UNIVERSITY OF CALIFORNIA SAN DIEGO

Multi-omic characterization of *E. coli* for the purpose of microbial-based production

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

in

Bioengineering

by

Wun Kiat Justin Tan

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Simpson Joseph
Professor Milton H. Saier
Professor Kun Zhang
Professor Brian M. Zid

2019

The dissertation of Wun Kiat Justin Tan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To Arden, my light, my motivation, and my inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr Bernhard Ø. Palsson for his unwavering and at many times, unexpected, support. Throughout the various highs and lows of my PhD research, my (extremely) brief foray into startups, and the birth of my son, you have always been understanding, generous with advice, and most importantly, on my side.

I would also like to thank my wife, Maxine Tan. You have seen me through my bests and my worsts, my triumphs and my failures. You are my wife, my best friend, and my support. You've never failed to provide a listening ear, help me brainstorm for ideas, and when all else fails, make me go get you a glass of milk.

To the many current and previous members of the Systems Biology Research Group, I thank you. Mallory Embree, Haythem Latif, Ali Ebrahim, Joanne Liu, Janna Tarasova, Karsten Zengler, Teddy and Josh, you took me in when I first joined the lab and gave me a home. To my mentors, Laurence Yang, Elizabeth Brunk, Daniel Zelienski, thank you so much for all the long discussions and guidance. My batchmates Colton Lloyd, Bin Du, James Yurkovich, Jared Broddrick, as well as all the other members of SBRG I have had the priviledge of interacting with in the past 6 years, I would like to thank you for all your help, your advice, and most of all, your friendship. In particular I would like to mention Anand Sastry for all your help with ICA, Connor Olson for your tremendous knowledge about ALE, Pman for always being ready to give ideas, joke around, and being a 30 year old grown man, Richard Szubin for your amazing understanding of molecular biology and all your assistance in the wetlab, Amitesh Anand, for always being ready to lend a hand and give suggestions.

I would also like to thank the members of Center for Biosustainability at Technical University of Denmark. Karoline Fremming, Bjorn Voldborg, Sara Bjorn, Anna Koza, Alexandra

Hoffmeyer, Stefan Kol. Thank you so much for the warmth and hospitality you extended me during my stay there, I will always remember Denmark fondly.

To my ruggers, I am who I am today because of you. I hope we never change.

Lastly, I would like to express my appreciation to my parents, for all the love and support that you have provided me during my 10 years of studies abroad. Thank you for putting me through college, being there every Saturday afternoon to hear about my week, and always welcoming me home with open arms. My siblings, Joshua and Christine, you two have been one of the few constants in my life, and it fills me with joy to know that no matter where life takes us we will always be siblings. I love our conversations and discussions, it constantly amazes me how much the two of you have grown up and developed.

I would also like to thank my funding sources that have supported this work. These include the Novo Nordisk Foundation (NNF10CC1016517), the Technical University of Denmark (2011-3780), the National Science Foundation (EFRI-1332344), the U.S. Department of Energy (DE-SC0008701) and the National Institute of General Medical Science (GM057089)

Chapter 2 is a reprint of material published in: A Ebrahim*, E Brunk*, **J Tan***, EJ O'Brien, DH Kim, R Szubin, JA Lerman, A Lechner, A Sastry, A Bordbar, AM Fiest and BO Palsson. 2016. "Multi-omic data integration enables discovery of hidden biological regularities" *Nature Communications* 7 (13091). The dissertation author was one of three primary authors.

Chapter 3 is a reprint of the material in: **J Tan**, AV Sastry, KS Fremming, SP Bjorn, A Hoffmeyer, SW Seo, BG Voldborg and BO Palsson. 2019. "Independent component analysis of *E. coli*'s transcriptome reveals the cellular processes that respond to heterologous gene expression"

Submitted. The dissertation author is the primary author.

Chapter 4 is a reprint of the material: **J Tan**, CA Olson, JH Park, AV Sastry, PV Phaneuf, L Yang, R Szubin, Y Hefner, AM Feist, BO Palsson "Experimental evolution reveals the genetic basis and systems biology of superoxide stress tolerance" *Submitted*. The dissertation author is the primary author.

VITA

2013	Bachelor of Science in Bioengineering, Biotechnology, University of California San Diego
2019	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

E. R. Pfeiffer, A. T. Wright, A. G. Edwards, J. C. Stowe, K. Mcnall, **J. Tan**, I. Niesman, H. H. Patel, D. M. Roth, J. H. Omens, A. D. McCulloch. 2014. "Caveolae in Ventricular Myocytes are Required for Stretch-Dependent Conduction Slowing." *Journal of Molecular and Cellular Cardiology* 76:265-74.

Chen, P., Torralba, M., **Tan, J.**, Embree, M., Zengler, K., Starkel, P., Pijkeren, JP., DePew, J., Loomba, R., Ho, SB., Bajaj, JS., Mutlu, EA., Keshavarzian, A., Tsukamoto, H., Nelson, KE., Fouts, DE., Schnabl, B. 2015. "Supplementation of Saturated Long-chain Fatty Acids Maintains Intestinal Eubiosis and Reduces Ethanol-induced Liver Injury in Mice." *Gastroenterology* 148(1):203-214.

Haythem Latif, Richard Szubin, **Justin Tan**, Elizabeth Brunk, Anna Lechner, Karsten Zengler, and Bernhard O. Palsson. 2015. A streamlined ribosome profiling protocol for the characterization of microorganisms. *BioTechniques* 58:329-332.

Tan, J., Zuniga, C., Zengler, K. 2015. Unraveling interactions in microbial communities – from co-cultures to microbiomes. *Journal of Microbiology* 53(5):295-305.

Yang, L.*, **Tan, J.***, O Brien, E.J., Monk, J., Kim, D., Li, H.J., Churasanti, P., Ebrahim, A., Lloyd, C.J., Yurkovich, J.T., Du, B., Drager, A., Thomas, A., Sun, Y., Saunders, M.A., Palsson, B.O. 2015. A systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proceedings of the National Academy of Sciences* 112(34):10810-10815.

Bin Du, Daniel C. Zielinski, Erol S. Kavvas, Andreas Drager, **Justin Tan**, Zhen Zhang, Kayla E. Ruggiero, Garri A. Arzumanyan and Bernhard O. Palsson. 2016. Evaluation of rate law approximations in bottom-up kinetic models of metabolism. *BMC Systems Biology* 10:40.

Ali Ebrahim*, Elizabeth Brunk*, **Justin Tan***, Edward J. O Brien, Donghyuk Kim, Richard Szubin, Joshua A. Lerman, Anna Lechner, Anand Sastry, Aarash Bordbar, Adam M. Feist, Bernhard O. Palsson. 2016. Multi-omic data integration enables discovery of hidden biological regularities. *Nature Communications* 7:13091.

Fang X, Sastry A, Mih N, Kim D, **Tan J**, Yurkovich JT, Lloyd CJ, Gao Y, Yang L, Palsson BO. 2017. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proc Natl Acad Sci USA* 114(38): 10286-10291.

L Yang, Mih N, Anand A, Park JH, **Tan J**, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Seo SW, Kim D, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Szubin R, Hefner Y, Feist AM, Palsson BO. 2019. Cellular responses to reactive oxygen species are predicted from molecular mechanisms. *Proceedings of the National Academy of Sciences*.

J Tan, AV Sastry, KS Fremming, SP Bjørn, A Hoffmeyer, SW Seo, BG Voldborg, BO Palsson. Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. *Submitted*.

J Tan, CA Olson, JH Park, AV Sastry, PV Phaneuf, L Yang, R Szubin, Y Hefner, AM Feist, BO Palsson. Experimental evolution reveals the genetic basis and systems biology of superoxide stress tolerance. *Submitted*.

* equal contribution

ABSTRACT OF THE DISSERTATION

Multi-omic characterization of *E. coli* for the purpose of microbial-based production

by

Wun Kiat Justin Tan

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2019

Professor Bernhard Ø. Palsson, Chair

E. coli has been highly favored as a model organism and platform production strain because of its high rate of growth under simple culture conditions and its genetic tractability allowing introduction of foreign genes to expand its native metabolic capabilities. However, we lack understanding of how sequence features affect effective protein expression in *E. coli* , as well as the burden on the host cell during heterologous protein expression. In this dissertation we make use of next-generation sequencing to peer into the cell at the genomic, transcription, and translation level. We integrate the data across multiple scales in order to better understand

protein expression in *E. coli* under normal growth conditions, whilst expressing heterologous proteins, and after adaptation to oxidative stress. Firstly, we examine the translation dynamics of native proteins in the cell under normal growth conditions to reveal the causes and functions of programmed translational pauses along the transcript. Secondly, we investigate the transcriptome of *E. coli* whilst expressing a large library of heterologous proteins to identify 4 major host cell responses which vary widely across these proteins, Fear vs Greed, Metal Homeostasis and Respiration, Protein Folding, and Amino Acid and Nucleotide Biosynthesis. Lastly, we use Adaptive Laboratory Evolution to increase tolerance to oxidative stress, commonly found to be generated during high levels of protein expression. We make use of genomic resequencing, transcriptomics, and ribosome profiling to achieve a systems level understanding of the adaptations which occur in response to oxidative stress. As a whole, this work improves our understanding of *E. coli* as a platform production strain through A) identifying fundamental constraints on translation rates in native proteins, B) classifying host cell responses during expression of a variety of heterologous proteins to identify target areas for further research, and C) elucidating tolerance adaptations and mechanisms to oxidative stress, a common endogenous and exogenous stress during industrial biotechnology.

