the genotype-to-phenotype relationship associated with genetic changes observed as a result of adaptive evolution (see Chapter 4). Interestingly, one of the genomic regions most impacted by the evolution is contained in a cassette that was not included in the original reference genome assembly. This, as well as another example of incomplete genome assemblies, are discussed in Chapter 2) to emphasize the importance of an accurate genome sequence.

Another major benefit of the integrated workflow presented here is in the quality of the datasets that are utilized. Previous multi-omic approaches relied on different platforms to elucidate various levels of cellular information. Sequencing was used for

determining the genome composition. Microarrays were then constructed based on the genome sequence to study the transcriptome and transcription factor regulatory networks. Highly sophisticated and expensive mass spectrometry approaches were needed to study the proteome. These methods, while highly informative, have their limitations. Many of these limitations are overcome by recent improvements and expansions of next-generation sequencing methodologies.

Perhaps the most logical experimental improvement is in the use of RNA-seq for quantitative studies of gene expression instead of microarray technology. RNA-seq produces single nucleotide resolution data and provides quantitative information on transcript abundance over a broad dynamic range that is orders of magnitude greater than that obtained from microarrays. The resolution of RNA-seq allows for identification of small RNAs, anti-sense RNAs, previously unidentified transcripts, and the determination of TUs with precise bounds. Furthermore, arrays required the design and synthesis of organism specific chips based on *a priori* knowledge of the genome sequence. Given the current costs of sequencing and time needed to produce custom arrays, it is without question advantageous to utilize sequencing. RNA-seq is used in multi-omic data integration studies located in Chapters 3, 4, and 5.

Characterization of protein/DNA interactions has also benefited from advances in next-generation sequencing. ChIP-chip was the staple methodology for studying protein/DNA interactions *in vivo* and at genome-scale. Subsequently, a next-generation sequence approach (ChIP-seq) was devised that provided improved resolution of ChIP peak regions (≈200-300 bp). Recently a derivative of ChIP-seq was developed that improves the identification of ChIP binding regions to ≈30 bp, has increased signal-to-noise, and has reduced rates of false positives compared to ChIP-chip and ChIP-seq. This method, ChIP-exonuclease (ChIP-exo) [32], uses two nuclease digestion steps to produce a high resolution footprint of protein protected regions of DNA. The value of this methodology is highlighted in Chapter 5 where a multi-scale study of the *E. coli* Crp regulon was conducted. Using ChIP-exo, one can begin to deconvolve the mechanistic aspects of promoter architecture and gene expression. Ultimately, this can one day be connected with systems approaches to provide a mechanistic model of transcription.

Less apparent have been the advancements made using next-generation approaches to characterize the proteome. Recently, an experiment protocol was devised that sequences ribosome protected fragments of mRNA. This gives an *in vivo*, genome-scale snapshot of actively translated proteins. Like shotgun proteomics, ribosome profiling can identify the start codon but has the added benefits of locating the RBS. Furthermore, ribosome profiling yields information along an entire coding region whereas peptide sequences are often sparsely detected along a coding region. Therefore, ribosome profiling offers a cheaper alternative to costly shotgun proteomics. Ribosome profiling has also been used to expand the annotation of genomic features. For instance, this protocol was used to determine translation pause sites associated with anti-Shine Dalgarno sequences in *E. coli* [33]. In addition to the annotation benefits and detailed molecular mechanisms revealed using this protocol, it enables quantitative measure of the proteome and has been used for predictive modeling [34, 35, 36]. In Chapter 6 a detailed ribosome profile protocol is described that increases the accessibility of this method to the greater community of microbiologists by streamlining laborious steps.

Collectively, the updates and improvement described here to the systems approach to multi-omic data integration workflows results in the easier, faster, and more accurate annotation of a broader set of cellular features. This is achieved using sequencing technology and bioinformatics approaches which are becoming more ubiquitous and cost effective for researchers to access. Therefore, we can propel microbial characterization studies beyond the mere sequencing and annotation of genomes towards obtaining a more sophisticated, comprehensive view of the genome organization.

1.6 Bibliography

[1] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage ΦX174 DNA. Nature 265: 687-95.

[2] Sanger F, Coulson AR, Hong G, Hill D, Petersen G (1982) Nucleotide sequence of bacteriophage λ DNA. Journal of Molecular Biology 162: 729–773.

[3] Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457–465.

[4] Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota Si, Inokuchi H, Ozeki H (1986) Chloroplast gene organization deduced from complete sequence of liverwort marchantia polymorpha chloroplast DNA. Nature 322: 572–574.

[5] Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessières P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Cordani JJ, Connerton IF, Cummings NJ, Daniel RA, Denziot F, Devine KM, Düsterhöft A, Ehrlich SD, Emmerson PT, Entian KD, Errington J, Fabret C, Ferrari E, Foulger D, Fritz C, Fujita M, Fujita Y, Fuma S, Galizzi A, Galleron N, Ghim SY, Glaser P, Goffeau A, Golightly EJ, Grandi G, Guiseppi G, Guy BJ, Haga K, Haiech J, Harwood CR, Hènaut A, Hilbert H, Holsappel S, Hosono S, Hullo MF, Itaya M, Jones L, Joris B, Karamata D, Kasahara Y, Klaerr-Blanchard M, Klein C, Kobayashi Y, Koetter P, Koningstein G, Krogh S, Kumano M, Kurita K, Lapidus A, Lardinois S, Lauber J, Lazarevic V, Lee SM, Levine A, Liu H, Masuda S, Mauël C, Médigue C, Medina N, Mellado RP, Mizuno M, Moestl D, Nakai S, Noback M, Noone D, O'Reilly M, Ogawa K, Ogiwara A, Oudega B, Park SH, Parro V, Pohl TM, Portelle D, Porwollik S, Prescott AM, Presecan E, Pujic P, Purnelle B, Rapoport G, Rey M, Reynolds S, Rieger M, Rivolta C, Rocha E, Roche B, Rose M, Sadaie Y, Sato T, Scanlan E, Schleich S, Schroeter R, Scoffone F, Sekiguchi J, Sekowska A, Seror SJ, Serror P, Shin BS, Soldo B, Sorokin A, Tacconi E, Takagi T, Takahashi H, Takemaru K, Takeuchi M, Tamakoshi A, Tanaka T, Terpstra P, Togoni A, Tosato V, Uchiyama S, Vandebol M, Vannier F, Vassarotti A, Viari A, Wambutt R, Wedler H, Weitzenegger T, Winters P, Wipat A, Yamamoto H, Yamane K, Yasumoto K, Yata K, Yoshida K, Yoshikawa HF, Zumstein E, Yoshikawa H, Danchin A (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249-56.

[6] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.

[7] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274: 546, 563-7.

[8] Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Analytical Biochemistry 242: 84-9.

[9] Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Research 15: 1767–1776.

[10] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Research 18: 1051–1063.

[11] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138.

[12] Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1135–1145.

[13] Metzker ML (2009) Sequencing technologies — the next generation. Nature Reviews Genetics 11: 31–46.

[14] Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387–402.

[15] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends in Genetics 24: 133–141.

[16] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends in Genetics 30: 418–426.

[17] Wetterstrand KA (2014). DNA sequencing costs: Data from the nhgri genome sequencing program (gsp). URL www.genome.gov/sequencingcosts.

[18] Consortium GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

[19] Reddy T, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA, Kyrpides NC (2014) The Genomes OnLine

Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification. Nucleic Acids Research : gku950.

[20] Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO (2009) The transcription unit architecture of the *Escherichia coli* genome. Nature Biotechnology 27: 1043–1049.

[21] Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326: 1268–1271.

[22] Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Hartig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RA, Nannapaneni P, Noone D, Pohl S, Rinn B, Rugheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stulke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Volker U, Bessieres P, Noirot P (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science 335: 1103–1106.

[23] Qiu Y, Cho BK, Park YS, Lovley D, Palsson BO, Zengler K (2010) Structural and operational complexity of the *Geobacter sulfurreducens* genome. Genome Research 20: 1304–1311.

[24] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464: 250–255.

[25] Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. Genome Biology 12: R47.

[26] Qiu Y, Nagarajan H, Embree M, Shieu W, Abate E, Juárez K, Cho BK, Elkins JG, Nevin KP, Barrett CL, Lovley DR, Palsson BO, Zengler K (2013) Characterizing the interplay between multiple levels of organization within bacterial sigma factor regulatory networks. Nature Communications 4: 1755.

[27] Cho BK, Kim D, Knight EM, Zengler K, Palsson BO (2014) Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states. BMC Biology 12: 4.

[28] Buescher JM, Liebermeister W, Jules M, Uhr M, Muntel J, Botella E, Hessling B, Kleijn RJ, Le Chat L, Lecointe F, Mader U, Nicolas P, Piersma S, Rugheimer

F, Becher D, Bessieres P, Bidnenko E, Denham EL, Dervyn E, Devine KM, Doherty G, Drulhe S, Felicori L, Fogg MJ, Goelzer A, Hansen A, Harwood CR, Hecker M, Hubner S, Hultschig C, Jarmer H, Klipp E, Leduc A, Lewis P, Molina F, Noirot P, Peres S, Pigeonneau N, Pohl S, Rasmussen S, Rinn B, Schaffer M, Schnidder J, Schwikowski B, Van Dijl JM, Veiga P, Walsh S, Wilkinson AJ, Stelling J, Aymerich S, Sauer U (2012) Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. Science 335: 1099–1103.

[29] Oltvai ZN, Barabási AL (2002) Lifes complexity pyramid. Science 298: 763–764.

[30] Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nature Chemical Biology 6: 787–789.

[31] Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K (2013) The genome organization of *Thermotoga maritima* reflects its lifestyle. PLoS Genetics 9: e1003485.

[32] Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147: 1408–1419.

[33] Li GW, Oh E, Weissman JS (2012) The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541.

[34] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223.

[35] Li GW, Burkhardt D, Gross C, Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624–635.

[36] Subramaniam AR, Zid BM, OShea EK (2014) An integrated approach reveals regulatory controls on bacterial translation elongation. Cell 159: 1200–1211.

Chapter 2

Genome assembly improvements using next-generation sequencing technology.

2.1 Abstract

The importance of a pristine reference genome sequence cannot be understated. Advances made in next-generation sequencing have enabled researchers to rapidly elucidate the genome sequence and subsequently annotate protein coding regions and structural RNAs. Here, two examples of genome assemblies are presented that illustrate the improvements resulting from next-generation sequencing. The first example details an ≈9 kb gap present in the *Thermotoga maritima* reference assembly which was detected as part of an multi-omic characterization study to elucidate the genome organization of this organism (see Chapter 3). Whole-genome, paired-end sequencing was conducted using Illumina's technology platform and then annotated which revealed that the gap region encodes key metabolic and regulatory genes. The importance of the genes in this region is revisited in Chapter 4. The second example covers the assembly of a susbtantially more complicated bacterial genome. *Escherichia coli* EDL933 is the prototypic strain of enterohemorrhagic *E. coli* serotype O157:H7 and is associated with deadly foodborne outbreaks. The publicly available

reference assembly for EDL933 has >6,000 ambiguous base calls and has large gaps as a result of several long phage related repeat regions. Through PacBio long read sequencing and Illumina short read sequencing, the assembly was updated to produce a complete, high-quality, unambiguous genome sequence. Both genome sequences are made publicly available to researchers under accessions NC_021214 for *T. maritima* and CP008957.1 (chromosome) and CP008958.1 (plasmid) for *E. coli* O157:H7 EDL933.

2.2 Introduction

The genome sequence is vital to our understanding of living systems. It is essential to a myriad of analyses spanning multiple scales ranging from biomolecular characterization to systems approaches, comparative analysis, and evolutionary studies to name a few. Our understanding of the molecular composition of a macromolecules that participate and carry out complex cellular processes such as transcription, translation, and regulation is contingent on the availability of the genome sequence and annotation [1, 2, 3, 4, 5, 6, 7, 8]. Bottom-up reconstruction of genome-scale models and predictive modeling is founded upon the genetic makeup of an organism [9, 10]. Furthermore, the source and spread of pathogenic outbreaks has been traced using whole-genome sequencing of isolates from infected patients [11, 12, 13, 14, 15]. Evolution studies use similar sequencing approaches followed by comparative analysis to unveil the genotype-to-phenotype relationship [16]. Though only a small sampling of the analyses reliant on genome sequence information, they are all united in the need for a highly accurate, comprehensive genome assembly.

Next-generation sequencing has enabled researchers to sequence entire microbial genomes in a matter of days, with greater accuracy, and at a fraction of the cost of Sanger sequencing [17, 18, 19, 20, 21]. Of equal importance is the development of bioinformatics approaches that rapidly assemble sequenced reads into large contigs (e.g., Velvet [22], SOAPdenovo [23], SPADES [24]) and annotation pipelines for functional prediction of genes (e.g., GeneMark [25], RAST [26, 27]). A challenge for assembly of any genome is the presence of repeat regions that cannot be resolved

using short read technologies. Recently, microbial genomes were classified according to their assembly complexity based on their composition of repeat regions [28]. The simplest of the three classes, Class I, contains genomes with few, short repeats followed by Class II genomes which have longer repeats (e.g., IS elements) but the rRNA operon is the largest repeat region. Lastly, Class III genomes contain large repeats (e.g., phage regions, duplications, large tandem repeats) that significantly exceed the rRNA operon. These classes also correspond with the level of difficulty in producing a finished genome assembly with Class III being the most difficult to complete.

Here, a Class I (*Thermotoga maritima*) genome and a Class III (*Escherichia coli* serotype O157:H7) genome are sequenced to illustrate that recent advances in next-generation sequencing platform read lengths, quality, and costs enable researchers to rapidly produce pristine fished microbial genomes. Both genomes previously had reference assemblies generated via Sanger sequencing and were found to have large gaps. All gaps, errors, ambiguities, and misassemblies were resolved for these genomes to produce new reference assemblies. Furthermore, the genomes were annotated to reflect the updated genome sequences. The genome sequence and functional annotation are deposited in publicly available repositories.

2.3 Results and Discussion

2.3.1 *Thermotoga maritima* ATCC genomovar sequence and gene annotation for a missed ≈9 kb region

T. maritima is a phylogenetically deep-branching, hyperthermophilic bacterium with a compact 1.86 Mb genome first reveal by Nelson et al. in 1999 [29]. Originally isolated from geothermally heated marine sediment, *T. maritima* grows between 60-90 °C with an optimal growth temperature of 80 °C [30]. This species belongs to the order Thermotogales that have, until recently, been exclusively comprised of thermophilic or hyperthermophilic organisms. Organisms in hydrothermal vent communities, where many Thermotogales have been isolated, are thought to harbor traits of early microorganisms [31]. Phylogenetic analysis of 16S rRNA sequences

place the Thermotogae at the base of the bacterial phylogenetic tree [32, 33]. Members of this phylum are among the deepest branching bacterial species and, as such, are prime candidates for evolutionary studies. *T. maritima* has one of the smallest genomes in this order and maintains one of the most compact genomes among all sequenced bacterial species (<5% noncoding DNA) [29, 34]. The short intergenic regions in the *T. maritima* genome (5 bp median) resemble those in the genome of *Pelagibacter ubique*, a bacterium that has undergone genome streamlining and has the shortest median intergenic space (3 bp) among free-living bacteria [34]. *T. maritima* has only a few short repeat regions and a single rRNA operon which is often the largest repeat region in bacterial genomes. Therefore, this genome is of Class I.

Recently, an ≈9 kb chromosomal region in the DSMZ derived strain (DSMZ genomovar, Genbank Accession AGIJ00000000.1) was identified that was absent in the original genome sequence which was derived from a strain of *T. maritima* grown at TIGR (Genbank Accession AE000512.1, NC_000853) [35]. Therefore, as part of a larger multi-omic characterization study of the genome organization of *T. maritima* (see Chapter 3), whole genome resequencing of the ATCC derived strain was performed. Briefly, paired-end resequencing libraries were prepared from genomic DNA following standard Illumina protocols and sequenced on an Illumina GAIIx platform. The updated genome sequence was assembled as follows: (1) Reads were aligned to the 8.9 kb region identified in the *T. maritima* MSB8 DSMZ genomovar [35] and the TIGR genomovar sequence using SHOREmap [36] and MosaikAligner (http://bioinformatics.bc.edu/marthlab/Mosaik). (2) Unaligned reads were *de novo* assembled using Velvet [22] to ensure no additional assemblies were present. (3) The sequence was corrected for SNPs and indels detected during read alignment. A greater discussion of the genome improvements that resulted from resequencing and updating the annotation are reserved for Chapter 3. Here, only the chromosomal gap region in the TIGR genomovar is discussed.

Resequencing results indicate that the ATCC genomovar (NC_021214) contains a genomic region that aligns well to the DSMZ genomovar further indicating that the TIGR strain is divergent from the strains used by most researchers studying *T. maritima*. Though the discovery of the gap region is important, the functional and

regulatory content encoded therein is of greater significance. This was revealed for the ATCC genomovar by annotating the updated genome sequence using the RAST pipeline with the default parameters [26]. Predicted gene sequences were mapped to the AE000512.1 annotation using a bidirectional Smith-Waterman alignment to identify the corresponding locus tags from the TIGR genomovar. In agreement with the findings of Boucher et al. [35], this region in the ATCC genomovar was found to contain 7 protein-coding regions including two ABC importer operons and beta-glucosidase (Tmari_1856 thru Tmari_1862) (Figure 2.1). Furthermore, annotation revealed that a truncated ROK family transcriptional regulator in the TIGR genomovar (TM1847) is not truncated in the ATCC genomovar (Tmari_1855). As part of a parallel effort, functional testing of the ATCC genomovar showed that the ABC transporters primarily import glucose (*gluEFK*) and trehalose (*treEFG*) which are both locally regulated in part by Tmari_1855 (*gluR*) [37, 38].

The physiological impact of the genes present in the gap region is reserved for detailed discussion in Chapter 4. However, the existence of the genomic gap described here negatively impacts efforts to fully capture the organism's metabolic and regulatory capabilities.

2.3.2 *Escherichia coli* serotype O157:H7 EDL933 assembly.

E. coli serotype O157:H7, a causative agent in food poisoning outbreaks leading to hemorrhagic colitis or hemolytic uremic syndrome, gained public attention following its association with an outbreak in 1993 related to the U.S. fast-food chain Jack-in-the-Box [39, 40] and another large outbreak among school children in Sakai, Japan in 1996 [41]. Moreover, *E. coli* O157:H7 strains are known for their prophage-rich genomes and are currently considered the bacterial genomes with the largest number of integrated phages [42, 43]. Strain EDL933 (ATCC 43895), isolated from ground beef linked to a massive hamburger outbreak in Michigan, USA in 1982 [44], is the prototypic reference strain representing this pathotype.

Although the full genome of EDL933 was sequenced and published in 2001 [43], the deposited assembled genome has 6,000 ambiguous base calls and a chromosomal gap of 4,000 bp. While the utility of this reference genome, cited in 3,200 publications,

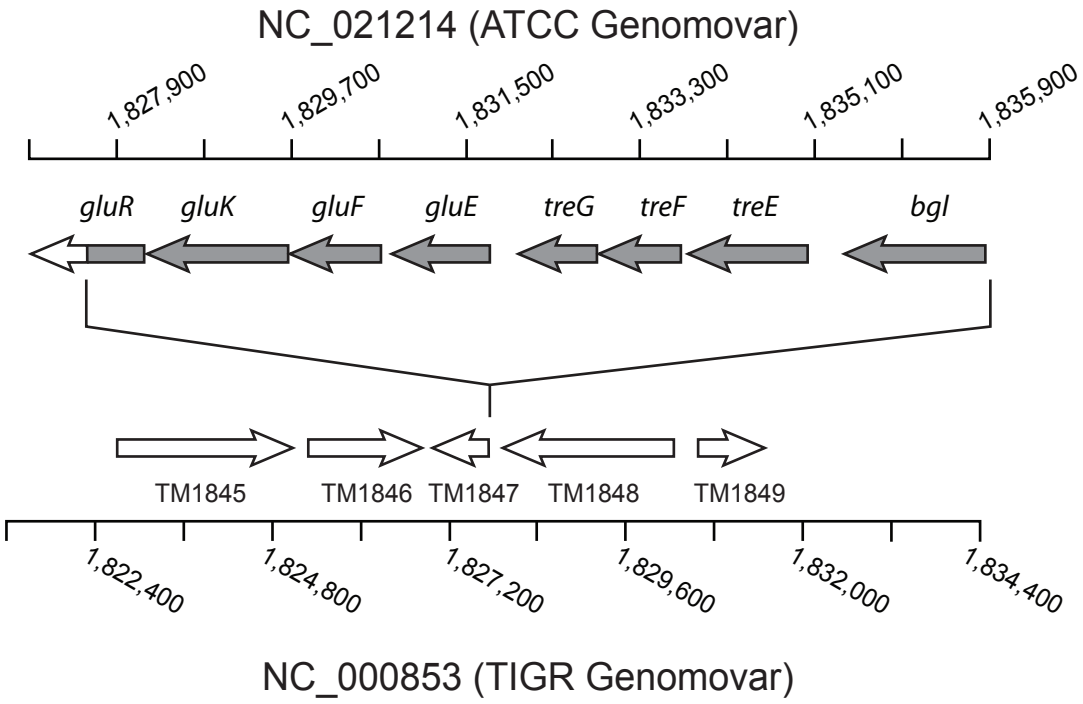


Figure 2.1: ***T. maritima* genome gap revealed.** Sequencing of the ATCC genomovar (NC_021214) revealed 7 protein coding genes and the correction to a truncated gene. These genes encode two ABC transporters primarily responsible for glucose (*gluEFK*) and trehalose (*treEFG*) transport, a beta-glucosidase gene, and a local acting transcription factor (*gluR*). Newly identified genes are shown in gray arrows while white arrows signify genes present in the TIGR genomovar (NC_000853).

is indisputable, several analyses reliant on a pristine reference (e.g., single nucleotide polymorphism studies) are hindered by those ambiguities and gaps. EDL933 has long phage-associated repeat regions 7 kb. Microbial genomes with these characteristics are the most complex to assemble [28], so we resorted to single-molecule sequencing using PacBio followed by polishing using Illumina short reads to complete the EDL933 sequence. This produced a gapless genome assembly, with no ambiguous base calls, and an updated genome annotation.

Genomic DNA from the EDL933 strain was prepared for PacBio and Illumina sequencing. PacBio libraries were prepared according to standard library preparation procedures with Blue Pippen size selection for 20-kb fragments and sequenced using P5/C3 chemistry and 3 h movies on the RS II system at the UCSD Genomics Core, San Diego, CA. Illumina libraries were prepared according to the TruSeq DNA PCR-

Free sample preparation kit protocol (Illumina) and paired-end sequenced (2x250) on a MiSeq. SMRTAnalaysis 2.2.0 HGAP v2 assembly of PacBio reads (66,927) produced three polished contigs: two corresponding to the chromosome and one the plasmid. When compared with the reference, NC_002655, a region of high read density within the two chromosomal contigs was shown to be a large duplication that unites the two contigs. After the plasmid and chromosome were circularized, reads were mapped back to the assembled sequences to check for variants by first using Bridge Mapper (RS_BridgeMapper.1) with PacBio reads and then Breseq v0.24rc6 [45] with Illumina short reads. Coverage was ≈100X for PacBio data and ≈300X for Illumina data. The final assembled genome was automatically annotated, then manually corrected, through the RAST server using SEED annotation tools [27, 26].

The updated EDL933 genome consists of a 5,547,323-bp chromosome and a 92,076-bp plasmid compared with 5,528,445 and 92,077-bp in the current EDL933 assembly (Figure 2.2). This gapless assembly eliminates 6,641 ambiguous base calls in the current EDL933 chromosome including 2,413 non-N ambiguous bases and 4,000 Ns belonging to a chromosomal gap. The updated genome has 5,675 and 97 annotated coding sequences (CDSs) compared with 5,286 and 99 CDSs found in the current reference chromosome and plasmid, respectively.

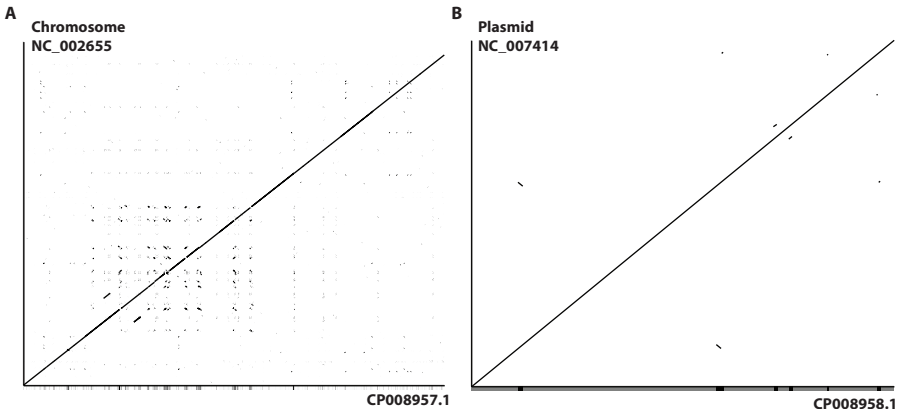


Figure 2.2: **Updated *E. coli* serotype O157:H7 EDL933 assembly.** Synteny plot comparing genome sequences for the EDL933 chromosome (A) and plasmid (B). The x-axis is the updated genome and the y-axis shows the original reference assembly.

2.4 Conclusion

Here, we have demonstrated how microbial genome reference sequences can suffer from gaps and errors. Next-generation sequencing was leveraged to help refine the genome assemblies of a Class I genome (*T. maritima*) and a substantially more complicated Class III genome (*E. coli* O157:H7 EDL933). The former is the model organism of its phylum and is of evolutionary and biotechnological interest. Resequencing of the *T. maritima* genome corroborates findings that the TIGR genomovar sequence contains an ≈9 kb gap compared compared with the ATCC genomovar. The gap region encodes two ABC transporter cassettes, a local acting transcriptional regulator, and a beta-glucosidase gene. These genes reveal key metabolic and regulatory insights into glucose and trehalose uptake for this organism. The updated genome sequence of the phage-rich *E. coli* O157:H7 EDL933 genome yields an important reference for pathogenic research. We applied PacBio technology to generate a broad range of read lengths spanning several kb's and then complemented that with high sequencing depth Illumina short reads for polishing. The lineage of outbreaks related to this strain can now be studied using a complete and unambiguous reference genome. Furthermore, the approach used here can be applied to characterize other Class III microorganisms by, at a minimum, utilizing long PacBio reads (or equivalent). Collectively, these example illustrate the importance of having a pristine genome sequence and that high-fidelity assemblies can be obtained through current next-generation sequencing technologies.

2.5 Acknowledgements

For the work presented on *T. maritima* we acknowledge Dmitry Rodionov and Andrei Osterman of the Sanford-Burnham Medical Research Institute, La Jolla, CA, for providing TF binding site loci, and Daniela Bezdan for assistance with the genome annotation. We also thank Bernhard Palsson, Pep Charusanti, and Ramy K. Aziz for guidance with manuscript preparation. Funding: Funding for this work was provided by the Office of Science of the U.S. Department of Energy (DOE) under grants DE-FG02-08ER64686 and DE-FG02-09ER25917. HL is supported through the National

Science Foundation Graduate Research Fellowship under grant DGE1144086. Proteomics capabilities were developed under support from the DOE Office of Biological and Environmental Research (BER) Pan-omics Project and the NIH National Center for Research Resources (RR018522), and a significant portion of this work was performed in the Environmental Molecular Sciences Laboratory (EMSL), a DOE-BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RLO 1830. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Conceived and designed the experiments: HL JAL VAP KZ. Performed the experiments: HL VAP YT D-HL KZ. Analyzed the data: HL JAL HN ACS-R JNA YQ. Contributed reagents/materials/analysis tools: RDS JNA. Wrote the paper: HL JAL HN KZ.

The work presented on *E. coli* O157:H7 was supported by the National Institute of Health under grant GM098105. Haythem Latif is supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086.

Chapter 2 is in part adapted from two published manuscripts: (1) Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K. (2013) The Genome Organization of *Thermotoga maritima* Reflects Its Lifestyle. PLoS Genet 9(4): e1003485. doi:10.1371/journal.pgen.1003485. * indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Joshua A. Lerman (equal contributor), Vasiliy A. Portnoy, Yekaterina Tarasova, Harish Nagarajan, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Dae-Hee Lee, Yu Qiu, and Karsten Zengler. (2) Latif H, Li HJ, Charusanti P, Palsson BØ., Aziz RK. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 Strain EDL933. Genome Announc. 2014 Aug 14;2(4). pii: e00821-14. doi: 10.1128/genomeA.00821-14. The dissertation author was the primary author of this paper responsible for the research. The other authors were Howard J Li, Pep Charusanti, Bernhard Ø. Palsson, and Ramy K Aziz.

2.6 Bibliography

[1] Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nature Chemical Biology 6: 787–789.

[2] Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326: 1268–1271.

[3] Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castano-Diez D, Chen WH, Devos D, Guell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Bottcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin AC (2009) Proteome organization in a genome-reduced bacterium. Science 326: 1235–1240.

[4] Qiu Y, Cho BK, Park YS, Lovley D, Palsson BO, Zengler K (2010) Structural and operational complexity of the *Geobacter sulfurreducens* genome. Genome Research 20: 1304–1311.

[5] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464: 250–255.

[6] Yoon SH, Reiss DJ, Bare JC, Tenenbaum D, Pan M, Slagel J, Moritz RL, Lim S, Hackett M, Menon AL, Adams MW, Barnebey A, Yannone SM, Leigh JA, Baliga NS (2011) Parallel evolution of transcriptome architecture during genome reorganization. Genome Research 21: 1892–1904.

[7] Buescher JM, Liebermeister W, Jules M, Uhr M, Muntel J, Botella E, Hessling B, Kleijn RJ, Le Chat L, Lecointe F, Mader U, Nicolas P, Piersma S, Rugheimer F, Becher D, Bessieres P, Bidnenko E, Denham EL, Dervyn E, Devine KM, Doherty G, Drulhe S, Felicori L, Fogg MJ, Goelzer A, Hansen A, Harwood CR, Hecker M, Hubner S, Hultschig C, Jarmer H, Klipp E, Leduc A, Lewis P, Molina F, Noirot P, Peres S, Pigeonneau N, Pohl S, Rasmussen S, Rinn B, Schaffer M, Schnidder J, Schwikowski B, Van Dijl JM, Veiga P, Walsh S, Wilkinson AJ, Stelling J, Aymerich S, Sauer U (2012) Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. Science 335: 1099–1103.

[8] Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Hartig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat

L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RA, Nannapaneni P, Noone D, Pohl S, Rinn B, Rugheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stulke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Volker U, Bessieres P, Noirot P (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science 335: 1103–1106.

[9] Joyce AR, Palsson BØ (2006) The model organism as a system: integrating 'omics' data sets. Nature Reviews Molecular Cell Biology 7: 198–210.

[10] Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nature Reviews Genetics 7: 130–141.

[11] Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISCCSPG, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Science Translational Medicine 4: 148ra116.

[12] Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med 10: e1001387.

[13] Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. New England Journal of Medicine 364: 730–739.

[14] Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R (2011) Identification of a salmonellosis outbreak by means of molecular sequencing. New England Journal of Medicine 364: 981–982.

[15] Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, Wain J (2013) Public health value of next-generation dna sequencing of enterohemorrhagic *Escherichia coli* isolates from an outbreak. Journal of Clinical Microbiology 51: 232–237.

[16] Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243-7.

[17] Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1135–1145.

[18] Metzker ML (2009) Sequencing technologies—the next generation. Nature Reviews Genetics 11: 31–46.

[19] Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387–402.

[20] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends in Genetics 24: 133–141.

[21] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends in Genetics 30: 418–426.

[22] Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research 18: 821–829.

[23] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J (2012) Soapdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1: 18.

[24] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) Spades: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19: 455–477.

[25] Lukashin A (1998) Genemark.hmm: new solutions for gene finding. Nucleic Acids Research 26: 1107–1115.

[26] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The rast server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

[27] Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Stevens RL, Vonstein V, Xia F (2012) Seed servers: high-performance access to the seed genomes, annotations, and metabolic models. PLoS One 7: e48053.

[28] Koren S, Harhay GP, Smith T, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology 14: R101.

[29] Nelson KE (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399: 323–329.

[30] Huber R, Langworthy TA, Konig H, Thomm M, Woese CR, Sleytr UB, Stetter KO (1986) *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 °C. Archives of Microbiology 144: 324–333.

[31] Martin W, Baross J, Kelley D, Russell MJ (2008) Hydrothermal vents and the origin of life. Nature Reviews Microbiology 6: 805–814.

[32] Achenbach-Richter L, Gupta R, Stetter KO, Woese CR (1987) Were the original eubacteria thermophiles? Systematic and Applied Microbiology 9: 34–39.

[33] Munoz R, Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R (2011) Release LTPs104 of the All-Species Living Tree. Systematic and Applied Microbiology 34: 169–170.

[34] Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242–1245.

[35] Boucher N, Noll KM (2011) Ligands of thermophilic ABC transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8 screened by differential scanning fluorimetry. Applied and Environmental Microbiology 77: 6395–6399.

[36] Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nature Methods 6: 550–551.

[37] Kazanov MD, Li X, Gelfand MS, Osterman AL, Rodionov DA (2013) Functional diversification of rok-family transcriptional regulators of sugar catabolism in the thermotogae phylum. Nucleic Acids Research 41: 790-803.

[38] Rodionov DA, Rodionova IA, Li X, Ravcheev DA, Tarasova Y, Portnoy VA, Zengler K, Osterman AL (2013) Transcriptional regulation of the carbohydrate utilization network in *Thermotoga maritima*. Frontiers in Microbiology 4: 244.

[39] Chen J, Griffiths M (1999) Cloning and sequencing of the gene encoding universal stress protein from *Escherichia coli* o157: H7 isolated from jack-in-a-box outbreak. Letters in Applied Microbiology 29: 103–107.

[40] Pennington H (2010) *Escherichia coli* o157. The Lancet 376: 1428–1435.

[41] Michino H, Araki K, Minami S, Takaya S, Sakai N, Miyazaki M, Ono A, Yanagawa H (1999) Massive outbreak of *Escherichia coli* o157: H7 infection in school children in sakai city, japan, associated with consumption of white radish sprouts. American Journal of Epidemiology 150: 787–796.

[42] Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H (2003) Prophage genomics. Microbiology and Molecular Biology Reviews 67: 238–276.

[43] Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* o157:h7. Nature 409: 529–533.

[44] Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT, Blake PA, Cohen ML (1983) Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. New England Journal of Medicine 308: 681–685.

[45] Deatherage DE, Barrick JE (2014) Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. In: Engineering and Analyzing Multicellular Systems, Springer. pp. 165–188.

Chapter 3

The genome organization of *Thermotoga maritima* reflects its lifestyle

3.1 Abstract

The generation of genome-scale data is becoming more routine, yet the subsequent analysis of omics data remains a significant challenge. Here, an approach that integrates multiple omics datasets with bioinformatics tools was developed that produces a detailed annotation of several microbial genomic features. This methodology was used to characterize the genome of *Thermotoga maritima*—a phylogenetically deep-branching, hyperthermophilic bacterium. Experimental data were generated for whole-genome resequencing, transcription start site (TSS) determination, transcriptome profiling, and proteome profiling. These datasets, analyzed in combination with bioinformatics tools, served as a basis for the improvement of gene annotation, the elucidation of transcription units (TUs), the identification of putative non-coding RNAs (ncRNAs), and the determination of promoters and ribosome binding sites. This revealed many distinctive properties of the *T. maritima* genome organization relative to other bacteria. This genome has a high number of genes per TU (3.3), a paucity of putative ncRNAs (12), and few TUs with multiple TSSs (3.7%). Quan-

titative analysis of promoters and ribosome binding sites showed increased sequence conservation relative to other bacteria. The 5′UTRs follow an atypical bimodal length distribution comprised of 'Short' 5′UTRs ($11 - 17$ nt) and 'Common' 5′UTRs (26-32 nt). Transcriptional regulation is limited by a lack of intergenic space for the majority of TUs. Lastly, a high fraction of annotated genes are expressed independent of growth state and a linear correlation of mRNA/protein is observed (Pearson r = 0.63, $p < 2.2 \times 10^{-16}$ t-test). These distinctive properties are hypothesized to be a reflection of this organism's hyperthermophilic lifestyle and could yield novel insights into the evolutionary trajectory of microbial life on earth.

3.2 Author Summary

Genomic studies have greatly benefited from the advent of high-throughput technologies and bioinformatics tools. Here, a methodology integrating genome-scale data and bioinformatics tools is developed to characterize the genome organization of the hyperthermophilic, phylogenetically deep-branching bacterium *Thermotoga maritima*. This approach elucidates several features of the genome organization and enables comparative analysis of these features across diverse taxa. Our results suggest that the genome of *T. maritima* is reflective of its hyperthermophilic lifestyle. Ultimately, constraints imposed on the genome have negative impacts on regulatory complexity and phenotypic diversity. Investigating the genome organization of Thermotogae species will help resolve various causal factors contributing to the genome organization such as phylogeny and environment. Applying a similar analysis of the genome organization to numerous taxa will likely provide insights into microbial evolution.

3.3 Introduction

A fundamental step towards obtaining a systems-level understanding of organisms is to obtain an accurate inventory of cellular components and their interconnectivities [1, 2, 3]. The genome sequence and *in silico* predictions of gene an-

notation are the starting points for assembling a network. For prokaryotes, these *in silico* approaches detect open reading frames and structural RNAs with varying degrees of accuracy [4]. Recently, multi-omic data generation and analysis studies [5, 6, 7, 8, 9, 10, 11] have revealed an abundance of genomic features that are not detected computationally such as transcription start sites (TSSs), promoters, untranslated regions (UTRs), non-coding RNAs, ribosome binding sites (RBSs) and transcription termination sites [12]. However, the rate at which multi-omic datasets are being generated is substantially outpacing the development of analysis workflows for these inherently dissimilar data types [13]. Here, multi-omic experimental data is generated and analyzed in conjunction with bioinformatics tools to annotate numerous bacterial genomic features that cannot accurately be detected using *in silico* approaches alone. This methodology was applied to study the genome organization of *Thermotoga maritima*—a phylogenetically deep-branching, hyperthermophilic bacterium with a compact 1.86 Mb genome.

Originally isolated from geothermally heated marine sediment, *T. maritima* grows between $60 - 90°$C with an optimal growth temperature of 80°C [14]. This species belongs to the order Thermotogales that have, until recently, been exclusively comprised of thermophilic or hyperthermophilic organisms. Compared to most bacteria, Thermotogales are capable of sustaining growth over a remarkably wide range of temperatures. For instance, *Kosmotoga olearia* can be cultivated between $20 - 80°$C [15]. Recently, the existence of mesophilic Thermotogales [16, 17] was confirmed with the description of *Mesotoga prima*, which grows from $20 - 50°$C with an optimum at 37 °C [18]. Sequencing of *M. prima* revealed that it has the largest genome of all the Thermotogales at 2.97 Mb with ≈15% noncoding DNA [19]. *T. maritima*, which grows at the upper-limit known for Thermotogales, has one of the smallest genomes in this order and maintains one of the most compact genomes among all sequenced bacterial species (<5% noncoding DNA) [20, 21]. The short intergenic regions in the *T. maritima* genome (5 bp median) resemble those in the genome of *Pelagibacter ubique*, a bacterium that has undergone genome streamlining and has the shortest median intergenic space (3 bp) among free-living bacteria [20]. Although it remains unclear whether *T. maritima* has also undergone streamlining, both organisms en-

code only a few global regulators (four sigma factors in *T. maritima* versus two in *P. ubique*) and carry just a single rRNA operon. In contrast with *P. ubique*, *T. maritima* displays more metabolic diversity through its ability to ferment numerous mono- and polysaccharides [14, 22].

Thermotogales have been the focus of many evolutionary studies [23, 24, 25]. Organisms in hydrothermal vent communities, where many Thermotogales have been isolated, are thought to harbor traits of early microorganisms [26]. Phylogenetic analysis of 16S rRNA sequences place the Thermotogae at the base of the bacterial phylogenetic tree [27, 28]; however, Zhaxybayeva et al. [25] determined through analysis of 16S rRNA and ribosomal protein genes that Thermotogae and Aquificales (a hyperthermophilic order) are sister taxa. The authors also determined that the majority of Thermotogae proteins align best with those found in the order Firmicutes [25]; therefore, the exact phylogenetic position of Thermotogae is still unresolved. Nevertheless, members of this phylum are among the deepest branching bacterial species and, as such, prime candidates for evolutionary studies.

Thermophiles such as *T. maritima* implement numerous strategies at both the protein and nucleic acid levels to support growth at high temperatures. For instance, intrinsic protein stabilization is achieved by utilizing more charged residues at the protein surface, encoding for a dense hydrophobic core, and increasing disulfide bond usage [29, 30]. DNA is typically kept from denaturing by introducing positive supercoils via reverse gyrase activity while phosphodiester bond degradation is prevented by stabilization through interaction with cations (e.g. K^+, Mg^{2+}) and polyamines [31, 32]. However, the impact of temperature on genome features essential to gene expression such as promoters and RBSs remains largely unexplored. Bacterial transcription initiation is governed by recognition of promoter sequences by sigma factors, which load the RNA polymerase holoenzyme upstream of the transcription start site (TSS). Translation initiation is predominantly reliant on base pairing between the anti-Shine-Dalgarno sequence found near the 3′-terminus of the 16S rRNA and the Shine-Dalgarno sequence (i.e. the RBS). Therefore, thermophilic macromolecular synthesis machinery must establish and retain contacts with nucleic acids while facing greater thermodynamic challenges.

The integrated approach described here enables an experimentally anchored annotation of several bacterial genomic features including protein-coding genes, functional RNAs, non-coding RNAs, transcription units (TUs), promoters, ribosome binding sites (RBSs) and regulatory sites such as transcription factor (TF) binding sites, 5′ and 3′ untranslated regions (UTRs) and intergenic regions. This is achieved through the simultaneous analysis of genomic, transcriptomic and proteomic experimental datasets with complementary bioinformatics approaches. In addition to providing a valuable resource to the research community, this analysis framework facilitates quantitative and comparative analysis of annotated features across microbial species. For the genome of *T. maritima*, several distinguishing characteristics were identified and their potential causal factors are discussed.

3.4 Results

3.4.1 An integrative, multi-omic approach for the annotation of the genome organization

An integrative workflow was developed to re-annotate the genome of *T. maritima*. The re-annotated genome is the result of the simultaneous reconciliation of multiple omics data sources (Figure 3.1, upper left) with bioinformatics approaches (Figure 3.1, upper right). Omics data generated included: (1) genome resequencing, (2) transcription start site (TSS) identification using a modified 5′ RACE (Rapid Amplification of cDNA Ends) protocol, (3) transcriptome profiling using both high-density tiling arrays and strand-specific RNA-seq, and (4) LC-MS/MS shotgun proteomics. Transcriptome data were generated from cultures grown in diverse conditions including log phase growth, late exponential phase, heat shock, and growth inhibition by hydrogen (See Materials and Methods). Proteomic datasets include log phase growth and late exponential phase growth conditions. In combination with various bioinformatics approaches, integration of these omics datasets allowed for the definition of gene and transcription units (TU) boundaries with single base-pair resolution. The updated and expanded annotation served as the basis for genome-wide

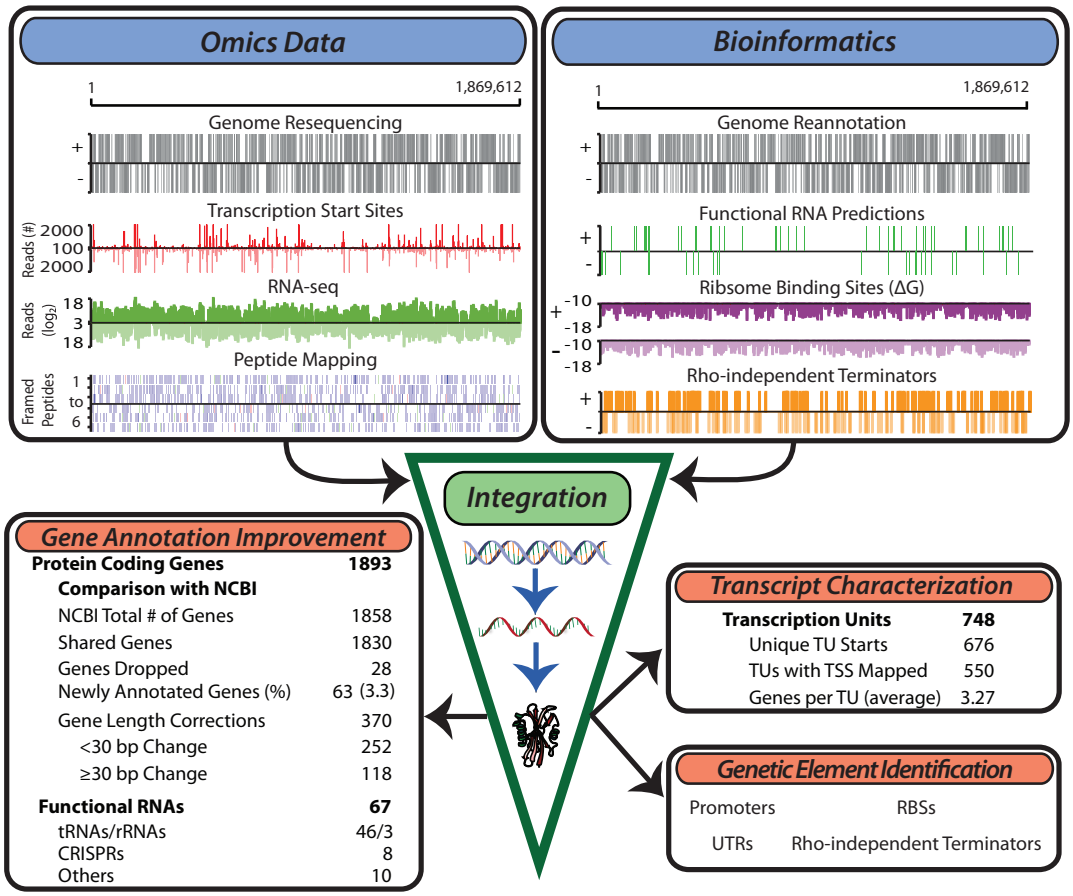


Figure 3.1: **Generation of multiple genome-scale datasets integrated with bioinformatics predictions reveals the genome organization.** Experimental data generated for the study of the *T. maritima* genome include genome resequencing, TSS determination, RNA-seq, tiling arrays (not shown) and LC-MS/MS peptide mapping (top left). Bioinformatics approaches used include genome re-annotation, functional RNA prediction, ribosome binding site energy calculations, and determination of intrinsic terminators (top right). Integration of these distinct datasets involves normalization and quantification to genomic coordinates. This experimentally anchors gene annotation improvements, defines the TU architecture, identifies non-coding RNAs and serves as a basis for the identification of additional genetic elements such as promoters and ribosome binding sites.

identification of promoters, ribosome binding sites (RBSs), intrinsic transcriptional terminators and UTRs.