Chapter 1

Introduction

Microorganisms have been used for millenia as a means of production, many foods such as alcohol, cheese and bread rely on the microbial production of various compounds to achieve their distinctive flavor and texture. By leveraging on the ability of microorganisms to produce a wide variety of chemical compounds, various societies developed ways to perform complex chemical synthesis at using readily available raw material and at readily achievable temperatures and conditions. The development of methods for genetic manipulation of microorganisms accelerated this field, allowing the migration of lowly expressed genes from hard to culture organisms into well understood, easily cultured organisms such as *E. coli* where they could be readily studied and expressed [1]. Today, microorganisms have been developed for the production of not just food, but biofuels [2], commodity chemicals [3, 4], drugs [5] and therapeutic proteins [6]. However, despite more than a decade worth of efforts, only relatively few successes have been reported [7], a testament to the difficulties underlying these ventures. A major hurdle to further development has been due to a large void in our understanding of fundamental biology and its processes. *Escherichia coli*, a model organism and commonly used platform production strain

has been studied for decades, yet to date up to 30% of its genes are unknown or uncharacterized. The central dogma of biology was first stated in 1957 by Francis Crick [8], yet to date mechanisms behind transcription, translation and replication are still being discovered [9–11]. In this disseration, we take a multi-omic approach to characterize *E. coli* in an effort to better understand its constraints and limitations during expression of proteins, both native and foreign, as well as under oxidative stress.

1.1 Dynamic translation rates are important for protein expression and folding

Translation of proteins sits at the foundation of almost all bioproduction. By using various techniques to introduce foreign proteins into *E. coli* , we are now able to expand its native metabolic capabilities to allow the production of various compounds and proteins. However, this process is far from perfect, under 20% of foreign genes introduced are found to be expressed immediately [12]. Expression yields are typically increased through trials and error and various optimizations to a host of factors such as codon usage [13], mRNA secondary structure [14] and GC content, making the field as much an art as a science. Recent developments have further complicated matters, with the discovery that translation speed along the transcript affects the final folding and solubility of the protein [15]. Various theories have emerged, linking improved protein activity and solubility to the decreased translation rates at the N-terminus [16] and between domains [17]. Whilst these optimization procedures have shown some measure of success, it has previously been difficult to capture differences in translation rate.

1.2 Ribosome profiling as a proxy for translation speed

The development of high-throughput ribosome profiling as a molecular biology technique has provided a window into codon or even nucleotide-level resolution of translation along a transcript [18]. Ribosome profiling works by stalling actively translating ribosomes in place on mRNA through flash freezing, GTP-analogues or drugs such as cycloheximide and chloramphenicol [18–20]. These regions of mRNA are thus protected from cleavage by an endonuclease, and are then sequenced. This provides a direct snapshot into not only the number of ribosomes translating each transcript, but also the positions of the ribosomes along the transcript. Regions of fast translation result in sparser ribosomes, while regions of slow translation become more densely populated [21]. This allows ribosome profiling data to be correlated against various protein features in order to determine optimal translation rates.

1.3 Stresses caused by heterologous protein expression

One of the major difficulties with developing a robust expression system is that not all heterologous proteins are created equal. Large screens of protein expression have been performed to correlate protein features to expression levels, implicating features such as mRNA secondary structure, amino acid composition, and codon usage bias [13, 22, 23]. Many of the results from these studies only serve as guidelines for protein optimization, and we have yet to establish clear rules for protein optimization. On the other hand, each of these features affect expression levels due to their impact on the host cell physiology, and we can make use of these cellular responses to better understand the bottlenecks faced during expression of each of these proteins. Some proteins are more prone to misfolding than others, and overexpression results in misfolded protein

stress and up-regulation of heat shock proteins, chaperones and proteases [24]. Translation of the protein itself consumes energy and varying amount of resources such as amino acids, imposing a metabolic burden on the cell [25]. Overusage of other metabolites can also lead to metabolic and redox imbalances within the cell [26], while the production of metabolically active proteins could result in the production of toxic compounds and intermediates. High levels of overexpression also result in competition for ribosomes between the heterologous transcript and the cells native proteins, ultimately resulting in ribosomal degradation and cell death [27].

1.4 ROS is a common stressor in industrial biotechnology

A common cause for loss of cellular viability during protein expression is oxidative stress. These could come from intrinsic sources such as the overusage of cofactors resulting in redox imbalances [26], or genetic modifications to improve product formation in the host cell might involve knock out of thioredoxin and gluthathione biosynthesis genes [28], reducing the ability to deal with endogenous oxidative stress. These could also have extrinsic source such as product toxicity or high oxygen partial pressures in fermentation vessels during industrial scale production [29]. Oxidative stress poses a major issue to consistent protein expression as it results in damage to almost all macromolecules within the cell. DNA damage could result in increase in mutation rates and genetic instability during the production phase, while protein damage would result in reduced yields and growth rates.

1.5 Adaptive Laboratory Evolution

One method which has been developed to increase the rate of evolution of bacteria to selective pressures is Adaptive Laboratory Evolution (ALE). This procedure, performed over a hundred years ago by William Dallinger [30] and recently accelerated and automated by the use of robotics and liquid handling devices [31, 32], leverages the rapid growth of microorganisms in a laboratory setting to select for increasingly fit individuals. This technique has successfully improved growth of organisms on various substrates [31, 32], and improved tolerance to various stresses such as high temperatures [33], antibiotics [34] and toxic chemicals [35]. A recent advancement in ALE technologies, Tolerization Adaptive Laboratory Evolution (TALE) allows developments of tolerance to stressors far beyond the initial lethal concentrations through stepwise increments of stressor concentrations over the course of growth. When paired with genome resequencing and transcriptomics, ALE allows identification of causal mutations and elucidation of the mechanisms underlying fitness gains.

1.6 References

1. Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E. J., Short, J. M. & Keller, M. Cultivating the uncultured. en. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 15681–15686. ISSN: 0027-8424 (Nov. 2002).

2. Rodionova, M. V., Poudyal, R. S., Tiwari, I., Voloshin, R. A., Zharmukhamedov, S. K., Nam, H. G., Zayadan, B. K., Bruce, B. D., Hou, H. J. M. & Allakhverdiev, S. I. Biofuel production: Challenges and opportunities. *International journal of hydrogen energy* **42,** 8450–8461. ISSN: 0360-3199 (Mar. 2017).

3. Nakamura, C. E. & Whited, G. M. Metabolic engineering for the microbial production of 1,3-propanediol. en. *Current opinion in biotechnology* **14,** 454–459. ISSN: 0958-1669 (Oct. 2003).

4. Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J. & Van Dien, S. Metabolic engineering

of *Escherichia coli* for direct production of 1,4-butanediol. en. *Nature chemical biology* **7,** 445–452. ISSN: 1552-4450, 1552-4469 (May 2011).

5. Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K. W., Fickes, S., Galazzo, J., Gaucher, S. P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Jiang, H., Kizer, L., Lieu, B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L. F., Xu, L., Yu, M., Zhang, Y., Zhao, L., Lievense, J., Covello, P. S., Keasling, J. D., Reiling, K. K., Renninger, N. S. & Newman, J. D. High-level semi-synthetic production of the potent antimalarial artemisinin. en. *Nature* **496,** 528–532. ISSN: 0028-0836, 1476-4687 (Apr. 2013).

6. Johnson, I. S. Human insulin from recombinant DNA technology. en. *Science* **219,** 632–637. ISSN: 0036-8075 (Feb. 1983).

7. Chubukov, V., Mukhopadhyay, A., Petzold, C. J., Keasling, J. D. & Martin, H. G. Synthetic and systems biology for microbial production of commodity chemicals. en. *NPJ systems biology and applications* **2,** 16009. ISSN: 2056-7189 (Apr. 2016).

8. Crick, F. H. On protein synthesis. en. *Symposia of the Society for Experimental Biology* **12,** 138–163. ISSN: 0081-1386 (1958).

9. Latif, H., Federowicz, S., Ebrahim, A., Tarasova, J., Szubin, R., Utrilla, J., Zengler, K. & Palsson, B. O. ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions. en. *PloS one* **13,** e0197272. ISSN: 1932-6203 (May 2018).

10. Zhang, G. & Ignatova, Z. Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PloS one* **4,** e5036. ISSN: 1932-6203 (Apr. 2009).

11. Ebrahim, A., Brunk, E., Tan, J., O'Brien, E. J., Kim, D., Szubin, R., Lerman, J. A., Lechner, A., Sastry, A., Bordbar, A., Feist, A. M. & Palsson, B. O. Multi-omic data integration enables discovery of hidden biological regularities. en. *Nature communications* **7,** 13091. ISSN: 2041-1723 (Oct. 2016).

12. Savitsky, P., Bray, J., Cooper, C. D. O., Marsden, B. D., Mahajan, P., Burgess-Brown, N. A. & Gileadi, O. High-throughput production of human proteins for crystallization: the SGC experience. en. *Journal of structural biology* **172,** 3–13. ISSN: 1047-8477, 1095-8657 (Oct. 2010).

13. Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K.-H., Su, M., Luff, J. D., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P. & Hunt, J. F. Codon influence on protein expression in *E. coli* correlates with mRNA levels. en. *Nature* **529,** 358–363. ISSN: 0028-0836, 1476-4687 (Jan. 2016).

14. Gaspar, P., Moura, G., Santos, M. A. S. & Oliveira, J. L. mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic Acids Research* **44,** 5490–5490 (2016).

15. Hess, A.-K., Saffert, P., Liebeton, K. & Ignatova, Z. Optimization of translation profiles enhances protein expression and solubility. *PloS one* **10,** e0127039. ISSN: 1932-6203 (May 2015).

16. Verma, M., Choi, J., Cottrell, K. A., Lavagnino, Z., Thomas, E. N., Pavlovic-Djuranovic, S., Szczesny, P., Piston, D. W., Zaher, H., Puglisi, J. D. & Djuranovic, S. *Short translational ramp determines efficiency of protein synthesis* en. Mar. 2019.

17. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology* **16,** 274–280. ISSN: 1545-9993, 1545-9985 (Mar. 2009).

18. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223. ISSN: 0036-8075 (2009).

19. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635. ISSN: 0092-8674, 1097-4172 (Apr. 2014).

20. Latif, H., Szubin, R., Tan, J., Brunk, E., Lechner, A., Zengler, K. & Palsson, B. O. A streamlined ribosome profiling protocol for the characterization of microorganisms. *BioTechniques* **58,** 329–332. ISSN: 0736-6205, 1940-9818 (2014).

21. Sharma, A. K., Sormanni, P., Ahmed, N., Ciryam, P., Friedrich, U. A., Kramer, G. & O'Brien, E. P. A chemical kinetic basis for measuring translation initiation and elongation rates from ribosome profiling data. en. *PLoS computational biology* **15,** e1007070. ISSN: 1553-734X, 1553-7358 (May 2019).

22. Sastry, A., Monk, J., Tegel, H., Uhlén, M., Palsson, B. O., Rockberg, J. & Brunk, E. Machine Learning in Computational Biology to Accelerate High-Throughput Protein Expression. en. *Bioinformatics*. ISSN: 1367-4803, 1367-4811. doi:10.1093/bioinformatics/btx207 (Apr. 2017).

23. Cambray, G., Guimaraes, J. C. & Arkin, A. P. *Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli* 2018. doi:10.1038/nbt.4238.

24. Schweder, T. & Jürgen, B. in *Recombinant Protein Production with Prokaryotic and Eukaryotic Cells. A Comparative View on Host Physiology: Selected articles from the Meeting of the EFB Section on Microbial Physiology, Semmering, Austria, 5th–8th October 2000* (eds Merten, O.-W., Mattanovich, D., Lang, C., Larsson, G., Neubauer, P., Porro, D., Postma, P., de Mattos, J. T. & Cole, J. A.) 359–369 (Springer Netherlands, Dordrecht, 2001). ISBN: 9789401597494. doi:10.1007/978-94-015-9749-4_27.

25. Wu, X., Altman, R., Eiteman, M. A. & Altman, E. Adaptation of *Escherichia coli* to elevated sodium concentrations increases cation tolerance and enables greater lactic acid production. *Applied and environmental microbiology* **80,** 2880–2888. ISSN: 0099-2240, 1098-5336 (May 2014).

26. De Ruijter, J. C., Koskela, E. V., Nandania, J., Frey, A. D., *et al.* Understanding the metabolic burden of recombinant antibody production in *Saccharomyces cerevisiae* using a quantitative metabolomics approach. *Yeast.* ISSN: 0749-503X (2018).

27. Dong, H., Nilsson, L. & Kurland, C. G. Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. en. *Journal of bacteriology* **177,** 1497–1504. ISSN: 0021-9193 (Mar. 1995).

28. Bessette, P. H., Aslund, F., Beckwith, J. & Georgiou, G. Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 13703–13708. ISSN: 0027-8424 (Nov. 1999).

29. Baez, A. & Shiloach, J. *Escherichia coli* avoids high dissolved oxygen stress by activation of SoxRS and manganese-superoxide dismutase. *Microbial cell factories* **12,** 23. ISSN: 1475-2859 (Mar. 2013).

30. Bennett, A. F. & Hughes, B. S. Microbial experimental evolution. en. *American journal of physiology. Regulatory, integrative and comparative physiology* **297,** R17–25. ISSN: 0363-6119, 1522-1490 (July 2009).

31. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Applied and environmental microbiology* **81,** 17–30. ISSN: 0099-2240, 1098-5336 (Jan. 2015).

32. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Applied and environmental microbiology* **83.** ISSN: 0099-2240, 1098-5336. doi:10.1128/AEM.00410-17 (July 2017).

33. Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., Palsson, B. O., Sommer, M. & Feist, A. M. Evolution of *Escherichia coli* to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. en. *Molecular biology and evolution* **31,** 2647–2662. ISSN: 0737-4038, 1537-1719 (Oct. 2014).

34. Jahn, L. J., Munck, C., Ellabaan, M. M. H. & Sommer, M. O. A. Adaptive Laboratory Evolution of Antibiotic Resistance Using Different Selection Regimes Lead to Similar Phenotypes and Genotypes. en. *Frontiers in microbiology* **8,** 816. ISSN: 1664-302X (May 2017).

35. Mohamed, E. T., Wang, S., Lennen, R. M., Herrgaard, M. J., Simmons, B. A., Singer, S. W. & Feist, A. M. Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution. *Microbial cell factories* **16,** 204. ISSN: 1475-2859 (Nov. 2017).

Chapter 2

Multi-omic data integration enables discovery of hidden biological regularities

The rapid growth in size and complexity of biological data sets has led to a grand challenge referred to as Big Data to Knowledge. Here we address a critical need for the development of advanced data integration methods to enable multi-level analysis of genomic, transcriptomic, ribosomal profiling, proteomic, and fluxomic data across multiple experimental conditions [1]. First, we show that pairwise integration of primary omics data reveals biological regularities that tie certain cellular processes together in Escherichia coli: the number of protein molecules made per mRNA transcript and the number of ribosomes required per translated protein molecule. Second, we show that genome-scale models, which are based on genomic and bibliomic data, enable the quantitative synchronization of disparate omics data types [2]. Integrating omics

data with models enabled the discovery of two novel regularities: condition invariant in vivo turnover rates of enzymes and the correlation of protein structural motifs and translational pausing. How these regularities relate to one another mechanistically is formally represented in a computable knowledge base, which allows for the coherent interpretation and prediction of fitness and selection underlying cellular physiology.