Annotation of open reading frames (ORFs)

Reannotation of the *T. maritima* MSB8 genome began with whole genome resequencing of the ATCC derived strain. Genome resequencing was prompted by the recent identification of a ≈9 kb chromosomal region in the DSMZ derived strain (DSMZ genomovar, Genbank Accession AGIJ00000000.1) that is not present in the original genome sequence derived from a TIGR strain (TIGR genomovar, Genbank Accession AE000512.1) [33]. Resequencing the ATCC derived strain (presented as the ATCC genomovar, Genbank Accession CP004077) ensured that subsequent analyses referenced an accurate genome sequence. The ATCC genomovar sequence consists of 1,869,612 bp and, like the DSMZ genomovar, carries an ≈9 kb chromosomal region found between TM1847 and TM1848 of the TIGR annotation. The draft genome was annotated using the RAST Pipeline [34] and was then reconciled with the existing TIGR genomovar annotation. The RAST draft annotation had 1,887 protein-coding sequences while the TIGR annotation contained 1,858. Comparison of these two annotations with transcriptome, proteome and bioinformatics datasets resulted in a final annotation containing 1,893 protein-coding sequences (Table S1 in [35]). The final gene annotation retained a total of 1,830 NCBI annotated genes while 28 NCBI annotated genes were dropped (or replaced) due to a lack of experimental support. An additional 63 genes were annotated based on evidence found in multiple data-types. Furthermore, 370 genes varied in length when comparing the final gene annotation to the NCBI annotation. These discrepancies in gene length were predominantly due to differences in the start codon assignment, thus changing the amino acid sequence at the N-terminus. Gene length annotation differences of less than 10 amino acids were not resolved using the generated datasets without the presence of direct proteomic evidence to support one annotation over the other. However, 118 of these 370 genes (32%) had large discrepancies in their gene length annotation, equaling or exceeding 10 amino acids. For these cases, annotation conflicts were resolved using data from peptide mapping, transcript presence and bioinformatics tools.

Annotation of transcription units (TUs)

In addition to the annotation of ORFs, the genome annotation was expanded to include the TU architecture. The TU architecture is defined here to be the genomic coordinates of all RNA molecules in the transcriptome. To expand the annotation to include TUs, transcript bounds were resolved with single base pair resolution using data from RNA-seq and TSS determination. Definition of these bounds was facilitated by bioinformatics approaches; for example, the prediction of intrinsic transcriptional terminators was used to aid in assigning 3′ bounds of transcripts. This approach resulted in the assignment of 748 TUs with a total of 676 unique TSSs (Table S2 in [35]). The majority of TUs were found to be polycistronic (427, 57%) while the rest of the TUs contain only a single gene (321, 43%). The average TU contains 3.3 genes which is greater than the typical 1-2 genes per transcript observed in other bacteria [7, 36, 37] but similar to those found in archaea [9, 38]. Previous high-resolution studies of microbial transcriptomes have identified the transcription of suboperonic regions as a source of transcriptional complexity [5, 8, 36]. In *T. maritima* 165 TUs (22%) are suboperonic, having their initiation site within a longer TU. This fraction of suboperons observed in *T. maritima* is within the range observed in other bacteria; however, some organisms such as *Helicobacter pylori* have similarly sized genomes (1.67 Mb) but use suboperonic transcription much more frequently (47%, excluding antisense suboperons) [8]. Another source of transcriptional complexity comes from the use of multiple start sites, however, only a small number of *T. maritima* TUs (28, Table S3 in [35]) were observed to utilize them.

Annotation of non-coding RNAs

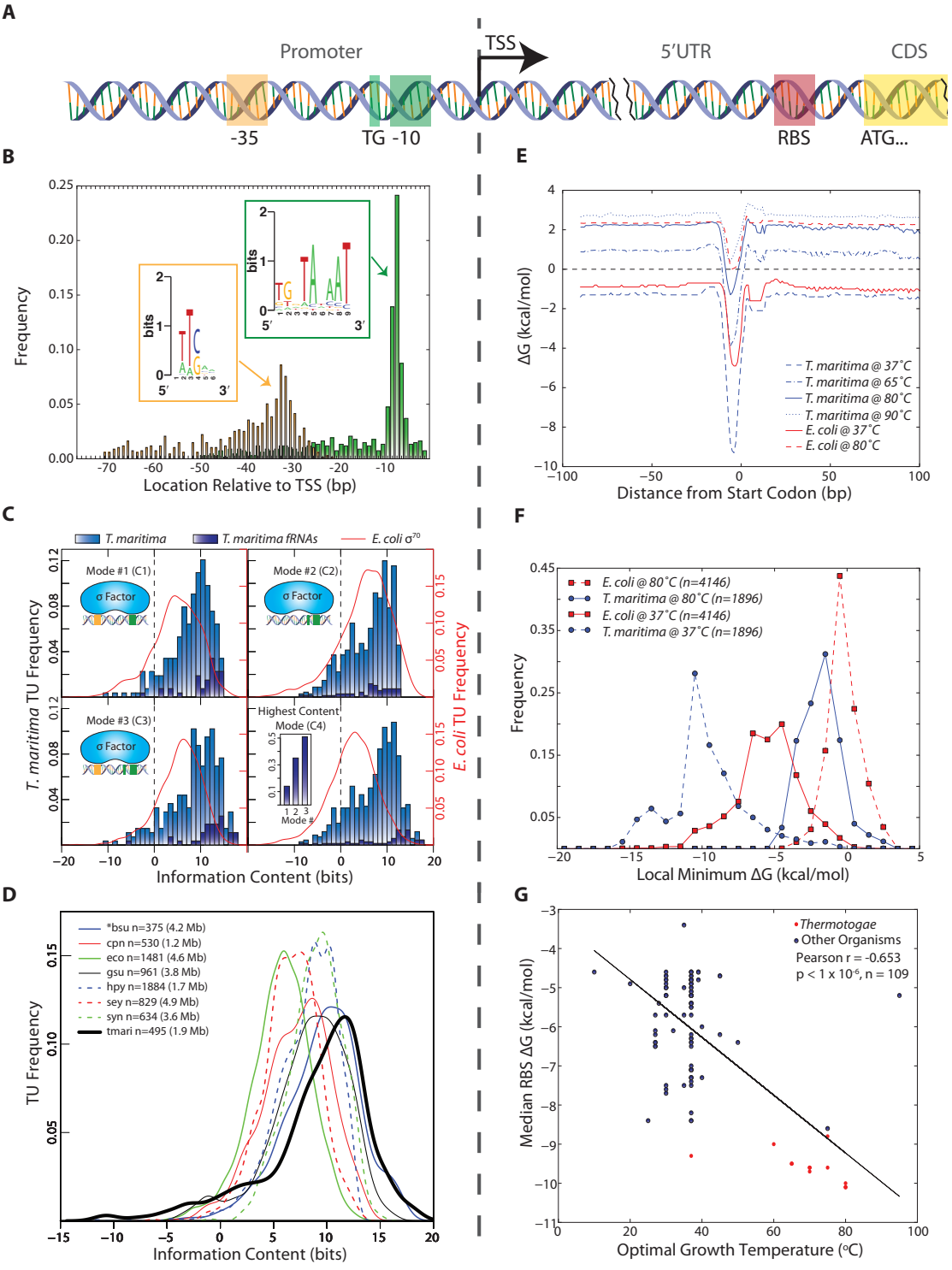
Beyond facilitating protein-coding gene annotation, transcriptome data provided experimental evidence supporting the bioinformatics prediction of 46 tRNAs, 3 rRNAs, 8 CRISPR cassettes and an additional 10 non-coding RNAs which include riboswitches, leader sequences, RNase P RNA, tmRNA and SRP RNA. These features are included in the final annotation presented here (CP004077, Table S1 in [35]). Transcription was detected antisense to 19% of annotated genes (Table S4 in [35]). However, 3′UTRs account for 52% of these antisense transcripts and only 62 anti-

sense transcripts have an experimentally identified TSS. Furthermore, the median log phase FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values are much lower for antisense transcripts (4.5) than those found for protein-coding genes (117). Transcriptome data also enabled identification of 13 putative non-coding RNAs (ncRNAs, Table S5 in [35]). No secondary structures could be defined for these putative ncRNAs using the prediction algorithms RNAfold [39] and Infernal [40] at 80°C. Four of these putative ncRNAs contain small ORFs (<40 amino acids) but no peptide evidence for these small ORFs was found in the proteomic datasets.

3.4.2 Identification of promoters and RBSs followed by quantitative intra- and interspecies analysis of binding free energies

The genome-wide identification of promoter and RBS sites was facilitated by the annotated TU start loci and protein start codons (Figure 3.2A). Promoter and RBS sequences were then quantitatively analyzed using thermodynamic principles. These same quantitative measures were applied to numerous organisms for interspecies comparison.

Figure 3.2: **Identification and quantitative comparison of genetic elements for transcription and translation initiation.** (A) Schematic showing the position of the promoter upstream of the TSS and the RBS upstream of the translation start codon. (B) The genomic position of the 3′ end of each promoter element is shown relative to the TSS for all *T. maritima* TUs. Promoter elements were identified using a gapped motif search for a −35 hexamer and a −10 nonamer. This revealed an *E. coli* σ^{70} promoter architecture for the housekeeping sigma factor of *T. maritima*, RpoD. The motif for both promoter elements is displayed as a sequence logo (insets). (C) The relative binding free energy of σ^{70} is captured using information content. Each panel shows the distribution of promoter information content for *T. maritima* and *E. coli*. Mode 1 (C1) calculates information content based on σ^{70} contacts with the −35 and −10 hexamer promoter elements n_{tmari} = 265, n_{tmari_fRNA} = 38, n_{eco} = 650). Mode 2 (C2) represents binding to the extended −10 promoter (n_{tmari} = 676, n_{tmari_fRNA} = 57, n_{eco} = 1,481). Mode 3 (C3) represents σ^{70}-binding to both the −35 and the extended −10 promoter elements (n_{tmari} = 274, n_{tmari_fRNA} = 37, n_{eco} = 657). (C4) shows the distribution of information content for all promoters when only the highest scoring mode is considered (n_{tmari} = 676, n_{tmari_fRNA} = 57, n_{eco} = 1,481). The inset shows the highest distribution of functional RNAs across the modes. (D) The σ^{70} binding modes from (C) were used to calculate the promoter information content for seven additional bacterial species. Analogous to (C4), the distribution of information scores when only the highest bit score mode is considered is shown. The organism abbreviations correspond to the following: bsu, *Bacillus subtilis*; cpn, *Chlamydophila pneumoniae* CWL029; eco, *Escherichia coli* K-12 MG1655; gsu, *Geobacter sulfurreducens* PCA; hpy, *Helicobacter pylori* 26695; sey, *Salmonella enterica* subsp. enterica serovar Typhimurium SL1344; syn, *Synechocystis* sp. PCC 6803; tmari, *T. maritima* MSB8. The genome size is given in paranthesis. *bsu data is extracted from a highly curated source that is a collection of small-scale experiments and, as such, this distribution is not a genome-scale assessment of promoter strength. (E) The calculated median RBS ΔG for all genes based on the position relative to the start codon. Temperature profiles are shown for *T. maritima* at 37°C (for comparison), 65°C (lower growth limit), 80°C (growth optimum) and 90°C (upper growth limit). Similar profiles are shown for *E. coli* at 37°C (optimal) and 80°C (for comparison). (F) The local minimum RBS ΔG for all genes in a 30 nt window upstream of the annotated start codon generated for *T. maritima* and *E. coli* at 37°C and 80°C. (G) Similar to (F), the median of the local minimum RBS ΔG was calculated and plotted for 109 bacteria against their optimal growth temperature. Species in the Thermotogae phylum (n = 15) are shown in red.


A
Promoter
TSS
5'UTR
CDS
-35
TG -10
RBS
ATG...
B
Frequency
Location Relative to TSS (bp)
bits
5'
3'
C
T. maritima
T. maritima fRNAs
E. coli σ70
Mode #1 (C1)
Mode #2 (C2)
Mode #3 (C3)
Highest Content Mode (C4)
σ Factor
T. maritima TU Frequency
E. coli TU Frequency
Information Content (bits)
D
*bsu n=375 (4.2 Mb)
cpn n=530 (1.2 Mb)
eco n=1481 (4.6 Mb)
gsu n=961 (3.8 Mb)
hpy n=1884 (1.7 Mb)
sey n=829 (4.9 Mb)
syn n=634 (3.6 Mb)
tmari n=495 (1.9 Mb)
TU Frequency
Information Content (bits)
E
ΔG (kcal/mol)
T. maritima @ 37°C
T. maritima @ 65°C
T. maritima @ 80°C
T. maritima @ 90°C
E. coli @ 37°C
E. coli @ 80°C
Distance from Start Codon (bp)
F
E. coli @ 80°C (n=4146)
T. maritima @ 80°C (n=1896)
E. coli @ 37°C (n=4146)
T. maritima @ 37°C (n=1896)
Frequency
Local Minimum ΔG (kcal/mol)
G
Thermotogae
Other Organisms
Pearson r = -0.653
p < 1 x 10-6, n = 109
Median RBS ΔG (kcal/mol)
Optimal Growth Temperature (°C)

Annotation-guided search for motifs reveals promoter structures that enable many contacts with RNA polymerase holoenzyme

Bacterial RNA polymerase is recruited predominantly through the binding of sigma factors to promoter regions. A promoter motif search was performed upstream of all unique *T. maritima* TU start sites. This revealed a strongly conserved, *E. coli* σ^{70}-like consensus sequence for the housekeeping sigma factor RpoD (Tmari_1457). No motifs were detected for the alternate sigma factors RpoE, SigH and FliA (See Materials and Methods). The RpoD motif has three distinct promoter elements: a -10 hexamer, a -35 hexamer and a 5′TGn element directly upstream of the -10 hexamer (Figure 3.2B). Individual promoters identified carried combinations of these three elements. The distance between the TSS and the 3′ end of the -10 element was found to be 7 bp (Figure 3.2B). This is in strong agreement with the expected spacing for the consensus sequence of *E. coli* σ^{70}. The same is true of the -35 element though the location of the -35 hexamer is more variable compared with the -10 hexamer partly due to the variability of the spacing between the -10 and -35 promoter elements. Plotting the spacer between the -10 and -35 promoter elements yields a distribution centered around 17 bp, which also is in agreement with the *E. coli* σ^{70} consensus (Figure S1 in [35]). Furthermore, plotting of genomic AT content upstream and downstream of aligned -10 promoter elements reveals an increase in AT content $\approx$75 bp upstream of the -10 promoter element (Figure S2 in [35]). This suggests the presence of UP elements for a subset of *T. maritima* promoters. The α-subunits of RNA polymerase bind to UP elements, facilitating initiation of transcription [41, 42].

Quantitative assessment of *T. maritima* promoters indicates high information content across multiple σ^{70} binding modes

The identification of σ^{70} promoter elements enabled the quantitative study of the relative binding free energy associated with individual promoters. The sequence conservation of an individual promoter element (i.e. the information content measured in bits [43]) can be computed through application of molecular information theory and is achieved through quantitative comparison of a given sequence to the average sequence conservation across the genome as measured through the position weight

matrix [44] (See Materials and Methods). Information content has been correlated to binding free energy (ΔG) through the second law of thermodynamics [45, 46, 47], where sequences with high information content are closer to consensus and, therefore, have stronger relative binding free energy (more negative ΔG). Experimental results, both *in vitro* and *in vivo*, have shown that information content is moderately predictive of promoter strength and activity [48].

The information content for individual *T. maritima* promoters was computed using a model of σ^{70} promoters that accounts for the information content of each promoter element and the variation in spacing between the -10 and -35 elements [47]. Using this approach, the information content of each *T. maritima* promoter was determined for three, σ^{70}-binding modes that represent the potential contacts between σ^{70} and the promoter elements (Figure 3.2$C1 - C3$). Plotting the maximum information carrying binding mode for all promoters (Figure 3.2C4) shows that the vast majority of promoters (90%) have information content greater than zero. This indicates that, for these TUs, σ^{70} binding and transcription initiation is thermodynamically favorable (ΔG$<$0). Furthermore, the distribution of information content indicates that the median *T. maritima* promoter has 8.7 bits compared to *E. coli* σ^{70} promoters whose median is 5.9 bits. Comparison of *T. maritima* promoters across all modes shows that the extended -10 promoter (-10 hexamer and upstream 5′TGn, Mode 2) provides the highest information for most TUs (63%). Furthermore, an extended -10 promoter combined with a -35 box (Mode 3) yields the highest information content in 25% of all promoters and 51% of functional RNA promoters (Figure 3.2C4 inset). These RNAs, which are among the most actively transcribed genes, encode promoters with exceptionally high information content (median 12.1 bits).

Interspecies comparative analysis reveals that *T. maritima* promoters have high relative sequence conservation

The surprisingly high sequence conservation of *T. maritima* promoters prompted a comparative analysis of information content across multiple bacterial species. The scope of the comparative analysis was limited by the lack of datasets detailing bac-

terial TSS locations and the association of those TSSs with σ^{70}. Publically available datasets for only seven additional, diverse microorganisms met this criteria. The organisms included in the analysis are the Gammaproteobacteria *Escherichia coli* K12 MG1655 [49] and *Salmonella enterica* subsp. enterica serovar Typhimurium SL1344 [50], the Deltaproteobacterium *Geobacter sulfurreducens* PCA [7], the Epsilonproteobacterium *Helicobacter pylori* 26695 [8], the Chlamydiae *Chlamydophila pneumoniae* CWL029 [51], the Cyanobacterium *Synechocystis* sp. PCC 6803 [52] and the Firmicute *Bacillus subtilis* [53]. Since these datasets contain only experimentally confirmed TSS loci, only *T. maritima* TUs with an experimentally confirmed TSS were included in this interspecies comparison (495 TUs out of 676). As before, the information content across all three σ^{70}-binding modes was calculated. The distribution of the highest information content mode (Figure 3.2D) indicates that *T. maritima* promoters are the strongest among all organisms studied, carrying a median of 10.2 bits of information. Thus, among bacteria, *T. maritima* promoter information content associated with σ^{70}-binding is relatively high.

Analysis of *T. maritima* RBS binding strength reveals strong binding free energies that support translation initiation at 80 °C

The RNA/RNA binding free energy of the Shine-Dalgarno with the anti-Shine-Dalgarno was calculated in a temperature-dependent manner using the gene annotation as a reference point. Across all protein coding genes, the median RBS ΔG was calculated $\pm$100 nucleotides (nt) from the start codon at temperatures ranging from 37 °C to 90 °C (Figure 3.2E). The position of the lowest ΔG is shown to be $4-6$ nt upstream of the start codon, which is in agreement with the optimal RBS location for translation initiation [54]. *T. maritima* is shown to maintain a thermodynamically favorable median ΔG up to its growth temperature maximum of 90 °C [14]. Plotting the distribution of local minimum ΔG's at 80 °C (Figure 3.2F) reveals that 93% of *T. maritima* protein-coding genes have a RBS with $\Delta G<0$. Calculating RBS free energy distributions at different temperatures (Figure 3.2F) reveals that at higher temperatures there is a narrowing in the range of observed free energies. *T. maritima* RBSs have a median absolute deviation of 1.30 kcal/mol at 37 °C compared to 0.87

kcal/mol at 80 °C ($p = 4.4 \times 10^{-33}$, Wilcoxon rank-sum test). Comparison of *E. coli* and *T. maritima* RBSs reveals that *T. maritima* RBSs are substantially weaker at their respective optimal growth temperatures (Figure 3.2F). A large fraction (36%) of *E. coli* genes have a $\Delta G>0$ at 80 °C and would not be capable of supporting hyperthermophilic life. When compared at equal temperatures (Figure 3.2F, 80 °C) *T. maritima* RBSs are stronger.

Interspecies analysis indicates that RBS binding strength is influenced by both optimal growth temperature and phylogeny

To more rigorously test for a relationship between RBS strength and optimal growth temperature, RBS ΔG's were calculated for all genes in 108 additional bacterial species spanning numerous phyla (including 14 members of the Thermotogae phylum). These organisms include psychrophilic, mesophilic, thermophilic and hyperthermophilic microorganisms. A significant linear correlation was found between optimal growth temperature and median RBS ΔG (Pearson r = -0.653, $p < 1 \times 10^{-6}$ random permutation test), where increasing optimal growth temperatures trend with a lower median RBS ΔG calculated at 37 °C (Figure 3.2G). However, the energetic analysis of RBSs applied here is based on the 16S rRNA sequence of the anti-Shine-Dalgarno and, as such, phylogeny is a potential contributing factor to this correlation. To test this, three distance matrices were constructed: (1) for local minimum median RBS ΔG (across all genes in a given genome), (2) for optimal growth temperatures, and (3) for phylogenetic distances determined from 16S rRNA sequences. The Mantel test was then applied to evaluate the correlations among the pairwise distance matrices (Figure S3 in [35]) allowing for the contribution of optimal growth temperature to be decoupled from phylogeny with respect to RBS strength. This test indicated that both phylogeny and optimal growth temperature impact median RBS strength, with temperature slightly more significant than phylogeny (Mantel Statistic r = 0.37 vs 0.35, $p = 1 \times 10^{-4}$ random permutation test).

3.4.3 *T. maritima* promoter-containing intergenic regions reveal a unique distribution of 5′UTRs and spatial limitations on regulation

Regulation in *T. maritima* was studied from the vantage point of an organism with extremely short intergenic regions. In both microbes [55] and higher organisms [56] it was shown that the regulatory complexity of an operon positively correlates with the amount of intergenic space found upstream of that operon. Promoter-containing intergenic regions (PIRs) served as well-defined genomic regions for this analysis (Figure 3.3A). PIRs contain target sites for transcriptional regulation (e.g. promoters and TF binding sites) as well as translational regulation (e.g. RBSs). Each PIR can be divided into two components in relation to the TSS: the sequence downstream of the TSS (the 5′UTR) and the sequence upstream of the TSS.

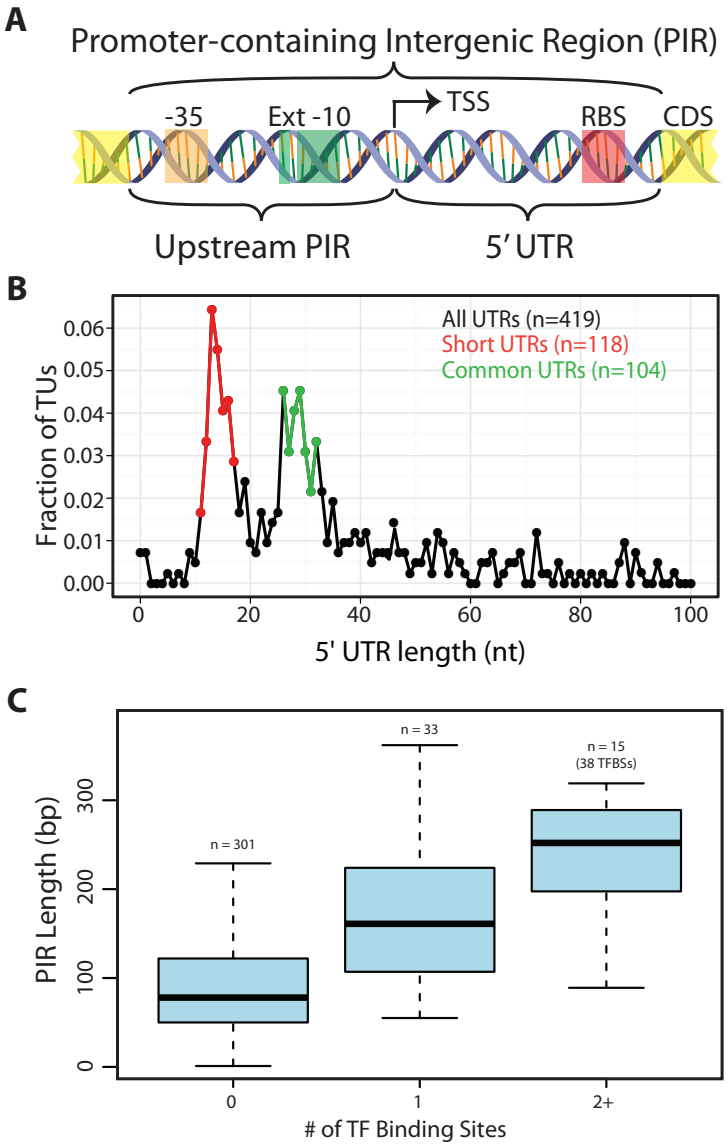


Figure 3.3: **Arrangement of genomic features contained within promoter-containing intergenic regions (PIRs).** (A) Schematic of the two subdivisions of the PIR and the genetic elements they typically carry. (B) The 5′UTR distribution is shown for all TUs with an experimentally identified TSS. The Short 5′UTR group (11−17 nt) is shown in red. The Common 5′UTR group (26−32 nt) is shown in green. Transcripts with an annotated functional RNA as the first feature were omitted from the analysis. Though only the first 100 nt are plotted, frequencies are based on the entire set of 5′UTR lengths. (C) A quartile plot of the length distribution of PIRs is shown. PIRs are grouped according to the number of TF binding sites they contain (no TF, a single TF or multiple TFs).

T. maritima has a bimodal distribution of 5′UTRs comprised of uncharacteristically 'Short' 5′UTRs and 'Common' 5′UTRs

T. maritima exhibits an unusual bimodal distribution with respect to the length of 5′UTRs (Figure 3.3B). To date, the 5′UTRs of all other microorganisms follow a unimodal distribution centered at approximately 30 nt [7, 8, 36, 37]. Though *T. maritima* has a distinct peak (local maxima) from 26-32 nt (Common 5′UTR Group), it has a second peak containing shorter 5′UTRs with lengths between $11-17$ nts (Short 5′UTR Group). Interestingly, there is underrepresentation of 5′UTRs with lengths between $18-25$ nt. Leaderless transcripts were not detected in *T. maritima*, echoing the RNA/RNA binding energy analysis that indicated exclusive use of RBSs for translation initiation.

To better understand the bimodal nature of the 5′UTR distribution, various factors were tested that could differentiate the Short 5′UTR Group from the Common 5′UTR Group and provide insights into the lack of 5′UTRs between $18-25$ nt. Factors tested for over- or underrepresentation of the different 5′UTR groups included: (1) gene expression level (both mRNA and protein levels), (2) protein expression normalized to mRNA expression, (3) phylogenetic origin of genes, (4) RBS and promoter strengths, (5) divergent vs. convergent operons, and (6) cellular functional categorization. These factors yielded no discrimination between the Short 5′UTR Group and the Common 5′UTR Group and could not explain the bimodal nature of the 5′UTR length distribution.

T. maritima PIRs are predominantly too short to permit transcription factor regulation

To enable regulation of transcription, space in the genome must be dedicated to operator sites, which serve as docking locations for TF recruitment. Typically, these sites reside upstream of the TSS, but can also be found downstream of the TSS (in the 5′UTR). An analysis centered on PIRs was chosen to capture the potential for TF binding sites both upstream and downstream of the TSS. A total of 31 TF regulons with a combined total of 91 genomic binding sites were extracted from the RegPrecise database [57]. Mapping of the TF binding sites to the *T. maritima* genome showed

that 71 were within PIRs, 12 mapped to intergenic regions not carrying a promoter and the remaining 8 were within or overlapped an annotated gene (Table S6 in [35]). The length distribution of PIRs without a TF binding site was compared to that of PIRs with TF binding sites (Figure 3.3C). The median length of PIRs that do not contain a TF binding site is 78 bp. This is significantly shorter than the length of PIRs that carry a single TF binding site (median = 161 bp, Wilcoxon rank-sum test $p = 6.9 \times 10^{-8}$) or multiple TF binding sites (median = 252 bp, Wilcoxon rank-sum test $p = 2.8 \times 10^{-7}$). Thus, the majority of *T. maritima* PIRs do not contain the typical space required to encode a TF binding site.

3.4.4 *T. maritima* has an actively transcribed genome that is tightly correlated to protein abundances

Transcriptome data indicate that the genome of *T. maritima* is exceptionally active irrespective of growth condition (Figure 3.4A) with 91-96% of genes expressed above an FPKM threshold of 8. This fraction of genes transcribed is uncharacteristically high compared to other free-living bacteria (see Table S7 in [35]). Furthermore, translational evidence supporting the high gene expression activity of *T. maritima* is found in the proteomic datasets. In each condition tested, peptide evidence was detected for 74% of the annotated proteins. It is also found that mRNA and protein abundances are tightly linked (Pearson r = 0.63, $p < 2.2 \times 10^{-16}$ t-test) (Figure 3.4B). This correlation is stronger and more significant than those reported in comparable studies for other bacteria [58, 59].

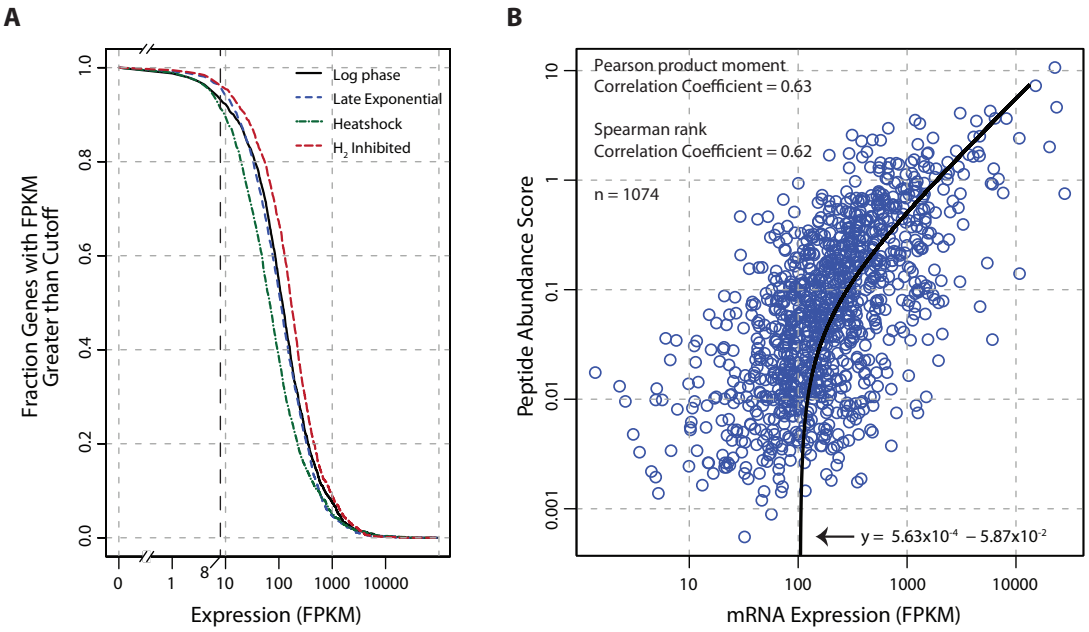


Figure 3.4: **Global analysis of mRNA and protein expression levels.** (A) The fraction of transcribed genes as a function of the FPKM threshold. Under growth promoting conditions (log-phase) and early in the transition to stressed conditions (carbon-limited late exponential phase, heat shock, and hydrogen inhibition), 91-96% of the genome is expressed using a conservative FPKM threshold of $\geq$8. (B) Correlation of mRNA expression and protein abundance. The line of best fit indicates a strong linear relationship (Pearson r = 0.63, $p < 2.2 \times 10^{-16}$ t-test) between transcription and translation. The peptide abundance score for each protein was derived by dividing the total spectral count by the number of possible tryptic peptides (400-2000 m/z up to a charge state (z) of 3, hence a maximum fragment mass of 6000). Abbreviations: FPKM, Fragments Per Kilobase of transcript per Million mapped reads; m/z, mass-to-charge ratio.

3.5 Discussion

Genome-scale technologies have provided researchers unprecedented access to large volumes of data detailing the composition of a cell. However, approaches for data analysis and interpretation have lagged behind due to the scope and complexity of these data types. Here, we present a framework for multi-omic data analysis that annotates genomic features involved in transcription, translation and regulation. This methodology integrates genome-scale datasets with bioinformatics predictions to produce 1) an improvement of the gene annotation, 2) an experimentally validated TU architecture and 3) the identification of putative antisense, non-coding transcripts and alternative TSSs. Using these annotated genomic features enabled the genome-wide identification of promoters and RBSs, which are difficult to identify solely using *in silico* approaches [60, 61]. Furthermore, the relative binding strength of individual promoters and RBSs was quantitatively measured using thermodynamic principles enabling multi-species comparison of these sequence features. The annotated genome organization served as a scaffold for analyzing regulatory features. Transcription factor regulation was examined with respect to promoter containing intergenic regions while the translational impact of the 5′UTR distribution was considered. The multi-omic data generation and analysis demonstrated here is applicable to many microbial species.

Applying this methodology to study the genome organization of *T. maritima* revealed that it has many distinctive properties compared to other organisms. Genome-scale analysis of promoters showed that *T. maritima* encodes a highly conserved, robust architecture that ensures transcription initiation. Similarly, RBS sequence conservation was shown to be thermodynamically sufficient for translation initiation for almost all *T. maritima* genes at 80°C compared with only a fraction of *E. coli* genes. The distinctive properties of the *T. maritima* genome extend beyond sequence composition and are apparent at the organizational level. The high protein-coding density and minimal intergenic space found in this organism have resulted in a high number of genes per TU, a paucity of putative ncRNAs and few TUs with multiple start sites. Furthermore, transcriptional regulation appears to be limited to a few TUs due to a lack of genomic space in PIRs. Interestingly, the 5′UTR compo-

nent of the PIR was found to be uncharacteristically bimodal and was comprised of an unusually short grouping of 5′UTRs. Lastly, the constrained genome organization of *T. maritima* is reflected in the physiological state of the cell. Transcription of the vast majority of genes is detected independent of culture condition and the correlation between protein and mRNA is stronger than previously observed in other bacteria.

We hypothesize that the hyperthermophilic lifestyle of *T. maritima* could potentially explain the distinctive characteristics of this organism's genome organization. For instance, the increased sequence conservation of promoter elements and RBSs throughout the *T. maritima* genome may be attributed to the need to maintain gene expression under extreme temperature conditions. Macromolecular interactions (e.g. protein/protein, protein/DNA and RNA/RNA) are intrinsically harder to maintain at higher temperatures. In the case of TF binding sites, it has been shown that each nucleotide deviation from consensus results in a ≈2kbT penalty to the maximum binding free energy for a given TF (where kb is Boltzmann's constant and T is temperature) [62]. Increasing the temperature amplifies the binding free energy penalty for every non-conserved base pair. Therefore at 80°C, mismatches between the Shine-Dalgarno and anti-Shine-Dalgarno sequence are especially severe. Thus, *T. maritima* must overcome the intrinsic challenge of recognizing and retaining contact at the initiation site for both transcription and translation. Our data suggests that high sequence conservation of promoter and RBS sequences is one of the mechanisms used by *T. maritima* to ensure sufficient gene expression. This sequence-level adaptation could be analogous to many others observed in thermophilic organisms such as the amino acid composition of proteins [29, 30] and the GC content of structural RNAs [63].

The minimal intergenic space found in the *T. maritima* genome is reminiscent of a streamlined genome, which could explain the limited regulatory capacity observed in this organism. Inflexibility of metabolic regulons has been previously alluded to for other Thermotogales [64]. Here it is demonstrated that, for most TUs, a lack of physical space exists for transcriptional regulation by TFs. Furthermore, the Short 5′UTR group carries the minimum number of nucleotides needed to recruit the ribosome based on Shine-Dalgarno/anti-Shine-Dalgarno interactions [54]. Further

reduction in 5′UTR length would abolish translation. Short 5′UTRs also reduce the capacity to regulate by limiting 5′UTR interactions [65, 66].

Though thermodynamics and physical space are hypothesized to contribute to the characteristic features of the *T. maritima* genome, the phylogenetic contribution cannot be dismissed. These potential causal factors are difficult to decouple. For RBSs, we were able to determine the impact of phylogeny and optimal growth temperature on RBS binding strength. By analyzing RBSs from 109 bacterial species spanning many phyla and having a diverse range of optimal growth temperatures we were able to demonstrate that both phylogeny and optimal growth temperature were significant determinants of RBSs sequence composition. However, a recent analysis of genome size among species of the order Thermotogales could not resolve the impact of phylogeny from optimal growth temperature [19]. The authors found that a negative correlation between genome size and optimal growth temperature exists within this order but the correlation did not hold when phylogeny was accounted for in the analysis. Interestingly, this study also found that the number of predicted transcriptional regulators and intergenic space is higher in *Mesotoga prima*, a mesophilic member of the Thermotogales. Thus, the relationship between phylogeny and the genome organization is difficult to elucidate without the generation of more datasets similar to the one presented here.

Thermotogae are an ideal phylum for future investigations on the causal impact of factors such as temperature, intergenic space and phylogeny on genome organization. This phylum contains organisms that are found in many diverse environments with a wide range of optimal growth temperatures. Generating multi-omic datasets and analyzing them using an integrated, quantitative workflow for numerous Thermotogae species would enable assessment of various environmental factors in the context of phylogenetic distance. Furthermore, given their phylogenetic depth, characterization of the Thermotogae will also provide insights in the evolutionary trajectory of microbial life on earth.

3.6 Materials and Methods

3.6.1 Culture conditions and physiology

T. maritima MSB8 ATCC derived cultures were grown at 80°C under anoxic conditions in a chemically defined, minimal medium [67]. Cultures were maintained in either serum bottles or pH-controlled (6.5) fermenters with continuous 80% N2, 20% CO_2 sparging. Maltose and acetate concentrations were measured using an HPLC. HPLC parameters were previously described [68]. The following growth conditions were used for omics analysis: 1) log phase, 2) carbon-limited late exponential phase, 3) heat shock and 4) H_2 inhibition. Log phase samples were collected from mid-exponential phase cultures grown in 125 mL serum bottles with 50 mL working volume of media and 10 mM maltose as the sole carbon source. Carbon-limited late exponential phase cultures were grown in pH controlled fermenters with pH control and continuous stripping of evolved hydrogen. Cultures were monitored for OD and maltose concentration and samples were collected upon depletion of maltose. The heat shock condition was achieved by rapidly heating mid-exponential phase cultures grown in serum bottles (similar to the log phase condition) to 90°C and sampled after 10 minutes for transcriptome analysis. This has been shown to result in the heat shock response [69]. H_2 inhibition was achieved by allowing the native evolution of hydrogen to accumulate in serum bottles (similar to the log phase condition). Arrested growth was indicated by successive OD readings that showed no change measured every 30 minutes. Growth profiles for these conditions are shown in Figure S4 in [35].

3.6.2 Genome resequencing and annotation updates

The recent identification of a 9 kb gap in the *T. maritima* MSB8 genome [33] prompted genome resequencing. Genomic DNA was isolated using Promega's Wizard Genomic DNA Purification Kit. Paired-end resequencing libraries were generated following standard Illumina protocols and sequenced on an Illumina GAIIx platform. The updated genome sequence was assembled as follows: (1) Reads were aligned to the 8.9 kb region identified in the *T. maritima* MSB8 DSMZ genomovar

(AGIJ00000000.1) [33] and the TIGR genomovar (AE000512.1) sequence using SHOREmap [70] and MosaikAligner (http://bioinformatics.bc.edu/
marthlab/Mosaik). (2) Unaligned reads were *de novo* assembled using Velvet [71] to ensure no additional assemblies were present. (3) The sequence was corrected for SNPs and indels detected during read alignment.

An updated genome annotation was generated using the RAST pipeline with the default parameters [34]. Predicted gene sequences were mapped to the AE000512.1 annotation using a bidirectional Smith-Waterman alignment to identify the corresponding locus tags. Instances where $\geq$30 bp separated the predicted gene length between annotations were reconciled through manual inspection of gene expression data and bioinformatics predictions. Gene length differences <30 bp could not be reconciled (unless peptide data supported only one annotation). In these cases, the updated sequence annotation was retained.

3.6.3 Transcription start site determination

Total RNA was isolated from log phase cultures using the hot SDS/phenol approach as previously described (http://www.bio.davidson.edu/projects/GCAT/
protocols/ecoli/RNApurification.pdf). DNase-treated total RNA samples were recovered using Fisher SurePrep TrueTotal RNA columns. Two biological replicate TSS sequencing libraries were constructed as previously described [7]. Illumina reads were aligned to the updated *T. maritima* genome using the Mosaik Aligner. The number of sequenced reads and the number of aligned reads can be found in Table S10 in [35]. Only uniquely mapped 5′ ends with $\geq$5 reads were retained as potential TSSs.

3.6.4 Transcriptome characterization and gene expression

Tiling array and RNA-seq data were generated under log phase growth, carbon-limiting late exponential phase, heat shock and hydrogen inhibited conditions. Total RNA was isolated using the TRIzol (Invitrogen) extraction procedure followed by DNase treatment and purification using either the Qiagen RNeasy Mini Kit (Tiling Arrays) or the SurePrep TrueTotal RNA columns (RNA-seq).

Custom tiling arrays were synthesized based on the AE000512.1 genome sequence by Roche Nimblegen to carry 71,548 probes with a mean interval of 25 bp. Probe information was remapped to the updated genome sequence. Of the original 71,548 probes, only 125 did not map. Labeled cDNA was generated and processed as previously described [7]. The Transcription Detector algorithm [72] determined probes expressed above background at a FDR = 0.05.

Paired-end, strand-specific RNA-seq was performed using the dUTP method [73] with the following modifications. rRNA was removed with Epicentre's Ribo-Zero rRNA Removal Kit. Subtracted RNA was fragmented for 3 min using Ambion's RNA Fragmentation Reagents. cDNA was generated using Invitrogen's SuperScript III First-Strand Synthesis protocol with random hexamer priming. Illumina reads were aligned to the updated *T. maritima* genome using Bowtie [74] with up to 2 mismatches per read alignment. The number of sequenced reads and the number of aligned reads can be found in Table S10 in [35]. FPKM values were calculated using Cufflinks [75]. Functional RNA transcripts were excluded from FPKM determination.

3.6.5 Proteomics, peptide mapping, and protein abundance quantitation

Proteomics samples and data were generally prepared as previously described [76]. In summary, triplicate samples of both log phase and late exponential phase culture were lysed by French press, and proteins were extracted into global, soluble, and insoluble fractions. The three protein fractions were digested with trypsin (Promega) for 4 h at 37°C and then cleaned-up using C18 or SCX SPE columns (Supelco), as appropriate. Resulting peptide samples were separated in the first dimension by high pH HPLC (Agilent) and then analyzed by LC-MS/MS using C18 resin (Phenomenex) with an expontial gradient on a custom built LC platform coupled to a linear ion trap (LTQ) or a Velos Orbitrap mass spectrometer (Thermo Scientific) operated in data dependent mode. Peptides were identified by SEQUEST (Thermo Scientific) against a six-frame translation of the *T. maritima* genome with no protease specified in the search. Xcorr values were refined to conform to generally accepted criteria and were applied to result in a false discovery rate of 0.16% at the peptide

level. Non-quantitative peptide-level data can be found in Table S8 in [35].