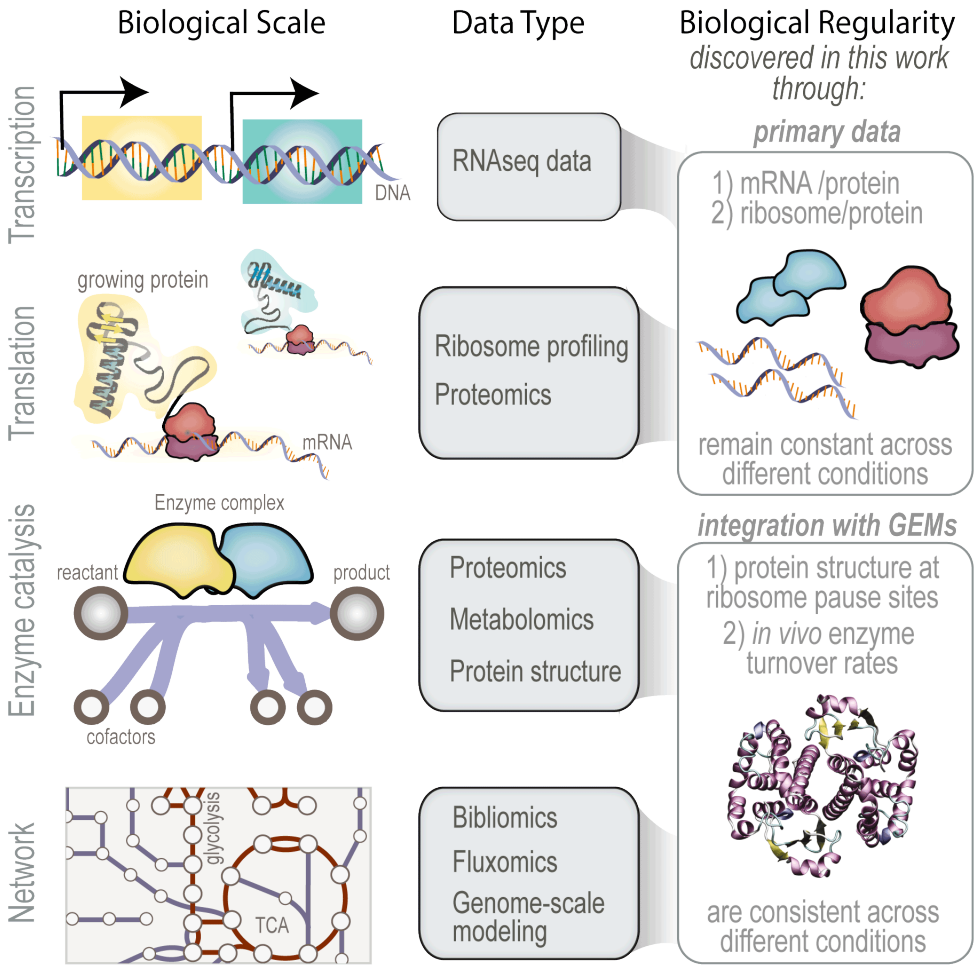

Figure 2.1: A multi-scale, multi-omics framework detects significant biological regularities in E. coli. Tracing the central dogma of biology (left column), we can link specific data types (middle column) to explain each of these biological processes. In this work, novel biological regularities that relate these processes are discovered through: (i) primary omics data (top box, right column) and (ii) integration with genome-scale models of metabolism (GEMs; bottom box, right column).

2.1 Background

Progress of the biological sciences in the era of big data will depend on how we address the following question: "How do we connect multiple disparate data types to obtain a meaningful understanding of the biological functions of an organism?" Owing to large-scale improvements in omics technologies, we can now quantitatively track changes in biological processes in unprecedented detail [3, 4]. While such measurements span a diverse range of cellular activities, developing an understanding of how these data types quantitatively relate to one another and to the phenotypic characteristics of the organism remains elusive. This issue is central to the so-called Big Data to Knowledge (BD2K) grand challenge, which aims to integrate multiple disparate data types into a biologically meaningful, multi-level structure [1, 2].

Interpretation of disparate data requires understanding how the primary measurements of different omics data are quantitatively coupled to one another [5]. We approach this task by identifying regularities (relationships between biological data types that remain relatively constant across conditions) between pairwise omics data types. While some regularities can readily be discovered through direct pairwise omics data comparisons, we find that other regularities emerge only through more intricate analysis leveraged by mechanistically-based network reconstructions6. Such reconstructions can be used as a context for poly-omic data integration and analysis [6, 7] and, when combined with constraint-based modeling approaches [8, 9], provide important links between omics data and phenotypic characteristics of the organism.

As we will show, this approach leads to a comprehensive synchronization of poly-omic data with computed growth states. The approach directly addresses the BD2K grand challenge and is made conceptually accessible by tracing the 'information flow' through the familiar 'central dogma' to establish relationships between measurements and cell physiology (Figure 2.1).

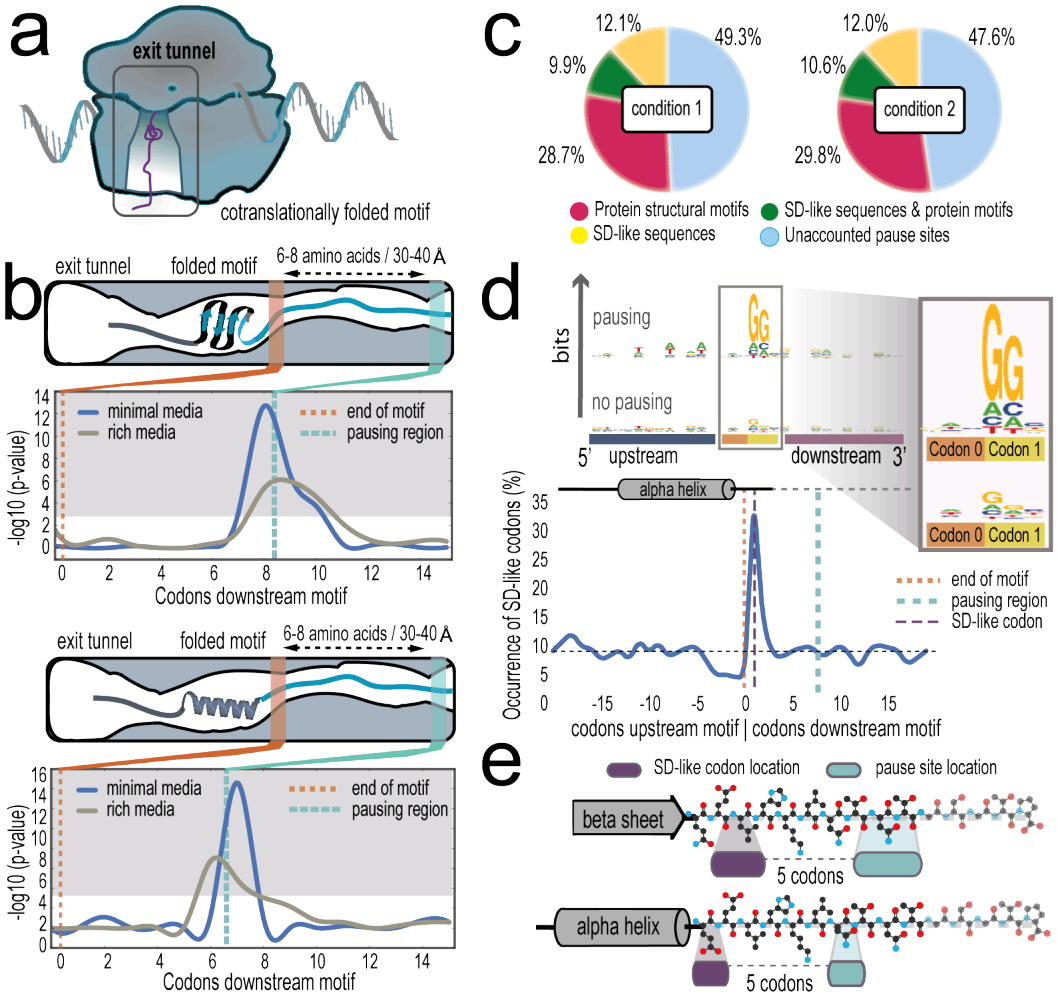

Figure 2.2: Regularities in translational pausing and structural motifs. (a) Cartoon depiction of co-translational folding intermediates, such as secondary structure motifs, inside the ribosome exit tunnel. (b) Analysis of ribosome profiling and translational pausing in conjunction with protein structure properties in E. coli grown under MOPS Rich and MOPS Minimal media, taken from Li et. al. 2014 [10]. Pausing is enriched at positions downstream of protein secondary structures (top: beta sheets, bottom: alpha helices, p-value $<$ 6.67x10-3). These correlations are consistent across conditions (e.g. minimal and rich nutrient conditions). (c) Coverage of specific secondary structure elements and sequence elements that account for increased ribosome occupancy. Condition 1 refers to minimal media and condition 2 refers to rich media. (d) Protein structure motifs that exhibit pausing have increased propensity for SD-like sequences compared to those which do not exhibit pausing or the global background existence, 35% SD-like codons for alpha helices, 18% SD-like codons for beta sheets, compared to 9% global average. (c) A cartoon depiction of the relationship between structure, translation and sequence.

2.2 Results

2.2.1 Gene specific translation efficiency is consistent across conditions

First, we examine the information flow from transcription to translation to protein production by identifying correlations across primary omics data types, such as RNAseq [11], ribosome profiling [10, 12, 13] and proteomics [14], collected for E. coli batch growth on glucose, fumarate, pyruvate, and acetate (Figure 2.1, "primary data box"). We found relatively poor correlations of mRNA to protein across conditions ($r^2 < 0.4$), consistent with previous studies [15, 16]. Stronger correlations ($r^2 > 0.8$) emerge when analyzing the ratio of protein per mRNA (ρPM) on a per-gene basis (the difference between peptide abundance and relative mRNA read counts per gene for multiple growth conditions; Supplementary Figure A.1(a)). Computing the median coefficient of variation shows that changes in ρPM across conditions are relatively invariant. In addition, we find the number of ribosomes required (ribosome occupancy of mRNA) per protein translated is also relatively invariant across all four conditions ($r^2 > 0.7$; Supplementary FigureA.1(b)).

Second, we examined pairwise relationships between other omics data types, such as ribosome profiling, proteomics and fluxomics, by integrating these data types into next generation genome-scale models (Figure 2.11, "integration with GEMs box"). Genome-scale models of metabolism (GEMs) are based on the annotated sequence and analysis of the bibliome for functionally annotated gene products6. The most recent generations of genome-scale models incorporate protein structural information [17] and allow for the computation of the synthesis of the entire proteome of a cell in addition to the balanced use of its metabolic network [18]. These models can integrate multiple layers of biological organization to balance the use of all cellular

components to achieve a cellular state. It can thus extend our understanding of how information flows from translation to protein folding and catalysis, and its role in producing whole cell functions.

2.2.2 Translation pausing is correlated with protein secondary structure

We examined how information flows during protein translation, which includes protein folding. Recent studies indicate a possible link between translation speed and proper folding [12, 19]. Analysis of translational pausing has typically been approached from a sequence-based viewpoint [19]. Here, we approach this analysis from a different perspective, by correlating the occurrence of translational pausing on a transcript to the location of nearby protein secondary (2^o) structure motifs (Figure 2.2). The establishment of this correlation is based on; 1) ribosome profiling [10, 12, 13], which provides ample information on the queuing of ribosomes along mRNA transcripts, and 2) a recent network reconstruction that contains comprehensive protein structural information linked to the translated protein at the proteome-scale [17].

Several striking regularities in translational pausing and protein structure are consistently observed across multiple growth conditions in E. coli, which suggest the co-translational folding of intermediate secondary structure motifs inside the ribosome exit tunnel (Figure 2.2(a)). We find that pause sites are enriched (pvalue <0.01 using a hypergeometric test) downstream of specific secondary structure motifs, such as -helices and -sheets (Figure 2.2(b), Supplementary Figure A.2) yet are not significantly enriched at the termini of domains (See Supplementary Information). On average, pausing becomes most substantial six to eight amino acids downstream of -helices and -sheets, which, in the majority of cases, fall either on disordered regions of the protein or on helical residues. Such instances consistently account for more than 35-40% of pause sites

across different conditions (Figure 2.2(c), Supplementary Figure A.2). These findings strongly corroborate with a growing theory that partially folded intermediate protein structures begin to immediately fold inside the ribosome exit tunnel, following polypeptide-chain synthesis. Several previous studies have shown that partially folded protein structures, such as small domains, can be detected within the exit tunnel. [20–22] More recently, Nilsson et al. demonstrated the co-translational folding of small zinc finger-like domain deep within the ribosome exit tunnel using arrest-peptide mediated force measurements in conjunction with cryo-electron tomography. [23]

2.2.3 Translation pausing is encoded at the sequence level

Do sequence-specific motifs drive co-translational pausing to ensure proper protein folding? We find that Shine-Dalgarno (SD) like sequences account for 20-22% of ribosome density at pause sites (2.2(c), Supplementary Methods: Identification of Shine-Dalgarno-like codons), which is consistent with recent studies [24] and four times less frequent than what is found previous studies [19]. Of the pausing instances linked to SD-like sequences, we find that, on average, nearly half of these pausing regions also fall in the nearby vicinity (five to ten codons) of helices or sheets. The link between pausing, SD-like sequence and protein secondary structure becomes clear when comparing the average occurrence of SD-like sequence genome-wide (9%) with their occurrence directly downstream of -helices (35%) and -sheets (18%, 2.2(d)). Together, these sequence and structure motifs account for the majority of pause sites (60%) or nearly half of the total ribosome occupancy (Supplementary Figure A.4). These findings suggest that co-translational pausing occurs for distinct secondary structural elements and supports the potential role of sequence-specific factors to drive pausing for ensuring proper protein folding (2.2(e)).

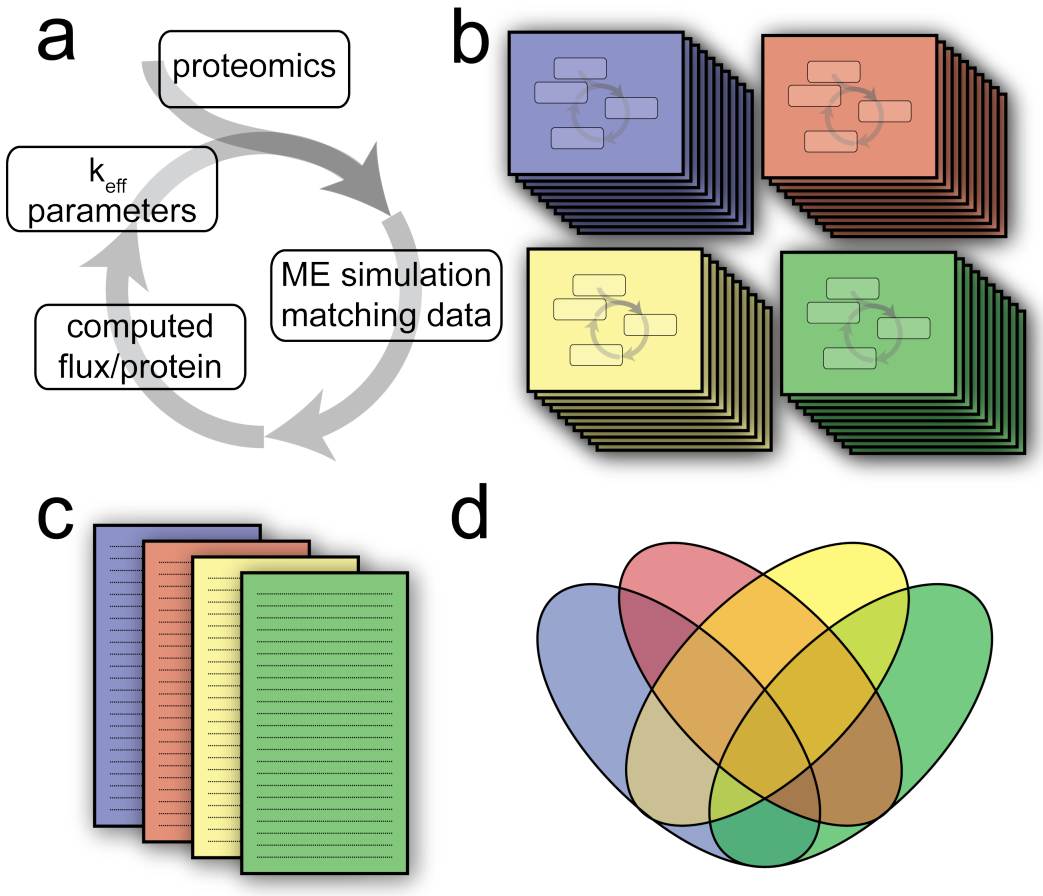

Figure 2.3: Effective enzyme turnover rates (k_{eff}) as regularities emerging from coupling quantitative in vivo proteomic data with genome scale modeling. (a) Iterative workflow for generating turnover rate values from different nutrient conditions. This panel is a schematic of the overall workflow. A detailed version is found in the Supplement. (b) Venn diagram of calculated turnover rates shows all four conditions share 90% of the same estimates (Pearson correlations below). (c) Pairwise comparisons across four conditions for calculated turnover rate parameters demonstrate 94% are within one order of magnitude. The upper inset show the parameter estimation for the 10% most variable components of the proteome between the four conditions examined. The lower inset show a histogram of the distances of every point from the diagonal line. The gray box contains the 94% of the values that deviate from one another within an order of magnitude. A more detailed version is found in Supplementary Figure 5

2.2.4 Predicting model parameters by integrating proteomics data

How does information flow between an individual enzyme's catalytic activity and the activity of an entire network? To evaluate the effective turnover rate of enzymes, reaction flux

per enzyme can be directly computed using experimental values for both flux (the rate of reactions) and enzyme abundance [25] on a small scale (mainly for central carbon metabolism). To assess enzyme turnover on a genome scale, we computed the ratio of an enzyme's abundance (measured from proteomics data) and its corresponding flux derived from network-based analyses using the iOL1650-ME model (Figure 2.3(a)). Because the iOL1650-ME model directly relates enzyme synthesis and metabolic flux, we were able to develop a method which uses the model to extrapolate the most likely flux state from a proteomic data set (Supplementary Methods: Computational Method for Predicting k_{eff}parameters). These ratios quantitatively couple experimentally-derived flux estimates and protein abundances to make a quantitative connection between data types.

Estimates of enzyme turnover rates (k_{eff}), which represent coupling coefficients between the fluxome and the proteome, were analyzed across four nutrient conditions to understand the effect that carbon uptake has on metabolic enzyme turnover rates. We find that these parameters show considerable regularity in relating flux to protein abundance, which suggests that in vivo turnover rate for most enzymes does not strongly depend on growth in diverse batch culture settings. For high-flux metabolic reactions, the estimated turnover rates were consistent across all four conditions (a total of 284 turnover rate values; Figure 2.3(b)), with high correlation between any two conditions (Figure 2.3(c) and Supplementary Figure A.5). The computed turnover rates were averaged across experimental conditions to give the largest set of flux-per-enzyme parameters estimated computationally to date under in vivo conditions. It is important to note that these estimated turnover rates do not have a direct relationship with fundamental enzyme kinetic parameters obtained in vitro but can be viewed as an in vivo data-driven estimate of the enzyme turnover rate.