Normalized protein abundances can be found in Table S9 in [35]. Quantitative Peptide-level data was extracted from Lerman et al. [77] and mapped to the CP004077 genome annotation. The following criteria were used to filter proteins for quantitative analysis: 1) the protein has a total spectral count $\geq$2 across all conditions (minimum of two unique peptides or a single unique peptide with two observations), 2) the protein has $\geq$1 observed peptide under log phase since our data was correlated against log phase transcriptome data. Redundant peptides (i.e. peptides mapping to multiple protein entries) were excluded from the analysis to minimize potential ambiguity. For quantitative analysis, we normalized the observed spectral counts for each ORF by the number of possible fully tryptic peptides in the ORF. The number of possible fully tryptic peptides for each ORF was determined using the Protein Digestion Simulator (http://omics.pnl.gov
/software/ProteinDigestionSimulator.php). Default settings were used, except the parameter Max Missed Cleavages was set to 0 and Minimum Residue Count was set to 6. These options require fully tryptic peptides of at least length 6. This program only considers peptides 400-2000 m/z up to a charge state (z) of 3, hence a maximum fragment mass of 6000.

3.6.6 Promoter element motif analysis and position weight matrix (PWM) generation

The process of determining individual σ^{70} promoter elements upstream of each unique TU start in *T. maritima* was an iterative process, involving two software packages: BioProspector [78] and MEME [79]. BioProspector is able to identify gapped motif elements so it was used to initially identify *T. maritima* motifs. In BioProspector, sequences 75 bp upstream of TU starts were searched for bipartite elements (6 and 9 bp in width) with a 10-25 bp allowable gap and visualized through WebLogo [80]. MEME provides deterministic position-weight matrices appropriate for information content calculations. The -10 and extended -10 boxes were searched [-1 to -18] upstream of the TSS while the -35 box was searched [-20 to -44]. *E. coli* TUs annotated with σ^{70} promoters and experimentally validated TSSs in the

EcoCyc Database (version 15.0) [49] were extracted for comparative analysis.

A similar approach was applied to identify promoter motifs for alternative sigma factors. *T. maritima* has three annotated alternative sigma factors: RpoE (Tmari_1606), SigH (Tmari_0531) and FliA (Tmari_0904). For RpoE and SigH, the upstream region of TUs having genes showing high differential expression under a given stress condition (heat shock, hydrogen inhibited and carbon-limited late exponential phase) were searched for motif elements. The upstream regions of flagellar gene encoding TUs were searched for a FliA motif. However, no sequence motif could be detected for any of the three alternate sigma factors.

3.6.7 Information content calculations

Position weight matrices (PWMs) for each promoter element were converted to individual information weight matrices using the following formula established in the field of molecular information theory [44]: Riw(b, i) = 2-(-log2f(b, i)), where f(b, i) is taken to be the probability of observing base b at position i. The individual information of a sequence, Iseq, was calculated by summing the relevant entries of Riw. For any particular sequence, only one entry of Riw is relevant among 4 bases for each position i in the sequence. Iseq is measured throughout in bits since the log was base 2 in converting the PWM to Riw.

Iseq reflects sequence conservation for a single sequence, but natural promoters are often formed by multiple promoter elements, each with their own sequences and corresponding Iseq values. When multiple elements are present, variable length spacers are frequently found between the elements. We applied an approach previously described by Shultzaberger et al. [47] to properly account for all possible promoter elements and the variation in their spacing. This allowed us to assess total sequence conservation for an entire promoter. For each promoter, the information content for a particular binding mode was calculated based on the formulas: (1) Mode 1: Iseq_whole_promoter = Iseq(-10 element)+Iseq(-35 element)-GS(d); (2) Mode 2: Iseq_whole_promoter = Iseq(extended -10 element); (3) Mode 3: Iseq_whole_promoter = Iseq(extended -10 element)+Iseq(-35 element)-GS(d). GS(d) is gap surprisal accounting for variable spacing (of length d) between the

-10 and -35 elements. GS(d) penalizes for unexpected spacing given the major groove accessibility of B-form DNA and was defined as in equation (3) in Shultzaberger [49] with no small-sample correction factor as the analysis here is performed at genome scale. In accordance with the Shultzaberger model, the space between the -10 and -35 elements was restricted to $15 - 20$ bp as measured from the $3'$ end of the -35 element and the $5'$ end of the -10 element. This limit on the spacer distance Iseq_whole_promoter is measured in bits.

3.6.8 Ribosome binding site energy calculations

The anti-RBS sequence $5'$-UCACCUCCUU-$3'$ ($3'$ end of the 16S rRNA) was selected for this study. The hybrid-2s program in the UNAFold software package [81] was used to compute hybridization energies (ΔG) for all possible 10-mers over the temperature range 20-100°C. This dictionary was mined for three applications: (1) binding energy values for all 10-mer sequences in the updated *T. maritima* genome were computed to aid in annotation improvement, (2) the median positional ΔG for all CDSs $\pm$100 bp from the start codon, and (3) the local minimum ΔG for all CDSs 30 bp upstream of the start codon. RBS binding energies across 109 organisms were calculated using this dictionary. Optimal growth temperatures for all non-Thermotogae bacteria were collected from Takemoto et al. [82] and the protein coding gene annotation for each bacterium was extracted from NCBI. CDS data for all Thermotogae with a complete genome sequence were extracted from NCBI with the exception of *T. maritima* for which the annotation generated in this study was used. For each organism, the median RBS ΔG was calculated from the set of minimum RBS ΔG's found for each CDS 30 bp upstream of the annotated start codon. Three distance matrices were constructed for analysis of the 109 bacterial species for which optimum growth temperatures were found. The matrices included are as follows: (1) the absolute difference of median RBS strength values, (2) the absolute difference of optimal growth temperatures and (3) the distance matrix generated by aligning full-length 16S rRNA gene sequences using ClustalW2 (slow mode) followed by the phylogenetic tree generation script (http://www.ebi.ac.uk/Tools/phylogeny/) with default settings. Next, the Mantel test, which tests the correlation between two

distance matrices, was applied to compute the significance of various correlations. The vegan package of R was used with its default settings.

3.6.9 Rho-independent terminator site determination

Intrinsic terminators were predicted using the TransTermHP program [83]. To avoid bias introduced by annotation, no genome annotation was used in prediction of Rho-independent terminators. Only terminator structures predicted with a 100% confidence score were included in the curation of TUs.

3.6.10 Prediction of small RNAs

Small RNAs were predicted with Infernal [40] using cmsearch with default settings against the Rfam 10.0 Database [84] of small RNA families. sRNAs with an E-value<0.01 were manually curated to verify expression. These sRNAs were checked against the sRNA predictions from Rfam and fRNA-DB (http://www.ncrna.org) based on the AE000512.1 genome sequence.

3.6.11 Transcription unit assembly

TU assembly was accomplished through an iterative procedure beginning with tiling array expression data. Tiling array data was processed with two Bioconductor packages for transcript segmentation based on change point analysis: tilingArray (http://www.bioconductor.org/
packages/2.2/bioc/html/tilingArray.html) and DNAcopy
(http://www.bioconductor.org/packages/2.3/bioc/html/DNAcopy.html).
Manual comparison of the output from both packages with array data was used to refine the automated set of transcriptional segments. Additional datasets and bioinformatics predictions were added and manually curated to fully characterize the TU assembly. TSS and RNA-seq data provided single-base pair resolution of segment boundaries. Intrinsic terminator predictions were also used for $3'$ boundary definition. ncRNAs were identified using the transcript segments. Transcribed regions not associated with a TU and with length exceeding 68 nt (the combined length of

the paired end reads with no insert separating them) were quantified using Cufflinks to generate FPKM values across all RNA-seq conditions. Regions with at least two conditions showing FPKM values >8 were retained as putative ncRNAs.

3.6.12 Transcription factor binding site mapping

TF binding sites were extracted from RegPrecise [57] and coordinates were mapped to the updated genome. Table S6 in [35] has the TF binding sites used in Figure 3.3C.

3.6.13 Data deposition

The *T. maritima* MSB8 ATCC (genomovar) genome and annotation are found under Genbank Accession CP004077. RNA-seq, TSS, and tiling array datasets are available in the Gene Expression Omnibus under Accession GSE37483. Proteogenomic data are made available through PNNL (http://omics.pnl.gov) and in Table S8 in [35].

3.7 Acknowledgments

We acknowledge Dmitry Rodionov and Andrei Osterman of the Sanford-Burnham Medical Research Institute, La Jolla, CA, for providing TF binding site loci, and Daniela Bezdan for assistance with the genome annotation. We also thank Bernhard Palsson, Pep Charusanti, and Ramy K. Aziz for guidance with manuscript preparation. Funding: Funding for this work was provided by the Office of Science of the U.S. Department of Energy (DOE) under grants DE-FG02-08ER64686 and DE-FG02-09ER25917. HL is supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086. Proteomics capabilities were developed under support from the DOE Office of Biological and Environmental Research (BER) Pan-omics Project and the NIH National Center for Research Resources (RR018522), and a significant portion of this work was performed in the Environmental Molecular Sciences Laboratory (EMSL), a DOE-BER national scientific user facility at Pacific

Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RLO 1830. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Conceived and designed the experiments: HL JAL VAP KZ. Performed the experiments: HL VAP YT D-HL KZ. Analyzed the data: HL JAL HN ACS-R JNA YQ. Contributed reagents/materials/analysis tools: RDS JNA. Wrote the paper: HL JAL HN KZ.

Chapter 3 in full is a reprint of a published manuscript: Latif H*, Lerman JA*, Portnoy VA, Tarasova Y, Nagarajan H, et al. (2013) The Genome Organization of *Thermotoga maritima* Reflects Its Lifestyle. PLoS Genet 9(4): e1003485. doi:10.1371/journal.pgen.1003485. * indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Joshua A. Lerman (equal contributor), Vasiliy A. Portnoy, Yekaterina Tarasova, Harish Nagarajan, Alexandra C. Schrimpe-Rutledge, Richard D. Smith, Joshua N. Adkins, Dae-Hee Lee, Yu Qiu, and Karsten Zengler.

3.8 Bibliography

[1] Kitano H (2002) Systems biology: a brief overview. Science 295: 1662–1664.

[2] Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology 7: 129–143.

[3] Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nature Reviews Genetics 7: 130–141.

[4] Overbeek R, Bartels D, Vonstein V, Meyer F (2007) Annotation of bacterial and archaeal genomes: improving accuracy and consistency. Chemical Reviews 107: 3431–3447.

[5] Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326: 1268–1271.

[6] Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castano-Diez D, Chen WH, Devos D, Guell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Bottcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin AC (2009) Proteome organization in a genome-reduced bacterium. Science 326: 1235–1240.

[7] Qiu Y, Cho BK, Park YS, Lovley D, Palsson BO, Zengler K (2010) Structural and operational complexity of the *Geobacter sulfurreducens* genome. Genome Research 20: 1304–1311.

[8] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464: 250–255.

[9] Yoon SH, Reiss DJ, Bare JC, Tenenbaum D, Pan M, Slagel J, Moritz RL, Lim S, Hackett M, Menon AL, Adams MW, Barnebey A, Yannone SM, Leigh JA, Baliga NS (2011) Parallel evolution of transcriptome architecture during genome reorganization. Genome Research 21: 1892–1904.

[10] Buescher JM, Liebermeister W, Jules M, Uhr M, Muntel J, Botella E, Hessling B, Kleijn RJ, Le Chat L, Lecointe F, Mader U, Nicolas P, Piersma S, Rugheimer F, Becher D, Bessieres P, Bidnenko E, Denham EL, Dervyn E, Devine KM, Doherty G, Drulhe S, Felicori L, Fogg MJ, Goelzer A, Hansen A, Harwood CR, Hecker M, Hubner S, Hultschig C, Jarmer H, Klipp E, Leduc A, Lewis P, Molina F, Noirot P, Peres S, Pigeonneau N, Pohl S, Rasmussen S, Rinn B, Schaffer M, Schnidder J, Schwikowski B, Van Dijl JM, Veiga P, Walsh S, Wilkinson AJ, Stelling J, Aymerich S, Sauer U (2012) Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. Science 335: 1099–1103.

[11] Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Hartig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RA, Nannapaneni P, Noone D, Pohl S, Rinn B, Rugheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stulke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Volker U, Bessieres P, Noirot P (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science 335: 1103–1106.

[12] Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nature Reviews Genetics 11: 9–16.

[13] Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nature Chemical Biology 6: 787–789.

[14] Huber R, Langworthy TA, Konig H, Thomm M, Woese CR, Sleytr UB, Stetter KO (1986) *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 °C. Archives of Microbiology 144: 324–333.

[15] Dipippo JL, Nesbo CL, Dahle H, Doolittle WF, Birkland NK, Noll KM (2009) *Kosmotoga olearia* gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. International Journal of Systematic and Evolutionary Microbiology 59: 2991–3000.

[16] Nesbo CL, Dlutek M, Zhaxybayeva O, Doolittle WF (2006) Evidence for existence of "mesotogas," members of the order Thermotogales adapted to low-temperature environments. Applied and Environmental Microbiology 72: 5061–5068.

[17] Nesbo CL, Kumaraswamy R, Dlutek M, Doolittle WF, Foght J (2010) Searching for mesophilic Thermotogales bacteria: "mesotogas" in the wild. Applied and Environmental Microbiology 76: 4896–4900.

[18] Nesbo CL, Bradnan DM, Adebusuyi A, Dlutek M, Petrus AK, Foght J, Doolittle WF, Noll KM (2012) *Mesotoga prima* gen. nov., sp. nov., the first described mesophilic species of the Thermotogales. Extremophiles 16: 387–393.

[19] Zhaxybayeva O, Swithers KS, Foght J, Green AG, Bruce D, Detter C, Han S, Teshima H, Han J, Woyke T, Pitluck S, Nolan M, Ivanova N, Pati A, Land ML, Dlutek M, Doolittle WF, Noll KM, Nesbo CL (2012) Genome sequence of the mesophilic Thermotogales bacterium *Mesotoga prima* MesG1.Ag.4.2 reveals the largest Thermotogales genome to date. Genome Biology and Evolution 4: 700–708.

[20] Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242–1245.

[21] Nelson KE (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399: 323–329.

[22] Conners SB, Mongodin EF, Johnson MR, Montero CI, Nelson KE, Kelly RM (2006) Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. FEMS Microbiology Reviews 30: 872–905.

[23] Mongodin EF, Hance IR, Deboy RT, Gill SR, Daugherty S, Huber R, Fraser CM, Stetter K, Nelson KE (2005) Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. Journal of Bacteriology 187: 4935–4944.

[24] Nesbo CL, Dlutek M, Doolittle WF (2006) Recombination in Thermotoga: implications for species concepts and biogeography. Genetics 172: 759–769.

[25] Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbo CL, Doolittle WF, Gogarten JP, Noll KM (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proceedings of the National Academy of Sciences of the United States of America 106: 5865–5870.

[26] Martin W, Baross J, Kelley D, Russell MJ (2008) Hydrothermal vents and the origin of life. Nature Reviews Microbiology 6: 805–814.

[27] Achenbach-Richter L, Gupta R, Stetter KO, Woese CR (1987) Were the original eubacteria thermophiles? Systematic and Applied Microbiology 9: 34–39.

[28] Munoz R, Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R (2011) Release LTPs104 of the All-Species Living Tree. Systematic and Applied Microbiology 34: 169–170.

[29] Fields PA (2001) Review: Protein function at thermal extremes: balancing stability and flexibility. Comparative Biochemistry and Physiology Part A, Molecular and Integrative Physiology 129: 417–431.

[30] Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? Cellular and Molecular Life Sciences 58: 1216–1233.

[31] Gerday C, Glansdorff N, American Society for Microbiology CN - Jefferson or Adams Building Reading Rooms QR1009; P59 2007 Reference - Science Reading Room (Adams tFQP (2007) Physiology and biochemistry of extremophiles. Washington, D.C.: ASM Press, xvi, 429 p. pp. URL http://www.loc.gov/catdir/toc/ecip077/2006102166.html.

[32] Robb F, Antranikian G, Grogan D, Driessen A (2007) Thermophiles: biology and technology at high temperatures. CRC Press.

[33] Boucher N, Noll KM (2011) Ligands of thermophilic ABC transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8 screened by differential scanning fluorimetry. Applied and Environmental Microbiology 77: 6395–6399.

[34] Aziz RK (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

[35] Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K (2013) The genome organization of *Thermotoga maritima* reflects its lifestyle. PLoS Genetics 9: e1003485.

[36] Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO (2009) The transcription unit architecture of the *Escherichia coli* genome. Nature Biotechnology 27: 1043–1049.

[37] Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. Genome Biology 12: R47.

[38] Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo FY, Pratap A, Deutsch EW, Peterson A, Martin D, Baliga NS (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. Molecular Systems Biology 5: 285.

[39] Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL (2008) The Vienna RNA websuite. Nucleic Acids Research 36: W70–4.

[40] Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.

[41] Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, Severinov K, Gourse RL (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. Science 262: 1407–1413.

[42] Blatter EE, Ross W, Tang H, Gourse RL, Ebright RH (1994) Domain organization of RNA polymerase alpha subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. Cell 78: 889–896.

[43] Schneider TD (1996). New Approaches in Mathematical Biology: Information Theory and Molecular Machines. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.8584.

[44] Schneider TD (1997) Information content of individual genetic sequences. Journal of Theoretical Biology 189: 427–441.

[45] D'Haeseleer P (2006) What are DNA sequence motifs? Nature Biotechnology 24: 423–425.

[46] Schneider TD (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. Journal of Theoretical Biology 148: 125–137.

[47] Shultzaberger RK, Roberts LR, Lyakhov IG, Sidorov IA, Stephen AG, Fisher RJ, Schneider TD (2007) Correlation between binding rate constants and individual information of *E. coli* Fis binding sites. Nucleic Acids Research 35: 5275–5283.

[48] Rhodius VA, Mutalik VK (2010) Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, σ^E. Proceedings of the National Academy of Sciences of the United States of America 107: 2854–2859.

[49] Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, Karp PD (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. Nucleic Acids Research 39: D583–90.

[50] Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC (2012) The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proceedings of the National Academy of Sciences of the United States of America 109: E1277–86.

[51] Albrecht M, Sharma CM, Dittrich MT, Muller T, Reinhardt R, Vogel J, Rudel T (2011) The transcriptional landscape of *Chlamydia pneumoniae*. Genome Biology 12: R98.

[52] Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J, Hess WR (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. Proceedings of the National Academy of Sciences of the United States of America 108: 2124–2129.

[53] Sierro N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. Nucleic Acids Research 36: D93–6.

[54] Chen H, Bjerknes M, Kumar R, Jay E (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. Nucleic Acids Research 22: 4953–4957.

[55] Molina N, van Nimwegen E (2008) Universal patterns of purifying selection at noncoding positions in bacteria. Genome Research 18: 148–160.

[56] Nelson CE, Hersh BM, Carroll SB (2004) The regulatory content of intergenic DNA shapes genome architecture. Genome Biology 5: R25.

[57] Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I, Rodionov DA (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. Nucleic Acids Research 38: D111–8.

[58] Maier T, Schmidt A, Guell M, Kuhner S, Gavin AC, Aebersold R, Serrano L (2011) Quantification of mRNA and protein and integration with protein turnover in a bacterium. Molecular Systems Biology 7: 511.

[59] Nie L, Wu G, Zhang W (2006) Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: a multiple regression to identify sources of variations. Biochemical and Biophysical Research Communications 339: 603–610.

[60] Towsey M, Hogan JM, Mathews S, Timms P (2007) The *in silico* prediction of promoters in bacterial genomes. Genome Informatics 19: 178–189.

[61] Rangannan V, Bansal M (2011) PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes. BMC Research Notes 4: 257.

[62] Gerland U, Moroz JD, Hwa T (2002) Physical constraints and functional characteristics of transcription factor-DNA interaction. Proceedings of the National Academy of Sciences of the United States of America 99: 12015–12020.

[63] Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. Journal of Molecular Evolution 44: 632–636.

[64] Frock AD, Gray SR, Kelly RM (2012) Hyperthermophilic Thermotoga species differ with respect to specific carbohydrate transporters and glycoside hydrolases. Applied and Environmental Microbiology 78: 1978–1986.

[65] Darfeuille F, Unoson C, Vogel J, Wagner EG (2007) An antisense RNA inhibits translation by competing with standby ribosomes. Molecular Cell 26: 381–392.

[66] Waters LS, Storz G (2009) Regulatory RNAs in bacteria. Cell 136: 615–628.

[67] Rinker KD, Kelly RM (1996) Growth physiology of the hyperthermophilic Archaeon *Thermococcus litoralis*: development of a sulfur-free defined medium, characterization of an exopolysaccharide, and evidence of biofilm formation. Applied and Environmental Microbiology 62: 4478–4485.

[68] Portnoy VA, Herrgard MJ, Palsson BO (2008) Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient *Escherichia coli strain*. Applied and Environmental Microbiology 74: 7561–7569.

[69] Pysz MA, Ward DE, Shockley KR, Montero CI, Conners SB, Johnson MR, Kelly RM (2004) Transcriptional analysis of dynamic heat-shock response by the hyperthermophilic bacterium *Thermotoga maritima*. Extremophiles 8: 209–217.

[70] Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nature Methods 6: 550–551.

[71] Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research 18: 821–829.

[72] Halasz G, van Batenburg MF, Perusse J, Hua S, Lu XJ, White KP, Bussemaker HJ (2006) Detecting transcriptionally active regions using genomic tiling arrays. Genome Biology 7: R59.

[73] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Methods 7: 709–715.

[74] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.

[75] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511–515.

[76] Schrimpe-Rutledge AC, Jones MB, Chauhan S, Purvine SO, Sanford JA, Monroe ME, Brewer HM, Payne SH, Ansong C, Frank BC, Smith RD, Peterson SN, Motin VL, Adkins JN (2012) Comparative omics-driven genome annotation refinement: application across Yersiniae. PLoS One 7: e33903.

[77] Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, Palsson BO (2012) *In silico* method for modeling metabolism and gene product expression at genome scale. Nature Communications 3: 929.

[78] Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing : 127–138.

[79] Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology 2: 28–36.

[80] Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Research 14: 1188–1190.

[81] Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods in Molecular Biology 453: 3–31.

[82] Takemoto K, Nacher JC, Akutsu T (2007) Correlation between structure and temperature in prokaryotic metabolic networks. BMC Bioinformatics 8: 303.

[83] Kingsford CL, Ayanbule K, Salzberg SL (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biology 8: R22.

[84] Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A (2011) Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Research 39: D141–5.

Chapter 4

Adaptive evolution of *Thermotoga maritima* reveals plasticity of the ABC transporter network

4.1 Abstract

Thermotoga maritima is a phylogenetically deep branching hyperthermophilic anaerobe that efficiently ferments a variety of carbon sources to produce hydrogen. *T. maritima* utilizes a vast and diverse network of ABC transporters to metabolize these carbon sources. However, the organism does not metabolize glucose as readily as di- and polysaccharides. After a 25-day glucose laboratory evolution, phenotypes were observed with 60-143% faster growth rates and 26-57% increased glucose utilization rates compared with wild type. Genome resequencing and gene expression analysis showed characteristic mutation across evolved replicates that impacted the glucose responsive ABC transporters. The native glucose ABC transporter, GluEFK, has more abundant transcripts as a result of either gene duplication-amplification or through mutations to the operator sequence regulating this operon. Conversely, BglEFGKL—a transporter of beta-glucosides—is substantially down-regulated either due to a nonsense mutation to the solute binding protein gene or due to a deletion of the upstream promoter. Analysis of the ABC2 uptake porter families for carbohydrate

and peptide transport reveals that the solute binding protein, often among the highest detected transcripts, are predominantly down-regulated in the evolved cultures while the membrane spanning domain and nucleotide binding components are less varied. Similar trends were observed in evolved strains grown on glycerol, a substrate that is not dependent on ABC transporters. Therefore, improved growth on glucose is achieved through mutations favoring *gluEFK* expression over *bglEFGKL* and that in lieu of carbon catabolite repression the ABC transporter network is modulated to achieve improved growth fitness.

4.2 Introduction

Cells are continually changing and adapting in an effort to find an optimal state given the environmental conditions they encounter [1, 2, 3, 4]. The dynamic nature of regulatory networks [5, 6, 7, 8] and the evolution of genomes in response to prolonged exposure to a given environment [4] are manifested in changes in the cells phenotype. Thus, the genotype, the phenotype, and the environment the microbes inhabit are intricately connected [1, 2, 3, 4, 9, 10]. A fundamental step in understanding the behavior of living systems is to reveal the connections underpinning the genotype-phenotype relationship. Between the genotype and phenotype lie many cellular processes such as transcription, translation, RNA degradation, and protein turnover, all of which are governed by complex regulatory networks.

An approach to better define the genotype-phenotype relationship is to track evolutionary changes in a laboratory setting [11, 12, 13]. Laboratory evolution is a systematic approach to examine the dynamic response cells undergo to a changed environmental state. Prolonged exposure to a perturbed environment such as elevated growth temperatures [14, 15] or exposure to a non-native carbon source [16] may result in adapted regulation and/or genetic changes that result in improved phenotypic properties. Throughout the laboratory evolution physiological properties are measured and related back to the original phenotype. This can then be coupled with multiple genome-scale approaches to identify the underlying changes that produced the new phenotype [12, 13]. Integrated analysis of genome resequencing data and

gene expression profiling has been demonstrated to reveal causal mutations and the downstream impact of those on regulation, transcription, and protein functionality [11, 12, 13]. The simplicity of laboratory evolution experiments and the advent of next-generation sequencing makes any culturable microorganism a candidate for an evolution study, providing insight into the different mechanisms underlying adaptation and evolution.

Here, laboratory evolution is applied to study the genotype-phenotype relationship in *Thermotoga maritima*. *T. maritima* is the best characterized species of the Thermotogae phylum. Thermotogae are found at the base of the bacterial 16S rDNA-based phylogenetic tree [17, 18]. Although the exact depth of the phylum has been a matter of debate [19, 20], the phylum is consistently considered deep-branching and, as such, has been the focus of several evolutionary studies [19, 21, 22, 23, 24, 25, 26]. Furthermore, organisms from hydrothermal vent communities, such as *T. maritima*, are thought to harbor traits of early life [27]. *T. maritima* grows anaerobically and has an optimal growth temperature of 80 °C [28]. The microorganism is strictly fermentative and capable of converting a large variety of mono-, di- and polysaccharides to produce hydrogen with high stoichiometric efficiency. As a source of thermostable enzymes and as an efficient producer of hydrogen, *T. maritima* has garnered interest for many biotechnology applications [29, 30].

The genome sequence of *T. maritima* was first completed in 1999 and revealed that *T. maritima* has no phosphotransferase system and is heavily reliant on ATP-binding cassette transporters (ABC transporters) for the import of carbon sources [19]. Recently, this genome has been updated to include a 10kb gap missing in the initial assembly [31, 32]. This cassette encodes two ABC transporters primarily responsible for the transport of glucose (*gluEFK*) and trehalose (*treEFG*) [32, 33, 34]. ABC transport genes account for nearly 60% of all classified transporter proteins in this organism and nearly 10% of all protein coding genes. Of the 173 ABC transporters found in *T. maritima*, 139 belong to the ABC2 uptake family. It is this vast network of ABC transporters that enables *T. maritima* to metabolize such a diverse set of carbon sources. However, growth on di- and polysaccharides like maltose and starch, is typically faster than that of monosaccharides [34]. The average doubling

time for *T. maritima* grown on glucose, mannose, and xylose was found to be more than double that for various polysaccharides (200 min and 90 min, respectively) [35]. It has been suggested that glucose is not readily metabolized due to its thermolability at the physiological growth temperature of *T. maritima* [29]. No carbon catabolite repression system is known to exist in *T. maritima* [35, 36, 34] and the bacterium is known to metabolize multiple carbon sources simultaneously [36].

Recent advances in genome characterization of *T. maritima* have significantly enhanced our understanding of regulation and gene expression in this microorganism. A comprehensive and detailed reconstruction of the carbohydrate utilization regulatory network revealed that each of the 17 local transcription factor regulons controls at most seven operons [33, 34]. Most of the *T. maritima* ABC2 uptake transporters are controlled via transcription factors in this network. Complementary to these efforts, the transcription start sites, transcription units, and the σ^{70} promoter composition were defined for *T. maritima* in a multi-omic data integration effort [31].

Here, the genotype-phenotype relationship is studied in *T. maritima* through a multi-omic, genome-scale characterization of laboratory evolved cultures grown with glucose as the sole carbon source. Wild type cultures were continually passaged until no additional growth improvements were observed. These cultures were then characterized physiologically, genetically, as well as through gene expression profiling and the results were analyzed in comparison with the wild type strain. In doing so, the inherent growth limitations for *T. maritima* growing on glucose are revealed and the underlying genetic modifications and regulatory changes that alleviate this limitation are interrogated.

4.3 Materials and Methods

4.3.1 Culture conditions and physiology.

Thermotoga maritima MSB8 ATCC cultures were grown anaerobically at 80 °C in chemically defined, minimal medium [37]. For glucose laboratory evolution, cultures were serially passaged on 10 mM glucose in serum bottles daily. Evolutions were terminated upon observation of a plateau in the calculated growth rate. Glyc-

erol adaptation was performed by an initially prolonged incubation (2-3 weeks) until growth was detected using optical density measurements (OD_{600}). The adaptation was continued for an additional month with the time decreasing between passages. The physiology of evolved cultures was assessed using a batch bioreactor setup with continuous N_2/CO_2 gas sparging and pH control set to 6.5 as previously described [31]. Growth in the bioreactors was assessed using OD_{600} measurements. Simultaneously, samples were collected for extracellular glucose and acetate levels using HPLC (Aminex HPX-87H Column #125-0140). Cells were collected during exponential phase for total RNA isolation and, subsequently, gene expression via RNA-seq.

4.3.2 Genomic DNA sequencing and variant analysis.

Genomic DNA was isolated from evolved cultures using standard phenol chloroform isoamyl alcohol extraction techniques. Sequencing libraries were constructed using Nextera XT DNA Sample Preparation Kit (Illumina) following vendor instructions. Paired-end libraries were then sequenced (2x250) on an Illumina MiSeq platform. Genetic variants were detected using the Breseq software package [4] using the NC_021214 reference genome sequence and annotation. This reference genome was constructed using Illumina paired-end reads. The raw reads used to assemble this genome (SRX233319) were also analyzed using Breseq to exclude potential assembly errors from variant analysis. Breseq flags for copy number variation and polymorphism prediction were turned on. Mutations with a frequency exceeding 30% were considered significant in this study.

4.3.3 RNA-seq and transcript abundance estimation.

Total RNA was isolated from exponentially growing cells harvested during bioreactor experiments. Biological replicates of each mutant were enzymatically lysed as previously described [31]. Crude lysate was then purified for total RNA using the RNeasy Mini Kit (Qiagen) with on-column DNase treatment. Total RNA was quantified using a Nanodrop (Thermo Scientific) and the quality was checked using a Bioanalyzer (Agilent). Purified total RNA was then prepared for Illumina sequenc-

ing as previously described [31]. Paired-end, strand-specific RNA-seq was performed using the dUTP method [38]. rRNA was depleted using the Ribo-Zero rRNA Removal Kit for Bacteria (Epicentre). rRNA depleted RNA was then fragmented using sing Ambion's RNA Fragmentation Reagent. Reverse transcription was primed using random hexamers. Completed libraries were sequenced using an Illumina MiSeq platform. Paired-end reads were mapped against the NC_021214 genome using bowtie2 with default settings [39]. Subsequent alignment files were processed using Cuffdiff from the Cufflinks suite of tools for FPKM (Fragments Per Kilobase of transcript per Million mapped reads) determination and comparative analysis [40].

4.3.4 Gene expression analysis.

Regulon information on *T. maritima* was obtained from Regpricise [41] and from the publications defining the sugar regulons in *T. maritima* [33, 34]. Hypergeometric enrichment was performed using a custom python script using the scipy function hypergeom. Heatmaps were generated using the R package gplots function heatmap.2. Clusters of Orthologous Groups (COGs) were extracted from the Integrated Microbial Genomes (IMG) database [42]. Genes belonging to the Transporter Classification Database (TCDB) [43] ABC transporter superfamily (3.A.1) were identified in IMG. Family predictions were performed using TransporterTP [44]. TCDB families belonging to the ABC2 uptake family are defined in the following publication [45]. RNAFold from the ViennaRNA Package [46] was used for prediction of hairpin structures using default settings and a temperature set point of 80 °C.

4.3.5 Data Deposition.

Data is made publically available through the Gene Expression Omnibus under Series GSE63141.

4.4 Results

4.4.1 Glucose evolution and evolved phenotypic properties.

Glucose evolution was conducted under batch conditions with daily serial passaging. Cultures were maintained in exponential phase using a variable seed inoculum calculated from growth rate estimates. Figure 4.1 shows the increase in estimated growth rate for the three evolved cultures generated in this study and the number of generations needed to achieve these end points. The evolution was completed within 25 days and 240 generations. Beyond this number of generations no observed improvements in estimated growth rates were observed. The evolutionary end points generated here have a cumulative number of cell divisions (CCD) of $4.3 - 6.0 \times 10^9$ which is three orders of magnitude lower than that observed for *E. coli* evolution experiments [14, 47, 48].

Further characterization of the phenotypic properties of the evolved cultures was conducted using a series of batch bioreactor experiments where pH was maintained at a set point of 6.5 and the accumulation of inhibitory levels of hydrogen was prevented by implementing a continuous gassing strategy with 80:20 N_2/CO_2 mix. Triplicate biological replicates for the three evolved cultures and the wild type (TM-wt) were grown on 10 mM glucose in bioreactors. The following designations will be used to reference the evolved cultures: eTMglc1, eTMglc2, eTMglc3. Growth rates, glucose uptake rates, acetate production rates and other physiological metrics are presented in Table 4.1. Other than acetate, no other organic metabolic products were found in significant quantities. All three strains show improved growth on glucose with relative fitness improvements ranging from 60-143%. They also show improved glucose utilization rates (26-57%) but acetate production rates did not significantly change. Therefore, the improved growth fitness does not maintain the same conversion efficiency of glucose to acetate compared with wild type.

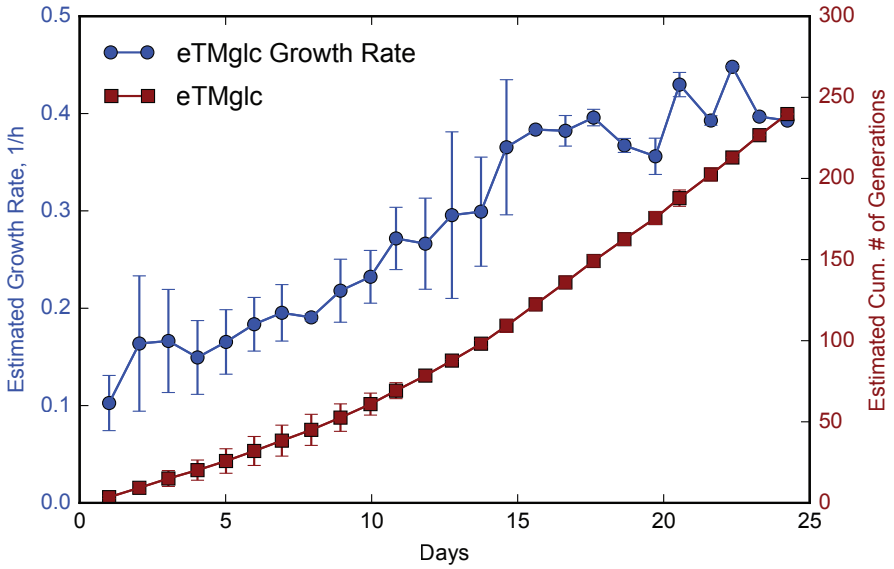


Figure 4.1: **Glucose evolution time course** This plot shows the estimated growth rate (blue) and cumulative number of generations (red) needed to achieve the evolved phenotype for *T. maritima* grown on glucose as the sole carbon source. Evolutions were completed in under 25 days and 250 generations.

Table 4.1: Physiological properties of glucose evolved cultures

	Growth Rate (1/h)	Doubling Time (h)	Relative Fitness	Glc Utilization Rate mM/ (gDCW*h)	Ace Production Rate mM/ (gDCW*h)	Ace:Glc
TM-wt	0.095 ± 0.027	7.8 ± 2.6	1.00	6.1 ± 1.5	10.1 ± 2.9	1.7 ± 0.11
eTMglc1	0.230 ± 0.036	3.1 ± 0.5	2.43	9.6 ± 1.5	12.5 ± 2.3	1.3 ± 0.05
eTMglc2	0.151 ± 0.004	4.6 ± 0.1	1.60	8.6 ± 1.7	11.7 ± 2.3	1.4 ± 0.11
eTMglc3	0.214 ± 0.004	3.2 ± 0.1	2.26	7.7 ± 1.7	10.2 ± 2.3	1.3 ± 0.11

4.4.2 Genetic variants in evolved cultures on glucose.

The evolved cultures (eTMglc) were sequenced to determine possible causal mutations for the observed growth phenotype. Genomic DNA was isolated and sequenced on an Illumina MiSeq with the mean fit coverage exceeding 250X for all cultures. Genetic variants were detected using the Breseq software package. Table S1 (see the Supplemental Material of [49]) summarizes the genetic variants detected. A total of 21 mutations were present in at least one of the replicates. Of the 14 unique gene products associated with mutations, 10 are membrane associated/membrane spanning proteins. ABC transporter genes, their associated transcription factor, or the intergenic region upstream of ABC transporter genes represent 7 of the 14 unique mutations.

Furthermore, we examined mutations for potential impact on glucose metabolism. It is known that glucose is an effector for three locally acting transcription factors — GluR, BglR, XylR [33, 34]. The operons encoding *gluR* and *bglR* are found to have mutations in all three evolved replicate cultures. For eTMglc1, the entire glucose ABC transporter cassette (*gluEFK*) and *gluR* are contained within a large gene duplication-amplification mutation that spans 17.8 kb from Tmari_1847 to Tmari_1860 (Figure 4.2A). Genes in this region have a copy number estimated at 11X with the exception of Tmari_1852 (Maltodextrin glucosidase) and Tmari_1853 (Pullulanase), which have a copy number of around 27X. eTMglc1 also has a non-sense mutation in the sugar binding protein, *bglE* (Tmari_0028), of the *bglEFGKL* ABC transporter (Figure 4.2B). This transporter recognizes and imports beta-glucosides such as cellobiose. However, this mutation to *bglE* is near the center of the gene and eliminates key tryptophan residues (W381, W384, and W536) responsible for forming Van der Waals interactions with cellobiose [50].

Similarly, eTMglc2 and eTMglc3 carry mutations associated with *gluEFK* and *bglEFGKL* ABC transporter operons. These strains carry two point mutations in the intergenic region upstream of the solute binding protein of the *gluEFK* transporter (Tmari_1858, *gluE*). One of these mutations occurs in the GluR operator sequence and the second is slightly upstream of the operator (Figure 4.2A inset). The *bglEFGKL* operon is directly affected by a large deletion that is upstream of the *bglR* gene, it

spans Tmari_0030 and ends upstream of Tmari_0031 — a gene coding for a ferredoxin (Figure 4.2B). This mutation retains only four base pairs upstream of *bglR*, thereby eliminating the native promoter region and transcription start site of the *bglEFGKL* operon.

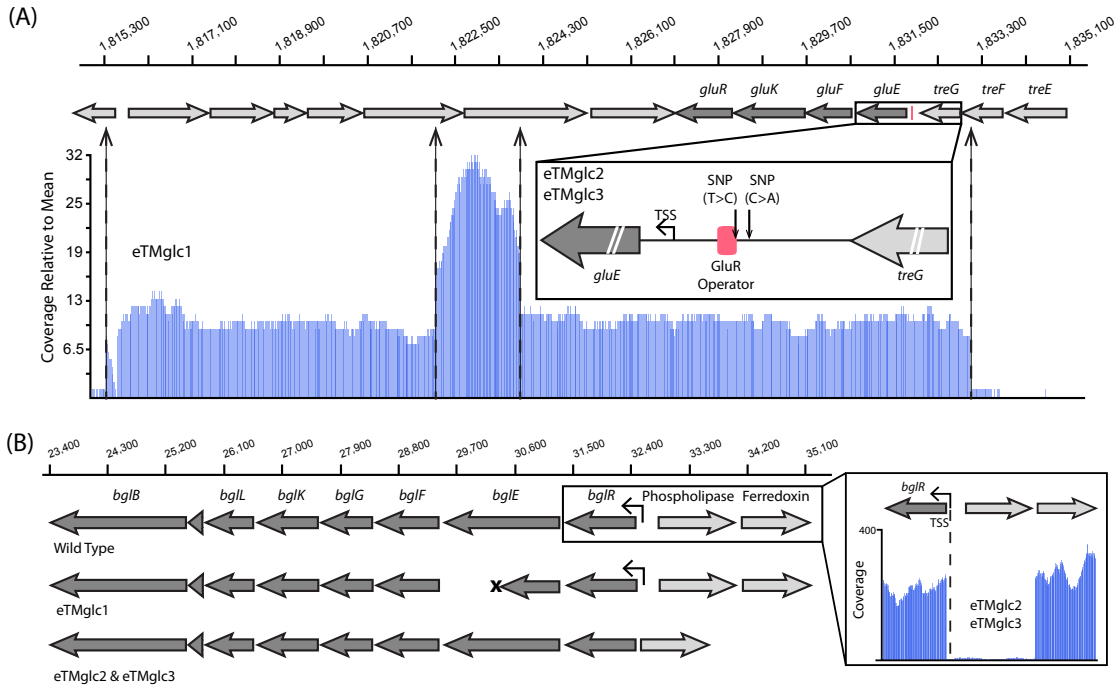


Figure 4.2: **Mutations to the *gluEFK* and *bglEFGKL* ABC transporter operons.** (A) Coverage is shown relative to the mean coverage across the entire genome for the genes spanning Tmari_1847 and Tmari_1861 to illustrate the gene duplication-amplification event observed in eTMglc1. The callout magnifies the intergenic region upstream of *gluE* to illustrate the position of the point mutations relative to the GluR operator sequence (red) found in eTMglc2 and eTMglc3. The *gluEFK*/*gluR* operon is highlighted in dark gray. (B) Genes Tmari_0020-Tmari_0031 are shown for the TM-wt, eTMglc1, and eTMglc2 and eTMglc3 from top to bottom. The *bglEFGKL*/*bglR* genes are highlighted in dark gray. A nonsense mutation resulting in a truncated *bglE* gene is shown for eTMglc1. eTMglc2 and eTMglc3 carry a deletion that spans the intergenic regions upstream of *bglR* and ferredoxin and eliminates Tmari_0030 (call out).

4.4.3 Gene expression analysis of eTMglc mutant cultures.

To examine the potential impact of genetic mutations on gene expression, RNA-seq libraries were constructed from samples collected at mid-log during growth experiments performed in bioreactors. Data was generated for the three evolved replicates and compared to wild type using the Cuffdiff package [40]. A hypergeometric enrichment for significantly overrepresented Clusters of Orthologous Groups (COGs) showed that only the carbohydrate transport and metabolism category (G) was significantly enriched in all three evolved cultures (p-value < 0.05, see Figure S1 in the Supplemental Material of [49]). This result prompted us to further examine the sugar regulons defined for *T. maritima* [33, 34]. Hypergeometric enrichment of the sugar regulons (Figure 4.3A) shows that genes regulated by GluR, BglR, AraR, IolR, and UgpR are overrepresented among the differentially expressed genes in all three evolved cultures, whereas XylR, KdgR, TreR, GalR, and UctR regulons are only enriched in one or two of the evolved cultures.

Of the enriched regulons, GluR and BglR are of particular interest because of the presence of mutations that involve glucose transport and metabolism pathways. For GluR, there is up-regulation of the *gluEFK* ABC transporter genes and *gluR* (Figure 4.3B), whereas the BglR regulon is dramatically down-regulated (Figure 4.3C). The first gene in the *gluEFK* operon, *gluE*, is not differentially expressed but the subsequent genes are up-regulated. *gluE* has the highest FPKM (Fragments Per Kilobase per Million reads mapped) in the operon, as is common for solute binding proteins [51, 52, 53], and has a predicted secondary structure in the intergenic region between *gluE* and *gluF* that could be a target for post-transcriptional regulation (see Figure S2 in the Supplemental Material of [49]). Expression of the *bglEFGKL* ABC transporter genes is nearly abolished in eTMglc2 and eTMglc3 where the upstream promoter is deleted. Similarly, the *bglE* gene carrying the non-sense mutant in eTMglc1 and all downstream genes are down-regulated with *bglE* having the largest fold change (1.98 $\log_2$). Glucose is a potential effector for XylR in *T. maritima*. Wild type XylR is activated *in vitro* in the presence of glucose [33] but uninduced *in vivo* [34]. XylR regulated genes are predominantly up-regulated in eTMglc1 and eTMglc3 (Figure 4.3D). The enrichment of the KdgR regulon is likely due to fact that XylR

and KdgR regulate many of the same genes. The three genes regulated solely by KdgR, Tmari_0060-Tmari_0062, are not differentially expressed, further supporting XylR as the transcription factor causing up-regulation in the evolved cultures.