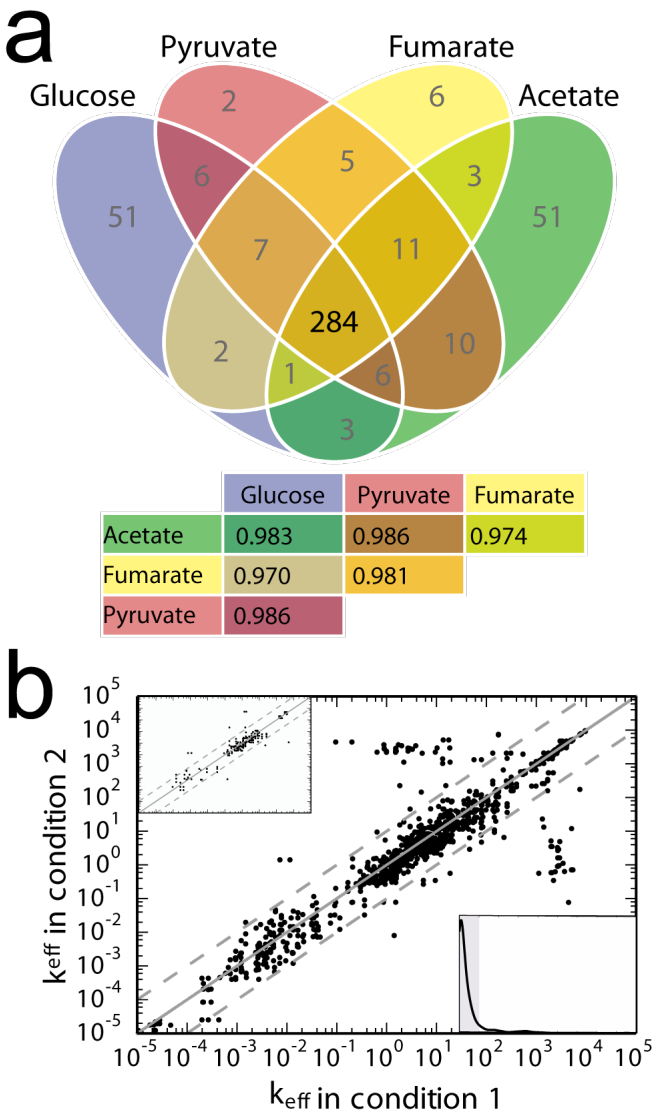

Figure 2.4: Predicting the results of perturbation from a parameterized homeostatic state. (a) Using a cross-validation approach, protein abundance is predicted by mRNA levels using information (ρPM) obtained from other conditions (r2 > 0.75). Condition-specific mRNA and protein levels show little correlation (inset) (b) Accuracy of predicting differential expression is significantly enhanced using k_{eff}parameters. (c) Changes in gene expression and protein abundance predicted in different media supplementation. Accuracies range between 56-100% and specific genes are significantly enriched ($p < 0.05$ using a hypergeometric distribution).

While these correlations provide information about relationships between biological components and, in some cases, take on predictive value (Figure 2.4(a)), understanding their collective influence on cell physiology is harder to decipher. This issue can be addressed using a

genome-scale model that assesses cost-benefit tradeoffs from a cell-centric perspective [9, 26]. Genome-scale models (iOL1650-ME) compute the value of cellular components relative to the function of all other cellular components. To this end, the turnover rate values provide the minimum 'capital expenditure' for protein synthesis required to achieve a unit of flux through a given reaction. Thus as a group, the calculated turnover rates provide coupling between proteome allocation and achievement of a physiological state.

The knowledge of the biological regularities identified in this work enables the parameterization of coupling constraints used in a genome-scale model of metabolism and gene expression (ME). A parameterized model allows for prediction of responses to environmental perturbations. We tested the predictive capacity of a model containing parameter values derived from multiple conditions described above (Supplementary Methods: Predicting Differential Gene Expression with iOL1650-ME) to compute optimal cellular composition under new environmental conditions where we did not have proteomics data available. We perturbed a reference growth state through the addition of nutrients to the medium: batch growth on glucose was supplemented with adenine, glycine, tryptophan or threonine. We collected omics data sets under these four perturbed conditions to compare gene expression changes to the computated responses.

Using the parameterized model, we predicted the enzymes that would be differentially used in the supplemented condition (Figure 2.4(b)). When validating our predictions using experimentally measured differential gene expression, we find high predictive accuracies of significant changes in gene expression (p-values ranging 0.04 to 4e-6 using a hypergeometric test). Using the parameterized model, we are able to predict the regulation of genes that accompany changes in supplementation to a new growth environment. Such environmental changes oftentimes causes non-intuitive shifts in what precursors the cell uses to synthesize amino acid molecules (Fig-

ure 2.4(c); Supplementary Discussion).

Taken together, we demonstrate an ability to systematically integrate multi-omic data to enable discovery of multiple hidden biological regularities. These regularities take on biological meaning when put into the context of a network reconstruction that is comprised of fundamentally structured relationships between cellular components. We have shown that this contextualization leads to: (i) insights into underlying biological mechanisms during protein translation and (ii) predictive computations based on cellular-econometric cost-benefit ratios associated with the function of the cell as a whole. Thus both multi-omic data analysis and genome-scale models will play an important role in establishing big data analysis frameworks to explain and predict cellular physiology.

Acknowledgements

Conceptualization, A.E., E.B., J.T., and B.O.P.; Methodology, (ME modeling: A.E., J.L., E.J.O., and A.M.F.) (Ribosome profiling analysis: E.B., J.T.); Investigation, J.T., A.E., and E.B.; Writing – Original Draft, E.B., A.E., and B.O.P.; Writing – Review and Editing, E.B., A.E., J.T., J.L., A.M.F. and B.O.P.; Discussion, All authors; Funding Acquisition, B.O.P.; Resources, B.O.P.; Supervision, B.O.P.

This work was funded from a generous gift from the Novo Nordisk Foundation to the Center for Biosustainability. It was also supported by DE-FOA-000014 from the U.S. Department of Energy, NIH R01 GM057089 from the National Institutes of Health, and by the Swiss National Science Foundation (grant p2elp2_148961 to E.B). This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. The authors

gratefully acknowledge Dr. Mahmoud Al-Bassam, and Dr. Jinwoo Kim for scientific discussions on ribosome profiling.

Chapter 2 in full is a reprint of material published in: A Ebrahim*, E Brunk*, **J Tan***, EJ O'Brien, D Kim, R Szubin, JA Lerman, A Lechner, A Sastry, A Bordbar, AM Feist, BO Palsson. 2016. "Multi-omic data integration enables discovery of hidden biological regularities" *Nature Communications* 7: 13091. The dissertation author was one of the primary authors.

2.3 References

1. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nature reviews. Genetics* **14,** 333–346. ISSN: 1471-0056, 1471-0064 (May 2013).

2. Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology* **7,** 198–210. ISSN: 1471-0072 (2006).

3. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207. ISSN: 0028-0836 (Mar. 2003).

4. De Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C. & Mann, M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–1254. ISSN: 0028-0836, 1476-4687 (Oct. 2008).

5. Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A. & Tagkopoulos, I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Molecular systems biology* **10,** 735. ISSN: 1744-4292, 1744-4292 (July 2014).

6. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5,** 93–121. ISSN: 1754-2189 (2010).

7. Hyduke, D. R., Lewis, N. E. & Palsson, B. Ø. Analysis of omics data with genome-scale models of metabolism. *Molecular bioSystems* **9,** 167–174. ISSN: 1742-206X, 1742-2051 (Feb. 2013).

8. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature biotechnology* **28,** 245–248. ISSN: 1087-0156 (2010).

9. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161,** 971–987. ISSN: 0092-8674, 1097-4172 (May 2015).

10. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635. ISSN: 0092-8674, 1097-4172 (Apr. 2014).

11. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature* **473,** 337–342. ISSN: 0028-0836, 1476-4687 (May 2011).

12. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223. ISSN: 0036-8075 (2009).

13. Latif, H., Szubin, R., Tan, J., Brunk, E., Lechner, A., Zengler, K. & Palsson, B. O. A streamlined ribosome profiling protocol for the characterization of microorganisms. *BioTechniques* **58,** 329–332. ISSN: 0736-6205 (2015).

14. Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R. & Heinemann, M. The quantitative and condition-dependent Escherichia coli proteome. *Nature biotechnology.* ISSN: 1087-0156, 1546-1696. doi:10.1038/nbt.3418 (Dec. 2015).

15. Laurent, J. M., Vogel, C., Kwon, T., Craig, S. A., Boutz, D. R., Huse, H. K., Nozue, K., Walia, H., Whiteley, M., Ronald, P. C. & Marcotte, E. M. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10,** 4209–4212. ISSN: 1615-9853, 1615-9861 (Dec. 2010).

16. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology* **4,** 117. ISSN: 1465-6906 (Aug. 2003).

17. Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A. & Palsson, B. O. Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. *Science* **340,** 1220–1223. ISSN: 0036-8075, 1095-9203 (June 2013).

18. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. eng. *Molecular systems biology* **9,** 693. ISSN: 1744-4292 (2013).

19. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484,** 538–541. ISSN: 0028-0836, 1476-4687 (Apr. 2012).

20. Mingarro, I., Nilsson, I., Whitley, P. & von Heijne, G. Different conformations of nascent polypeptides during translocation across the ER membrane. *BMC cell biology* **1,** 3. ISSN: 1471-2121 (Dec. 2000).

21. Bhushan, S., Gartmann, M., Halic, M., Armache, J.-P., Jarasch, A., Mielke, T., Berninghausen, O., Wilson, D. N. & Beckmann, R. [alpha]-Helical nascent polypeptide chains vi-

sualized within distinct regions of the ribosomal exit tunnel. *Nature structural & molecular biology* **17,** 313–317. ISSN: 1545-9993 (2010).

22. Tu, L., Khanna, P. & Deutsch, C. Transmembrane segments form tertiary hairpins in the folding vestibule of the ribosome. *Journal of molecular biology* **426,** 185–198. ISSN: 0022-2836, 1089-8638 (Jan. 2014).

23. Nilsson, O. B., Hedman, R., Marino, J., Wickles, S., Bischoff, L., Johansson, M., Müller-Lucks, A., Trovato, F., Puglisi, J. D., O'Brien, E. P., Beckmann, R. & von Heijne, G. Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell reports* **12,** 1533–1540. ISSN: 2211-1247 (Sept. 2015).

24. Mohammad, F., Woolstenhulme, C. J., Green, R. & Buskirk, A. R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell reports* **14,** 686–694. ISSN: 2211-1247 (Feb. 2016).

25. Arike, L., Valgepea, K., Peil, L., Nahku, R., Adamberg, K. & Vilu, R. Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. *Journal of proteomics* **75,** 5437–5448. ISSN: 1874-3919 (Sept. 2012).

26. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nature reviews. Genetics* **15,** 107–120. ISSN: 1471-0056 (Feb. 2014).

Chapter 3

Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression

Achieving the predictable expression of heterologous genes in a production host has proven difficult. Each heterologous gene expressed in the same host seems to elicit a different host response governed by unknown mechanisms. Historically, most studies have approached this challenge by manipulating the properties of the heterologous gene through methods like codon optimization. Here we approach this challenge from the host side. We express a set of 45 heterologous genes in the same Escherichia coli strain, using the same expression system and culture conditions. We collect a comprehensive RNAseq set to characterize the host's transcriptional

response. Independent Component Analysis of the RNAseq data set reveals independently modulated gene sets (i-modulons) that characterize the host response to heterologous gene expression. We relate 55% of variation of the host response to: Fear vs Greed (16.5%), Metal Homeostasis (19.0%), Respiration (6.0%), Protein folding (4.5%), and Amino acid and nucleotide biosynthesis (9.0%). If these responses can be controlled, then the success rate with predicting heterologous gene expression should increase.

3.1 Understanding heterologous protein expression from the host perspective

E. coli is extensively used as a production strain for the production of chemicals and pharmaceuticals in a metabolically engineered strain, or for the production of proteins as a product [1]. Most of these proteins and pathways are not naturally found in E. coli, and therefore host strain design requires the introduction of foreign genes and proteins in order to expand its native capabilities. Various approaches have been developed to enable researchers to clone foreign genes into E. coli thus allowing the expression of both natural and engineered heterologous proteins [2].

The successful expression of many proteins in E. coli requires some form of manipulation of their properties and their optimization; less than 20% of all foreign genes introduced into E. coli are able to be expressed immediately [3]. Considerable effort has been devoted towards determining protein feature determinants of protein expression [4–7]. The most successful efforts have been achieved through the use of machine learning approaches on large, extensive libraries of proteins. Various factors such as codon usage, mRNA secondary structure, and amino acid

composition have been found to have statistically significant effects on protein expression, solubility, and fitness of the cells [8–10]. Unfortunately, we are still unable to generate clear rules on the design of heterologous proteins for proper expression and solubility.

Fewer studies have looked at the host strain's reaction to the expression of heterologous proteins. Some have shown that overexpression of heterologous genes result in the upregulation of heat shock proteins [11, 12], destruction of ribosomes [13] downregulation of amino acid biosynthesis and TCA cycle genes [14], triggering of SOS response [15], and the downregulation of the ArcA/ArcB two-component system [16]. Most of these studies have been performed with only a small number of proteins, resulting in a limited view of the host's response to heterologous protein expression. In this study, we will focus on the host's response to the induction of a larger set of 45 heterologous proteins at the transcriptional level. We have assembled one of the largest collections of transcriptomic datasets of heterologous gene expressing lines to date. We then make use of Independent Component Analysis, a method recently shown to be able to identify clear transcriptional regulatory responses in large datasets [17–19], to identify the major modes of the host's cell response during heterologous protein expression.

3.2 Results

3.2.1 Expression of different heterologous genes elicit variable host response under the same induction conditions

We made use of 3' LIC [3] to clone a library of 45 heterologous genes from various organisms (Supp Table 1) into a plasmid which was transformed into a strain of E. coli evolved for optimal growth on glycerol (See Methods in SI)[20]. Several factors, such as the ribosomal binding

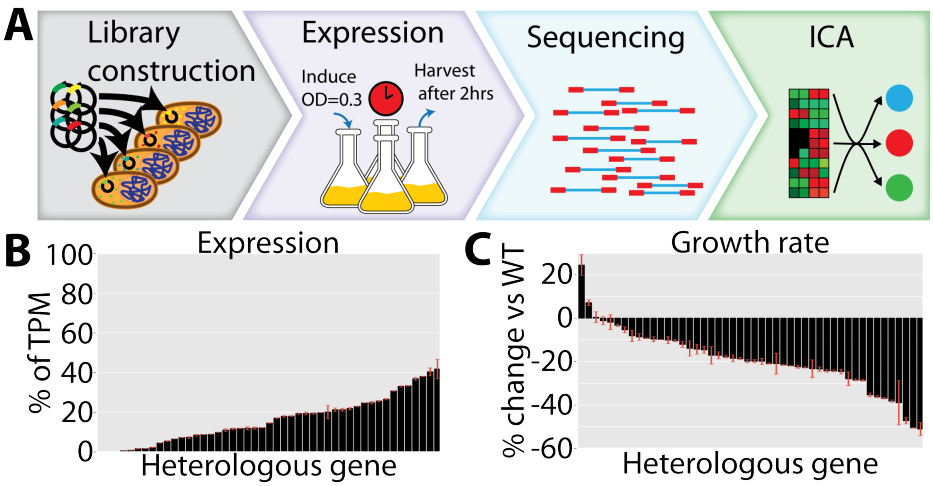

Figure 3.1: Workflow and physiological characterization. A: Experimental workflow for library of 40 heterologous genes. B: Heterologous gene expression varied between <1% and 42% of the total transcripts. C: Induction resulted in a reduction in growth rate compared to wild type for almost all strains. Error bars for duplicates are shown in red.

site and the secondary structure of the 5' end of the mRNA, are known to have a large impact on the final expression level of heterologous proteins [5, 21–23]. In order to control for initiation rate, we designed our plasmids with a bi-cistronic design (BCD) region upstream of the heterologous protein [24, 25], and all heterologous proteins were fused with a His-TEV tag at the 5' end to normalize 5' mRNA secondary structure. A medium copy plasmid backbone (pNic28) [2, 3] was used in combination with a rhamnose inducible promoter. Rhamnose induction concentration was set to 1mM in order to obtain a lower expression of protein compared to typical T7 induced plasmids (70-80% of total proteome). Despite our efforts to control heterologous protein levels in the cell, RNAseq measurements of heterologous gene expression varied between <1% and 42% of total transcripts (Figure 3.1B). This variance in expression levels was not found to correlate well with traditional measures of protein success such as codon adaptation index (CAI) [26] (Supp. Fig. B.1, B.2).

The expression of heterologous protein places a variety of stresses on the host, the most

fundamental of which is the added burden of metabolic precursors such as amino acids and the energy required for the production of unnecessary heterologous proteins, reducing the allocation of cellular resources to growth [27]. In our study, the addition of rhamnose during induction provides an additional carbon source, resulting in a spike in wild type growth rate post- induction. To account for this spike, growth rate was normalized to wild type cultures. We find that the introduction of an empty plasmid control resulted in an increase in growth rate (24.5% $\pm$ 6.75), whereas heterologous protein production negatively affects growth rate (Figure 3.1C).