Many of the *T. maritima* sugar regulons regulate genes encoding transporters belonging to the ABC2 uptake family [43, 45]. TransporterTP predicts that *T. maritima* encodes 139 genes that fall into 14 different ABC2 uptake families with the two carbohydrate uptake transporter families (3.A.1.1 & 3.A.1.2) and the peptide/opine/nickel uptake transporter family (3.A.1.5) accounting for 107 (77%) of these genes. Of these 107 genes, 38, 33, and 39 were differentially expressed for eTMglc1, eTMglc2, and eTMglc3 respectively representing 13-14% of all differentially expressed genes. Genes in these categories were characterized based on their predicted function using the categorization provided by ABCdb [54]. This divided the cohort into three groups: solute binding proteins, 'S'; proteins with membrane spanning-domains, 'M'; and proteins with nucleotide-binding domains that hydrolyze ATP to ADP, 'N'. Genes from these categories for a given ABC transporter are often operonic and expressed in a single transcription unit [52, 55, 56].

Comparison of the absolute transcript levels (FPKMs) between genes in these categories shows that the solute binding proteins are higher than those observed seen in the membrane spanning and ATP hydrolyzing groups (Figure 4.4A). However, examination of the evolved strains reveals a drop in the global FPKM distribution for the solute binding proteins. Differential expression analysis of the evolved strains relative to the wild type strain confirms that the global drop in the solute binding proteins is significant greater than for other constituents of ABC transports (Figure 4.4B). To examine this further, RNA-seq was performed on three independently adapted cultures on glycerol. Unlike most carbon sources, glycerol enters *T. maritima* through passive or facilitated diffusion rather than through ABC mediated transport [19]. Therefore, there is no obvious selective advantage to overexpressing ABC importer genes. Analogous to the trends seen with the evolved cultures on glucose, the solute binding proteins had the highest expression levels but were relatively more down-regulated in comparison with the wild type strain (see Figure S3 in the Supplemental Material of [49]). This indicates that the evolution on glucose results in

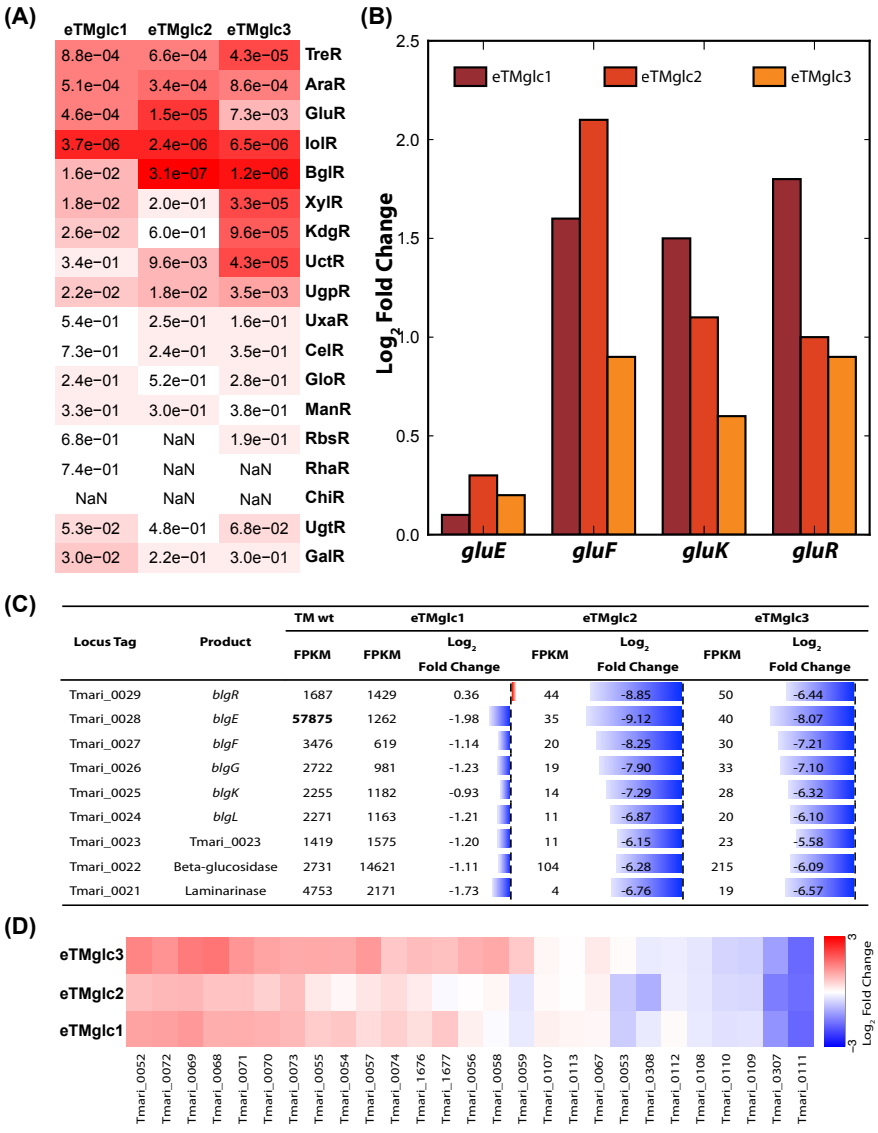


Figure 4.3: **Gene expression analysis of eTMglc cultures.** (A) Heatmap illustrating the significance of the hypergeometric enrichment test performed on sugar regulons. Values in the blocks of the heatmap correspond to their respective p-values. (B) Increased gene expression of the *gluEFK* and *gluR* regulon in eTMglc cultures compared with wild type. Bars represent $\log_2$ fold change for each gene in the operon. (C) Down regulation of the *bglEFGKL* operon for the eTMglc cultures. For each evolved culture, the FPKM value for each gene is shown as is the $\log_2$ fold change with respect to wild type *T. maritima* grown on glucose. (D) Heatmap of the $\log_2$ fold change for the XylR regulon.

a global regulatory response that streamlines the transcript levels of ABC2 uptake transporters and that this streamlining is primarily achieved by modulating the solute binding protein transcript levels.

4.5 Discussion

This study reveals a clear connection between genotype and phenotype through the application of multi-omic data integration and physiological characterization following laboratory evolution experiments. *T. maritima* is capable of overcoming deficiencies in glucose metabolism in relatively short time frames. This organism achieved an enhanced phenotype within 25 days of the study onset with a CCD that is 1,000 times lower than that typically observed in *E. coli* [47, 48, 14]. The eTMglc phenotypes are capable of growing twice as fast as the wild type and utilizes upwards of 57% more glucose. The rapid evolution and the vastly improved phenotype achieved on glucose suggest that the thermolability of glucose is unlikely to be the root cause for the poor growth achieved by *T. maritima* as has been previously suggested [29]. Such a biophysical limitation in the sole carbon source would present a formidable challenge for the cells to overcome. Therefore, the poor growth on glucose is likely an inherent limitation in glucose uptake and metabolism.

This inherent limitation is further supported by the multi-omic characterization of the genome and transcriptome of the evolved cultures. In all evolved cultures, the genes regulated in the GluR and BglR regulons are also associated with genetic mutations. These local acting transcription factors directly bind glucose as an effector [33, 34] and regulate operons encoding ABC2 uptake transporters. In the case of GluR, the gene expression of the glucose ABC transporter, *gluEFK*, was enhanced. The *gluEFK*/*gluR* operon was observed to undergo point mutations in the intergenic region upstream of *gluE*, one of which directly impacts the GluR operator sequence in two evolved cultures. The third culture was found to have repetitively undergone a gene duplication-amplification event that increased the gene dosage of the *gluEFK*/*gluR* operon. Interestingly, *gluEFK* and *treEFG* were only recently discovered in *T. maritima* after an initial omission of an ≈9 kb region in the reference

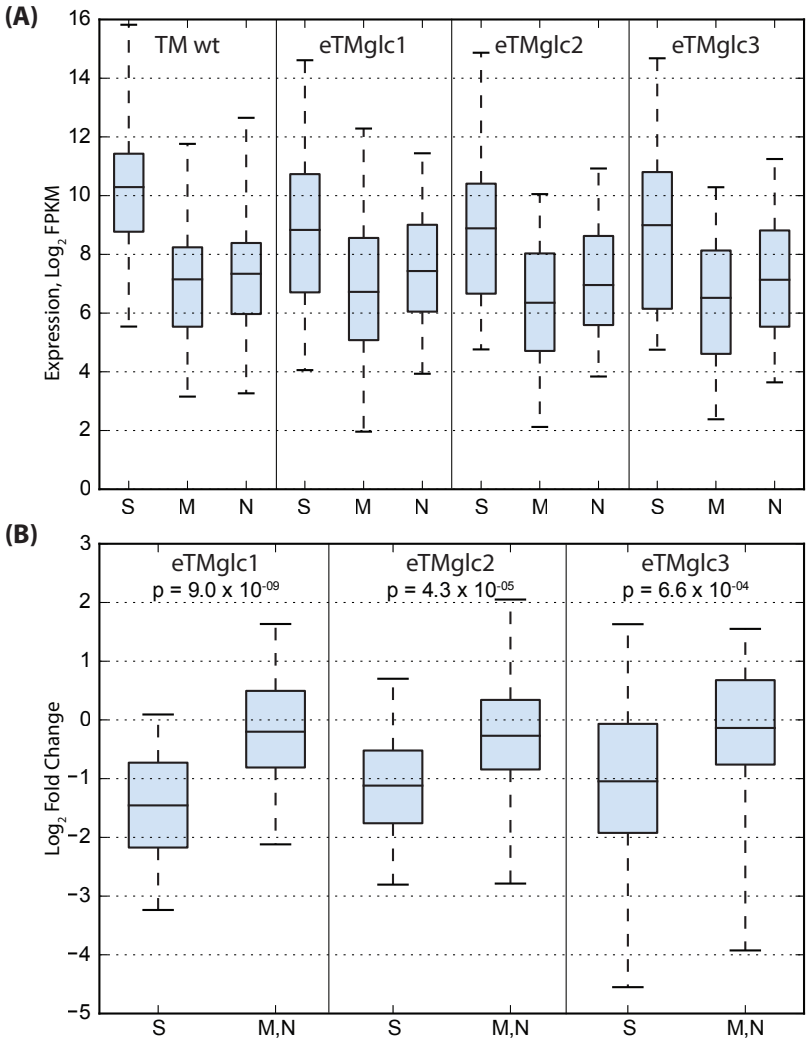


Figure 4.4: **Gene expression analysis of the different functional categories of proteins found in the ABC2 importer families for carbohydrate uptake (3.A.1.1 & 3.A.1.2) and the peptide/opine/nickel uptake family (3.A.1.5) for cultures grown on glucose.** (A) Boxplot showing the absolute transcript abundance measures (FPKMs) for the different ABC transporter protein components across all cell lines. (B) Boxplot showing the distribution of the $\log_2$ fold change for the different ABC transporter protein components for all evolved cell lines relative to wild type. Box plot shows values falling within 95% confidence intervals with the box comprising of the interquartile range and the horizontal line within the box representing the mean. Two tailed p-values were determined using the Students t-test. ABC transporter proteins are categorized as and 'S' for solute binding protein, 'M' for membrane spanning domain, 'N' for ATP hydrolyzing protein.

genome [32]. In fact, the large duplication events discussed here spans the entire ≈9 kb region. It is thought that this region was omitted from the original genome assembly as a result of a deletion event that occurred during early sub-culturing [32]. Furthermore, the most upstream gene of this duplication event, Tmari_1847, is part of one of two maltose ABC transporters in *T. maritima* that are thought to have arisen via a duplication event [21]. Therefore, this segment of the genome appears to be rather unstable but further characterization of this region may provide valuable insights into the evolutionary trajectory of ABC transporters.

Unlike the GluEFK transporter, the evolved glucose strains substantially down-regulate the *bglEFGKL* operon. *bglE*, the solute binding component, is the second highest detected transcript in the wild type culture with a $\log_2$ FPKM of 15.8 but this is reduced to nearly zero in two of the evolved glucose cultures. This is the result of a chromosomal gene deletion to Tmari_0030 but, more importantly, the promoter governing expression of the operon carrying *bglEFGKL*, *bglR* and other pathway genes is also eliminated. The third evolved culture harbors a nonsense mutation that truncates key amino acids necessary for binding of beta-glucosides [50]. These changes to the functionality of BglEFGKL reveal a potential regulatory inefficiency due to effector overlap between GluR and BglR. Glucose induces a strong transcriptional response for *bglEFGKL*, which is known to only transport beta-glucoside polymers comprised of 2-5 monomers [50, 51, 57]. Therefore, while GluR senses glucose and induces a beneficial transcriptional response, BglR produces a greater transcriptional response that is ineffectual in the uptake of glucose monomers. Consequently, cellular resources are diverted from producing the appropriate GluEFK protein complex to the transcription and translation of BglEFGKL ABC transporter complexes creating a substantial deficiency in the uptake rate of glucose.

Furthermore, interrogating the global transcriptional response of ABC2 uptake porters revealed a global modulation of the transcript levels of the solute binding domain component in response to glucose evolution. The transcript abundance for the solute binding protein is greater than that found for the membrane spanning domains and ATP-hydrolyzing proteins. This pattern of transcript abundance has been observed previously in ABC transporters [51, 52, 53] and it is thought to be

due to a stabilizing hairpin in the intergenic region downstream of the solute binding protein that hinders 3'-5' exonucleases activity [55, 52, 56]. It has been demonstrated that elimination of the stabilizing hairpin of the solute binding protein greatly reduces transcript abundance in the *E. coli malEFG* operon and the *Bacillus subtilis pst* operon [55, 52, 56]. In fact, all of the *T. maritima* solute binding proteins in the ABC2 uptake families for carbohydrate and peptide transport contain a predicted hairpin structure (see Figure S4 in the Supplemental Material of [49]). Therefore, the higher transcript abundance for the solute binding protein genes is likely due to increased mRNA half-lives and one would subsequently expect reduced fluctuations in the abundance measures for these highly stable transcripts. Yet, the evolved strains on glucose demonstrate greater down-regulation of these genes compared with the other components of ABC2 uptake porters. A similar observation was made for cultures adapted to grow on glycerol, a carbon source transported independent of ABC transporters. While the absolute transcript levels of the solute binding protein are high, they are lower in the evolved cultures relative to wild type and the other ABC import complex proteins are relatively unchanged.

Overall, the evolved cultures primarily modulate the transcript abundance of the solute binding proteins within the ABC2 uptake transporter network to achieve a more efficient phenotype streamlined for glucose utilization and metabolism. The evolved cultures expend less energy on expressing solute binding proteins that recognize unavailable carbon sources and redirect their efforts to increased uptake of glucose. This drastic change in the solute binding protein is not shared across the other ABC transporter proteins. Though transcriptional regulation is likely a contributing factor to this effect, the possibility of post-transcriptional control of mRNA stability cannot be ruled out. The latter better explains our observation that the solute binding protein is down-regulated while maintaining comparable levels of membrane spanning and ATP-hydrolyzing components. Implementing a post-transcriptional mechanism that has limited impact on the membrane spanning and ATP-hydrolyzing transcript abundance could also result in a faster response to a change in carbon source. The execution of additional laboratory evolution experiments on different carbon sources will help further unravel the regulatory mechanisms that govern the ABC transporter

network and lead to deeper insights into the evolutionary trajectory of this ubiquitous class of proteins.

4.6 Acknowledgments

Funding for this work was provided by the Office of Science of the U.S. Department of Energy (DOE) under grants DE-FG02-08ER64686 and DE-FG02-09ER25917. HL is supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086.

Chapter 4 has been submitted for publication of the material as it may appear in Latif H, Sahin M, Tarasova J, Tarasova Y, Portnoy VA, Zengler K. Adaptive evolution of *Thermotoga maritima* reveals plasticity of the ABC transporter network. Submitted to Appl. Environ. Microbiol. 2014. The dissertation author was the primary author of this paper responsible for the research. The other authors are Merve Sahin, Janna Tarasova, Yekaterina Tarasova, Vasiliy A. Portnoy, and Karsten Zengler.

4.7 Bibliography

[1] Lynch M (2006) Streamlining and simplification of microbial genome architecture. Annual Review of Microbiology 60: 327-49.

[2] Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401-4.

[3] Andersson DI, Hughes D (2009) Gene amplification and adaptive evolution in bacteria. Annual Review of Genetics 43: 167-95.

[4] Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243-7.

[5] van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends in Genetics 19: 479-84.

[6] Janga SC, Collado-Vides J (2007) Structure and evolution of gene regulatory networks in microbial genomes. Research in Microbiology 158: 787-94.

[7] McAdams HH, Srinivasan B, Arkin AP (2004) The evolution of genetic regulatory systems in bacteria. Nature Reviews Genetics 5: 169-78.

[8] Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proceedings of the National Academy of Sciences of the United States of America 101: 3160-5.

[9] Via S, Lande R (1985) Genotype-environment interaction and the evolution of phenotypic plasticity. Evolution : 505-522

[10] Koonin EV, Wolf YI (2010) Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics 11: 487-98.

[11] Portnoy VA, Bezdan D, Zengler K (2011) Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering. Current Opinion in Biotechnology 22: 590-4.

[12] Conrad TM, Lewis NE, Palsson BØ (2011) Microbial laboratory evolution in the era of genome-scale science. Molecular Systems Biology 7: 509.

[13] Hindré T, Knibbe C, Beslon G, Schneider D (2012) New insights into bacterial adaptation through *in vivo* and *in silico* experimental evolution. Nature Reviews Microbiology 10: 352-65.

[14] Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM (2014) Evolution of *Escherichia coli* to 42 °c and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Molecular Biology and Evolution 31: 2647-62.

[15] Goodarzi H, Bennett BD, Amini S, Reaves ML, Hottes AK, Rabinowitz JD, Tavazoie S (2010) Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in *E. coli*. Molecular Systems Biology 6: 378.

[16] Lee DH, Palsson BØ (2010) Adaptive evolution of Escherichia coli k-12 mg1655 during growth on a nonnative carbon source, l-1,2-propanediol. Applied and Environmental Microbiology 76: 4158-68.

[17] Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer KH, Glöckner FO, Rosselló-Móra R (2011) Release ltps104 of the all-species living tree. Systematic and Applied Microbiology 34: 169-70.

[18] Achenbach-Richter L, Gupta R, Stetter KO, Woese CR (1987) Were the original eubacteria thermophiles? Systematic and Applied Microbiology 9: 34-9.

[19] Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA,

Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399: 323-9.

[20] Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbø CL, Doolittle WF, Gogarten JP, Noll KM (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the thermotogales. Proceedings of the National Academy of Sciences of the United States of America 106: 5865-70.

[21] Noll KM, Lapierre P, Gogarten JP, Nanavati DM (2008) Evolution of *malABC* transporter operons in the thermococcales and thermotogales. BMC Evolutionary Biology 8: 7.

[22] Mongodin EF, Hance IR, Deboy RT, Gill SR, Daugherty S, Huber R, Fraser CM, Stetter K, Nelson KE (2005) Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. Journal of Bacteriology 187: 4935-44.

[23] Ben Hania W, Postec A, Aüllo T, Ranchou-Peyruse A, Erauso G, Brochier-Armanet C, Hamdi M, Ollivier B, Saint-Laurent S, Magot M, Fardeau ML (2013) *Mesotoga infera* sp. nov., a mesophilic member of the order thermotogales, isolated from an underground gas storage aquifer. International Journal of Systematic and Evolutionary Microbiology 63: 3003-8.

[24] Nesbo CL, L'Haridon S, Stetter KO, Doolittle WF (2001) Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Molecular Biology and Evolution 18: 362-75.

[25] Nesbø CL, Dlutek M, Doolittle WF (2006) Recombination in thermotoga: implications for species concepts and biogeography. Genetics 172: 759-69.

[26] Nesbø CL, Doolittle WF (2003) Targeting clusters of transferred genes in *Thermotoga maritima*. Environmental Microbiology 5: 1144-54.

[27] Martin W, Baross J, Kelley D, Russell MJ (2008) Hydrothermal vents and the origin of life. Nature Reviews Microbiology 6: 805-14.

[28] Huber R, Langworthy TA, König H, Thomm M, Woese CR, Sleytr UB, Stetter KO (1986) *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 °c. Archives of Microbiology 144: 324–333.

[29] Conners SB, Mongodin EF, Johnson MR, Montero CI, Nelson KE, Kelly RM (2006) Microbial biochemistry, physiology, and biotechnology of hyperthermophilic thermotoga species. FEMS Microbiology Reviews 30: 872-905.

[30] Frock AD, Notey JS, Kelly RM (2010) The genus thermotoga: recent developments. Environmental Technology 31: 1169-81.

[31] Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K (2013) The genome organization of *Thermotoga maritima* reflects its lifestyle. PLoS Genetics 9: e1003485.

[32] Boucher N, Noll KM (2011) Ligands of thermophilic abc transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* msb8 screened by differential scanning fluorimetry. Applied and Environmental Microbiology 77: 6395-9.

[33] Kazanov MD, Li X, Gelfand MS, Osterman AL, Rodionov DA (2013) Functional diversification of rok-family transcriptional regulators of sugar catabolism in the thermotogae phylum. Nucleic Acids Research 41: 790-803.

[34] Rodionov DA, Rodionova IA, Li X, Ravcheev DA, Tarasova Y, Portnoy VA, Zengler K, Osterman AL (2013) Transcriptional regulation of the carbohydrate utilization network in *Thermotoga maritima*. Frontiers in Microbiology 4: 244.

[35] Chhabra SR, Shockley KR, Conners SB, Scott KL, Wolfinger RD, Kelly RM (2003) Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. Journal of Biological Chemistry 278: 7540-52.

[36] Frock AD, Gray SR, Kelly RM (2012) Hyperthermophilic thermotoga species differ with respect to specific carbohydrate transporters and glycoside hydrolases. Applied and Environmental Microbiology 78: 1978-86.

[37] Rinker KD, Kelly RM (1996) Growth physiology of the hyperthermophilic archaeon *Thermococcus litoralis*: Development of a sulfur-free defined medium, characterization of an exopolysaccharide, and evidence of biofilm formation. Applied and Environmental Microbiology 62: 4478-85.

[38] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific rna sequencing methods. Nature Methods 7: 709-15.

[39] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nature Methods 9: 357-9.

[40] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with rna-seq. Nature Biotechnology 31: 46-53.

[41] Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, Kazanov MD, Riehl W, Arkin AP, Dubchak I, Rodionov DA (2013) Regprecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics 14: 745.

[42] Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012) Img: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research 40: D115-22.

[43] Saier MH Jr, Reddy VS, Tamang DG, Västermark A (2014) The transporter classification database. Nucleic Acids Research 42: D251-8.

[44] Li H, Benedito VA, Udvardi MK, Zhao PX (2009) Transporttp: a two-phase classification approach for membrane transporter prediction and characterization. BMC Bioinformatics 10: 418.

[45] Zheng WH, Västermark Å, Shlykov MA, Reddy V, Sun EI, Saier MH Jr (2013) Evolutionary relationships of atp-binding cassette (abc) uptake porters. BMC Microbiology 13: 98.

[46] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) Viennarna package 2.0. Algorithms for Molecular Biology 6: 26.

[47] LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM (2014) Discovery of key mutations enabling rapid growth of *Escherichia coli* k-12 mg1655 on glucose minimal media using adaptive laboratory evolution. Applied and Environmental Microbiology .

[48] Lee DH, Feist AM, Barrett CL, Palsson BØ (2011) Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of *Escherichia coli*. PLoS One 6: e26172.

[49] Latif H, Sahin M, Tarasova J, Tarasova Y, Portnoy V, Zengler K (2014) Adaptive evolution of *Thermotoga maritima* reveals plasticity of the ABC transporter network. Submitted to Applied and Environmental Microbiology .

[50] Cuneo MJ, Beese LS, Hellinga HW (2009) Structural analysis of semi-specific oligosaccharide recognition by a cellulose-binding protein of thermotoga maritima reveals adaptations for functional diversification of the oligopeptide periplasmic binding protein fold. Journal of Biological Chemistry 284: 33217-23.

[51] Nanavati DM, Nguyen TN, Noll KM (2005) Substrate specificities and expression patterns reflect the evolutionary divergence of maltose abc transporters in *Thermotoga maritima*. Journal of Bacteriology 187: 2002-9.

[52] Newbury SF, Smith NH, Higgins CF (1987) Differential mrna stability controls relative gene expression within a polycistronic operon. Cell 51: 1131-43.

[53] Stern MJ, Prossnitz E, Ames GF (1988) Role of the intercistronic region in post-transcriptional control of gene expression in the histidine transport operon of *Salmonella typhimurium*: involvement of rep sequences. Molecular Microbiology 2: 141-52.

[54] Fichant G, Basse MJ, Quentin Y (2006) Abcdb: an online resource for abc transporter repertories from sequenced archaeal and bacterial genomes. FEMS Microbiology Letters 256: 333-9.

[55] Allenby NEE, O'Connor N, Prágai Z, Carter NM, Miethke M, Engelmann S, Hecker M, Wipat A, Ward AC, Harwood CR (2004) Post-transcriptional regulation of the *Bacillus subtilis pst* operon encoding a phosphate-specific abc transporter. Microbiology 150: 2619-28.

[56] Newbury SF, Smith NH, Robinson EC, Hiles ID, Higgins CF (1987) Stabilization of translationally active mrna by prokaryotic rep sequences. Cell 48: 297-310.

[57] Conners SB, Montero CI, Comfort DA, Shockley KR, Johnson MR, Chhabra SR, Kelly RM (2005) An expression-driven approach to the prediction of carbohydrate transport and utilization regulons in the hyperthermophilic bacterium *Thermotoga maritima*. Journal of Bacteriology 187: 7267-82.

Chapter 5

Integrated analysis of molecular and systems level function of Crp using ChIP-exo.

5.1 Abstract

A fundamental challenge of systems biology is to unify detailed molecular mechanisms with genome-scale predictive models. Meeting this challenge for transcriptional regulatory networks has proven difficult. Here, ChIP-exo, a derivative of ChIP-seq, is applied to help bridge this gap for transcription initiation in bacterial systems. σ^{70} ChIP-exo profiles revealed patterns of DNA protection at genome-scale under *in vivo* conditions that strongly corroborate *in vitro* DNA footprinting studies. Aligning and orienting the ChIP-exo data relative to the TSS identified a unimodal distribution centered at +20 on the template strand and multimodal distributions located between −35 and +5. Both strands indicate the capture of stable intermediates of transcription initiation that occur post-closed complex formation. Similar profiles are observed for ChIP-exo performed on activators but not for repressors. Activators are found to provide little protection at the operator site whereas repressors are centered on them. Furthermore, genetic perturbations of RNAP/Crp interactions highlight the importance of RNAP/Crp interactions for stabilization of the ternary

complex. This suggests that post-recruitment, Crp is associated with RNAP but independent of the operator site. Finally we are able to confirm recent physiological models of Crp regulation at the genome-scale and provide an initial link between promoter level mechanisms and systems level regulation.

5.2 Introduction

A longstanding goal of molecular biology is to link structure with function. This structure-function relationship plays out at many levels with one of the most complex being the link between the structural attributes of proteins and the overall functions of pathways or systems level features of a cell. This link proves especially difficult because it comprises many smaller, yet equally important links. One crucial smaller link is the relation between a proteins structure and its individual function [1] and another is the link between the structure of a biomolecular network and its corresponding function [2]. Thus, attempting to provide cohesion between the structure of individual components of a cell and the complexes of those components with the overall function of the cellular system as a whole will continue to be a grand challenge for many years to come. Here, we seek to provide an initial link in that direction by synthesizing protein structural level detail of transcriptional initiation complexes with systems level features and physiological models of cellular behavior.

One crucial technique in biological sciences is chromatin immunoprecipitation (ChIP) paired with microarrays (ChIP-chip) [3] and next-generation sequencing (ChIP-seq) [4] to comprehensively unravel the molecular networks of protein binding to DNA. This approach has proved to be highly successful and has resulted in an explosion of knowledge about transcriptional regulation. However, both ChIP-chip and ChIP-seq are limited to the size of the DNA fragments that are produced. ChIP-exonuclease (ChIP-exo) [5] was developed in order to significantly increase this resolution, reaching the level of a dozen or so nucleotides. By utilizing a $5'$ exonuclease to digest the ends of DNA fragments unprotected by a cross-linked transcription factor, it is possible to achieve single nucleotide resolution binding profiles. While ChIP-exo has been extensively deployed in mammalian systems [5], it has yet to be applied in

detail to transcriptional regulation in bacteria. Since much of the foundation of our knowledge of transcriptional regulation comes from bacteria, and *Escherichia coli* in particular, we sought to assess whether or not ChIP-exo would enable new or complementary insights that could be based off the rich body of detailed structural and biochemical work in bacteria.

One transcription factor in particular, Crp, has long been studied as the canonical transcriptional activator in bacteria. Many detailed studies have shown precisely how Crp carries out regulation at three separate classes of promoters [6]. Similarly, recent work has also shown the precise physiological mechanisms by which Crp regulates the entirety of metabolism and the cell [7]. We thus sought to provide a link between this detailed structural knowledge and detailed systems knowledge by fully elucidating the Crp regulatory network in precise chemical detail. Previous studies have been carried out with transcriptomics [8, 9] ChIP-chip [10, 11], and genomic SELEX [12]. However, none of these studies were able to assess a full range of physiologically relevant conditions or connect *in vivo* ChIP measurements with expression profiling to generate a full set of regulatory events. Similarly a vast array of studies has detailed the individual architecture and structure-function relationship at a number of classically studies promoters in *E. coli* [6, 13, 14, 15, 16, 17]. However, these studies did not enjoy the benefit of massively parallel assessment of detailed structural binding information from ChIP-exo.

In this study, we carry out ChIP-exo for both σ^{70} and Crp for wild type cells under glucose, fructose, and glycerol conditions. In addition we study the effects of canonical Crp mutation in two crucial regions, Ar1 and Ar2, on ChIP-exo binding profiles under glycerol conditions. We also carry out paired RNA sequencing across wild type, Δcrp, and strains harboring Ar1 and Ar2 mutations to elucidate the full Crp regulon.

5.3 Results

5.3.1 ChIP-exo data provides genome-scale, *in vivo* mechanistic insights into bacterial transcription activation.

Applying the ChIP-exo technique to interrogate transcriptional mechanisms illustrates the utility of this approach in providing genome-scale, *in vivo* insights into promoter activity. By capturing actively transcribing complexes, stable intermediates for a given promoter are revealed on the path to transcriptional elongation. ChIP-exo results for σ^{70} and for Crp, the classic model for transcription factor mediated activation, are discussed.

σ^{70} ChIP-exo serves as a method for locating the TSS.

The predecessors of ChIP-exo (i.e. ChIP-chip and ChIP-seq) are enriched at bacterial promoters when the σ factor is targeted [18, 19, 20, 21, 22, 23, 24, 25]. Ecocyc annotated TSSs where checked for σ^{70} ChIP-exo enrichment within $\pm$200 bp for data generated on three different substrates—glucose, fructose, and glycerol (Figure 5.1A and Supplementary Figure S1 in [26]). Like its predecessors, ChIP-exo peaks are consistently found near promoters but only the single-nucleotide resolution of ChIP-exo provides the exact genomic location bound by RNAP holoenzyme. Surprisingly, ChIP-exo data generated on σ^{70} is found to be a better proxy for the TSS than it is for the -35 and -10 promoter recognition sequence elements. The median position of the σ^{70} peak center is 5 bp downstream of the TSS for all three carbon sources (Figure 5.1A and Supplementary Figure S1 in [26]). The spatial consistency of the σ^{70} peak center demonstrates the utility of ChIP-exo to approximate TSSs to within base pairs from where they exist and provides an orthogonal method to complement $5'$ RACE-based TSS detection.

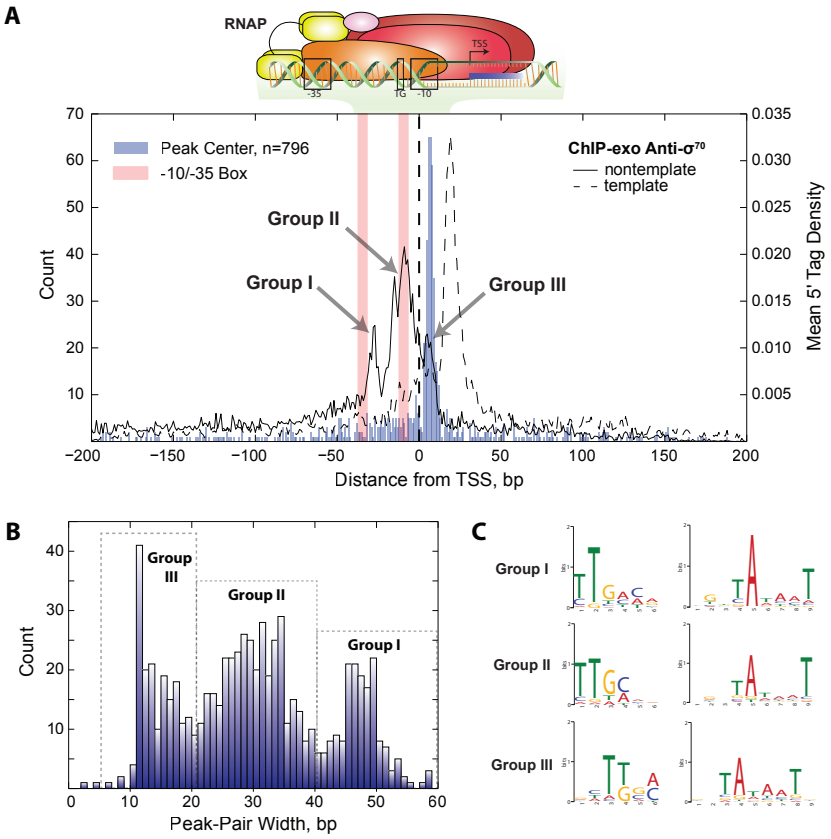


Figure 5.1: **TSS aligned and oriented σ^{70} ChIP-exo data reveals DNA footprint patterns consistent with stable transcription initiation intermediates.** (A) ChIP-exo peak regions were aligned and oriented relative to the TSS. The peak center (blue bars) is shown to be to consistently downstream of the TSS with a median of 5 bp. The mean distribution of the 5′ position of reads (5′ tags) is shown for both the template and nontemplate. The template strand distribution shows a unimodal profile that spans +20±7 bp and is consistent with *in vitro* footprinting studies characterizing the RP_O, the ITC, and the TEC stable intermediates. The nontemplate strand shows a multimodal distribution with modes centered approximately +5 relative to the TSS (Group III), upstream and over the −10 promoter element (Group II), and slightly downstream of the −35 promoter element (Group I). (B) Examination of the distance between template and nontemplate strand peak maximum shows that the footprint lengths are >40 bp, 21 to 40, <20 and for Group I, Group II, and Group III respectively. (C) A motif search was performed for the −10 and −35 promoter elements for Group I, Group II, and Group III promoters. All three show σ^{70}-like promoter sequences but with slight differences. Group I has a −35 motif that most closely resembles the consensus (TTGACA), has a highly conserved -11A, and a partial TGn motif. Group III has the least conserved −35 promoter element and no extended −10 promoter element.

Strand oriented σ^{70} ChIP-exo peak distributions reveal stable intermediates in transcription initiation.

The σ^{70} ChIP-exo peak distribution provides the bounds of protected DNA regions on the template and nontemplate strand. By aligning to the TSS as a reference position, we are able to determine the strand orientation for the σ^{70} ChIP-exo peak regions. ChIP-exo profiles across all binding sites were calculated for the template and nontemplate strand by first calculating the density of the $5'$ end of tags for each individual peak region spanning 400 bp centered on the TSS. The strand orientated ChIP-exo profiles for σ^{70} reveal significant distinctions between the template strand and the nontemplate strand (Figure 5.1A). The binding profiles show a unimodal distribution on the template strand whereas a multimodal distribution is seen on the nontemplate strand. The width of the peak regions was determined by calculating the distance between the position of maximum $5'$ tag count on the template strand relative to position of maximum $5'$ tag count on the nontemplate strand (Figure 5.1B). This corroborates the results seen on the mean $5'$ tag density profile on the two strands and indicates that most promoters have a σ^{70} ChIP-exo profile that falls into one of these distances.

The activity of lambda exonuclease is $5'$ to $3'$ [27] and, as such, the protected region on the template strand is found downstream of the TSS. The unimodal ChIP-exo distribution on the template strand has a maximum $5'$ tag density +20 bp downstream of the TSS and approximately 30% of the mean $5'$ tag density is found between 20±7 bp. The position of the unimodal distribution on the template strand is in strong agreement with numerous *in vitro* footprinting studies in model promoter constructs characterizing the open complex (RP_O) and the stable intermediates leading to RP_O formation, the initial transcribing complex (ITC) and the transition to the ternary elongation complex (TEC) [28, 29, 30, 31]. However, these results are not reflective of footprinting studies capturing the closed promoter complex (RP_C), which typically protect promoter DNA from -5 to $+5$ [28, 29, 30, 31].

Unlike the template strand, the ChIP-exo $5'$ tag distribution for the nontemplate strand is multimodal. This distribution marks the upstream boundary relative to the TSS. The dominant mode accounts for 28% of the $5'$ tag density and is found

between −18 and −1. Therefore, promoters that belong to this mode have partial to complete protection of the discriminator sequence, the −10 promoter element, and the TGn extended −10 element but offer no protection to the −35 promoter element or any upstream promoter elements (e.g., UP element). The −35 promoter element is partially protected by the mode farthest upstream which accounts for 9% of the 5′ tag density profile and spans −34 to −23 with a maximum located at −28. The upstream boundary, −3, is located in the center of the −35 element. The downstream mode accounts for 8% of the 5′ tag density and is located downstream of the TSS. The boundaries of this mode are between +4 and +12 with a local maximum at +6. The multimodal distribution on the nontemplate strand also reflects *in vitro* footprinting studies, which find greater variability in the upstream-protected region [28, 30, 31]. Like the template strand, the DNA protected regions of the different modes on the nontemplate strand provide little to no support that recruitment and RP_C complex formation is being captured by these ChIP studies. The upstream boundary from footprinting studies conducted on RP_C show periodic protected regions extending upstream of the −35 promoter element typically −55 to −12 but is highly dependent on the involvement of transcription factors and the binding of the α-subunit [32, 33, 31, 34]. There is, however, evidence to support that these modes are reflective of stable intermediates that occur post-recruitment and, in particular, RP_O, the ITC and transition to the early TEC [35, 36, 37]. The differentiation between the ITC and TEC can further be seen in the total length of the protected region (Figure 5.1B). Early TECs have been found to have footprint regions spanning ≈30 bp whereas the ITC has a longer footprint seen to be 50+ bp in length [38, 39, 36].

Promoter motif analysis of the σ^{70} ChIP-exo distributions.

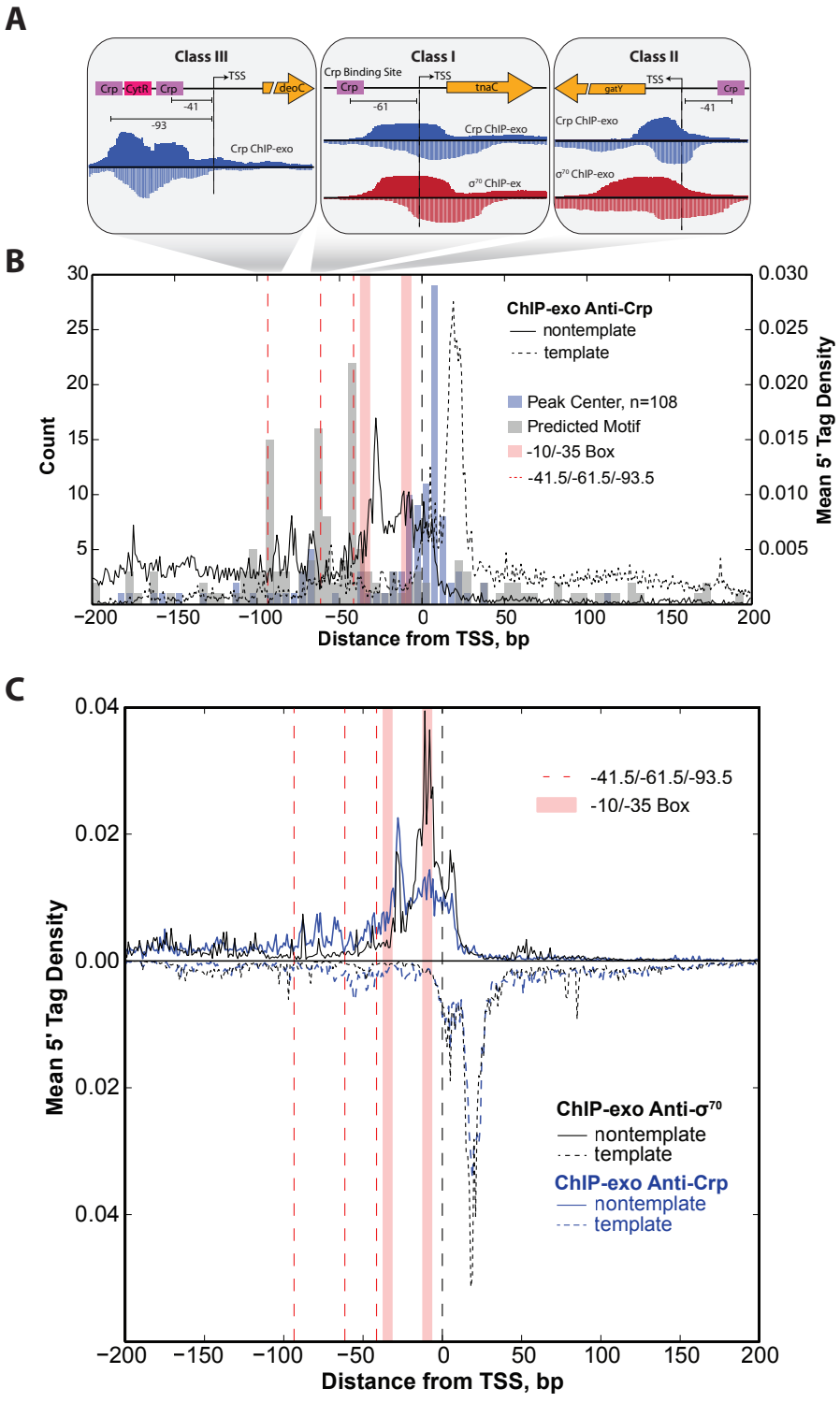
It is known that promoter sequence elements involved with RNAP holoenzyme recruitment contribute to the post-recruitment kinetics of transcription initiation [28, 30, 31]. Thus we examined the −10 and −35 promoter elements for the different σ^{70} groups (Figure 5.1C) as determined by the difference in peak-pairs (Figure 5.1B). σ^{70}-like promoter motifs were found in all three groups. Though a detailed analysis of additional DNA sequence elements (e.g., UP elements, transcription fac-

tor binding sites, nucleoid-associated protein binding sits) could be more revealing, examination of just these two promoter elements revealed subtle difference between the promoter elements of each groups. Group I, having the longest distance between peak-pairs, has a motif that most resembles the −35 consensus sequence (TTGACA). Furthermore, the −10 promoter element has near perfect consensus at the critical −11-A position and a partial TGn motif characteristic of the extended −10 promoter element. Conversely, Group III has the most divergent −35 motif from consensus and no appreciable motif for the extended −10 promoter element.

ChIP-exo profiling of a canonical transcriptional activator: Crp.

Transcription activation in bacteria was further studied by conducting ChIP-exo studies on Crp, the most studied and best characterized transcription factor [6, 15]. These profiles revealed differences in the DNA protection patterns observed among the different classes of Crp activators. Three representative examples of ChIP-exo profiles generated on glycerol are shown for each of the three Crp Classes (Figure 5.2A). The *deoC* promoter is a Class III promoter with two Crp binding sites flanking a CytR regulatory site that represses the activating action of Crp [40]. The ChIP-exo protected regions are in close proximity with the three operator sequences with protected regions near −40 and −90 as previously seen *in vitro* [40]. However, markedly different profiles are observed in the Class I (*tnaC*) and Class II (*gatY*) promoters that often have no exonuclease protection to the Crp binding site, but instead, have strong protection of the region surrounding the TSS. In fact, these regions correspond greatly with the ChIP-exo profiles generated for σ^{70} under the same condition but no observed σ^{70} ChIP-exo peak was detected for the repressed *deoC* promoter.

Figure 5.2: **Crp promoter classes have unanticipated ChIP-exo footprint regions.** (A) Gene tracks are shown that exemplify the different Crp ChIP-exo footprint profiles observed for the three different classes of Crp promoters. Crp and CytR regulate the Class III promoter *deoC*. ChIP-exo reads are found at both Crp operators and for CytR which binds between them resulting in repression of Crp mediated activation. Subsequently, no σ^{70} ChIP-exo peak region was detected for *deoC*. However, under the activating Class I and Class II promoters there are few observed reads over the Crp operator indicating a lack of exonuclease protection over the Crp binding site. Instead, the ChIP peak is centered on the TSS and the footprint region co-occurs with that found for σ^{70}. Examples of this are shown for *tnaC* (Class I) and *adhE* (Class II). (B) Shown is the mean $5'$ tag density ChIP-exo profile aligned and oriented relative to the TSS generated for Crp grown on glycerol minimal media, an activating carbon source. The distribution of the center position across all Crp peak regions (blue bars) shows close proximity to the TSS. The template strand distribution (dashed black trace) corresponds with the downstream region centered at +20 that is associated with stable intermediates of the RP_O, the ITC, and the TEC as was observed for σ^{70}. The nontemplate strand distribution indicates protection of DNA predominantly occurs downstream of the −35 element with little protection at the predicted binding sites (gray bars). However, the nontemplate strand has higher density between the marker for −93.5 and −61.5, which indicate the typical motif center of Class III and Class I Crp promoters respectively. (C) An overlay of the mean $5'$ tag density profile of all Crp peak regions (blue traces) and the associated σ^{70} mean $5'$ tag density profile in those same peak regions (black traces) illustrates the strong co-occurrence of Crp footprint regions with those determined for σ^{70}. Reference positions for the −10 and −35 promoter elements as well as the center position of Crp operator for Class I (−61.5), Class II (−41.5) and Class III (−93.5) are shown.


A
Class III
Crp
CytR
Crp
TSS
deoC
-41
-93
Crp ChIP-exo
Class I
Crp Binding Site
Crp
TSS
tnaC
-61
Crp ChIP-exo
σ70 ChIP-ex
Class II
gatY
TSS
Crp
-41
Crp ChIP-exo
σ70 ChIP-exo
B
ChIP-exo Anti-Crp
nontemplate
template
Peak Center, n=108
Predicted Motif
-10/-35 Box
-41.5/-61.5/-93.5
Count
Mean 5' Tag Density
Distance from TSS, bp
C
-41.5/-61.5/-93.5
-10/-35 Box
ChIP-exo Anti-σ70
nontemplate
template
ChIP-exo Anti-Crp
nontemplate
template
Mean 5' Tag Density
Distance from TSS, bp

The results for these individual promoters are consistent with those observed at the genome-scale. Examination of the mean 5′ tag distribution of Crp ChIP-exo data oriented relative to the TSS illustrates that Class I and Class II activators have a peak center that aligns greatly with the TSS and not the Crp binding site (Figure 5.2B). To confirm that ChIP-exo was enriching Crp regulated promoters, the predicted Crp binding sites were oriented and aligned relative to the TSS (Figure 5.2B). This shows three regions of elevated Crp operator sequence centered at −41.5, −61.5, and −93.5 bp upstream of the TSS corresponding with the expected positions of Class II, Class I and Class III promoters respectively [6, 15]. Similar ChIP-exo profiles were obtained when wild type *E. coli* was grown on fructose, a Crp activating condition, but when grown on glucose, a repressing condition for Crp, few binding sites were detected (Supplementary Figure S2 in [26]). We further verified that these results where not artifacts attributed to the anti-Crp antibody used to perform ChIP-exo by generating data on a Δ*crp* strain and no correlation was observed between biological replicate datasets indicating minimal impact due to non-specific binding (Supplementary Figure S3 in [26]).

The ChIP-exo 5′ tag density profile for Crp was also compared with that generated for σ^{70} (Figure 5.2C). The ChIP-exo 5′ tag density for all identified Crp peak regions for cultures grown on glycerol were processed and oriented as described previously for σ^{70}. These density profiles were then compared to the σ^{70} density profiles determined across the same set of peak regions. Strand orientated Crp density profiles reveal a unimodal distribution on the template strand and a multimodal distribution on the nontemplate strand analogous to those found for σ^{70}. The template strand strongly overlaps the one observed for σ^{70} with a downstream boundary of protected DNA centered on +20 accounting for 33% of the aggregate density profile. However, the Crp nontemplate density profile has distinctive features. First, there is increased DNA protection on the nontemplate strand between the −93.5 and −61.5 markers. This region encompasses 13% of the total 5′ tag density profile. These positions signify the center position of many Class III and Class I Crp operator sequences respectively [6, 15]. Furthermore, Group I, Group II, and Group III modes have relative ratios of 2.5:3.3:1 compared with 1.1:3.5:1 for the same regions in the σ^{70} ChIP-exo 5′ tag

density profile. The Crp promoter bound regions appear to have a larger relative fraction of DNA protected as far upstream as the center of the −35 box. However, none of these regions indicates protection of the Crp operator sequences found for Class I and Class III promoters and only partial protection for Class II promoters due to the overlap with the −35 box.

Rifampicin (rif) prevents transcription elongation beyond a length of 2-3 nt [41] and, in doing so, leaves the transcription machinery unable to advance beyond the ITC. Therefore, ChIP-exo was performed on cultures treated with rif prior to harvest followed by immunoprecipitation of Crp. The resulting mean $5'$ tag density profile generated on both the template and nontemplate strand closely resembles that obtained in the non-rif treated sample (Supplementary Figure S4 in [26]). Therefore, this chemical perturbation of the transcriptional state had no impact on the Crp ChIP-exo distribution and no additional upstream protection of the Crp binding site was observed. This indicates that the exonuclease footprints are occurring on initiation complexes occurring prior to the TEC. This observation coupled with the evidence against the short-lived RP_C complex strongly suggests that the Crp promoters studied here are being captured at stable intermediates associated with RP_O formation or the ITC.

Distinct ChIP-exo profiles for transcriptional activators and repressors.

Previously, applications of ChIP-exo have demonstrated binding events centered on the binding motifs in eukaryotic systems [42, 5, 43]. We have applied the ChIP-exo protocol to characterizing the transcriptional repressor Fur in *E. coli* [44] and observed similar binding profiles. To further examine the distinctions between ChIP-exo profiles of transcriptional activators and repressors, ChIP-exo was performed on c-Myc tagged strains of ArcA (repressor) and Fnr (activator) grown anaerobically on glucose minimal media. The data generated was then processed, aligned, and oriented relative to the nearest TSS (Figure 5.3). ArcA which typically binds near the TSS [45] has no defined ChIP-exo $5'$ tag distribution on either strand though there is a noticeable increase in the $5'$ tag density around the TSS (Figure 5.3A). In contrast, Fnr demonstrates a similar $5'$ tag density profile as was seen for Crp and σ^{70}

with a strong unimodal distribution on the template strand at +20 and a less defined modal distribution on the nontemplate strand (Figure 5.3B). The ArcA ChIP peak regions were aligned relative to the peak center position (Figure 5.3C). This yields a uniform distribution of $5'$ tag density with sharp peaks on the forward (+) strand and the reverse strand (−). Furthermore, plotting the predicted binding sites shows that the protected regions are centered on the ArcA motif. Lastly, the peak-pair differences for ChIP-exo profiles of ArcA and Fnr are shown (Figure 5.3D). This reveals that the footprint obtained for the repressor is ≈30 bp while the activator has a much broader footprint distribution extending to ≈80 bp.

Genetic perturbation of RNAP holoenzyme/Crp interactions.

The activating properties of Crp and many transcription factors is through stabilizing interactions with RNAP holoenzyme at the promoter [46, 47]. Molecular characterization studies and mutational analysis of Crp has revealed three activating regions (Ar's) that make protein/protein interactions at specific positions in RNAP holoenzyme [6, 15]. The first, Ar1, interacts with either of the α-subunits at the C-terminus. The HL159 mutation to Crp prevents this interaction from forming [48, 49]. This region is involved with activation at Class I and Class II promoters [6, 15]. The second region, Ar2, is only associated with Class II promoters and binds to the N-terminus of the α-subunit. This bond was shown to be severely disrupted by introduction of two mutations to Crp, KE101 and HY19 [48]. Lastly, a weaker interaction was found to occur at Ar3 between Crp and the σ factor.

We next sought to determine the impact of genetic perturbations to the RNAP holoenzyme/Crp interactions at Ar1 and Ar2. HL159, KE101+HY19, and HL159+KE101 mutation were introduced to create an Ar1, Ar2, and Ar1+Ar2 deficient mutant in Crp respectively (Figure 5.4A). From this point forward the mutants will be referred to as delAr1, delAr2, and delAr1delAr2. ChIP-exo was performed on these mutant strains with glycerol as the sole carbon source. In comparison with the wild type, each mutant resulted in the loss of peak regions (Figure 5.4B). The most drastic effect was observed in the delAr1delAr2 mutant which retained ≈40% of the peaks in the wild type strain. This result indicates the importance of these Ar regions on the

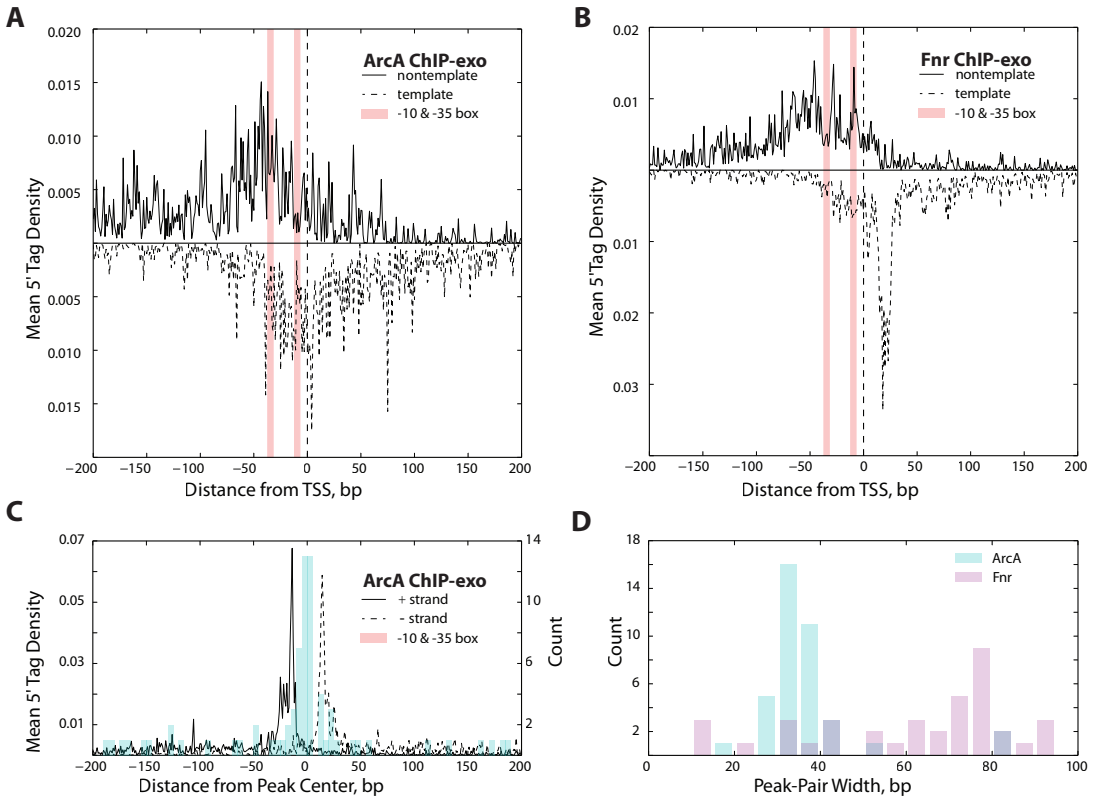
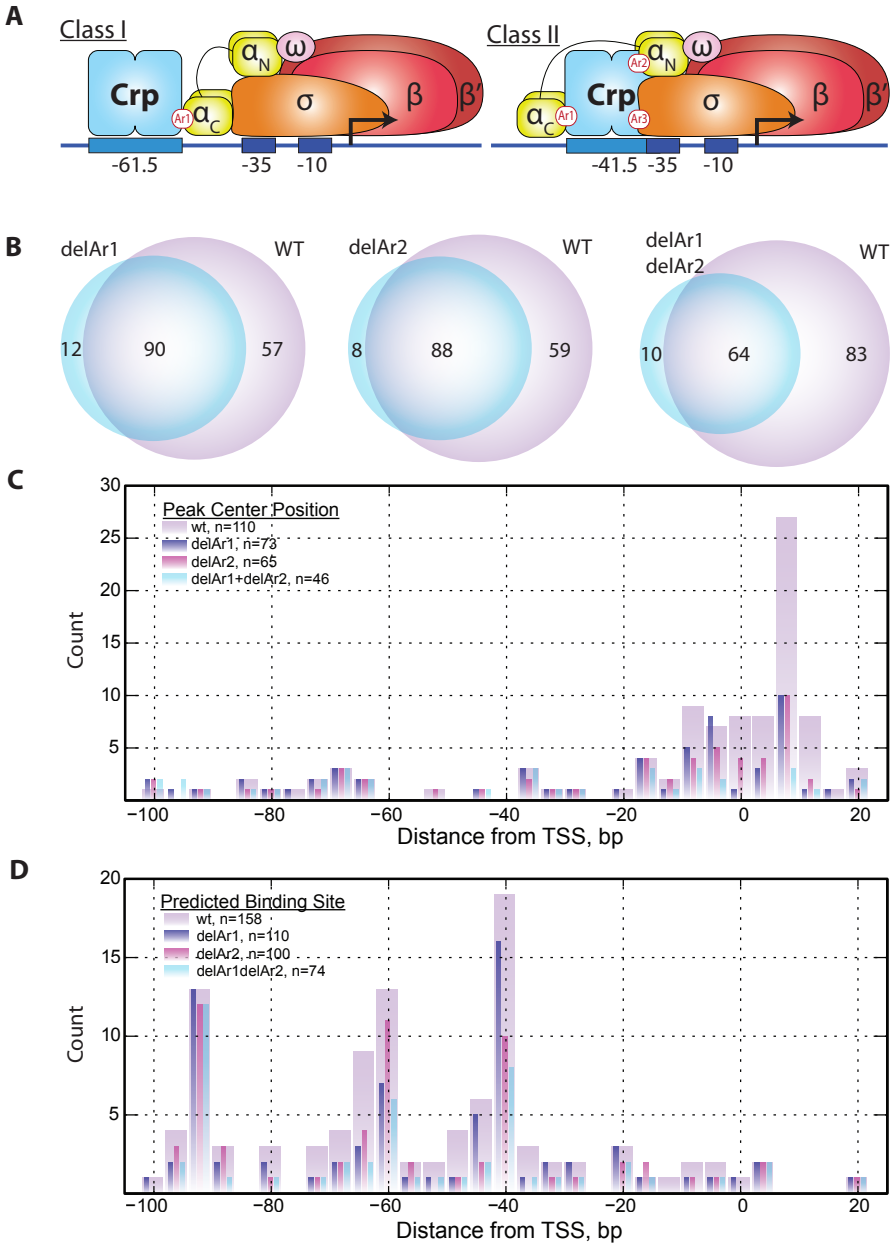


Figure 5.3: **Contrasting ChIP-exo profiles of repressors and activators.** (A) The TSS aligned ChIP-exo profile for ArcA, a predominantly repressive transcription factor, is shown to lack the characteristic distribution of mean 5′ tag density observed on both the template and nontemplate strand. (B) The TSS aligned mean 5′ tag density profile for Fnr, typically an activator, resembles the profile found for Crp and σ^{70}. (C) The ArcA ChIP-exo profile is shown for all peak regions aligned to the peak center position. Also shown is a histogram of the center of the predicted ArcA binding site relative to the peak center position. This illustrates that the ChIP-exo profile is centered on the predicted binding site. (D) A comparison of the peak-pair distance is shown to illustrate the difference in resolution observed between ArcA and Fnr. ArcA, the repressor, is revealed to have shorter footprints compared with Fnr, the activator.

stabilization of both Crp and RNAP holoenzyme at the promoter site. Furthermore, the characteristic ChIP-exo 5′ tag density profiles (see Figure 5.2C) on both strands were systematically degraded with each mutation resulting in profiles that no longer aligned well to the TSS (Supplementary Figure S5 in [26]). To determine which peak regions were lost as a result of these genetic perturbations, the distribution of peak region centers was analyzed (Figure 5.4C). The mutations predominantly result in a

loss of peak-regions where the peak center was located near the TSS (−10 to +20 bp) and peak centers farther away from the TSS were less impacted. Lastly, the distribution of predicted binding sites were examined in the context of the different mutant strains (Figure 5.4D). In agreement with expectation, modulation of Ar1 results in a drop in the predicted binding sites observed near −61.5, the typical Class I promoter distance from the TSS. This drop near −61.5 was partially recovered in the Ar2 mutant but a severe drop in the −41.5 centered binding sites occurred. This distance upstream of the TSS is associated with Class II promoters. The delAr1delAr2 mutant has a loss in peak regions with Crp binding sites matching those of Class I and Class II promoters. However, the peak regions of Class III found near the −93.5 position are unaffected by mutations in Ar1, Ar2, or both.

Figure 5.4: **The effect of genetic perturbation on Crp/RNAP interactions.** (A) Cartoon illustrating the interactions between activating regions (Ars) and RNAP for Class I and Class II activators. Class I activation has a single contact between Crp and RNAP (Ar1) whereas Class II activators can make upwards of three contacts (Ar1, Ar2, and Ar3). Deletion strains were constructed that modify specific amino acids to reduce functionality of Ar1, Ar2, or Ar1 and Ar2. Ar3 was not modified as it is thought to form weaker interactions than Ar1 and Ar2. (B) Venn diagram showing the pairwise comparison of peaks regions detected for delAr1, delAr2, and delAr1delAr2 with wild type Crp. All cultures were grown with glycerol as the carbon source. The mutations to Crp result in fewer detected peaks relative to wild type Crp with delAr1delAr2 retaining ≈40% of the wild type peak regions. (C) Histogram showing the peak center position relative to the TSS for wild type Crp, delAr1, delAr2, and delAr1delAr2 mutants. This plot illustrates that for the peak centers nearest the TSS (-15 to $+20$) are predominantly affected by mutations to Ar1 and Ar2, whereas peak regions centered upstream of the TSS (< -15) are largely unaffected. (D) An alternative view to the histogram shown in (C) that shows the distribution of predicted Crp binding sites relative to the TSS for wild type Crp, delAr1, delAr2, and delAr1delAr2 mutants. The peak regions detected in the delAr1 strain shows a reduced number of -61.5 motifs compared with wild type and delAr2. Conversely, delAr1 shows only a modest loss in -41.5 centered binding sites. This suggests that Class I promoters, dependent solely on Ar1 interactions, are most impacted by deletion of the Ar1 interactions with RNAP. Similarly, the delAr2 strain shows a substantial loss of Class II associated peak regions (-41.5 binding sites) compared with Class I (-61.5). The delAr1delAr2 mutant shows reductions in both -41.5 and -61.5 binding sites compared with the wild type. None of the mutant strains showed a reduction in the peak regions with Class III binding sites (e.g., -93.5 binding sites).


A
Class I
Crp
αN
ω
αC
Ar1
σ
β
β′
-61.5
-35
-10
Class II
Ar2
Ar3
-41.5
B
delAr1
WT
12
90
57
delAr2
8
88
59
delAr1
delAr2
10
64
83
C
Peak Center Position
wt, n=110
delAr1, n=73
delAr2, n=65
delAr1+delAr2, n=46
Count
Distance from TSS, bp
D
Predicted Binding Site
wt, n=158
delAr1, n=110
delAr2, n=100
delAr1delAr2, n=74
Count
Distance from TSS, bp

5.3.2 ChIP-exonuclease coupled with gene expression delineates the full Crp regulon

In addition to the ChIP-exo assays, we also performed paired RNA sequencing for the wild type and a Δ*crp* strain under batch glucose, fructose, and glycerol conditions. We further performed RNA sequencing for the delAr1, delAr2, and delAr1delAr2 on glycerol conditions. An overview and compilation of this data is shown in Figure 5.5. We show a 91% (21/23) overlap with experimentally validated Crp binding sites [11] and a 79% (23/29) overlap with previous ChIP-chip measurements that occurred in characterized Crp binding sites [10]. We further see a 65% overlap (317/486) with all reported Crp targets in RegulonDB [50] and a conservatively estimated 75% overlap (317/421) with binding sites that are found under similar environmental conditions. As shown in Figure 5.5, many of the genes that Crp regulates are carbon catabolic operons that are each specific to different utilized carbon sources. Since many of these specific operons are only activated under the presence of the exact carbohydrate that they degrade, they cannot be expected to be active or regulated by Crp under the experimental conditions used here. We thus conservatively removed only 65 of the 486 total Crp targets in RegulonDB to determine that 317/(486-65) ≈ 75% of comparable targets in RegulonDB are found in this study. The organization of Figure 5.5 is assembled based upon the two principal dimensions of the metabolic network functions that Crp regulates [51, 45, 52], that is, production of biomass (growth) and the production of energy (chemiosmosis). This organization is nearly identical to the classic and familiar categories of catabolism and anabolism. Each of the three subcategories of catabolism, anabolism, and chemiosmosis are then broken down further into the specific functions of which they are composed. Catabolism thus contains the transport genes to bring in catabolites, the recycling or secondary metabolic enzymes to break them down, and finally the central catabolic machinery to complete the degradation process. Anabolism contains both macromolecular synthetic reactions composed of transcription and translation along with biosynthetic reactions composed of specific pathways for carbohydrates, amino acids, and lipids. Chemiosmosis contains any gene responsible for maintaining the electrochemical gradient and subsequent energy supply of the cell. Chemiosmosis is

thus composed of the electron transport chain (ETC), fermentation machinery, and any ion pumps.

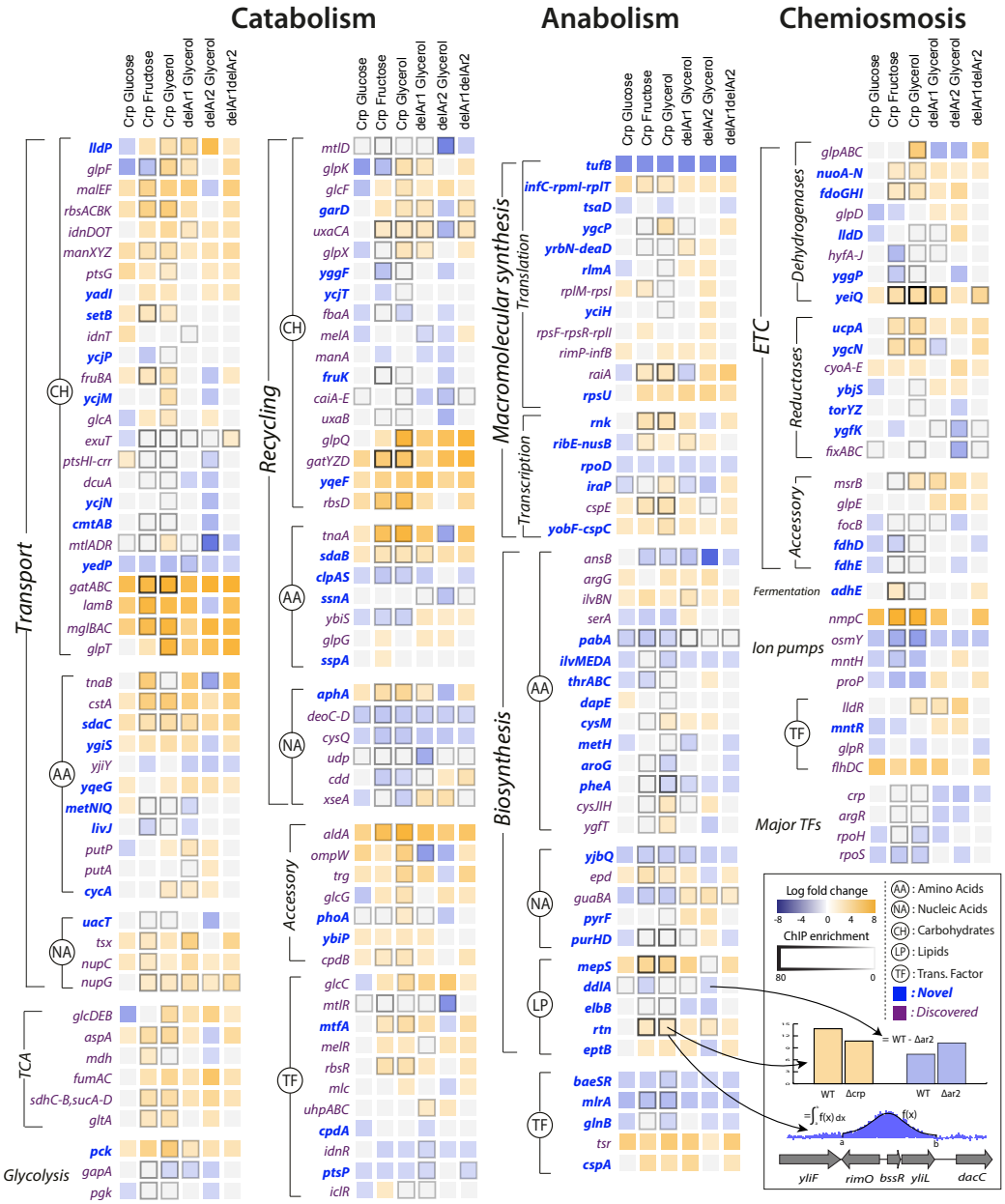
The most well known and considered regulatory targets of Crp lie in the catabolic sector of metabolism and the target list contains a large number of transport genes and recycling or secondary metabolic enzymes. We found a total of 97 regulated catabolic genes of which 30 represented novel findings. While we primarily confirm Crp regulation of transport genes, we also discover 15 novel transport gene regulation events. One such discovery is *ptsP*, the nitrogen PTS system, which exists as a completely independent PTS [53] to the main *ptsGHI* that is also shown to be regulated by Crp. Interestingly, *ptsGHI* is seen to be up regulated whereas *ptsP* is repressed, identifying an alternative pathway tradeoff enforced by Crp regulation. Much of the rest of the Crp regulation fills gaps in our knowledge for carbohydrate transporters but also highlights Crp regulation of amino acid transporters, in particular *metNIJ* for methionine, *sdaC* for serine, and *livB* for branched chain amino acids. We also confirm Crp regulation of recycling and secondary metabolic enzymes but again add 10 novel regulatory targets. In particular, regulation of the *clpAS* protease, *sdaC* serine degradation enzyme, and *fruK* fructose kinase broaden the scope of Crp regulation.

A less thoroughly studied but still crucially important swath of Crp regulatory targets lie in the anabolic sector of metabolism. Here we found 47 regulatory targets of which 33 represent novel discoveries. While many ribosomal genes were already known to be regulated by Crp, we discover 7 additional translation associated regulation targets. Of particular interest is *tufB* which carries aminoacylated tRNAs to the ribosome and is one of the most highly expressed [54] proteins in the cell. We also discover extensive regulation of amino acid, nucleotide, and lipid biosynthetic pathways. Activation of arginine and leucine production is complemented by activation of purine and pyrimidine production. One important novel finding is the repression of *glnB* that plays a crucial role in nitrogen metabolism by regulating the activity of glutamine synthetase. Crp repression of *glnB* is consistent with its role in regulating both catabolic production and anabolic demand. Additionally, Crp is known to regulate *rpoS*, *rpoH*, *argR*, and itself. However, we also show here regulation of *rpoD*

expression underlining the role of Crp as a globally decisive transcription factor.

Additionally, many Crp regulatory targets lie in the chemiosmotic sector of metabolism including key components of the ETC and fermentation machinery. Here we see 29 regulatory targets of which 14 represent novel findings. Two crucial novel regulatory targets are the primary NADH dehydrogenase, *nuoA-N*, along with one of the primary fermentation enzymes, *adhE*, which acts as an alcohol dehydrogenase. Activation of *nuoA-N* again underlines how Crp acts as a global regulator. Necessarily regulating the high flux backbone throughout catabolism, anabolism, and the energy generation pathways. Similarly, regulation of *adhE* can enable Crp to control one of the other major routes for generation of membrane potential, especially as Crp responds to the flow of carbon through central metabolism.

Figure 5.5: **The full Crp regulon defined through paired RNA-seq and ChIP-exonuclease data.** Specific regulation of each gene in the Crp regulon is shown as individual boxes across glucose, fructose, and glycerol environmental conditions with the addition of the Ar1, Ar2, and Ar1Ar2 genetic perturbations across glycerol conditions. Differential expression between a wild type strain and a full Δcrp strain is shown in the first three columns for glucose, fructose, and glycerol. The last three columns show differential expression between a wild type strain and a strain harboring either the Ar1, Ar2, or Ar1Ar2 mutations under glycerol conditions. ChIP-exonuclease peak density is indicated by border intensity for each gene in each column. Groupings of genes are performed similar to [45] and correspond to the catabolic, anabolic, and chemiosmotic sectors of metabolism. Novel discoveries are indicated and highlight discoveries in translation associated and amino acid biosynthetic gene products. Similarly, broad activation of catabolic genes and mixed regulation of anabolic genes is consistent with previously published models. Activation of chemiosmotic genes represents a novel addition to Crp regulatory knowledge.


Catabolism
Anabolism
Chemiosmosis
Crp Glucose
Crp Fructose
Crp Glycerol
delAr1 Glycerol
delAr2 Glycerol
delAr1delAr2
Transport
TCA
Glycolysis
Recycling
Accessory
Macromolecular synthesis
Translation
Transcription
Biosynthesis
ETC
Dehydrogenases
Reductases
Accessory
Fermentation
Ion pumps
Major TFs
Log fold change
ChIP enrichment
: Amino Acids
: Nucleic Acids
: Carbohydrates
: Lipids
: Trans. Factor
: Novel
: Discovered
= WT - Δar2
WT
Δcrp
Δar2
yliF
rimO
bssR
yliL
dacC

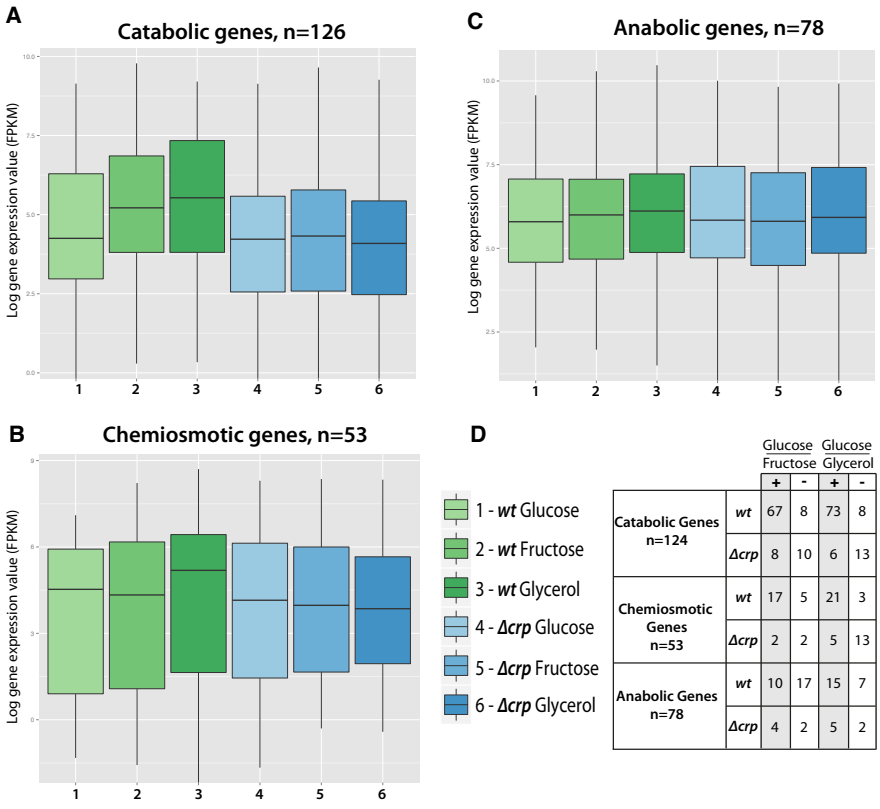
It is also important to note that another 60 regulatory events correspond to genes of unknown function that were not included in Figure 5.5. This brings the total of Crp active regulatory targets to 233 across the glycerol, fructose, and glucose conditions studied here. It also highlights that ≈25% of all active Crp regulatory targets occur on genes of completely unknown function, which concurs with previous reports [8].

Genome-scale analysis of the Crp regulon confirms physiological models

In order to understand systems level principles of Crp regulation we sought to characterize the Crp regulon in terms of recent physiological models of Crp regulation [7] and qualitative models of global regulation in microbes [45]. You et al. [7] recently elucidated the physiological function of Crp at the systems scale. In particular, this development shows how Crp senses carbon flow through central metabolism, and correspondingly up or down regulates the catabolic input or anabolic demand. A strong linear increase was observed between decreasing growth rates due to poorer carbon sources and the expression of catabolic genes. Similarly, a strong linear decrease was also observed among anabolic genes, presumably balancing proteome constraints among the catabolic and anabolic sectors. Here we sought to determine if the broad patterns described by the physiological model would be consistent at the genome-scale.

We first grouped genes based on their classification from Figure 5.5 and then plotted relative expression levels across each carbon source in Figure 5.6. The median expression levels and quartile distributions markedly shift from glucose to fructose to glycerol conditions for catabolic and chemiosmotic genes. The same plots for strains lacking Crp show unaffected expression levels across the same three substrates. This strong positive relation is also observed on historical microarray data across glucose, fructose, and acetate conditions (Supplementary Figure S6 in [26]). Further, the total number of genes which are differentially expressed at an FDR p-value $< .05$ are shown in Figure 5.6D. A total of 73 catabolic genes are significantly activated and only 8 are significantly repressed between glucose and glycerol conditions. Similarly 68 catabolic genes are significantly activated and only 8 are repressed between glucose and fructose

conditions.



		Glucose Fructose +	Glucose Fructose -	Glucose Glycerol +	Glucose Glycerol -
Catabolic Genes n=124	*wt*	67	8	73	8
	Δcrp	8	10	6	13
Chemiosmotic Genes n=53	*wt*	17	5	21	3
	Δcrp	2	2	5	13
Anabolic Genes n=78	*wt*	10	17	15	7
	Δcrp	4	2	5	2

Figure 5.6: **Distribution of expression levels for catabolic, anabolic, and chemiosmotic genes reflects physiological models of Crp regulation.** Boxplots display the quartiles and median distributions of all genes in each of the categories defined in Figure 5.5. (A) Overall expression levels are seen to increase from glucose to fructose to glycerol conditions in wild type strains. The same transition from glucose to fructose to glycerol conditions in Δcrp strains results in constant expression levels. (B) A similar trend as observed for catabolic genes is also observed for chemiosmotics genes. This trend is driven in part by novel discovery of the regulation of the NADH dehydrogenase, *nuoA-N* that exists as a high flux backbone for the *E. coli* electron transport chain. (C) Anabolic genes do not exhibit the same increase in expression across conditions. This is in line with previous observations and physiological models. (D) The number of differentially expressed genes between glucose and fructose or glucose and glycerol are shown for wild type and Δcrp strains. In general, the observation that a vast majority (e.g., 67/75 for catabolism glucose/fructose) of the differentially expressed genes at an FDR p-value $< .05$ between glucose and fructose are being activated indicates that the differences in the boxplot medians are significant. Similarly, this fraction drops off, e.g., 10/27 are activated for anabolic genes between glucose and fructose conditions.

This analysis provides strong confirmation for the physiological model of You et al. [7] by recreating their observed C-line at the genome-scale and extends its scope to include energy generation pathways. One key set of genes driving the chemiosmotic trend is the NADH dehydrogenase *nuoA-N*. This activation of the NADH dehydrogenase is consistent with a strategy in which Crp up regulates both catabolism and energy generation on poor carbon sources to maintain normal cell function at the expense of a higher growth rate via anabolic activity. Further, given its role as the primary dehydrogenase and a high flux backbone for energy generation, this discovery highlights the value in unbiased genome-scale assays for elucidating comprehensive regulons at the genome-scale.

Anabolic genes, in line with the same physiological model, do not increase in expression levels in response to poorer carbon sources. In the physiological model, a negative linear relation is observed for the anabolic gene sector. While we do not see a negative linear relation, we do see a clear flat lining and lack of clear response in the Δcrp strain. Some anabolic genes likely do respond along the C-line, but many do not. Part of the reason for this may lie in compensation by other transcription factors or cellular machinery at key promoters. Notably, *rplM*, *rlmA*, *yrbN*, and *yciH* are all important ribosome related genes that display very large ChIP-binding peaks, and differential expression under a delAr2 strain. However, these genes do not exhibit differential expression under the full Δcrp strain. In cases where full *crp* deletion does not affect expression levels, but the Ar2 genetic perturbation clearly does, we can conclude that some form of compensation must have occurred at those promoters in the Δcrp strain. In addition, biosynthetic genes across the amino acid, nucleotide, and lipid categories which include *thrABC*, *metH*, *pyrF*, *purHD*, *ddlA*, and *elbB* all indicate a form of compensation and help explain the flat lining of overall expression for anabolic genes regardless of Crp activity.

Genetic perturbations provide link between promoter mechanisms and systems level regulatory features.

Mutations in the functional units of Crp at the Ar1 and Ar2 regions have clear effects on Crp regulation at both the individual promoter level, and overall systems

level. We showed in Figure 5.3 how open promoter complexes are destabilized at individual promoters resulting in fewer binding peaks when the functional Ar1 and Ar2 contacts are perturbed. Interestingly, these same Ar1 and Ar2 genetic perturbations also exhibit systems level effects via the overall lessening of Crps ability to regulate the catabolic, anabolic, and chemiosmotic sectors of metabolism Figure 5.7. This clear linear decrease across the catabolic and chemisomotic sectors again highlights and solidifies Crps role in activating catabolic genes along the C-line [7]. It also provides a link between the promoter level mechanisms and the systems level regulatory features.

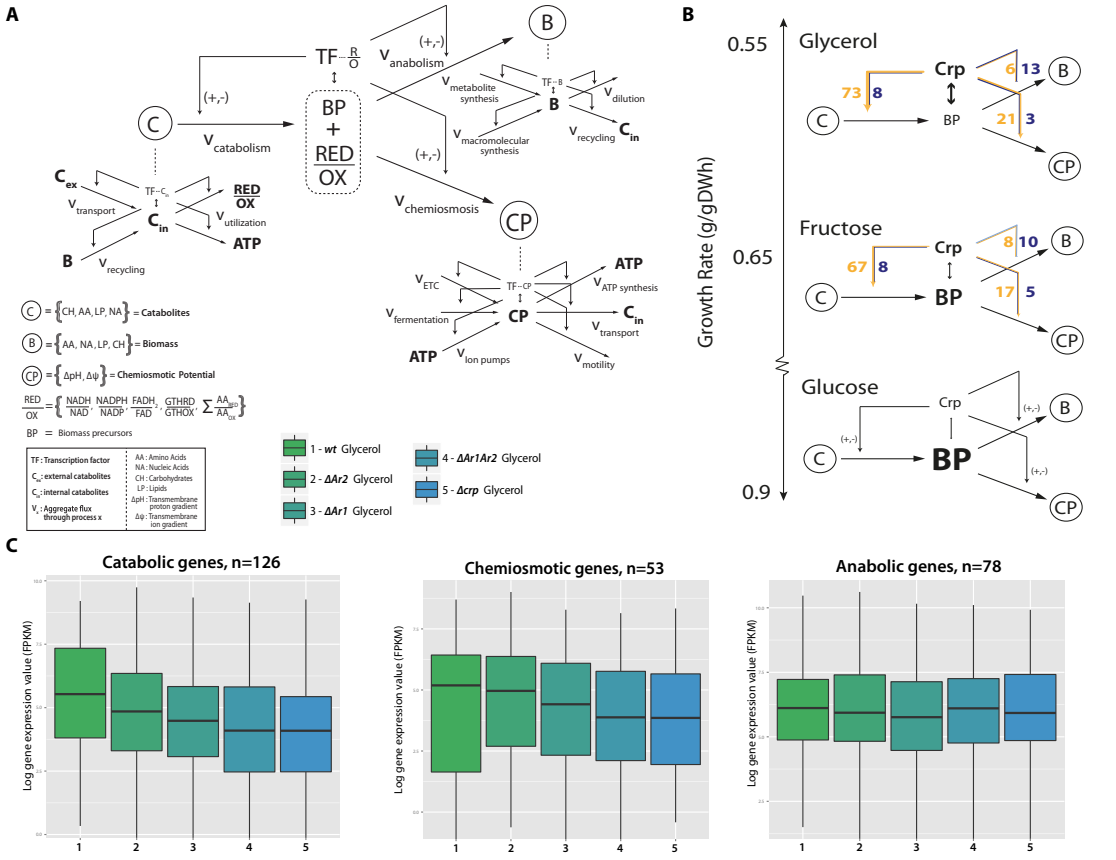


Figure 5.7: **Systems level circuitry of Crp regulation is in line with genetic perturbations.** (A) Systems level model shows the general relation between the anabolic, catabolic, and chemiosmotic metabolic-regulatory networks. Crp is able to sense multiple key biomass precursors (BP) in the form of alpha-keto acids flowing through central carbon metabolism along with cAMP and thereby regulate input catabolic fluxes and consumption anabolic and chemiosmotic fluxes. (B) Throughout the shift from glucose to fructose to glycerol conditions, carbon flux is diminished along with Crp regulation. Crp generally de-activates the catabolic and anabolic sectors in the transition from glucose to glycerol and does not heavily affect the overall expression of anabolic genes. (C) Distributions of expression levels ranging across wild type, delAr1, delAr2, delAr1delAr2, and a full Δcrp strain under glycerol conditions. As can be seen, a clear linear decrease from the wild type strain with an intact Crp regulatory protein and the full Δcrp provides additional external validation for the existence of the C-line.

5.4 Discussion

A fundamental goal of systems biology is to obtain a comprehensive, predictive understanding of promoters that regulate transcriptional output. This work illustrates the potential of ChIP-exo to help bridge the gap between the mechanistic processes of transcription initiation and the systems-level analysis of regulons. ChIP-exo datasets were generated for σ^{70} which revealed characteristic properties of transcriptional activation that were subsequently extended to transcription factor mediated activation. The ChIP-exo profiles of Crp and Fnr, commonly activating transcription factors, are found to be distinct compared with those found for microbial repressors such as ArcA (this study) and Fur [44]. Activating transcription factors are found to have ChIP-exo profiles that resemble those observed at σ^{70} promoters. The ChIP-exo profile for Crp provided little protection to the Crp operator sequence but rather protected DNA located +20 downstream of the TSS. This contrasts repressors which have ChIP-exo profiles centered on the transcription factor binding sites. Furthermore, genetic perturbations of the protein-protein interactions between RNAP holoenzyme and Crp where found to result in a systematic loss of Crp peak regions indicating the importance of these Ar regions to stabilizing Crp at these promoter regions. The loss of Crp peak regions was found be a result of a loss in peak-regions centered on the TSS. At the systems level, we present a significant expansion of the Crp regulon in addition to strongly confirming the multitude of previously known regulatory targets. We are also able to provide strong confirmation of recent physiological models at the genome-scale and extend the systems level understanding of Crp to C-line regulation of chemiosmotic gene products. Finally, we are able to show that the same Crp-RNAP holoenzyme stabilizing Ar interactions have a clear systems level effect by linearly diminishing C-line regulation across each of the catabolic, anabolic, and chemiosmotic categories.

Collectively, the ChIP-exo distribution of the mean $5'$ tag density on both strands indicates that σ^{70} is being capturing in complex with RNAP when assessed in the context of *in vitro* footprinting studies performed on model promoter constructs. Furthermore, both the template and nontemplate strands provide evidence that σ^{70} ChIP studies identify stable intermediates during transcription initiation

that occur after the recruitment of RNAP and RP_C formation. The template strand protected regions in the σ^{70} ChIP-exo profiles provide the strongest evidence that post-recruitment stable intermediates are being captured. The template strand distribution protection extends to +20 relative to the TSS, which agrees with numerous *in vitro* studies that have demonstrated a clear transition in the downstream protected boundary from −5 to +5 in RP_C to +20 for RP_O, ITC, and early TEC complexes [30, 28, 29, 31]. Hydroxyl radical footprinting studies on RP_C formation in the T7A1 promoter showed that the short-lived RP_C complex protects DNA to approximately −5 bp [33, 55, 34]. Similar results were observed in DNase footprinting of the T7A3 promoter, *lac*UV5, and *rrnB*P1 [56, 57]. Furthermore, an RNAP mutant with deficient open complex formation was found to have DNase footprints that extend to just +1 at the λP_R promoter [58]. However, the RP_C complex was only observed when the temperature was dropped in most of these studies. The temperature dependent capture of early closed complexes has been shown to be a result greater RP_O abundance at physiological temperatures [32, 57, 55, 34]. Conversely, the advanced downstream boundaries centered on +20 has been observed in studies performed on the intermediates leading to RP_O and the RP_O complex for the T7A1 promoter [33, 55, 34], the T7A3 promoter [57], the *rrnB*P1 promoter [56, 59, 60, 61], the λP_R promoter [62, 32], and the *lac*UV5 promoter [57, 63]. Furthermore, the ITC and the transition to the TEC also have a downstream footprint boundary of +20-25. DNase footprinting of T7A1, *tac*, and *lac*UV5 promoters showed that the ITC has a slightly advanced footprint at +25 compared with +20 for RP_O and the early TEC had a footprint at +30 [64, 39, 36].

ChIP-exo mean 5′ tag density profiles for σ^{70} on the nontemplate strand show a multimodal distribution with regions that provide protection to different components of σ^{70} promoter elements. These were found to form three modes spanning −34 to −23, −18 to −1, and +4 to +12 with the region spanning −18 to −1 accounting for the largest fraction of the 5′ tag density profile. While periodic patterns of DNA protected regions at the upstream boundary are common [30, 28, 31], the location of the boundaries is supportive of post-recruitment intermediates [35, 36, 37]. A detailed study on the *lac*UV5 promoter using DNase I, methylation protection, and

exonuclease III protection across transcription intermediates showed that transitions from RP_O to ITC undergoing abortive initiation retained strong protection of a region between -24 and -6 to exonuclease III digestion that was reduced to protection of the region downstream of -6 after escaping the abortive transcription phase to produce longer transcripts [36]. This is further corroborated by a recent study that showed the *lac*UV5 promoter has an upstream footprint boundary at -23 in the presence of σ^{70} compared with $-13/-14$ for the σ^{70} lacking transcribing complex [37]. A study of the T7A1 promoter using exonuclease III showed a drastic movement in the upstream-protected region from -43 to -3 in the transition from RP_O to early transcribing complexes (ITC or TEC) [35]. Furthermore, the width of the protected region agrees with studies examining RP_O, ITC, and TEC. Early TECs have been found to have footprint regions spanning $\approx$30 bp whereas RP_O and ITC have longer footprint seen to be 50+ bp in length [38, 39, 36].

Kinetic studies also support the notion that the ChIP-exo data presented here is reflective of stable intermediates occurring post-recruitment of RNAP and RP_C formation. Genome-scale characterization studies of bacterial transcription have shown the rate-limiting step in transcription predominantly occurs post-recruitment of RNAP [24, 65, 66]. For example, the λP_R promoter is limited at the opening of the transcription bubble marked by a slow transition from the closed intermediate to the open intermediate [62, 32, 67, 68, 69]. Furthermore, the promoter λP_R encodes a promoter-proximal pause site induced by a -10 like sequence downstream of the TSS [70]. A similar pause occurs in the *lac* promoter [71, 72] and, in fact, it is estimated that the occurrence of promoter-proximal pausing is upwards of 20% in *E. coli* [73, 74, 72]. However, numerous additional processes along the trajectory of RP_O to TEC formation have been found to be rate determining including scrunching [75], and promoter escape [29, 65]. Therefore, ascertaining the potential bottlenecks in transcription initiation at genome-scale and under *in vivo* conditions would be of value to genome-scale models of promoter kinetics.

The observations made for σ^{70} footprint profiles extend to those generated for transcriptional activators but not to repressors. The activators Crp and Fnr have profiles that closely resembled that of σ^{70}, in particular the downstream exonuclease

protected boundary. In most cases, Crp and Fnr were found to have protection on the template strand centered at +20 relative to the TSS. Furthermore, upstream protected regions for the activating Class I and Class II Crp operator sites are largely devoid of protection while suppressed Class III Crp operator sites, such as *deoC*, are protected. This indicates that, like σ^{70}, the ChIP-exo profile is reflective of RNAP holoenzyme captured at a stable intermediate that occurs post-recruitment. Therefore, these results potentially indicate that Crp interactions with the operator site are short-lived compared with the interactions formed with RNAP holoenzyme. Therefore, it is hypothesized that Crp dissociates from DNA after recruitment but remains bound to RNAP holoenzyme. The formation of a stressed intermediate through scrunching is though to provide the energy needed for promoter escape [76, 77]. It is plausible that scrunching provides sufficient energy to break the bonds formed between the transcription factor and DNA as well. However, confirmation of this hypothesis would require detailed molecular and structural studies that are beyond the scope of this work. The exact mechanisms associated with σ factor release have proven to be elusive [78, 79, 29, 80] and the release of transcriptional activators could be equally elusive.

At the level of regulons and specific targets of Crp regulation we provide the first comprehensive dataset to assess binding sites, regulatory outcomes, and compensation by other factors. This resulted in significant discovery of novel regulation events (77 total) along with extensive confirmation of previously discovered binding sites (96 total). Much of the novel regulation is discovered in the anabolic sector of catabolism and corresponds to translation associated or amino acid biosynthetic gene products. This regulation is clearly in line with known Crp regulatory targets [81, 82] but greatly expands on this area of regulation. Similarly regulation of chemiosmotic genes confirms targets like the *hyfA-G* [83] hydrogenase while discovering regulation of *adhE*, *fdoGHI*, and *nuoA-N*. Knowledge that Crp regulates these key fermentation routes and alternative dehydrogenases could prove crucial for many metabolic engineering strategies.

At the systems level we are able to confirm recent physiological studies that unraveled the systems level function of Crp. The model of You et al. [7] showed

a feedback mechanism from central metabolism to the catabolic and anabolic genes that Crp regulates. We first confirm this model at the genome-scale by showing that the distribution of expression values for catabolic, anabolic, and chemiosmotic genes in the Crp regulon obey the same trend as the C-line. We then extend this model by showing clear and significant regulation of the energy generating high flux backbone of the cell in the form of the NADH dehydrogenase. Finally we show that the C-line can also be reproduced via analysis of the Ar1 and Ar2 regions. This data shows that the same mutations that destabilize complexes at the level of an individual promoter are responsible for carrying out systems level regulatory features in a consistent and coherent manner.

5.5 Materials and Methods

5.5.1 Strains and Culturing Conditions

Escherichia coli K-12 MG1655 cells and derivatives thereof were used for all experiments performed in this study. Crp-8-Myc, Fnr-8-Myc, and ArcA-8-Myc tagged strains were those previously constructed and described in [84]. The Δcrp strain was generated using the method described [85]. Briefly, the *crp* gene was deleted from start codon to stop codon using the λ red, FLP-mediated site-specific recombination method and replaced with a gene conferring kanamycin resistance. The Δcrp was transformed with pKD46 and used as a basis for constructing the delAr1, delAr2 and delAr1delAr2 mutant strains using the λ red, FLP-mediated site-specific recombination method. Plasmids carrying the different Ar mutant sequences were *de novo* synthesized using GeneArt (Life Technologies) with restriction sites at the $5'$ and $3'$ end of the gene. The gene was digested and ligated into the pKD3 plasmid directly upstream of the chloramphenicol resistance gene. Linear PCR constructs carrying the Ar mutant *crp* gene, an FRT site, the chloramphenicol resistance gene, the second FRT site where PCR amplified using primers with overhangs with homology directly upstream of the start codon and downstream of the stop codon. This construct was transformed into electrocompetent Δcrp *E. coli* K-12 carrying the pKD46 plasmid. The chloramphenicol resistance gene was then removed from strains with a confirmed

insertion of the mutated *crp* gene by transformation of pCP20 as previously described [85]. The delAr1 mutant introduces a mutation to the Ar1 region, HL159, previously determined to break the contacts between Ar1 and the α-subunit of RNAP [48, 49]. The delAr2 mutant does the same for Ar2 but introduces two mutations, KE101 and HY19 [48]. The delAr1delAr2 strain carries the HL159 mutation and the KE101 mutation.

M9 minimal media was used for all cultures with 2 g/L of glucose, fructose, or glycerol depending on the conditions tested. For σ^{70}, Crp, Crp8myc, Δcrp, delAr1, delAr2, and delAr1delAr2 experiments, cultures were grown aerobically in shake flasks. Rifampicin conditions were incubated in the presence of rifampicin (50 μg/mL final concentration) for 20 min prior to crosslinking as previously described [20]. Fnr and ArcA experiments were conducted similarly but grown under anaerobic conditions.

5.5.2 ChIP Experiments

The ChIP-exo protocol was adapted based on the method described by Rhee et al. [5] and adapted for the Illumina platforms with the following modifications. DNA crosslinking, fragmentation, and immunoprecipitation were performed as previously described [86]. Briefly, cells were cross-linked in early exponential phase in 1% formaldehyde for 30 min at room temperature. This was followed by a 5 min quenching of the crosslinking reaction by addition of glycine to a final concentration of 125 mM. Cells were then washed 3X in ice-cold TBS. Cells were lysed as previously described in the presence of protease inhibitors. Clarified lysate was then continuously sonicated at 4 °C using a sonicator bell (6W) for 30 min. Cells were then immunoprecipitated with an appropriate antibody. Antibodies used in this study (all mouse derived) are: anti-Crp (Neoclone N0004), anti-σ^{70} (Neoclone WP004), anti-Myc (Santa Cruz Biotechnology sc-40). Immunocomplexes were captured using Pan Mouse IgG Dynabeads (Life Technologies). At this point the procedures for ChIP-chip and ChIP-seq deviate the ChIP-exo protocol. ChIP-chip was carried out as detailed in [86] while ChIP-seq undergoes end-repair, dA tailing, adaptor ligation, and PCR enrichment as is done for ChIP-exo. For ChIP-exo, the following steps were

performed while the protein/DNA/antibody complexes where bound to the magnetic beads: end repair (NEB End Repair Module), dA tailing (NEB dA-Tailing Module), adaptor 2 ligation (NEB Quick Ligase), nick repair (NEB PreCR Repair Mix), lambda exonuclease treatment (NEB), and RecJ_f exonuclease treatment (NEB). A series of step-down washes were performed between all steps using buffers previously described [86]. Strand regeneration and library preparation followed the approach of Rhee et al. [5] with the exception of a $3'$ overhang removal step after the first adaptor ligation and prior to PCR enrichment by treating with T4 DNA Polymerase for 20 min at 12 °C. Libraries were sequenced on an Illumina MiSeq. Reads were aligned to the NC_000913.2 genome using bowtie2 [87] with default settings. Peak calling was performed using GPS in the GEMS analysis package [88] with the ChIP-exo default read distribution file with the following parameter settings: mrc 20, smooth 3, no read filtering, and no filter predicted events. Note that GPS was used over GEMS because GEMS peak boundaries are influenced by motif identification whereas GPS is not. ChIP-peak calls were manually curated for anti-Crp conditions on all substrates and rifampicin treated cultures. A superset of GPS peak calls across all anti-Crp conditions was analyzed for presence/absence in each individual condition. Data for ArcA and Fnr were also manually curated.

5.5.3 Gene Expression

Gene expression analysis was performed on RNA-seq data. Briefly, total RNA was isolated and purified using the Qiagen Rneasy Kit with on-column DNase treatment. The quality of the total RNA was assessed using an Agilent Bioanalyzer. Paired-end, strand specific RNA-seq libraries were constructed using the dUTP method [89]. Briefly, total RNA was first depleted of ribosomal RNAs using Epicentres RiboZero rRNA removal kit for gram-negative bacterial. rRNA depleted RNA was then primed using random hexamers and reverse transcribed using SuperScript III (Life Technologies). Second strand was synthesized using *E. coli* DNA Polymerase, Rnase H, and *E. coli* DNA Ligase and with dNTPs with dUs replacing dTs. Sequencing library construction followed with end-repair, dA tailing, adaptor ligation, removal of the second strand carrying dUs, and PCR enrichment. Sequencing was

performed on an Illumina MiSeq. Reads were mapped to the NC_000913.2 reference genome using the default settings in bowtie2 [87].

5.6 Author Contributions

HL and SF conceived and planned the work presented here. HL, JT, and RS performed all of the experiments. HL and RS adapted the ChIP-exo protocol for microbial applications. SF and HL processed and analyzed all of the data. JUC provided guidance for construction of mutant Crp strains. HL, SF, BOP, and KZ wrote and edited the manuscript.

5.7 Acknowledgements

The Novo Nordisk Foundation and the NIH U01 grant GM102098-01 provided financial support for this work. HL was supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086.

Chapter 5 has been submitted for publication of the material as it may appear in Latif H*, Federowicz S*, Szubin R, Tarasova J, Utrilla J, Ebrahim A, Zengler K, Palsson BØ. Integrated analysis of molecular and systems level function of Crp using ChIP-exo. Submitted to Cell. 2014. *indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Stephen Federowicz (equal contributor), Richard Szubin, Janna Tarasova, Jose Utrilla, Ali Ebrahim, Karsten Zengler, and Bernhard Ø. Palsson.

5.8 Bibliography

[1] Hegyi H, Gerstein M (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. Journal of Molecular Biology 288: 147–164.

[2] Fraser JS, Gross JD, Krogan NJ (2013) From systems to structure: bridging networks and mechanism. Molecular Cell 49: 222–231.

[3] Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876–880.

[4] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.

[5] Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147: 1408–1419.

[6] Busby S, Ebright RH (1999) Transcription activation by catabolite activator protein (CAP). Journal of Molecular Biology 293: 199–213.

[7] You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang YP, Lenz P, Yan D, Hwa T (2013) Coordination of bacterial proteome with metabolism by cyclic AMP signalling. Nature 500: 301–306.

[8] Gosset G, Zhang Z, Nayyar S, Cuevas WA, Saier MH Jr (2004) Transcriptome analysis of Crp-dependent catabolite control of gene expression in *Escherichia coli*. Journal of Bacteriology 186: 3516–3524.

[9] Kao KC, Tran LM, Liao JC (2005) A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. Journal of Biological Chemistry 280: 36079–36087.

[10] Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJW (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. Proceedings of the National Academy of Sciences of the United States of America 102: 17693–17698.

[11] Zheng D, Constantinidou C, Hobman JL, Minchin SD (2004) Identification of the CRP regulon using *in vitro* and *in vivo* transcriptional profiling. Nucleic Acids Research 32: 5874–5893.

[12] Shimada T, Fujita N, Yamamoto K, Ishihama A (2011) Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. PLoS One 6: e20081.

[13] Germer J, Becker G, Metzner M, Hengge-Aronis R (2001) Role of activator site position and a distal UP-element half-site for sigma factor selectivity at a CRP/H-NS-activated sigma(s)-dependent promoter in *Escherichia coli*. Molecular Microbiology 41: 705–716.

[14] Kristensen HH, Valentin-Hansen P, Søgaard-Andersen L (1997) Design of CytR regulated, cAMP-CRP dependent class II promoters in *Escherichia coli*: RNA polymerase-promoter interactions modulate the efficiency of CytR repression. Journal of Molecular Biology 266: 866–876.

[15] Lawson CL, Swigon D, Murakami KS, Darst SA, Berman HM, Ebright RH (2004) Catabolite activator protein: DNA binding and transcription activation. Current Opinion in Structural Biology 14: 10–20.

[16] Pedersen H, Dall J, Dandanell G, Valentin-Hansen P (1995) Gene-regulatory modules in *Escherichia coli*: nucleoprotein complexes formed by cAMP-CRP and CytR at the *nupG* promoter. Molecular Microbiology 17: 843–853.

[17] Williams RM, Rhodius VA, Bell AI, Kolb A, Busby SJ (1996) Orientation of functional activating regions in the *Escherichia coli* CRP protein during transcription activation at class II promoters. Nucleic Acids Research 24: 1112–1118.

[18] Cho BK, Kim D, Knight EM, Zengler K, Palsson BO (2014) Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states. BMC Biology 12: 4.

[19] Cho BK, Knight EM, Palsson BØ (2008) Genomewide identification of protein binding locations using chromatin immunoprecipitation coupled with microarray. Methods in Molecular Biology 439: 131–145.

[20] Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ (2009) The transcription unit architecture of the *Escherichia coli* genome. Nature Biotechnology 27: 1043–1049.

[21] Herring CD, Raffaelle M, Allen TE, Kanin EI, Landick R, Ansari AZ, Palsson BØ (2005) Immobilization of *Escherichia coli* RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. Journal of Bacteriology 187: 6166–6174.

[22] Kröger C, Dillon SC, Cameron ADS, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JCD (2012) The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proceedings of the National Academy of Sciences of the United States of America 109: E1277–E1286.

[23] Qiu Y, Cho BK, Park YS, Lovley D, Palsson BØ, Zengler K (2010) Structural and operational complexity of the *Geobacter sulfurreducens* genome. Genome Research 20: 1304–1311.

[24] Reppas NB, Wade JT, Church GM, Struhl K (2006) The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. Molecular Cell 24: 747–757.

[25] Wade JT, Struhl K, Busby SJW, Grainger DC (2007) Genomic analysis of protein-DNA interactions in bacteria: insights into transcription and chromosome organization. Molecular Microbiology 65: 21–26.

[26] Latif H, Federowicz S, Tarasova J, Szubin R, Utrilla J, Ebrahim A, Zengler K, Palsson BØ (2014) Integrated analysis of molecular and systems level function of Crp using ChIP-exo. Submitted to Cell .

[27] Mymryk JS, Archer TK (1994) Detection of transcription factor binding *in vivo* using lambda exonuclease. Nucleic Acids Research 22: 4344–4345.

[28] Hook-Barnard IG, Hinton DM (2007) Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. Gene Regulation and Systems Biology 1: 275–293.

[29] Hsu LM (2002) Promoter clearance and escape in prokaryotes. Biochimica et Biophysica Acta 1577: 191–207.

[30] Record M, Reznikoff WS, Craig ML, McQuade KL, Schlax PJ (1996) *Escherichia coli* RNA polymerase (eσ^{70}), promoters, and the kinetics of the steps of transcription initiation. Escherichia coli and Salmonella Cellular and Molecular Biology Edited by Neidhardt FC et al ASM Press, Washington DC : 792–821.

[31] Saecker RM, Record MT Jr, Dehaseth PL (2011) Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. Journal of Molecular Biology 412: 754–771.

[32] Davis CA, Bingman CA, Landick R, Record MT Jr, Saecker RM (2007) Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by *Escherichia coli* RNA polymerase. Proceedings of the National Academy of Sciences of the United States of America 104: 7833–7838.

[33] Rogozina A, Zaychikov E, Buckle M, Heumann H, Sclavi B (2009) DNA melting by RNA polymerase at the T7A1 promoter precedes the rate-limiting step at 37 degrees C and results in the accumulation of an off-pathway intermediate. Nucleic Acids Research 37: 5390–5404.

[34] Sclavi B, Zaychikov E, Rogozina A, Walther F, Buckle M, Heumann H (2005) Real-time characterization of intermediates in the pathway to open complex formation by *Escherichia coli* RNA polymerase at the T7A1 promoter. Proceedings of the National Academy of Sciences of the United States of America 102: 4706–4711.

[35] Metzger W, Schickor P, Heumann H (1989) A cinematographic view of *Escherichia coli* RNA polymerase translocation. EMBO Journal 8: 2745–2754.

[36] Straney DC, Crothers DM (1987) A stressed intermediate in the formation of stably initiated RNA chains at the *Escherichia coli lac*UV5 promoter. Journal of Molecular Biology 193: 267–278.

[37] Zhilina E, Esyunina D, Brodolin K, Kulbachinskiy A (2012) Structural transitions in the transcription elongation complexes of bacterial RNA polymerase during σ-dependent pausing. Nucleic Acids Research 40: 3078–3091.

[38] Carpousis AJ, Gralla JD (1985) Interaction of RNA polymerase with *lac*UV5 promoter DNA during mRNA initiation and elongation. Footprinting, methylation, and rifampicin-sensitivity changes accompanying transcription initiation. Journal of Molecular Biology 183: 165–177.

[39] Krummel B, Chamberlin MJ (1989) RNA chain initiation by *Escherichia coli* RNA polymerase. Structural transitions of the enzyme in early ternary complexes. Biochemistry 28: 7829–7842.

[40] Valentin-Hansen P (1982) Tandem CRP binding sites in the *deo* operon of *Escherichia coli* K-12. EMBO Journal 1: 1049–1054.

[41] Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, Darst SA (2001) Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. Cell 104: 901–912.

[42] Rhee HS, Pugh BF (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 483: 295–301.

[43] Serandour AA, Brown GD, Cohen JD, Carroll JS (2013) Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. Genome Biology 14: R147.

[44] Seo SW, Kim D, Latif H, O'Brien EJ, Szubin R, Palsson BO (2014) Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. Nature Communcations 5: 4910.

[45] Federowicz S, Kim D, Ebrahim A, Lerman J, Nagarajan H, Cho Bk, Zengler K, Palsson B (2014) Determining the control circuitry of redox metabolism at the genome-scale. PLoS Genetics 10: e1004264.

[46] Browning DF, Busby SJ (2004) The regulation of bacterial transcription initiation. Nature Reviews Microbiology 2: 57–65.

[47] Lee DJ, Minchin SD, Busby SJW (2012) Activating transcription in bacteria. Annual Review of Microbiology 66: 125–152.

[48] Rhodius VA, West DM, Webster CL, Busby SJ, Savery NJ (1997) Transcription activation at class II CRP-dependent promoters: the role of different activating regions. Nucleic Acids Research 25: 326–332.

[49] West D, Williams R, Rhodius V, Bell A, Sharma N, Zou C, Fujita N, Ishihama A, Busby S (1993) Interactions between the *Escherichia coli* cyclic AMP receptor protein and RNA polymerase at class II promoters. Molecular Microbiology 10: 789–797.

[50] Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Research 41: D203–D213.

[51] Carlson R, Srienc F (2004) Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: identification of reactions. Biotechnology and Bioengineering 85: 1–19.

[52] Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional optimality of microbial metabolism. Science 336: 601–604.

[53] Rabus R, Reizer J, Paulsen I, Saier M Jr (1999) Enzyme I(Ntr) from *Escherichia coli*. A novel enzyme of the phosphoenolpyruvate-dependent phosphotransferase system exhibiting strict specificity for its phosphoryl acceptor, NPr. Journal of Biological Chemistry 274: 26185–26191.

[54] Weijland A, Harmark K, Cool RH, Anborgh PH, Parmeggiani A (1992) Elongation factor Tu: a molecular switch in protein biosynthesis. Molecular Microbiology 6: 683–688.

[55] Schickor P, Metzger W, Werel W, Lederer H, Heumann H (1990) Topography of intermediates in transcription initiation of *E.coli*. EMBO Journal 9: 2215–2220.

[56] Bartlett MS, Gaal T, Ross W, Gourse RL (1998) RNA polymerase mutants that destabilize RNA polymerase-promoter complexes alter NTP-sensing by *rrn* P1 promoters. Journal of Molecular Biology 279: 331–345.

[57] Kovacic RT (1987) The 0 degree C closed complexes between *Escherichia coli* RNA polymerase and two promoters, T7-A3 and *lac*UV5. Journal of Biological Chemistry 262: 13654–13661.

[58] Cook VM, Dehaseth PL (2007) Strand opening-deficient *Escherichia coli* RNA polymerase facilitates investigation of closed complexes with promoter DNA: effects of DNA sequence and temperature. Journal of Biological Chemistry 282: 21319–21326.

[59] Borukhov S, Sagitov V, Josaitis CA, Gourse RL, Goldfarb A (1993) Two modes of transcription initiation *in vitro* at the *rrnB* P1 promoter of *Escherichia coli*. Journal of Biological Chemistry 268: 23477–23482.

[60] Gourse RL (1988) Visualization and quantitative analysis of complex formation between *E. coli* RNA polymerase and an rRNA promoter *in vitro*. Nucleic Acids Research 16: 9789–9809.

[61] Newlands JT, Ross W, Gosink KK, Gourse RL (1991) Factor-independent activation of *Escherichia coli* rRNA transcription. II. characterization of complexes of *rrnB* P1 promoters containing or lacking the upstream activator region with *Escherichia coli* RNA polymerase. Journal of Molecular Biology 220: 569–583.

[62] Craig ML, Tsodikov OV, McQuade KL, Schlax P Jr, Capp MW, Saecker RM, Record M Jr (1998) DNA footprints of the two kinetically significant intermediates in formation of an RNA polymerase-promoter open complex: evidence that interactions with start site and downstream DNA induce sequential conformational changes in polymerase and DNA. Journal of Molecular Biology 283: 741–756.

[63] Spassky A, Kirkegaard K, Buc H (1985) Changes in the DNA structure of the *lac*UV5 promoter during formation of an open complex with *Escherichia coli* RNA polymerase. Biochemistry 24: 2723–2731.

[64] Krummel B, Chamberlin MJ (1992) Structural analysis of ternary complexes of *Escherichia coli* RNA polymerase. Deoxyribonuclease I footprinting of defined complexes. Journal of Molecular Biology 225: 239–250.

[65] Wade JT, Struhl K (2008) The transition from transcriptional initiation to elongation. Current Opinion in Genetics and Development 18: 130–136.

[66] Wade JT, Struhl K (2004) Association of RNA polymerase with transcribed regions in *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America 101: 17777–17782.

[67] Gries TJ, Kontur WS, Capp MW, Saecker RM, Record MT Jr (2010) One-step DNA melting in the RNA polymerase cleft opens the initiation bubble to form an unstable open complex. Proceedings of the National Academy of Sciences of the United States of America 107: 10418–10423.

[68] Saecker RM, Tsodikov OV, McQuade KL, Schlax PE Jr, Capp MW, Record MT Jr (2002) Kinetic studies and structural models of the association of *E. coli* σ^{70} RNA polymerase with the λP_R promoter: large scale conformational changes in forming the kinetically significant intermediates. Journal of Molecular Biology 319: 649–671.

[69] Tsodikov OV, Craig ML, Saecker RM, Record M Jr (1998) Quantitative analysis of multiple-hit footprinting studies to characterize DNA conformational changes in protein-DNA complexes: application to DNA opening by Eσ^{70} RNA polymerase. Journal of Molecular Biology 283: 757–769.

[70] Ring BZ, Yarnell WS, Roberts JW (1996) Function of *E. coli* RNA polymerase sigma factor sigma 70 in promoter-proximal pausing. Cell 86: 485–493.

[71] Brodolin K, Zenkin N, Mustaev A, Mamaeva D, Heumann H (2004) The sigma 70 subunit of RNA polymerase induces *lac*uv5 promoter-proximal pausing of transcription. Nature Structural and Molecular Biology 11: 551–557.

[72] Nickels BE, Mukhopadhyay J, Garrity SJ, Ebright RH, Hochschild A (2004) The σ^{70} subunit of RNA polymerase mediates a promoter-proximal pause at the *lac* promoter. Nature Structural and Molecular Biology 11: 544–550.

[73] Deighan P, Pukhrambam C, Nickels BE, Hochschild A (2011) Initial transcribed region sequences influence the composition and functional properties of the bacterial elongation complex. Genes and Development 25: 77–88.

[74] Hatoum A, Roberts J (2008) Prevalence of RNA polymerase stalling at *Escherichia coli* promoters after open complex formation. Molecular Microbiology 68: 17–28.

[75] Tang GQ, Roy R, Bandwar RP, Ha T, Patel SS (2009) Real-time observation of the transition from transcription initiation to elongation of the RNA polymerase. Proceedings of the National Academy of Sciences of the United States of America 106: 22175–22180.

[76] Kapanidis AN, Margeat E, Ho SO, Kortkhonjia E, Weiss S, Ebright RH (2006) Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. Science 314: 1144–1147.

[77] Revyakin A, Liu C, Ebright RH, Strick TR (2006) Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. Science 314: 1139–1143.

[78] Bai L, Santangelo TJ, Wang MD (2006) Single-molecule analysis of RNA polymerase transcription. Annual Review of Biophysics and Biomolecular Structure 35: 343–360.

[79] Herbert KM, Greenleaf WJ, Block SM (2008) Single-molecule studies of RNA polymerase: motoring along. Annual Review of Biochemistry 77: 149–176.

[80] Mooney RA, Darst SA, Landick R (2005) Sigma and RNA polymerase: an on-again, off-again relationship? Molecular Cell 20: 335–345.

[81] Bai G, Schaak DD, Smith EA, McDonough KA (2011) Dysregulation of serine biosynthesis contributes to the growth defect of a *Mycobacterium tuberculosis crp* mutant. Molecular Microbiology 82: 180–198.

[82] Shimada T, Yoshida H, Ishihama A (2013) Involvement of cyclic AMP receptor protein in regulation of the *rmf* gene encoding the ribosome modulation factor in *Escherichia coli*. Journal of Bacteriology 195: 2212–2219.

[83] Self WT, Hasona A, Shanmugam KT (2004) Expression and regulation of a silent operon, *hyf*, coding for hydrogenase 4 isoenzyme in *Escherichia coli*. Journal of Bacteriology 186: 580–587.

[84] Cho BK, Knight EM, Palsson BO (2006) PCR-based tandem epitope tagging system for *Escherichia coli* genome engineering. Biotechniques 40: 67–72.

[85] Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proceedings of the National Academy of Sciences of the United States of America 97: 6640–6645.

[86] Cho BK, Barrett CL, Knight EM, Park YS, Palsson BØ (2008) Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America 105: 19462–19467.

[87] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

[88] Guo Y, Mahony S, Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Computational Biology 8: e1002638.

[89] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Methods 7: 709–715.

Chapter 6

A streamlined ribosome profiling protocol for the characterization of microorganisms

6.1 Abstract

Ribosome profiling is a powerful tool for characterizing *in vivo* protein translation at the genome-scale with multiple applications ranging from detailed molecular mechanisms to systems-level predictive modeling. Though highly effective, the widespread use of this specialized protocol in the microbial research community has yet to occur. Here, we present a streamlined protocol with reduced barriers to entry for microbial characterization studies. The streamlined protocol was accomplished by avoiding specialized equipment during harvest, lysis, and recovery of monosomes and by eliminating time-consuming steps, in particular size-selection steps during library construction. Furthermore, this protocol drastically reduces contaminating rRNAs and tRNAs in the final library. Collectively, this streamlined protocol enables greater throughput, cuts the time from harvest to the final library in half, and generates a high fraction of informative reads all while retaining the high quality standards attributed to the existing protocol.

6.2 Introduction

Ribosome profiling is a novel, sequencing-based protocol that captures ribosomes as they traverse transcripts *in vivo* thereby revealing protein synthesis at the genome-scale [1, 2]. This protocol has produced numerous findings on the detailed molecular processes of translation such as the sequence-based prediction of translational pause sites [3], the characterization of the co-translational chaperone action of trigger factor [4], and revealed translational regulatory processes [2]. In addition to elucidating translation mechanisms, the quantitative nature of ribosome profiling has shown a strong correlation with quantitative proteomics [1, 5] and, more recently, been utilized in the systems-level modeling of translation elongation [6].

This protocol, detailed in [7, 8], applies endonuclease digestion to cell lysate thereby generating mRNA fragments protected by actively translating ribosomes which are then recovered and converted to a sequencing library. Though the original protocol is very robust, broad utilization of ribosome profiling for microbial applications is limited by the recommended use of specialized equipment associated with cell harvest, lysis, and ribosome recovery such as a 90-mm diameter filtration apparatus, a Retsch mill, an ultracentrifuge, and a fraction collector for recovery of monosomes [7, 8]. Furthermore, ribosome profiling is a laborious protocol taking 7-8 days to complete [7]. These libraries often contain a large fraction of contaminating rRNA and tRNA species [9] which produce uninformative sequencing reads. Many of these limitations have been addressed for eukaryotic applications in Epicentre's ARTseq protocol [9], but direct porting of the ARTseq protocol to bacteria is not currently possible.

Here, we sought to streamline this important protocol with the goal of making it available to the broader microbial research community by 1) avoiding the use of specialized equipment in the steps leading to footprint recovery, 2) enabling increased throughput and parallelization of samples, 3) reducing the time from harvest to library construction, and 4) eliminating rRNA and tRNA contaminating species in the final sequencing libraries.

6.3 Results and Discussion

The streamlined protocol we have developed is easier for microbial researchers to implement. This is achieved by first modifying the existing ribosome profiling protocol to perform harvest and lysis using short centrifugation followed by repeated freeze thaw lysis. This protocol uses antibiotic treatment prior to harvest as previously described [4, 7]. Lysate is then treated with micrococcal nuclease in a buffer that maintains ribosome integrity without sacrificing high nuclease activity. Monosomes are then recovered using a size exclusion spin-column analogous to those used by ARTseq [9]. Next, recovered ribosomes are treated with Qiazol (Qiagen) to recover RNA footprints. Footprints are then isolated using kit-based purification approaches followed by Ribo-Zero rRNA depletion (Epicentre). Library construction is then performed using a commercially available small RNA purification kit. For this study the NEBNext Small RNA Library Prep Kit for Illumina was used. After 3′ and 5′ adaptor ligation, additional steps are introduced to remove tRNAs by hybridization of custom anti-sense DNA oligos followed by treatment with a thermostable RnaseH. This effectively degrades tRNA contaminating species, leaving mRNA footprints unaffected. The anti-tRNA probes carry terminal dideoxynucleotides to avoid their participation during the final library amplification. Unlike, the existing ribosome profiling protocols, this procedure for library construction does not require any gel purification steps. Therefore, the time from harvest to finished library is reduced from 7-8 days to 3-4 days (Figure 6.1A).

We validated our streamlined protocol by generating ribosome profiling data replicating published conditions in *E. coli* K-12 MG1655 wild type cells [5] with the exception that chloramphenicol (CAM) was added to the cultures 2 min prior to harvest as previously described [7]. *E. coli* cultures were grown on MOPS rich media with 0.2% glucose and a full supplementation of amino acids (Teknova) aerobically to exponential phase. Libraries were sequenced using the Illumina MiSeq platform. (GEO accession GSE63858) and processed using the bioinformatics pipeline outlined previously [5, 7]. Examination of the mapped read length distribution produced using the streamlined protocol yields footprints in a size range comparable to that found using the original protocol (20-42 nt) [5] with 92% of mapped footprints falling

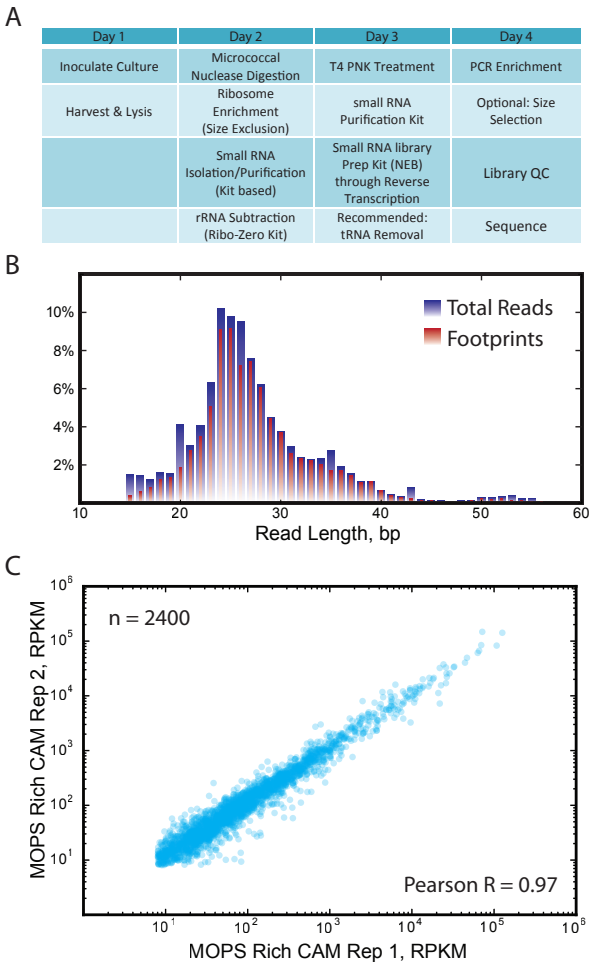


Figure 6.1: **Streamlined protocol for ribosome profiling of microorganisms.** (A) Overview of the major steps and associated timing of the streamlined protocol presented here. (B) The read length distribution for all reads (blue) and ribosomal footprints (red) generated using the streamlined protocol. (C) Correlation of RPKMs across all genes with RPKM>8 for biological replicates generated using the streamlined protocol. *E. coli* was grown on MOPS rich media and harvested at exponential phase with CAM addition prior to harvest.

within this range (Figure 6.1B). This plot also shows that 84% of reads mapping to genomic regions are not annotated as rRNAs, tRNAs, or other non-coding RNAs. Furthermore, the ribosome density was compared for biological replicates across all protein-coding genes with a Reads Per Kilobase per Million reads mapped (RPKM) exceeding 8 (Figure 6.1C). These datasets showed a strong linear correlation between biological replicates analogous to those generated by the original protocol [3, 1, 7].

Thus, the streamlined protocol is highly reproducible and yields a large fraction of informative reads in half the time of the original method.

Comparison with publically available data (GEO accession GSM1300279) further affirms that the streamlined protocol provides high-quality ribosome profiling data. Figure 6.2A compares one of the replicates generated using MOPS rich media with CAM added prior to harvest with data generated under similar conditions but without CAM added prior to harvest using the rapid harvest methodology [5]. This yields a linear correlation with a Pearson R-value of 0.89. Interestingly, a comparison of harvest approaches with and without CAM using *E. coli* MC4100 Δ*tig*::Kan + pTrc-*tig*-TEV-Avi cells showed a similar correlation with an R-value = 0.90 [7]. Likewise, examination of the meta-gene profile from the start codon to 300 amino acids downstream showed analogous density profiles between publically available MOPS rich data with the data generated here (Figure 6.2B). Lastly, the power of ribosome profiling was recently displayed in its ability to capture the stoichiometry of heteroprotein complexes using the calculated absolute synthesis rate [5]. This use of ribosomal profiling was perhaps best illustrated in the ability to predict the stoichiometry of the 8 proteins comprising the F_0F_1 ATP Synthase. Using the ribosome profiling data generated here we were able to accurately predict the stoichiometrys of this complex (Figure 6.2C).

6.4 Conclusion

Here, a modification to an increasingly important molecular biology approach is presented to help enable its availability to the greater microbial research community. This method allows for data to be generated in half the time needed for the original protocol, allows for multiple samples to be processed in parallel, and yields more usable data by eliminating undesirable library contaminants. These improvements are accomplished while producing high quality data, a hallmark of the original ribosome profiling protocol.

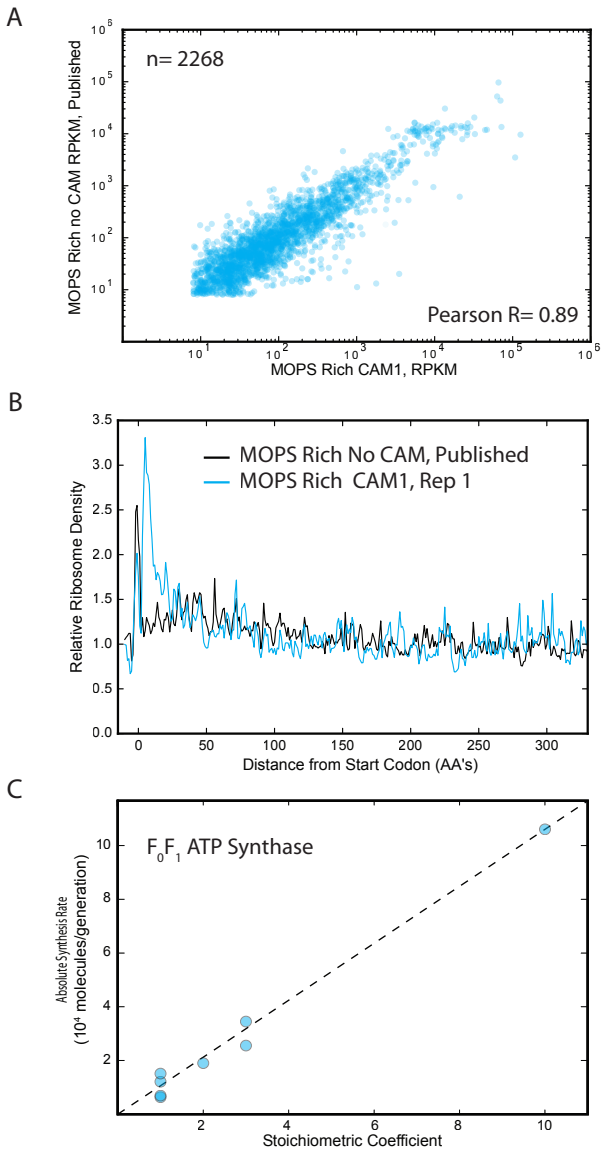


Figure 6.2: **Benchmarking the streamlined protocol against publically available data.** (A) Correlation of RPKMs across all genes with RPKM>8 for the streamlined protocol compared with publically available data. For the streamlined protocol *E. coli* was grown on MOPS rich media and harvested at exponential phase with CAM addition prior to harvest. The publically available data was generated under similar growth conditions but without the use of CAM prior to harvest. (B) Meta-gene analysis of the relative ribosome density profiles observed from the start codon to 333 amino acids downstream. (C) F_0F_1 ATP synthase stoichiometry determined by ribosome profiling.

6.5 Materials and Methods

6.5.1 Reagents

- 10 mM ATP
- 10 mM TE pH 7.5
- 10% Sodium Deoxycholate
- 5-guanylyl imidodiphosphate, GMPPNP (Sigma-Aldrich, St. Louis, MO, USA)
- $CaCl_2$, 500 mM
- Chloramphenicol
- EGTA 500 mM
- Hybridase Thermostable RNase H (Epicentre, Madison, WI, USA)
- $MgCl_2$, 1 M
- MgOAc, 1 M
- MNase, 500 Kunitz Units/μL (New England Biolabs, Ipswich, MA, USA)
- MOPS Rich Media 0.2% (Teknova, Hollister, CA, USA)
- NaCl, 5 M
- NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs, Ipswich, MA, USA)
- NH_4Cl, 1 M
- Nonidet P40, 100%
- Qiazol (Qiagen, Germantown, MD, USA)
- Qubit DNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA)

- Qubit RNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA)
- ReadyLyse Lysozyme (Epicentre, Madison, WI, USA)
- Ribo-Zero-rRNA Removal Kit (Epicentre, Madison, WI, USA)
- RNAse-Away (Molecular BioProducts, San Diego, CA, USA)
- RNase-free DNase I, 10 U/μL (Roche Life Sciences, Indianapolis, IN, USA)
- Sephacryl S400 MicroSpin Columns (GE Healthcare, Piscataway, NJ, USA)
- Superase-In 20 U/μL (Life Technologies, Carlsbad, CA, USA)
- T4 PNK (New England Biolabs, Ipswich, MA, USA)
- Tris pH 8.0, 1 M (Sigma-Aldrich, St. Louis, MO, USA)
- Triton X-100

6.5.2 Procedure

- USE ALL RNase-FREE SOLUTIONS
- ONLY USE LOW-RETENTION PLASTICWARE THROUGHOUT
- THE METHOD DESCRIBED HERE IS MODIFIED BASED ON THE PROTOCOL BY BECKER et al. [7]

Cell culture, harvest, and cell lysis

1. From an overnight culture of *E. coli*, inoculate a 200 mL culture in MOPS Rich Media with 0.2% glucose or media of choice (e.g. LB, M9 minimal).
2. Monitor cell growth using optical density (OD_{600}) measurements until the culture reaches an $OD_{600} = 0.4$.

3. Prior to harvest, cool the bench top centrifuge to 4 °C. Prepare 4 x 50 ml conical tubes filled to the 35 mL line with crushed ice and 50 μL Chloramphenicol at 50 mg/mL. Fill the dewar with liquid nitrogen (LN_2) and prepare ice water bath. Lastly, prepare buffers according to Recipe 1 and Recipe 2.

4. Add chloramphenicol 2 min prior to harvest to a final concentration of 100 μg/mL (i.e. add 400 μL chloramphenicol at 50 mg/mL to 200 mL culture).
   *This method is not validated for use without antibiotics introduced prior to harvest.

5. Distribute culture into 50 mL conical tubes prepared in Step 3, and spin at 5,000 x g for 1 min to generate a cell pellet.

6. Quickly decant media and briefly invert tube onto a paper towel to remove excess liquid.

7. Pipette 0.5 mL of ice cold Lysis Buffer into the 50 mL conical tube. Quickly vortex on high for ≈5 s or until cells are completely resuspended in the Lysis Buffer.

8. Plunge the tube into LN_2 for 2 min.
   *At this stage cells can be stored at -80 °C or one can proceed to lysis.

9. Transfer tubes to an ice water bath. Remove the ice that forms on the outside of the tube. Wait ≈20 minutes and vortex briefly. Return to ice water bath for another 10 minutes and repeat vortexing. Move between ice water bath and vortex until a slushy stage is reached.

10. Repeat Steps 8 and 9 two more times.
    *More freeze/thaw cycles may be necessary for other organisms

11. Thaw lysate on ice. Transfer the contents of the conical tube to 1.5 mL micro-centrifuge tubes.

12. Add 30 μL 10% Sodium Deoxycholate to each ≈1,000 μL sample and incubate for 3 minutes on ice.

13. Centrifuge for 10 minutes at 30,000 x g at 4 °C. If there is significant cloudiness present after the initial spin incubate on ice for 5 min and repeat spin.
    *Some cloudiness is acceptable.

14. Transfer cleared lysate to a new microcentrifuge tube.

15. Check the OD_{260} of 1:100 dilution in 10 mM TE pH 7.5 using a Nanodrop. Use lysis buffer diluted 1:100 with 10 mM TE pH 7.5 to blank the instrument prior to taking the sample reading.
    *At this stage the clarified lysate can be stored at -80 °C.

MNase Treatment

16. The input to MNase treatment is normalized based on the absorbance measure taken on Step 15 to calculate the 25 A_{260} units. The following formula should be used to determine the volume of cleared lysate to use in the next step: $Volume = 1,000 \times [25/(dilution factor \times OD_{600})]$

17. Prepare the mixture described in Recipe 3.

18. Incubate the mixture for 2 hr at 25 °C.

19. Quench the reaction by add 2.5 μL EGTA (500 mM stock)

Monosome Recovery, RNA Isolation, and rRNA Removal

20. Prepare Polysome buffer as described in Recipe 4.

21. Prepare 2 Sephacryl S400 MicroSpin columns for each sample. Shake Sephacryl S400 MicroSpin columns several times to resuspend the resin. Tap on bench top to remove bubbles that may form.

22. Open the columns and gently push through the buffer with an electronic serological pipettor (remove plastic nose-piece and form seal with rubber insert, pulse trigger). Stop as soon as froth appears.

23. Equilibrate the resin by passing 1.5 mL Polysome Buffer through each column in 3x500 μL aliquots emptying the collection tube after each wash. Remove the last wash by spinning for 4 minutes at 600 x g in a swinging-bucket tabletop microcentrifuge at 4 °C.
    *A fixed angle rotor can also be used as described by the vendor

24. Discard the flow through and attach a new collection tube to each column.

25. Apply half of each MNase-treated sample to a single Sephacryl S400 column (i.e. 100 μL per column) and centrifuge for 2 min at 4 °C and 600 x g.

26. Add 700 μL Qiazol to the flow-through and follow the protocol for isolation of small RNA as part of the miRNeasy Mini Kit (Qiagen). At this stage you will combine the samples processed using 2 Sephacryl S400 tubes. Load Qiagen column multiple times as needed.

27. Elute RNA in 15 μL of RNase-free water. Pass the elate back through the column a second time for greater recovery.

28. Measure the RNA concentration using the Qubit RNA HS Assay Kit.
    *At this stage isolated RNA can be stored at -80 °C.

29. Treat 1 μg Ribosomal Protected RNA (Step 28) with Ribo-Zero rRNA Removal Kit for bacteria (Epicentre) using the vendors instructions with the following modifications: Halve volumes of all components using 5 μL rRNA removal solution (i.e. 112.5 μL beads per sample, 20 μL reaction volume). For the RNeasy MiniElute Cleanup step bring the sample volume up to 50 μL with nuclease-free water, add 350 μL RLT buffer, vortex to mix, then add 600 μL 100% Ethanol and vortex again before applying to column. Elute with 19 μL RNase-free Water passed over the column two times.
    *At this stage rRNA depleted RNA can be stored at -80 °C.

Sequencing Library Construction (Illumina Compatible)

30. Prepare the T4 PNK reaction as described in Recipe 5.

31. Incubate at 37 °C for 30 minutes.

32. Add 1 μL 10 mM ATP and Incubate at 37 °C for 30 minutes.

33. Add 27 μL RNase-free water, 350 μL RLT buffer (Qiagen RNeasy Kit), and 600 μL 100% Ethanol. Apply to the Qiagen RNA mini-elute column. Wash with 650 μL of RPE and then with 500 μL of 80% Ethanol. Spin dry the column as per the manufacturers protocol.

34. Elute with 10 μL RNase-free Water. Pass eluate over the column a second time.

35. Check the concentration with Qubit RNA HS Assay Kit.

36. Follow NEBNext Small RNA Library Prep Set for Illumina protocol stopping after 5 adapter ligation.
    *The placement of tRNA depletion after adaptor ligation is crucial.

37. Purify ligated products using Qiagen RNeasy Kit: Adjust the volume to 100 μL with RNase-free water and add 350 μL RLT buffer. Vortex and then add 700 μL 100% Ethanol. Vortex to mix and apply to RNeasy MiniElute column. Wash with 650 μL RPE, then 500 μL 80% Ethanol. Spin dry the column as per the manufacturers protocol and elute with 11 μL Nuclease-free water.

38. tRNA removal is achieved by hybridizing DNA oligo probes specific for the 5′ and 3′ halves of tRNAs and digesting with a thermostable RNase H. Prepare the tRNA removal reaction as described in Recipe 6.

39. Heat the sample to 90 °C for 1 min. Ramp down to 65 °C over 2 min. While on the instrument, open the tube and add 1 μL Hybridase RNaseH enzyme. Close the tube and incubate for 30 min at 65 °C.

40. Prepare 1.5 mL tubes with 180 μL RLT and 70 μL RNase-free water during incubation.

41. Process one tube at a time: While in the thermocycler, open the tube and quickly add 170 μL of RLT buffer. Mix by pipetting, raising and lowering tip

as needed to avoid overflow. Transfer to 1.5 mL microcentrifuge tube prepared in Step 40. Mix by vortexing.

42. Add 700 μL of 100% Ethanol, vortex, and load onto a Qiagen RNeasy column.

43. Wash 2X with 650 μL RPE then use 500 μL of 80% Ethanol for the third wash. Spin dry the column as per the manufacturers protocol and elute with 30 μL RNase-free water.
    *At this stage tRNA depleted RNA can be stored at -80 °C.

44. Resume the NEBNext Small RNA Library Prep Set for Illumina protocol at reverse transcription as follows (performed in thermocycler): First, to 30 μL adaptor-ligated, tRNA-depleted product, add 1 μL of the reverse transcription primer. Next, denature the mixture at 65 °C for 5 min and then place the sample on ice. Add 8 μL of First Strand cDNA Synthesis buffer and 1 μL Murine RNAse inhibitor. Incubate at 25 °C for 5-10 minutes to anneal the RT primer. Then add 1 μL Protoscript II and incubate at 50 °C for 1 hour. Heat inactivate enzyme at 70 °C for 15 minutes.
    *Reaction can be allowed to hold overnight at 4 °C or stored at -20 °C.

Final Library Preparations

45. Perform PCR amplification as per the NEBNext Protocol with the addition of 1 μL 10X SYBR Green I Nucleic Acid Gel Stain (Life Technologies) in order to monitor progress in a real-time PCR instrument. Do not allow reaction to reach plateau phase.
    *12-16 cycles of amplification are typical.

46. Purify sample using QIAquick PCR purification kit (Qiagen) with the addition of two washes with 500 μL 35 Guanidine Hydrochloride (not included with kit) prior to wash with PE buffer. Elute sample with 100 μL nuclease-free water.

47. Quantitate using Qubit HS dsDNA assay.

48. Check the quality of the final library using a Bioanalyzer high sensitivity DNA assay. There should be a prominent single peak with average size of around 153 bases. This protocol does not completely eliminate minor peaks of lower sizes which probably represent unincorporated adapter and/or primer artifacts.

Sequencing and Analysis

49. Sequencing should be carried out on an Illumina platform with single end reads of lengths no less than 50 bp.

50. Data processing is done following a pipeline designed on those detailed in [3, 5, 7]. These scripts can be found at the following repository: https://github.com/SBRG/RibosomeProfiling

6.5.3 Recipes

Table 6.1: Recipe 1 Lysis Buffer (scale up as needed)

Components	**Volume**
Tris pH 8.0, 1 M	25 μL
NH_4Cl, 1 M	25 μL
MgOAc, 1 M	10 μL
Triton X-100, 100%	8 μL
RNase-free DNase I, 10 U/L	10 μL
Superase-In, 20 U/μL	15 μL
ReadyLyse Lysozyme	10 μL
Chloramphenicol, 50 mg/mL	10 μL
5-guanylyl imidodiphosphate (GMPPNP), 10 mg/mL	1 μL
RNase-free Water	886 μL
Total Volume	**1 mL**

-The GMPPNP stock solution is prepared at 10 mg/mL in water, aliquoted and stored at -80 °C as described in [7].
-The final concentrations of the Lysis Buffer components are 25 mM Tris pH 8.0, 25 mM NH_4Cl, 10 mM MgOAc, 0.8 % Triton X-100, 100 U/mL RNase-free DNase I, 0.3 U/μL Superase-In, 1.55 mM Chloramphenicol, and 17 μM 5-guanylyl imidodiphosphate (GMPPNP).

Table 6.2: Recipe 2 MNase Buffer (scale up as needed)

Components	**Volume**
Tris pH 8.0, 1 M	25 μL
NH_4Cl, 1 M	25 μL
MgOAc, 1 M	10 μL
Chloramphenicol, 50 mg/mL	10 μL
5-guanylyl imidodiphosphate (GMPPNP), 10 mg/mL	1 μL
RNase-free Water	886 μL
Total Volume	**1 mL**

-The GMPPNP stock solution is prepared at 10 mg/mL in water, aliquoted and stored at -80 °C as described in [7].
-The final concentrations of the Lysis Buffer components are 25 mM Tris pH 8.0, 25 mM NH_4Cl, 10 mM MgOAc, 0.8 % Triton X-100, 100 U/mL RNase-free DNase I, 0.3 U/μL Superase-In, 1.55 mM Chloramphenicol, and 17 μM 5-guanylyl imidodiphosphate (GMPPNP).

Table 6.3: Recipe 3 MNase Reaction Mix (200 μL)

Components	**Volume**
Clarified Lysate, 25 A_{260} Units	X μL
$CaCl_2$, 500 mM	2 μL
Superase-In, 20 U/μL	2 μL
NEB MNase 500 U/μL	12 μL
MNase Buffer (Recipe 2)	q.s. to 200 μL
Total Volume	**200 μL**

- X = the volume of lysate yielding 25 A_{260} units.

Table 6.4: Recipe 4 Polysome Buffer (10 mL)

Components	**Volume**
Nonidet P40, 100%	100 μL
$MgCl_2$, 1 M	500 μL
EGTA, 0.5 M	500 μL
Tris-HCl pH 8.0, 1 M	500 μL
NaCl, 5 M	500 μL
RNase-free water	7.9 mL
Total Volume	**10 mL**

- The final concentration of Polysome Buffer is 1% Nonidet P40, 50 mM $MgCl_2$, 25 mM EGTA, 50 mM Tris-HCl pH 8.0, 250 mM NaCl.

Table 6.5: Recipe 5 T4 PNK Reaction Mix (22.2 μL)

Components	**Volume**
rRNA-subtracted RNA	18 μL
10X T4 PNK Buffer	2.2 μL
Superase-In, 20 U/μL	1 μL
T4 PNK, 10 U/μL	1 μL
Total Volume	**22.2 μL**

Table 6.6: Recipe 6 tRNA Depletion (30 μL)

Components	**Volume**
Adapter-ligated small RNA	10 μL
Anti-tRNA Oligo Mix*	10 μL
Superase-In, 20 U/μL	1 μL
10X Hybridase Buffer	3 μL
RNase-free Water	6 μL
Total Volume	**10 mL**

- *The anti-tRNA oligo mix may be optimized for different experimental conditions. Here we found the following to be effective: 40 pmol of each anti-Asn oligo, 40 pmol of each anti-LeuT oligo, and 10 pmol of each remaining anti-tRNA oligo. These oligos anneal to 32 bases from the 5′ or 3′ ends of each tRNA (2 oligos per tRNA species). The sequences used here were obtained from the NC_000913.2 genome annotation. The oligo set can be ordered with the 3′ dideoxy modification or unmodified oligos can be treated with terminal transferase and purified using Zymos Oligo Clean & Concentrator Kit.

6.5.4 Troubleshooting

Problem

Cleared Lysate A_{260} too low

Possible causes and solutions

Lysis buffer:Cell Pellet ratio too high. Use 500 μL buffer per pellet from each 50 ml conical tube. Try using less ice in tubes to leave room for more culture. Consider growing cells to higher density if it will not adversely affect your experiment.

Incomplete lysis of cells. Be sure to allow cells to thaw sufficiently between freeze cycles. Try adding more freeze/thaw cycles. Do not leave out Sodium Deoxycholate step.

PROBLEM

Low or no yield of small RNA

POSSIBLE CAUSES AND SOLUTIONS

Degradation of ribosomes and/or mRNA. Follow good practices for handling RNA. When possible, purchase reagents certified nuclease-free by vendor. Home-made buffers can be passed through syringe filters having high non-specific affinity for proteins (e.g. nylon).

Poor performance of RNA purification kits. Follow Sample:RLT:Ethanol ratio recommendations carefully as they have been chosen to optimize binding of small RNA species to spin columns. Try double or triple loading samples onto column during binding and eluting steps (i.e. pass flow-through over column again). Consider using a microcentrifuge that allows for slow acceleration during sample binding and eluting steps.

PROBLEM

No amplification observed during final library preparation.

POSSIBLE CAUSES AND SOLUTIONS

Library construction failed. Assure that at least 100 ng of small RNA is input

into first step. During end repair step use fresh lot of ATP which has not undergone many freeze/thaw cycles. Consider running positive control such as total RNA as per kit instructions. Keep kit enzymes, especially ligases, chilled at all times (remove from -20 °C only briefly and as needed). Following denaturing steps consider quick chilling samples in second thermocycler preset to 4 °C and waiting for 1 min before proceeding. Check settings of real-time PCR instrument to assure detection of SYBR green in wells containing samples.

PROBLEM

Small DNA peaks in bioanalyzer trace (<100 bp) are very prominent.

POSSIBLE CAUSES AND SOLUTIONS

Carryover of oligos not incorporated into amplified footprint/adapter constructs. Reduce the amount of primers put into PCR amplification and/or run more cycles of PCR. PAGE based or Pippin Prep (Sage Science) size selection can be performed.

PROBLEM

Prominent tRNA contamination persists in sequenced library.

POSSIBLE CAUSES AND SOLUTIONS

Experimental conditions cause unique profile of tRNA contamination. Analyze the distribution of tRNA reads and adjust the anti-tRNA oligo mix accordingly.

The small RNA:anti-tRNA oligo ratio is too large. Reduce the amount of small RNA input into NEB kit.

6.5.5 Equipment

- Nanodrop (Thermo Scientific, Wilmington, DE)
- Qubit (Life Technologies, Carlsbad, CA, USA)
- Thermomixer (Eppendorf, Hauppauge, NY, USA)
- Thermocycler (Bio-Rad, Hercules, CA, USA)

- Real-Time PCR (Bio-Rad, Hercules, CA, USA)
- Bioanalyzer (PerkinElmer, Waltham, MA, USA)

6.6 Acknowledgements

We would like to thank Dr. Gene-Wei Li and Dr. Johnathan Weissman for their input and guidance. We also would like to thank Epicentre. The Novo Nordisk Foundation and the NIH U01 grant GM102098-01 provided financial support for this work. H.L. was supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086.

Chapter 6 has been submitted for publication of the material as it may appear in Latif H, Szubin R, Tan J, Brunk E, Lechner A, Zengler K, Palsson BØ. A streamlined ribosome profiling protocol for the characterization of microorganisms. Submitted to Biotechniques. 2014. The dissertation author was the primary author of this paper responsible for the research. The other authors are Richard Szubin, Justin Tan, Elizabeth Brunk, Anna Lechner, Karsten Zengler, and Bernhard Ø. Palsson.

6.7 Author Contributions

H.L. and R.S. conceived, developed, and executed the experiments. R.S. performed all experiments. H.L. processed and analyzed all datasets. H.L., J.T., E.B., A.L. developed the bioinformatics analytical pipeline. H.L., R.S., B.O.P., and K.Z. wrote and edited the manuscript.

6.8 Bibliography

[1] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223.

[2] Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nature Reviews Genetics 15: 205–213.

[3] Li GW, Oh E, Weissman JS (2012) The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541.

[4] Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, Nichols RJ, Typas A, Gross CA, Kramer G, Weissman JS, Bukau B (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor *in vivo*. Cell 147: 1295–1308.

[5] Li GW, Burkhardt D, Gross C, Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624–635.

[6] Subramaniam AR, Zid BM, OShea EK (2014) An integrated approach reveals regulatory controls on bacterial translation elongation. Cell 159: 1200–1211.

[7] Becker AH, Oh E, Weissman JS, Kramer G, Bukau B (2013) Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. Nature Protocols 8: 2212–2239.

[8] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mrna fragments. Nature Protocols 7: 1534–1550.

[9] Freeberg L, Kuersten S, Syed F (2013) Isolate and sequence ribosome-protected mrna fragments using size-exclusion chromatography. Nature Methods 10.

Chapter 7

Trash to treasure: Production of biofuels and commodity chemicals via syngas fermenting microorganisms

7.1 Abstract

Fermentation of syngas is a means through which unutilized organic waste streams can be converted biologically into biofuels and commodity chemicals. Despite recent advances, several issues remain which limit implementation of industrial-scale syngas fermentation processes. At the cellular level, the energy conservation mechanism of syngas fermenting microorganisms has not yet been entirely elucidated. Furthermore, there was a lack of genetic tools to study and ultimately enhance their metabolic capabilities. Recently, substantial progress has been made in understanding the intricate energy conservation mechanisms of these microorganisms. Given the complex relationship between energy conservation and metabolism, strain design greatly benefits from systems-level approaches. Numerous genetic manipulation tools have also been developed, paving the way for the use of metabolic engineering and systems biology approaches. Rational strain designs can now be deployed resulting

in desirable phenotypic traits for large-scale production.

7.2 Graphical Abstract

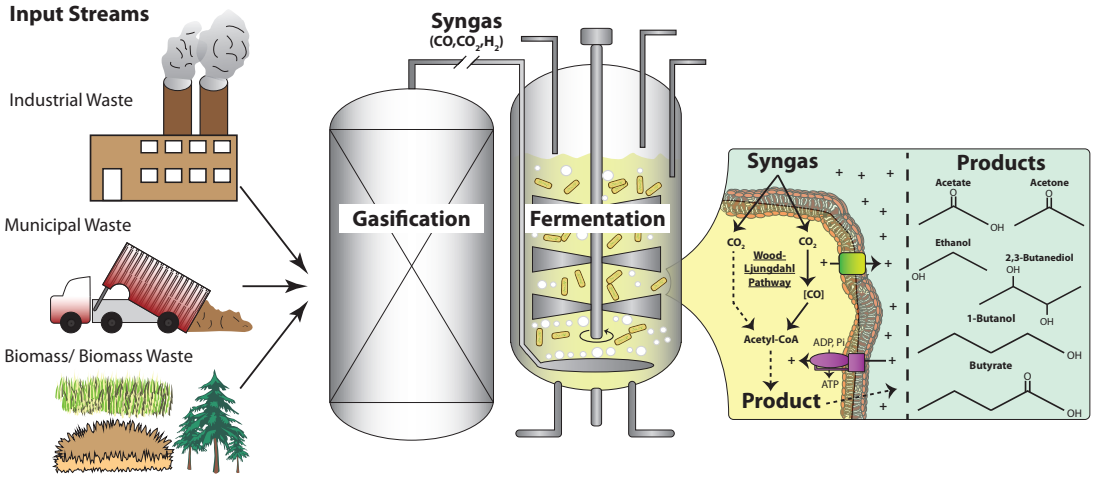


Figure 7.1: **Overview of syngas fermentation.** Syngas fermentation can be conducted using low cost waste streams of organic material. Waste matter can be converted to syngas through gasification. The syngas can then be metabolized by acetogens to produce high value commodity chemicals.

7.3 Introduction

Heightened concerns over global warming and fossil fuel supply and prices have led to a paradigm shift in perceived routes to chemical commodity production and energy generation. The majority of the world community has set challenging targets for reductions in greenhouse gas emissions to be achieved in part through the development of sustainable routes to chemicals, fuels, and energy production. The EU has targeted a 10% share of renewable biofuels in the transportation sector by 2020 [1] while the US has mandated the production of 36 billion gallons of biofuels by 2022 [2]. The biological conversion of renewable lignocellulosic biomass such as wheat straw, spruce, switchgrass, and poplar to biofuels is expected to play a prominent role in achieving these goals. These forms of biomass address many of the concerns associated with the production of first-generation biofuels [3, 4]. However, 10-35%

of lignocellulosic biomass is composed of lignin [5, 6, 7], which is highly resistant to breakdown by the vast majority of microorganisms [8]. Thus, if the EU and US cellulosic biofuel targets are realized, land allocation for biofuel production will increase and megatons of organic waste will be generated.

This organic waste provides a significant resource of biomass that can be utilized for producing biofuels as well as commodity chemicals. Through gasification, virtually any form of organic matter can be converted into a mixture of carbon monoxide (CO), carbon dioxide (CO_2), and hydrogen (H_2), referred to as synthesis gas or syngas. Gasification involves high temperature (usually 600-900 °C) partial oxidation of biomass in the presence of a gasifying agent (e.g., oxygen or steam) resulting in the production of gas with significant amounts of CO and H_2 [9]. Syngas can be metabolized by certain carbon-fixing microorganisms and converted to valuable multi-carbon compounds like acetate, ethanol, butanol, butyrate, and 2,3-butanediol [10, 11]. This process, known as syngas fermentation, provides an attractive means for converting low cost organic substrates and waste streams into valuable chemicals. Syngas fermentation has numerous advantages when compared to thermo-chemical processes such as Fischer-Tropsch synthesis. These include a higher tolerance for impurities such as sulfur compounds, a wider range of usable H_2, CO_2, and CO mixtures, a lower operating-temperature and -pressure, and higher product yield and uniformity. However, wide use of these syngas fermenting microorganisms as production hosts is currently hindered by several factors, including low volumetric product titers, product feedback inhibition, and low gas-liquid mass transfer coefficient (kLa) [12]. Though some of these challenges can be overcome, in part, through process improvements, a fundamental understanding of the biology enabling syngas fermentation is needed to guide those design strategies and to provide targets for cellular engineering. Thus, the biggest challenge facing process development for syngas fermentation may be the lack of tools and technologies that will further our understanding of the fundamental biology behind these versatile microorganisms. This review focuses on these unique microorganisms, their metabolic and energy conservation pathways, and the genetic engineering strategies that together will guide advances in the use of syngas fermentation for the production of biofuels and commodity chemicals.

7.4 Syngas fermenting microorganisms

A diverse range of microorganisms that can metabolize syngas and produce multi-carbon compounds have been identified (Figure 7.2). These organisms are ubiquitous in numerous habitats such as soils, marine sediments, and feces, exhibit various morphologies (e.g., rods, cocci, or spirochetes), have a wide range of optimal growth temperatures (psychrophilic, mesophilic, or thermophilic), and demonstrate different tolerance toward molecular oxygen [13]. Syngas fermenting microorganisms also have diverse metabolic capabilities, resulting in the formation of a variety of native products such as acetate, ethanol, butanol, butyrate, formate, H_2, H_2S, and traces of methane [13, 10, 14, 15, 11, 15]. Though some methanogens are known to synthesize multi-carbon organics from syngas, the vast majority of syngas fermenting organisms are acetogens (Figure 7.2). Acetogens are anaerobes that assimilate CO_2 via the Wood-Ljungdahl (WL) pathway, also referred to as the reductive acetyl-CoA pathway. Though we focus here mainly on autotrophic conversion of syngas in acetogens, the WL pathway is also active during heterotrophic growth.

7.5 The Wood-Ljungdahl pathway

The WL pathway is hypothesized to be the most ancient CO_2 fixation pathway [16]. However, the apparent simplicity of this linear pathway belies the complex, interconnected energy conservation mechanisms that enable growth on syngas [13, 17]. Only a short overview of the pathway will be given here since the WL pathway has been extensively reviewed elsewhere [13, 18, 19, 20]. Figure 7.3 shows the complete WL pathway with electron carrier proteins and cofactors as proposed for *Clostridium ljungdahlii* [21, 22].

The WL pathway consists of two branches: the methyl branch and the carbonyl branch. These two branches provide reduced, single-carbon molecules that contribute to the formation of acetyl-CoA, which can then be assimilated into cellular biomass or converted to acetate. For simplicity, the case of autotrophic growth on CO_2 and H_2 will be examined first. On the methyl branch, a six-electron reduction of CO_2 yields a methyl moiety while the carbonyl branch reduces CO_2 to CO

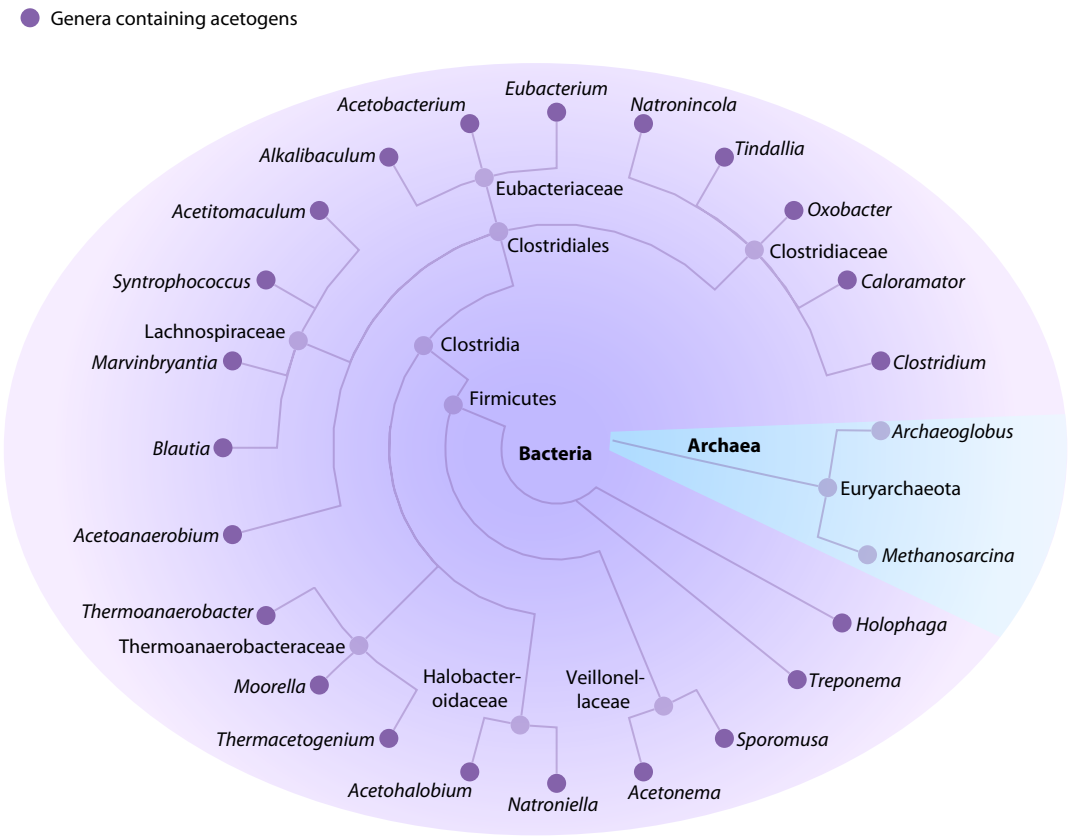
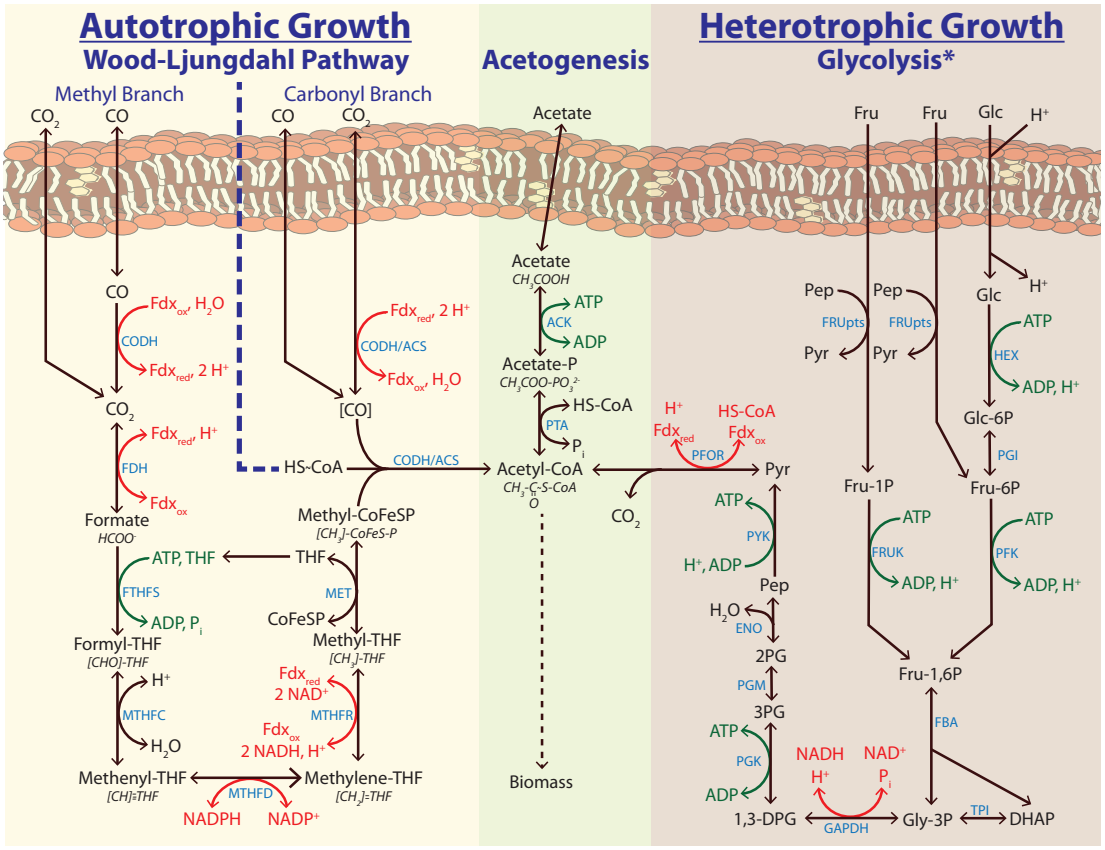


Figure 7.2: **Syngas fermenting organisms known to produce multi-carbon organic compounds.** Shown is the taxonomic classification of the genera capable of converting syngas to multi-carbon compounds based on organisms found in [13, 10, 14, 15, 11]. With the exception of two archaeal genera, all of the genera that produce multi-carbon organic compounds are considered acetogens. Genera are classified based on NCBIs current taxonomic nomenclature and categorization.

which is bound to the carbon monoxide dehydrogenase/acetyl-CoA synthase complex (CODH/ACS). CODH/ACS unites the two branches by reacting these products with coenzyme A to yield acetyl-CoA. Thus, a total of one ATP (needed for formate fixation) and eight electrons are needed to fix two molecules of CO_2. With H_2 as the sole electron donor, hydrogenases reduce cofactor intermediates (e.g., reduced ferredoxin, NAD(P)H) that are subsequently used to reduce CO_2. CO, typically the most abundant carbon species in syngas [23], is often able to sustain autotrophic growth through the WL pathway as the sole electron donor. In this case, CODH carries out a water-gas shift reaction where water and CO yield CO_2, two protons, and two elec-

trons providing the necessary inputs for the formate dehydrogenase reaction on the methyl branch. CO is bound directly by CODH/ACS on the carbonyl branch requiring no reduction. Thus, the conversion of two CO to acetyl-CoA requires one ATP and four electrons. The formation of acetate from acetyl-CoA yields one molecule of ATP through substrate level phosphorylation (SLP). As a whole, the WL pathway generates no net ATP through SLP and the transfer of four or eight electrons from reduced ferredoxin, NADH, or NADPH to fix syngas.

Figure 7.3: **The Wood-Ljungdahl pathway and its connection to heterotrophic metabolism.** Shown is the WL pathway for Clostridium ljungdahlii as assessed in [21, 22] with reduced ferredoxin, NADH, and NADPH serving as electron carriers during CO_2 fixation. The left panel shows the Methyl and Carbonyl branches of the WL pathway leading to either acetate formation or assimilation of acetyl-CoA into biomass. The right panel depicts fermentation through glycolysis for fructose and glucose. The electrons generated during glycolysis are used to produce an additional acetyl-CoA through the WL pathway. Redox reactions involving cofactor intermediates are shown in red with each cofactor reaction involving the transfer of two electrons. ATP consumption/generation reactions are shown in green. Reactants abbreviations: Acetate-P, acetate phosphate; CoA, coenzyme A; CoFeSP, corrinoid iron sulfur protein; THF, tetrahydrafolate; Fdx_{ox}, oxidized ferredoxin; Fdx_{red}, reduced ferredoxin; Pyr, pyruvate; Pep, phosphoenolpyruvate; 2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate; 1,3-DPG, 1,3-bisphosphoglycerate; DHAP, dihydroxyacetone phosphate; Gly-3P, glycerol 3-phosphate; Fru-1,6P, fructose 1,6-bisphosphate; Fru-6P, fructose 6-phosphate; Glc-6P, glucose 6-phosphate; Glc, glucose; Fru-1P, fructose 1-phosphate; Pyr, pyruvate; Pep, phosphoenolpyruvate; Fru, fructose. Enzyme abbreviations (in blue): ACK, acetate kinase; PTA, phosphotransacetylase; MET, methyl transferase; MTHFR, methylene tetrahydrofolate reductase; MTHFD, methylene tetrahydrofolate dehydrogenase; MTHFC, methenyl tetrahydrofolate cyclohydrolase; FTHFS, formyl tetrahydrofolate synthase; FDH, formate dehydrogenase; CODH/ACS, carbon monoxide dehydrogenase / acetyl-CoA synthase; PFOR, Pyruvate ferredoxin oxidoreductase; PGK, phosphoglycerate kinase; ENO, enolase; PGM, phosphoglycerate mutase; PGK, phosphoglycerate kinase; PFK, phosphofructokinase; TPI, triose phosphate isomerase; FBA, fructose-biphosphate aldolase; FRUK, fructokinase; PFK, phosphofructokinase; PGI, phosphoglucose isomerase; HEX, hexokinase; FRUT, fructose transporter. *The stoichiometry shown represents conversion of a single Gly-3P molecule, thus only half of the payoff phase of glycolysis is shown.


Autotrophic Growth
Wood-Ljungdahl Pathway
Acetogenesis
Heterotrophic Growth
Glycolysis*
Methyl Branch
Carbonyl Branch
CO2
CO
Acetate
Fru
Glc
H+
Fdxox, H2O
CODH
Fdxred, 2 H+
Fdxred, H+
FDH
Formate
HCOO-
ATP, THF
FTHFS
ADP, Pi
Formyl-THF
[CHO]-THF
MTHFC
Methenyl-THF
[CH]=THF
MTHFD
NADPH
NADP+
Methylene-THF
[CH2]=THF
MTHFR
Fdxred
2 NAD+
Fdxox
2 NADH, H+
Methyl-THF
[CH3]-THF
MET
CoFeSP
THF
Methyl-CoFeSP
[CH3]-CoFeS-P
CODH/ACS
[CO]
HS-CoA
Acetyl-CoA
CH3-C(=O)-S-CoA
ACK
ATP
ADP
Acetate-P
CH3COO-PO3 2-
PTA
Pi
Biomass
PFOR
Pyr
PYK
H+, ADP
Pep
ENO
H2O
2PG
PGM
3PG
PGK
1,3-DPG
GAPDH
NADH
NAD+
Gly-3P
TPI
DHAP
FBA
Fru-1,6P
FRUpts
FRUK
Fru-1P
PGI
Glc-6P
HEX
PFK
Fru-6P

During syngas fermentation electrons are donated by H_2 or CO. However, the WL pathway is capable of sourcing electrons from many compounds under heterotrophic conditions including alcohols, organic acids, and simple sugars [13]. This enables acetogens to have near complete stoichiometric conversion of hexoses to acetate. From glycolysis, acetogens typically synthesize two acetates, four ATPs (via SLP), and two CO_2's from one mole of hexose. The CO_2 generated can be fixed by the WL pathway using the eight electrons generated during glycolysis. Thus, a third mole of acetate is produced but no net ATP is synthesized via SLP. The overall conversion of hexose yields three moles of acetate and four moles of ATP via SLP.

7.6 Energy conservation in acetogens

Since the WL pathway yields no net ATP through SLP, a chemiosmotic gradient is necessary to drive ATP synthesis under autotrophic conditions, such as syngas fermentation. Some acetogens utilize anaerobic respiration (also known as electron transport phosphorylation, ETP), for energy conservation such as *Moorella thermoacetica* [24, 25, 26]. However, many acetogens do not possess electron transport chain proteins. Recently, flavin-based electron bifurcation (FBEB) has been described as an alternative means for energy conservation in certain bacteria [17]. FBEB couples exergonic redox reactions with endergonic redox reactions. Certain FBEB complexes couple this redox reaction with the translocation of cations thereby forming a chemiosmotic gradient. An example of this is the Rnf complex, which is found in many microorganisms. The Rnf complex is a membrane associated FBEB system that pumps Na^+ or H^+ ions out of the cell. This is driven by coupling the oxidation of reduced ferredoxin with the reduction of NAD^+ [27] (Figure 7.4). In the acetogen *Acetobacterium woodii*, a Na^+ gradient generated by the Rnf complex is converted to ATP by a Na^+ F_1F_0 ATP synthase [28, 29, 30, 31]. A proton translocating Rnf complex in *C. ljungdahlii* has recently been shown to be essential for autotrophic growth [32]. Studies have also revealed soluble FBEB complexes in acetogens. A [FeFe]-hydrogenase (Figure 7.4) that reduces ferredoxin and NAD^+ through the oxidation of H_2 has been identified in *A. woodii* [33] and *M. thermoacetica* [34, 35].

Recently, an NADP-specific [FeFe]-hydrogenase has been discovered in *Clostridium autoethanogenum* [36]. An Nfn complex similar to that of *Clostridium kluyveri* [37] (Figure 7.4) was identified in *M. thermoacetica* as well [34]. This complex reduces $NADP^+$ while oxidizing NADH and ferredoxin. Furthermore, it has been suggested that the WL reaction catalyzed by methylene tetrahydrofolate reductase (MTHFR) is also an electron bifurcation reaction in *A. woodii* [31] and *C. ljungdahlii* [21, 22] (Figure 7.3).

Together, membrane bound and cytoplasmic FBEB reactions create an intricate network of redox reactions that collectively contribute to energy conservation and redox homeostasis. For example, it is thought that energy conservation during hydrogen-dependent caffeate respiration in *A. woodii* is governed by three FBEB complexes working in concert; the membrane associated Rnf, the soluble [FeFe]-hydrogenase, and the soluble caffeyl-CoA reductase-Etf complex responsible for oxidation of NADH and reduction of ferredoxin [38]. FBEB reactions metabolically 'hard-wire' the redox state of the different cofactor pools and electron carrier proteins. Thus, these reactions can be envisioned as moving together in response to changes in metabolism and energy conservation.

7.7 Advances in genetic manipulation tools

A lack of versatile genetic tools for manipulating syngas fermenters has hindered the ability to engineer this group of microorganisms. However, there has been remarkable progress in the development of genetic systems for numerous industrially important anaerobes, especially clostridia, in recent years. Protocols for gene deletion via double crossover homologous recombination were recently developed for the syngas fermenting *C. ljungdahlii* [39] and *M. thermoacetica* [40, 41, 42]. Uracil-auxotrophic mutants of *M. thermoacetica* [40] and *C. thermocellum* [43] were constructed allowing for both positive- and counter-selection of desired recombinants. Moreover, methods employing bacterial group II intron for gene deletion [44, 45, 46] and a *Bacillus subtilis* resolvase [47, 48, 49] were developed as universal genetic tools for *Clostridium* species. Genetic systems were also established based on replicative plasmids capa-

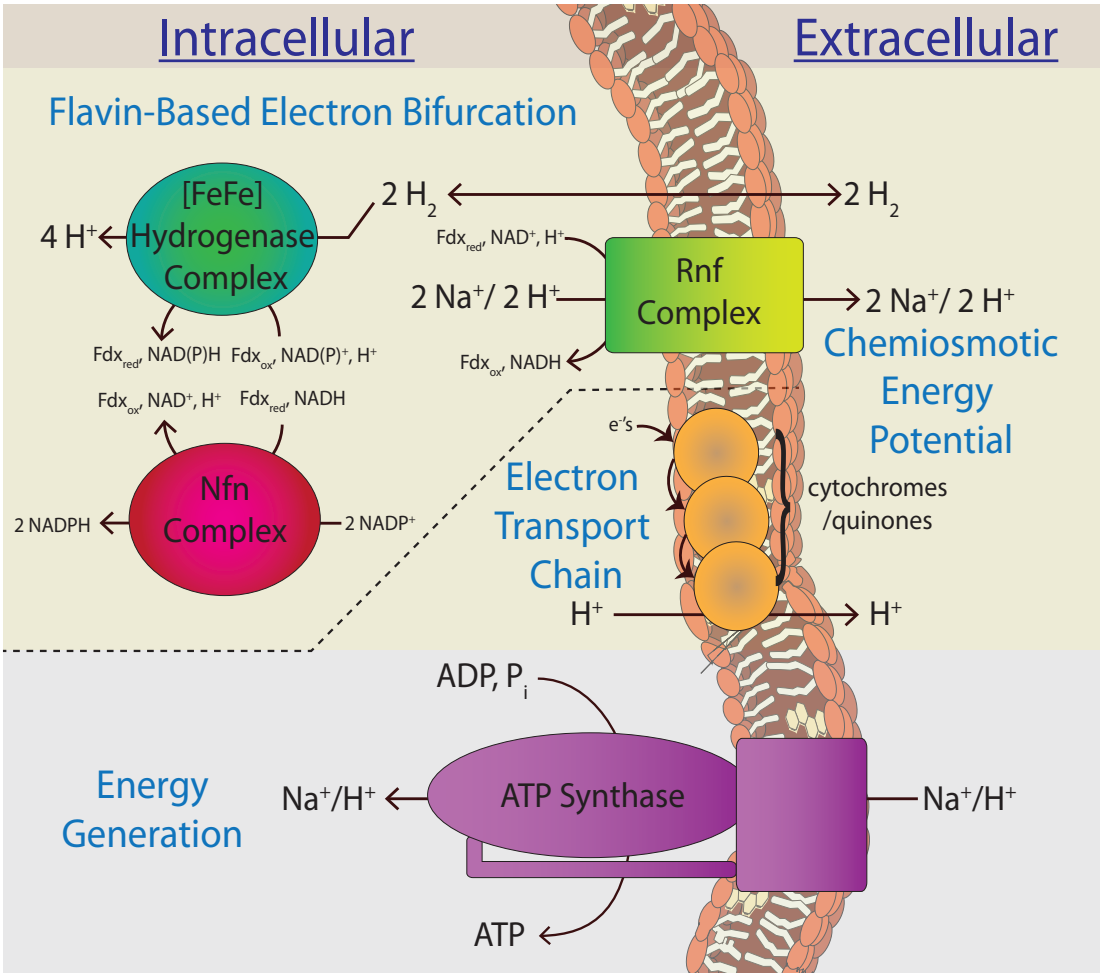


Figure 7.4: **Energy conservation during syngas fermentation.** Shown are examples of protein complexes found in acetogens that conserve energy during syngas fermentation. Syngas fermentation produces no net ATP through substrate level phosphorylation. Thus, a chemiosmotic gradient is utilized to drive ATP synthesis. This gradient is achieved predominantly through ion translocation via electron transport chain proteins or through membrane-bound flavin-based electron bifurcation (FBEB) complexes such as the Rnf. Soluble FBEB complexes like the [FeFe]-hydrogenase and Nfn complex have also been identified in acetogens. These, as well as other complexes (not shown) are important for maintaining cellular redox balance. Energy generation occurs through the ATP synthase. Acetogens such as *Clostridium ljungdahlii* and *Moorella thermoacetica* utilize a H^+ gradient while others like *Acetobacterium woodii* use Na^+.

ble of double crossover chromosomal integration [50], inducible counter-selection for markerless gene deletions/insertions at any desired genomic loci [51], coupled expression of heterologous selectable markers to a chromosomal promoter to select for double crossover events [52], and antisense RNA for protein down-regulation [53]. Reverse genetic tools in the form of transposon mutant libraries were generated using a *mariner*-based system in the autotrophic pathogen *C. difficile* and in *C. perfringens* [54, 55] and a *Tn1545*-based system in *C. cellulolyticum* [56].

7.8 Strain engineering to obtain desired production phenotypes

With the availability of such genetic tools and the recent advances in whole-genome sequencing, combined metabolic engineering and synthetic biology approaches can be applied to accelerate the development of syngas fermentation processes. Boosting the yield and productivity of syngas fermenters and broadening the spectrum of chemicals and fuels that can be produced are immediate goals. As a proof of concept, the deletion of the bifunctional aldehyde/alcohol dehydrogenase in *C. ljungdahlii* resulted in increased acetate yield on the expense of ethanol [39]. The functionality of a construct harboring the known acetone pathway of *C. acetobutylicum* was demonstrated in *C. aceticum*, enabling the latter to produce eight mg/L of acetone from a mixture of H_2 and CO_2 [11, 57]. Similarly, plasmids bearing heterologous genes for the butanol synthesis pathway of *C. acetobutylicum* were introduced into *C. ljungdahlii* allowing for low levels of butanol production [21]. Butanol was subsequently converted to butyrate at the end of the fermentation, indicating the need for further genetic modifications to prevent the loss of butanol. Yet, the reversibility of this and similar alcohol forming enzymatic reactions was exploited by providing external short-chain carboxylic acids and syngas to *C. ljungdahlii* for the production of alcohols [58].

In addition to improved yield and product-range of acetogens, engineering strategies targeting other limitations identified at industrial-scale may be considered. Among these limitations are low syngas kLa, low cell-biomass, sporulation, as

well as substrate and product inhibition. Engineering syngas fermenting microorganisms for enhanced biofilm development can help overcome their inherently low biomass yield while enabling the use of reactors with enhanced gas mass transfer rates such as air-lift reactors [59] or membrane biofilm reactors [60] as well as other types of biofilm-based reactors suitable for syngas fermentation [57, 15]. More effective biofilms can be formed by increased exopolysaccharide production [61] or by manipulating other factors that are known to modulate biofilm formation in *Clostridium* species [62]. Longer fermentation increases the possibility of spore formation. Inactivating Spo0A, the master regulator of sporulation in *Clostridium* species [63], was thought to be the obvious strategy for abolishing sporulation [64]. However, Spo0A's involvement in solvent production [65, 66] and biofilm formation [62, 63, 67] will necessitate tuned expression rather than complete inactivation. Alternatively, sporulation regulators downstream to Spo0A can be targeted for creating asporogenous strains, as was demonstrated for *C. acetobutylicum* [64] and *C. phytofermentans* [68].

Phenotypes with improved fitness for syngas fermentation can also be obtained by adaptive laboratory evolution (ALE) [69, 70, 71]. Various ALE approaches can be employed to obtain highly desirable industrial traits such as optimal growth during syngas fermentation and tolerance to high concentrations of substrates (e.g., CO) or products (e.g., ethanol). For instance, ALE approaches were used to generate strains of *Butyribacterium methylotrophicum* that were adapted to growth in a pure CO-headspace [72]. Ethanol tolerant strains of *Escherichia coli* and subsequent genome-scale analysis provided insights into the metabolic and regulatory mechanisms that caused that phenotype to emerge [73, 74]. A similar approach can be applied to syngas fermenting production strains.

7.9 Rational strain design and process optimization through a systems-level approach

Currently, little is known about the possibility of completely redirecting the metabolic fluxes from acetate to other products during syngas fermentation or its

effect on cellular energetics. Recently, Tyrurin and co-workers reported in a series of publications on completely abolishing acetate production in an undisclosed *Clostridium* strain, in favor of acetone [75], ethanol [76, 77], butanol [78], mevalonate [78], methanol, or formate [79] production during autotrophic growth on syngas. However, eliminating acetate synthesis by deletion of the phosphotransacetylase and/or acetate kinase (*ack*) genes presumably prevents the concurrent synthesis of ATP (via SLP) by Ack. The impact this change has on the growth energetics in these strains is not yet fully understood.

Perhaps the most effective way for rational strain design is through the use of genome-scale metabolic models. Genome-scale models have successfully been implemented in rational strain design [80, 81]. Recently, the first comprehensive reconstruction of metabolism in an acetogen has been generated [22]. Using this *C. ljungdahlii* model, simulation of a Δack mutant predicted conditional essentiality. *In silico* growth was observed under heterotrophic conditions and under autotrophic growth with CO as the electron donor. However, growth on H_2 as the sole electron donor was essential depending on the cofactor specificity of the [FeFe]-hydrogenase [22]. This conditional essentiality is attributed to differences in the redox state of the different electron carrier pools and their ability to contribute to FBEB-based energy conservation to compensate for the lost ATP generated during acetate synthesis. Thus, it is extremely important to detail the exact mechanisms of energy conservation, including the stoichiometry of the ATPase reaction, in the metabolic network reconstruction for more realistic phenotype predictions. Another model of the WL pathway in *C. ljungdahlii* was used to determine the ATP yield per mol CO consumed and the proton translocation per electron transfer of the Rnf complex [58]. Lastly, metabolic modeling has been used to optimize media formulations, for instance, based on energy demands [82]. These models can help optimize and reduce the cost of media, which is a challenge the needs to be addressed prior to commercial deployment of syngas fermentation [83, 84].

7.10 Summary and opportunities

Fermentation of syngas into biochemicals using acetogenic microorganisms offers an important economic potential for biofuel and commodity chemical production. Gasification allows for the processing of virtually all types of organic waste (e.g., industrial or municipal) into syngas. The potential of syngas fermentation is evident by the advent of large-scale projects. LanzaTech is working with steel manufactures [85] and coal producers [86] in China to make liquid fuels. BioMCN is converting glycerine to syngas which is fermented into bio-methanol and has constructed a 200,000 ton/year pilot scale production unit [87]. Coskata [88] is commercializing the production of fuels and chemicals using a wide variety of biomass sources through syngas fermentation. They have built a demonstration-scale production facility as part of a feasibility study. Lastly, SYNPOL [89], a large research project funded by the EU, is focused on the production of biopolymers via syngas fermentation. In order to fully utilize cheap carbon sources, however, construction of novel and optimized syngas fermenting strains for the production of biochemicals and biofuels is needed. Recent advances in our knowledge of energy conservation in acetogens and the development of new molecular biology tools provide a foundation for strain design strategies.

Recently developed genetic tools can be leveraged to achieve desirable phenotypic traits. Production of compounds of interest must be maximized under the conditions optimal for syngas fermentation. Further improvements will likely involve media optimization and evolution of strains for growth in minimal media. The formation of enhanced biofilm formers for some types of fermentations is highly desirable. Tolerance towards inhibitory substrate and product concentrations as well as other complex phenotypes may be achieved through ALE approaches. Lastly, genome-scale models for acetogens provide a valuable tool to account for the intricacies of acetogen metabolism when designing optimal production strains. With many of these enabling technologies now available, a significant growth in the research and metabolic engineering of acetogens can be expected to better utilize syngas for production of both biofuels and commodity chemicals.

7.11 Acknowledgements

AAZ, ATN, and KZ acknowledge support from The Novo Nordisk Foundation. HL is supported through the National Science Foundation Graduate Research Fellowship under grant DGE1144086.

Chapter 7 in full is a reprint of a published manuscript: Latif H, Zeidan AA, Nielsen AT, Zengler K. Trash to treasure: production of biofuels and commodity chemicals via syngas fermenting microorganisms. Curr Opin Biotechnol. 2014 Jun;27:79-87. doi: 10.1016/j.copbio.2013.12.001. The dissertation author was the primary author of this paper responsible for the research. The other authors are Ahmad A. Zeidan, Alex T. Nielsen, and Karsten Zengler.

7.12 Bibliography

[1] Commission EE (2009) Directive 2009/28/ec of the european parliament and of the council of 23 april 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing directives 2001/77/ec and 2003/30. Official Journal of the European Union Belgium .

[2] Law P (2007) Law 110-140, energy independence and security act of 2007. US Government Printing Office .

[3] Sims RE, Mabee W, Saddler JN, Taylor M (2010) An overview of second generation biofuel technologies. Bioresource Technology 101: 1570–1580.

[4] Havlk P, Schneider UA, Schmid E, Bttcher H, Fritz S, Skalsk R, Aoki K, Cara SD, Kindermann G, Kraxner F, Leduc S, McCallum I, Mosnier A, Sauer T, Obersteiner M (2011) Global land-use implications of first and second generation biofuel targets. Energy Policy 39: 5690–5702.

[5] Hamelinck CN, Hooijdonk Gv, Faaij AP (2005) Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle-and long-term. Biomass and Bioenergy 28: 384–410.

[6] Betts W, Dart R, Ball A, Pedlar S (1991) Biosynthesis and structure of lignocellulose. In: Biodegradation, Springer. pp. 139–155. URL http://link.springer.com/chapter/10.1007/978-1-4471-3470-1_7.

[7] Wang X, Padgett JM, De la Cruz FB, Barlaz MA (2011) Wood biodegradation in laboratory-scale landfills. Environmental Science and Technology 45: 6864–6871.

[8] Bugg TD, Ahmad M, Hardiman EM, Singh R (2011) The emerging role for bacteria in lignin degradation and bio-product formation. Current Opinion in Biotechnology 22: 394–400.

[9] Richardson Y, Blin J, Julbe A (2012) A short overview on purification and conditioning of syngas produced by biomass gasification: catalytic strategies, process intensification and new concepts. Progress in Energy and Combustion Science 38: 765–781.

[10] Henstra AM, Sipma J, Rinzema A, Stams AJM (2007) Microbiology of synthesis gas fermentation for biofuel production. Current Opinion in Biotechnology 18: 200–206.

[11] Schiel-Bengelsdorf B, Dürre P (2012) Pathway engineering and synthetic biology using acetogens. FEBS Letters 586: 2191–2198.

[12] Munasinghe PC, Khanal SK (2010) Syngas fermentation to biofuel: evaluation of carbon monoxide mass transfer coefficient (kla) in different reactor configurations. Biotechnology Progress 26: 1616–1621.

[13] Drake HL, Gössner AS, Daniel SL (2008) Old acetogens, new light. Annals of the New York Academy of Sciences 1125: 100–128.

[14] Mohammadi M, Najafpour GD, Younesi H, Lahijani P, Uzir MH, Mohamed AR (2011) Bioconversion of synthesis gas to second generation biofuels: a review. Renewable and Sustainable Energy Reviews 15: 4255–4273.

[15] Munasinghe PC, Khanal SK (2010) Biomass-derived syngas fermentation into biofuels: Opportunities and challenges. Bioresource Technology 101: 5013–5022.

[16] Martin WF (2012) Hydrogen, metals, bifurcating electrons, and proton gradients: the early evolution of biological energy conservation. FEBS Letters 586: 485–493.

[17] Buckel W, Thauer RK (2013) Energy conservation via electron bifurcating ferredoxin reduction and proton/na(+) translocating ferredoxin oxidation. Biochimica et Biophysica Acta 1827: 94–113.

[18] Ragsdale SW (2008) Enzymology of the wood-ljungdahl pathway of acetogenesis. Annals of the New York Academy of Sciences 1125: 129–136.

[19] Ragsdale SW, Pierce E (2008) Acetogenesis and the wood-ljungdahl pathway of co_2 fixation. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1784: 1873–1898.

[20] Diekert G, Wohlfarth G (1994) Metabolism of homocetogens. Antonie Van Leeuwenhoek 66: 209–221.

[21] Köpke M, Held C, Hujer S, Liesegang H, Wiezer A, Wollherr A, Ehrenreich A, Liebl W, Gottschalk G, Dürre P (2010) *Clostridium ljungdahlii* represents a microbial production platform based on syngas. Proceedings of the National Academy of Sciences of the United States of America 107: 13087–13092.

[22] Nagarajan H, Sahin M, Nogales J, Latif H, Lovley DR, Ebrahim A, Zengler K (2013) Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of *Clostridium ljungdahlii*. Microbial Cell Factories 12: 118.

[23] Van der Drift A, Van Doorn J, Vermeulen J (2001) Ten residual biomass fuels for circulating fluidized-bed gasification. Biomass and Bioenergy 20: 45–56.

[24] Das A, Ljungdahl LG (2003) Electron-transport system in acetogens. In: Biochemistry and physiology of anaerobic bacteria, Springer. pp. 191–204. URL http://link.springer.com/content/pdf/10.1007/0-387-22731-8_14.pdf.

[25] Müller V (2003) Energy conservation in acetogenic bacteria. Applied and Environmental Microbiology 69: 6345–6353.

[26] Pierce E, Xie G, Barabote RD, Saunders E, Han CS, Detter JC, Richardson P, Brettin TS, Das A, Ljungdahl LG, Ragsdale SW (2008) The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). Environmental Microbiology 10: 2550–2573.

[27] Biegel E, Schmidt S, González JM, Müller V (2011) Biochemistry, evolution and physiological function of the rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cellular and Molecular Life Sciences 68: 613–634.

[28] Imkamp F, Biegel E, Jayamani E, Buckel W, Müller V (2007) Dissection of the caffeate respiratory chain in the acetogen *Acetobacterium woodii*: identification of an rnf-type nadh dehydrogenase as a potential coupling site. Journal of Bacteriology 189: 8145–8153.

[29] Biegel E, Müller V (2010) Bacterial na^+-translocating ferredoxin:nad^+ oxidoreductase. Proceedings of the National Academy of Sciences of the United States of America 107: 18138–18142.

[30] Biegel E, Schmidt S, Müller V (2009) Genetic, immunological and biochemical evidence for a rnf complex in the acetogen *Acetobacterium woodii*. Environmental Microbiology 11: 1438–1443.

[31] Poehlein A, Schmidt S, Kaster AK, Goenrich M, Vollmers J, Thürmer A, Bertsch J, Schuchmann K, Voigt B, Hecker M, Daniel R, Thauer RK, Gottschalk G, Müller V (2012) An ancient pathway combining carbon dioxide fixation with the generation and utilization of a sodium ion gradient for atp synthesis. PLoS One 7: e33439.

[32] Tremblay PL, Zhang T, Dar SA, Leang C, Lovley DR (2012) The rnf complex of *Clostridium ljungdahlii* is a proton-translocating ferredoxin:nad$^+$ oxidoreductase essential for autotrophic growth. MBio 4: e00406–e00412.

[33] Schuchmann K, Müller V (2012) A bacterial electron-bifurcating hydrogenase. Journal of Biological Chemistry 287: 31165–31171.

[34] Huang H, Wang S, Moll J, Thauer RK (2012) Electron bifurcation involved in the energy metabolism of the acetogenic bacterium *Moorella thermoacetica* growing on glucose or h_2 plus co_2. Journal of Bacteriology 194: 3689–3699.

[35] Wang S, Huang H, Kahnt J, Thauer RK (2013) A reversible electron-bifurcating ferredoxin- and nad-dependent [fefe]-hydrogenase (hydabc) in *Moorella thermoacetica*. Journal of Bacteriology 195: 1267–1275.

[36] Wang S, Huang H, Kahnt J, Mueller AP, Köpke M, Thauer RK (2013) Nadp-specific electron-bifurcating [fefe]-hydrogenase in a functional complex with formate dehydrogenase in *Clostridium autoethanogenum* grown on co. Journal of Bacteriology 195: 4373–4386.

[37] Wang S, Huang H, Moll J, Thauer RK (2010) Nadp$^+$ reduction with reduced ferredoxin and nadp$^+$ reduction with nadh are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. Journal of Bacteriology 192: 5115–5123.

[38] Bertsch J, Parthasarathy A, Buckel W, Müller V (2013) An electron-bifurcating caffeyl-coa reductase. Journal of Biological Chemistry 288: 11304–11311.

[39] Leang C, Ueki T, Nevin KP, Lovley DR (2013) A genetic system for *Clostridium ljungdahlii*: a chassis for autotrophic production of biocommodities and a model homoacetogen. Applied and Environmental Microbiology 79: 1102–1109.

[40] Kita A, Iwasaki Y, Sakai S, Okuto S, Takaoka K, Suzuki T, Yano S, Sawayama S, Tajima T, Kato J, Nishio N, Murakami K, Nakashimada Y (2013) Development of genetic transformation and heterologous expression system in carboxydotrophic thermophilic acetogen *Moorella thermoacetica*. Journal of Bioscience and Bioengineering 115: 347–352.

[41] Kita A, Iwasaki Y, Yano S, Nakashimada Y, Hoshino T, Murakami K (2013) Isolation of thermophilic acetogens and transformation of them with the *pyrF* and *kan(r)* genes. Bioscience, Biotechnology, and Biochemistry 77: 301–306.

[42] Iwasaki Y, Kita A, Sakai S, Takaoka K, Yano S, Tajima T, Kato J, Nishio N, Murakami K, Nakashimada Y (2013) Engineering of a functional thermostable kanamycin resistance marker for use in *Moorella thermoacetica* atcc39073. FEMS Microbiology Letters 343: 8–12.

[43] Tripathi SA, Olson DG, Argyros DA, Miller BB, Barrett TF, Murphy DM, McCool JD, Warner AK, Rajgarhia VB, Lynd LR, Hogsett DA, Caiazza NC (2010) Development of *pyrF*-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a *pta* mutant. Applied and Environmental Microbiology 76: 6591–6599.

[44] Heap JT, Kuehne SA, Ehsaan M, Cartman ST, Cooksley CM, Scott JC, Minton NP (2010) The clostron: Mutagenesis in *Clostridium* refined and streamlined. Journal of Microbiological Methods 80: 49–55.

[45] Heap JT, Pennington OJ, Cartman ST, Carter GP, Minton NP (2007) The clostron: a universal gene knock-out system for the genus *Clostridium*. Journal of Microbiological Methods 70: 452–464.

[46] Kuehne SA, Heap JT, Cooksley CM, Cartman ST, Minton NP (2011) Clostron-mediated engineering of *Clostridium*. Methods in Molecular Biology 765: 389–407.

[47] Bi C, Jones SW, Hess DR, Tracy BP, Papoutsakis ET (2011) Spoiie is necessary for asymmetric division, sporulation, and expression of σ^f, σ^e, and σ^g but does not control solvent production in *Clostridium acetobutylicum* atcc 824. Journal of Bacteriology 193: 5130–5137.

[48] Jones SW, Tracy BP, Gaida SM, Papoutsakis ET (2011) Inactivation of σ^f in *Clostridium acetobutylicum* atcc 824 blocks sporulation prior to asymmetric division and abolishes σ^e and σ^g protein expression but does not block solvent formation. Journal of Bacteriology 193: 2429–2440.

[49] Papoutsakis ET, Tracy BP (2009). Methods and compositions for genetically engineering clostridia species. URL http://www.google.com/patents/US20100075424. US Patent App. 12/437,985.

[50] Harris LM, Welker NE, Papoutsakis ET (2002) Northern, morphological, and fermentation analysis of *spo0A* inactivation and overexpression in *Clostridium acetobutylicum* atcc 824. Journal of Bacteriology 184: 3586–3597.

[51] Al-Hinai MA, Fast AG, Papoutsakis ET (2012) Novel system for efficient isolation of *Clostridium* double-crossover allelic exchange mutants enabling markerless chromosomal gene deletions and dna integration. Applied and Environmental Microbiology 78: 8112–8121.

[52] Heap JT, Ehsaan M, Cooksley CM, Ng YK, Cartman ST, Winzer K, Minton NP (2012) Integration of dna into bacterial chromosomes from plasmids without a counter-selection marker. Nucleic Acids Research 40: e59–e59.

[53] Tummala SB, Welker NE, Papoutsakis ET (2003) Design of antisense rna constructs for downregulation of the acetone formation pathway of *Clostridium acetobutylicum*. Journal of Bacteriology 185: 1923–1934.

[54] Cartman ST, Minton NP (2010) A *mariner*-based transposon system for *n vivo* random mutagenesis of *Clostridium difficile*. Applied and Environmental Microbiology 76: 1103–1109.

[55] Liu H, Bouillaut L, Sonenshein AL, Melville SB (2013) Use of a *mariner*-based transposon mutagenesis system to isolate *Clostridium perfringens* mutants deficient in gliding motility. Journal of Bacteriology 195: 629–636.

[56] Blouzard JC, Valette O, Tardif C, de Philip P (2010) Random mutagenesis of *Clostridium cellulolyticum* by using a *Tn1545* derivative. Applied and Environmental Microbiology 76: 4546–4549.

[57] Daniell J, Köpke M, Simpson SD (2012) Commercial biomass syngas fermentation. Energies 5: 5372–5417.

[58] Perez JM, Richter H, Loftus SE, Angenent LT (2013) Biocatalytic reduction of short-chain carboxylic acids into their corresponding alcohols with syngas fermentation. Biotechnology and Bioengineering 110: 1066–1077.

[59] Merchuk JC, Siegel MH (1988) Air-lift reactors in chemical and biological technology. Journal of Chemical Technology and Biotechnology 41: 105–120.

[60] Nerenberg R, Rittmann BE (2004) Hydrogen-based, hollow-fiber membrane biofilm reactor for reduction of perchlorate and other oxidized contaminants. Water Science and Technology 49: 223–230.

[61] Leang C, Malvankar NS, Franks AE, Nevin KP, Lovley DR (2013) Engineering *Geobacter sulfurreducens* to produce a highly cohesive conductive matrix with enhanced capacity for current production. Energy & Environmental Science 6: 1901–1908.

[62] Dapa T, Leuzzi R, Ng YK, Baban ST, Adamo R, Kuehne SA, Scarselli M, Minton NP, Serruto D, Unnikrishnan M (2013) Multiple factors modulate biofilm formation by the anaerobic pathogen *Clostridium difficile*. Journal of Bacteriology 195: 545–555.

[63] Dawson LF, Valiente E, Faulds-Pain A, Donahue EH, Wren BW (2012) Characterisation of *Clostridium difficile* biofilm formation, a role for spo0a. PloS One 7: e50527.

[64] Tracy BP, Jones SW, Fast AG, Indurthi DC, Papoutsakis ET (2012) Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. Current Opinion in Biotechnology 23: 364–381.

[65] Deakin LJ, Clare S, Fagan RP, Dawson LF, Pickard DJ, West MR, Wren BW, Fairweather NF, Dougan G, Lawley TD (2012) The *Clostridium difficile spo0A* gene is a persistence and transmission factor. Infection and Immunity 80: 2704–2711.

[66] Sillers R, Chow A, Tracy B, Papoutsakis ET (2008) Metabolic engineering of the non-sporulating, non-solventogenic *Clostridium acetobutylicum* strain m5 to produce butanol without acetone demonstrate the robustness of the acid-formation pathways and the importance of the electron balance. Metabolic engineering 10: 321–332.

[67] Hamon MA, Lazazzera BA (2001) The sporulation transcription factor spo0a is required for biofilm development in *Bacillus subtilis*. Molecular Microbiology 42: 1199–1209.

[68] Blanchard J, Fabel J, Leschine S, Petit E (2009). Methods and compositions for regulating sporulation. URL http://www.google.com/patents/US20100105114. US Patent App. 12/483,118.

[69] Winkler J, Reyes LH, Kao KC (2013) Adaptive laboratory evolution for strain engineering. Methods in Molecular Biology 985: 211–222.

[70] Conrad TM, Lewis NE, Palsson BØ (2011) Microbial laboratory evolution in the era of genome-scale science. Molecular Systems Biology 7: 509.

[71] Portnoy VA, Bezdan D, Zengler K (2011) Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering. Current Opinion in Biotechnology 22: 590–594.

[72] Shen GJ, Shieh JS, Grethlein A, Jain M, Zeikus J (1999) Biochemical basis for carbon monoxide tolerance and butanol production by *Butyribacterium methylotrophicum*. Applied Microbiology and Biotechnology 51: 827–832.

[73] Goodarzi H, Bennett BD, Amini S, Reaves ML, Hottes AK, Rabinowitz JD, Tavazoie S (2010) Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in *E. coli*. Molecular Systems Biology 6: 378.

[74] Horinouchi T, Tamaoka K, Furusawa C, Ono N, Suzuki S, Hirasawa T, Yomo T, Shimizu H (2010) Transcriptome analysis of parallel-evolved *Escherichia coli* strains under ethanol stress. BMC Genomics 11: 579.

[75] Berzin V, Kiriukhin M, Tyurin M (2012) Selective production of acetone during continuous synthesis gas fermentation by engineered biocatalyst *Clostridium* sp. macet113. Letters in Applied Microbiology 55: 149–154.

[76] Berzin V, Kiriukhin M, Tyurin M (2012) Cre-lox66/lox71-based elimination of phosphotransacetylase or acetaldehyde dehydrogenase shifted carbon flux in acetogen rendering selective overproduction of ethanol or acetate. Applied Biochemistry and Biotechnology 168: 1384–1393.

[77] Berzin V, Kiriukhin M, Tyurin M (2012) Elimination of acetate production to improve ethanol yield during continuous synthesis gas fermentation by engineered biocatalyst *Clostridium* sp. mtetoh550. Applied Biochemistry and Biotechnology 167: 338–347.

[78] Berzin V, Tyurin M, Kiriukhin M (2013) Selective n-butanol production by *Clostridium* sp. mtbutoh1365 during continuous synthesis gas fermentation due to expression of synthetic thiolase, 3-hydroxy butyryl-coa dehydrogenase, crotonase, butyryl-coa dehydrogenase, butyraldehyde dehydrogenase, and nad-dependent butanol dehydrogenase. Applied Biochemistry and Biotechnology 169: 950–959.

[79] Tyurin M, Kiriukhin M (2013) Selective methanol or formate production during continuous co_2 fermentation by the acetogen biocatalysts engineered via integration of synthetic pathways using tn7-tool. World Journal of Microbiology & Biotechnology 29: 1611–1623.

[80] Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nature Biotechnology 26: 659–667.

[81] Lee JW, Na D, Park JM, Lee J, Choi S, Lee SY (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nature Chemical Biology 8: 536–546.

[82] Park JH, Kim TY, Lee KH, Lee SY (2011) Fed-batch culture of *Escherichia coli* for l-valine production based on *in silico* flux response analysis. Biotechnology and Bioengineering 108: 934–946.

[83] Gao J, Atiyeh HK, Phillips JR, Wilkins MR, Huhnke RL (2013) Development of low cost medium for ethanol production from syngas by *Clostridium ragsdalei*. Bioresource Technology 147: 508–515.

[84] Richter H, Martin ME, Angenent LT (2013) A two-stage continuous fermentation system for conversion of syngas into ethanol. Energies 6: 3987–4000.

[85] LanzaTech (2011). Chinese steel miller commercializing lanzatechs clean energy technology.

[86] LanzaTech (2011). Lanzatech joins chinas yankuang group on coal to fuel project.

[87] BioMCN (2013). http://www.biomcn.eu/. URL http://www.biomcn.eu/.

[88] Coskata (2013). www.coskata.com. URL www.coskata.com.

[89] SYNPOL (2013). The synpol project. URL http://www.synpol.org/.

Chapter 8

Conclusion

8.1 Model organisms and their knowledgebases.

Over the past century, biology has largely focused on characterizing a select set of model organisms [1]. These model organisms (e.g., *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, mouse, human) have been the beneficiaries of decades of detailed molecular, genetic, and functional characterization efforts. The wealth of information determined in these model systems has proven to be incredibly valuable to our understanding of the fundamentals of living systems and provided a basis for comparative analysis to non-model organisms.

Model organisms are the logical choice for the initial development of genome-scale methodologies since they can be benchmarked against numerous small-scale experiments. Therefore, the explosion of omics approaches has greatly expanded the information available for model organisms spanning various levels of cellular organization. This wealth of information is perhaps best captured through organized databases such as Ecocyc (*Escherichia coli*), Saccharomyces Genome Database (*Saccharomyces cerevisiae*), FlyBase (*Drosophila melanogaster*), and ENCODE (human). These knowledgebases are widely used by researchers in a variety of fields in the life sciences. However, the development of such databases is costly and time consuming. For example, the pilot and initial phases of the ENCODE project, whose goal is to annotate all human genomic features, took approximately $300 million, more than 400 researchers, and 9 years to complete [2].

8.2 Closing the knowledge gap using an integrated, multi-omic characterization workflow.

Microorganisms are the most abundant and diverse forms of life on our planet. Genome sequencing of closely related genomes has revealed a remarkable level of diversity with respect to their gene content [3]. The diverse physiological capabilities of microorganisms is only partially captured by characterization studies conducted in model organisms thus far. Until recently, the resources needed to elucidate non-model organisms with interesting physiological properties (such as the acetogens discussed in Chapter 7) have been prohibitive.

The multi-omic characterization approach outlined here overcomes the barriers preventing detailed molecular characterization of microorganisms. We have established a generalized framework in which the knowledge gap for non-model organisms can be closed at manageable costs and in a reasonable amount of time. This workflow is reliant solely on next-generation sequencing approaches to interrogate the genome, transcriptome, proteome, and regulatory networks at the transcriptional and post transcriptional levels. As described in Chapter 1, the multi-omic data integration process is rooted in four-steps: data generation, data processing, data integration, and data analysis. In comparison with previous multi-omic data integration efforts [4, 5] numerous modifications were made that enable an expanded array of cellular features to be elucidated including updating the experimental toolbox and complemented with bioinformatics approaches (Figure 8.1).

8.3 The benefits of a consolidated data generation platform.

Next-generation sequencing techniques are applicable to nearly all living systems not only in isolation, but even in complex communities such as the human microbiome and environmental samples. Next-generation sequencing does not require *a priori* knowledge of the molecular composition of the organisms of interest and the experimental protocols detailed in this workflow are readily ported from one microor-

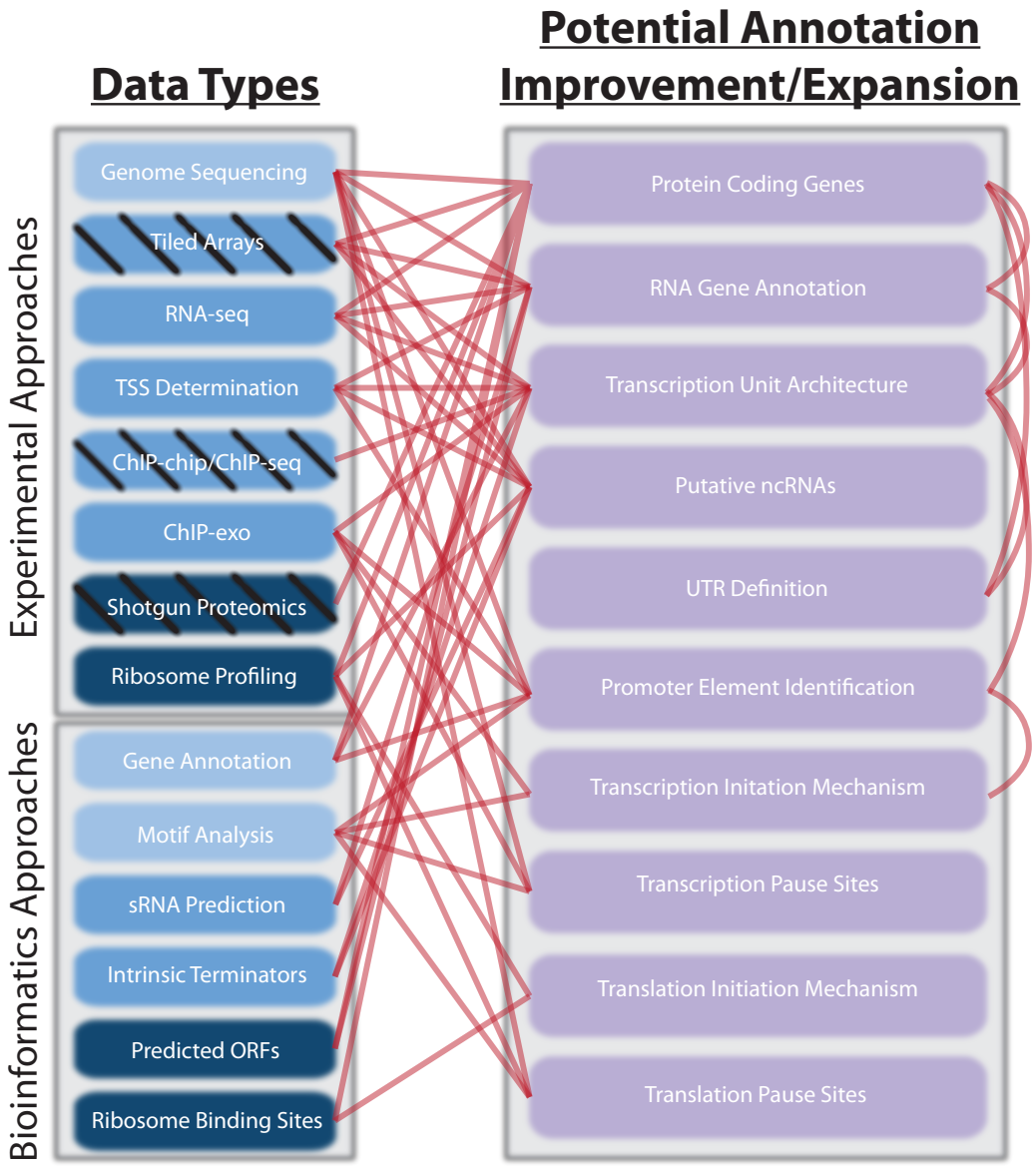


Figure 8.1: **Systems level workflow for multi-omic data integration.** The multi-omic workflow developed in this thesis elucidates an expanded list of annotated features. Data types generated include those that are experimental and those that are bioinformatic. The experimental toolbox has improved methodologies that leverage next-generation sequencing to elucidate the genome (light blue), transcriptome (blue), and proteome (dark blue). The experimental data and bioinformatics predictions are processed and, together, integrated to determine specific cellular features. Connections between data types and annotated features is shown in red.

ganism to another. Therefore, minimal customization is needed to tailor protocols for the study of different organisms.

Another advantage of using next-generation of sequencing is in the cost of data generation. As was detailed in Chapter 1, the cost of sequencing has dropped to the point where entire bacterial genomes can be sequenced for a few dollars as opposed to several thousands of dollars. Transcript sequencing on next-generation sequencing platforms eliminates the need for the design and synthesis of custom high-density microarrays. The effort needed to produce high-density arrays often is justifiable only when large, bulk orders are placed. Depending on the microorganisms being studied this can be easily achieved, but for others the conditions cannot be varied greatly and limit the scope of transcriptome studies. The same argument holds for the investigation of regulation using microarrays to elucidate protein/nucleic acid interactions. Lastly, whole-genome shotgun proteomics requires access to highly specialized equipment and is a relatively expensive assay to perform compared with ribosome profiling (see Chapter 6).

However, the main advantage of next-generation sequencing in the context of this workflow is in the quality and resolution of the datasets generated. Whole genome sequencing with great depth allows for identification of genetic variants with high statistical certainty. Determination of the bounds of transcripts and the precise location of the transcription start site at genome-scale can only be determined using next-generation sequencing approaches. The improvement of ChIP-exonuclease (ChIP-exo) over ChIP-chip and ChIP-seq are apparent in Chapter 5. ChIP-exo yields data with $\approx$30 bp resolution as opposed to 200-300 bp and $>$1 kb resolution for ChIP-seq and ChIP-chip respectively. This method also gives substantial improvements with regard to signal-to-noise and reduces the number of false positives compared with ChIP-seq. Ribosome profiling also provides greater information on the process of translation with data spanning the entire length of protein coding sequencing. This not only helps improve gene annotation but expands it by including potential translational pause sites. The method also predicts the stoichiometric ratios for heteroprotein complexes. This feature is extremely valuable when studying heteroprotein complexes. For instance, energy generation in acetogens under autotrophic conditions

is dependent on ATP synthase, Rnf, Nfn, and hydrogenases working in concert. Ribosome profiling can examine the functional state of these complexes and from this the interdependencies among them can be unveiled.

8.4 Applications of microbial knowledgebases.

Ultimately, the framework presented here for integration of multi-omic datasets produces a knowledgebase that is rooted in the annotation of cellular features projected onto the genome. Figure 8.2 illustrates some of the applications for which this knowledgebase can be used that have been applied in this thesis. An accurate and complete genome sequence is at the heart of our knowledgebase. Chapter 2 briefly highlights two cases where next-generation sequencing has been leveraged to improve genome assemblies. For one of these cases, *Thermotoga maritima*, the genome organization was then revealed using a multi-omic integration strategy that relied on experimental as well as bioinformatics approaches (Chapter 3). The characterization of this hyperthermophilic, phylogenetically deep-branching bacteria not only revealed the composition of cellular features necessary for gene expression but also yielded novel insights into the life style of this organism. Using cross-species comparative analysis, increased sequence conservation of promoter elements and ribosome binding sites was revealed. This potentially facilitates gene expression at the extremely high temperature environments inhabited by Thermotogales. *T. maritima* was also the subject of an adaptive evolution study (Chapter 4). The goal of this study was to understand the genotype-to-phenotype relationship with respect to metabolism of a suboptimal substrate—glucose. Applying laboratory evolution enabled for the selection of mutant cultures that were capable of superior growth on glucose relative to the wild type strain. Using a multi-omic approach, specific genetic mutations were identified to genes involved in glucose uptake that were then linked directly with a change in gene expression and ultimately phenotypic behavior. One of the two operons impact was located in the ≈9 kb gap region of the original *T. maritima* reference genome assembly. Therefore, the link between the genotype and phenotype was enabled by having both an accurate genome sequence and a well annotated genome

organization.

Next-generation sequencing methods that detail mechanisms of molecular interactions were conducted in Chapter 5 and Chapter 6. In Chapter 5, ChIP-exonuclease, a novel approach for obtaining high resolution protein/DNA interactions at genome-scale and *in vivo*, was conducted on the canonical transcriptional activator Crp. This study revealed distinct molecular states of transcription initiation. ChIP-exo data on σ^{70} suggests that the protein/DNA interactions captured by ChIP is predominantly of RNA polymerase stable intermediate that occurs after open complex formation but before the ternary elongation complex escapes the promoter. Crp binding profiles were found to closely match those obtained for σ^{70} suggesting that the transcription factor forms short-lived interactions with operator DNA but retains longer, more stable interactions with RNA polymerase post-recruitment. The molecular characterization of Crp, in conjunction with RNA-seq studies, facilitated elucidation of the Crp regulon. In addition to finding novel regulatory sites, the data helped expand a biological phenomena that correlated cellular growth rate with the quality of the carbon source presented, the so-called 'C-line' [6].

Chapter 6 provides a streamlined methodology for applying ribosome profiling to microorganisms. This methodology is critical to updating and centralizing the multi-omic workflow to be solely reliant on next-generation sequencing. Ribosome profiling gives detailed molecular information as to the exact position of actively translating ribosomes.

8.5 Perspective on the future of multi-omic data integration.

Collectively, the information revealed through sequencing the genome, revealing the genome organization, determining the genotype-to-phenotype relationship, understanding bacterial evolution, and characterizing interactions among cellular components produce valuable inputs into network reconstructions. Though often focused on metabolism, genome-scale network reconstructions are now capturing multiple scales of cellular activity and modeling them. Recently, a whole-cell model

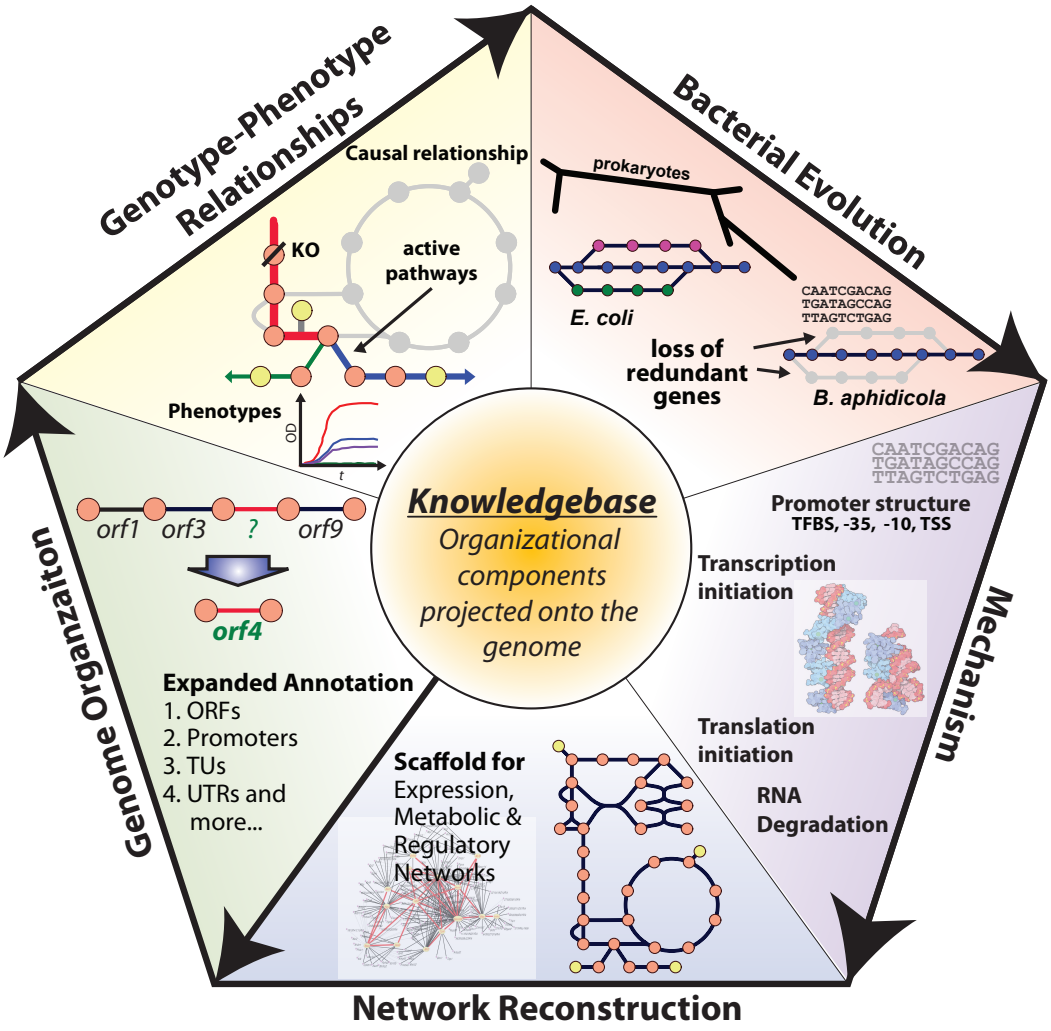


Figure 8.2: **Applications of genome-scale knowledge bases.** Illustrated are five areas where a knowledgebase rooted in the fundamental understanding of the genome organization are detailed. First, the elucidation of the genome organization results in the unveiling of genomic features beyond protein coding regions and structural RNAs including a plethora of elements integral to gene expression and regulation. The genotype-to-phenotype relationship can subsequently be explored and approaches like adaptive evolution can be applied to flush out its intricacies. High-quality omics datasets can be used to determine molecular mechanisms and interactions to describe cellular process. This ultimately can be represented through systems level network reconstructions.

was developed for *Mycoplasma genitalium* [7]. Furthermore, new models that link metabolic networks explicitly with expression have been developed for *T. maritima* [8] and *E. coli* [9]. These models account for the cost of macromolecule synthesis and

can determine optimal phenotypic states by minimizing costs. The reconciliation of the genome with transcription, translation, and regulation is on the horizon thanks in part to advances made in multi-omic data integration workflows like those presented here.

However, we must take caution to work smarter rather than working harder. As next-generation sequencing continues to evolve and gain a greater foothold in research, the focus will shift less on 'what can be sequenced' and more on 'how should we analyze what was sequenced'. Genome-scale data generation is becoming more and more routine. Therefore, efforts must be repurposed to developing novel data analysis approaches. A 2011 publication by Sboner et al. [10] revealed that, in comparison, the costs of sample collection and data analysis would greatly exceed those of sequencing and primary data processing. This is due to the fact that, while the reduction in costs for sequencing have declined at a pace that is faster than Moore's law, the costs associated with sample collection and data analysis have not dropped rapidly. Therefore, the multi-omic data integration pipeline developed here must be complemented with an array of tools to rapidly detect and automate the processes of cellular feature annotation. Currently, substantial manual effort is needed to analyze integrated datasets to validate bioinformatics approaches used. This bottleneck can be overcome by developing algorithms that can, with great precision, extract information from high resolution datasets and integrate them. For instance, transcription start site information can be coupled with ChIP-exo datasets for σ^{70} to reveal which stable intermediates are present in the initial transcribing complex. In parallel, promoter elements can be identified using predictive algorithms of sequence motifs and transcription factor binding sites. The promoter architecture can then be analyzed with gene expression data to reveal links between transcription initiation and transcription rate. The translational output from ribosome profiling can then be compared with transcript abundance. These, along with numerous other components and interactions can then serve as inputs for the construction of a whole-cell model. Automating these types of interconnections to extract physiological information will be the next great challenge for genome-scale science. When achieved, a new standard will be established for characterizing microbial systems to go beyond a one

dimensional view of genome sequences towards one of a multidimensional genome organization.

8.6 Bibliography

[1] Davis RH (2004) The age of model organisms. Nature Reviews Genetics 5: 69–76.

[2] Pennisi E (2012) Genomics. encode project writes eulogy for junk dna. Science 337: 1159, 1161.

[3] Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology .

[4] Qiu Y, Cho BK, Park YS, Lovley D, Palsson BO, Zengler K (2010) Structural and operational complexity of the *Geobacter sulfurreducens* genome. Genome Research 20: 1304–1311.

[5] Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO (2009) The transcription unit architecture of the *Escherichia coli* genome. Nature Biotechnology 27: 1043–1049.

[6] You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang YP, Lenz P, Yan D, Hwa T (2013) Coordination of bacterial proteome with metabolism by cyclic amp signalling. Nature 500: 301–306.

[7] Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-Garcia N, Glass JI, Covert MW (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150: 389–401.

[8] Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, Palsson BO (2012) *In silico* method for modelling metabolism and gene product expression at genome scale. Nature Communications 3: 929.

[9] O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular Systems Biology 9: 693.

[10] Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB (2011) The real cost of sequencing: higher than you think! Genome Biology 12: 125.