3.2.2 ICA elucidates the modes of host cell response

To understand the various modes of the host cell response, we performed Independent Component Analysis (ICA) on the transcriptomic dataset (See Methods). ICA decomposes a gene expression matrix into its independent components, each of which constitutes an independently modulated set of genes that have been shown to reflect the transcriptional regulatory network in E. coli [17]. ICA decomposition of the transcriptomic profiles resulted in 99 components, 69 of which were in common with i-modulons found from the PRECISE dataset in Sastry et. al. that represents 113 experimental conditions [17]. These results are summarized in Supp. Table 2.

3.2.3 RhaR i-modulon identifies failures of induction

The use of the rhamnose inducible promoter is reflected in the ICA decomposition, showing up as two new i-modulons which had not previously been detected in the PRECISE database [17]. The first of these two i-modulons is the Plasmid i-modulon, composed of two highly weighted genes, rhaR and rhaS, both of which are encoded on the plasmid (Figure 3.2A). This i-modulon is thus representative of the plasmid copy number. All the samples except wild type show a

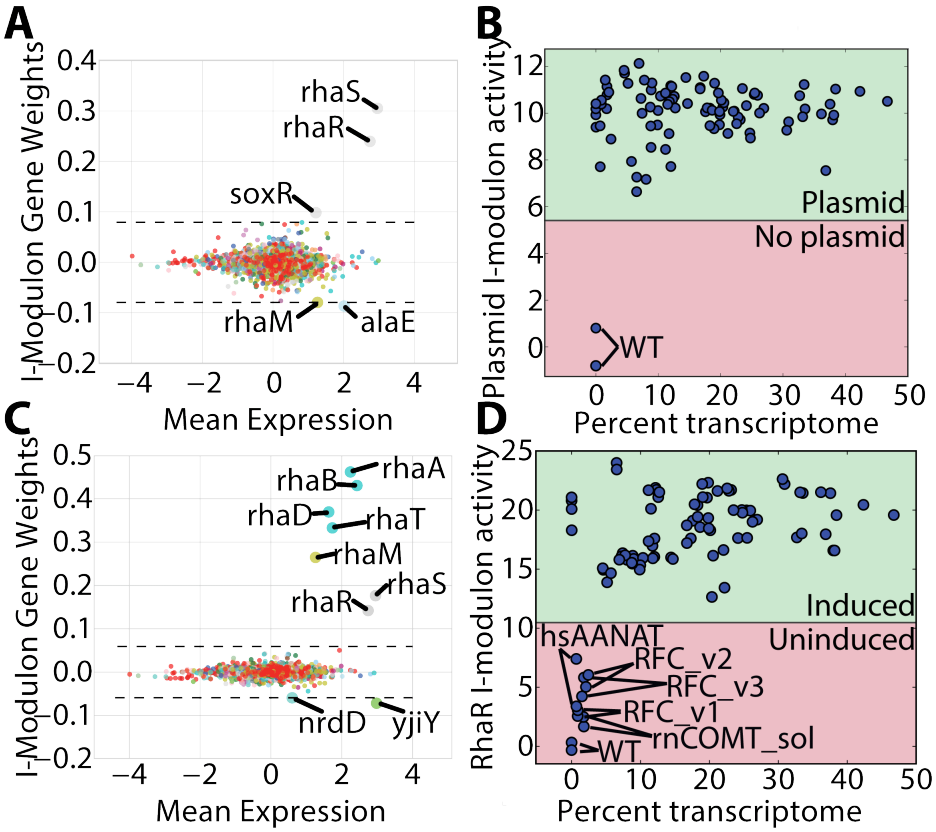

Figure 3.2: Distinguishing plasmid and induction through ICA. A: An i-modulon showed high weighting in rhaS and rhaR, but not the rest of the rhamnose metabolism genes. The rhamnose inducible plasmid coded for both the rhaS and rhaR genes, hence this i-modulon probably represented the plasmid copy number. B: All samples show a high level of this i-modulon except wild type, which did not contain the plasmid. C: Genes which show significant weights in the RhaR i-modulon are members of the rhaSR-rhaBAD operon: rhaA, rhaB, rhaD, rhaT, rhaM, rhaS, rhaR, with nrdD and yjiY showing marginal significance. D: While most samples show a high activity in the RhaR i-modulon implying a successful induction (green), several samples show failure of induction (red).

high activity level in this i-modulon, showing that the plasmid is present in all these samples (Figure 3.2B).

The second of these two i-modulons is the Rhamnose i-modulon. It consists of genes in the rhamnose operon which would be induced during growth on rhamnose as a carbon source such as *rhaA*, *rhaB*, *rhaD*, as well as the rhamnose regulatory genes *rhaS* and *rhaR* (Figure 3.2C). We found that excluding WT, five samples showed much lower activity levels in this i-modulon,

all of which displayed corresponding low transcript levels of the heterologous protein, leading us to conclude that these proteins suffered from a failure of induction for an unidentified reason (Figure 3.2D). Samples which showed this failure of induction were removed from subsequent analyses as they did not represent the true host cell response to the induction of heterologous protein expression, leaving us with 40 heterologous protein expressing lines. The detection of these two i-modulons demonstrates the ability of ICA to distinguish between two closely related signals.

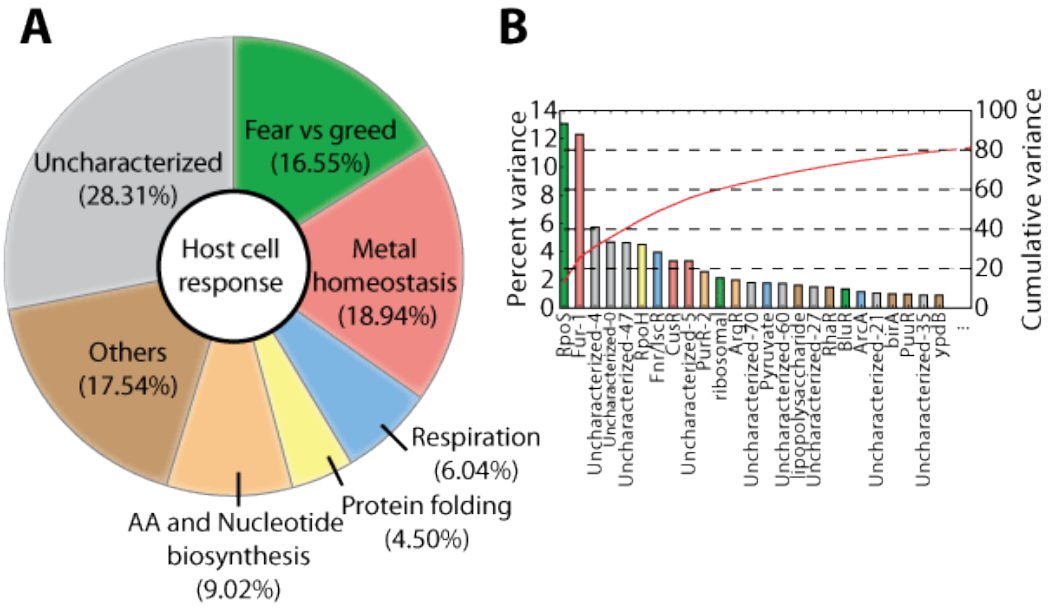

Figure 3.3: Changes in i-modulon activities during heterologous gene expression. A: An i-modulon showed high weighting in rhaS and rhaR, but not the rest of the rhamnose metabolism genes. The rhamnose inducible plasmid coded for both the rhaS and rhaR genes, hence this i-modulon probably represented the plasmid copy number. B: All samples show a high level of this i-modulon except wild type, which did not contain the plasmid. C: Genes which show significant weights in the RhaR i-modulon are members of the rhaSR-rhaBAD operon: rhaA, rhaB, rhaD, rhaT, rhaM, rhaS, rhaR, with nrdD and yjiY showing marginal significance. D: While most samples show a high activity in the RhaR i-modulon implying a successful induction (green), several samples show failure of induction (red).

Table 3.1: Major dimensions of host response and the i-modulons and genes that make up each of them.

Cellular response	i-modulon	Genes
Fear vs Greed	RpoS	*glsA, csgD, yedP, yhdW_2, mcbR, csiD, fic, yhjY, tam, ydcK, yfiL, gmr, aldB, yfdC, yqjE, yhbW, gabP, yebF, yohF, yccT, yphA, ytjA, ydcT, lhgO, yehX, osmE, rclA, mlrA, yeaQ, ylil, ybiO, hchA, curA, yqjK, ymgE, ybaY, gabT, gabD, yghA, ydcS, bfr, yhjG, ydaM, ygaU, yehW, psiF, dps, yjdN, fbaB, ecnB, yebV, yccJ, adhP, yahK, osmC, treA, ybhN, ahr, amyA, otsA, ybgA, yehY, yegS, osmF, ybdK, ydhS, yegP, ggt, ygaM, yfcG, sra, patA, msyB, yeaH, yeaG, otsB, wrbA, elaB, yahO, blc, clsB, talA, ycgB, tktB, ybhP, poxB, osmY, katE, yiaG, ycaC*
	Translation	*rpsI, rpmC, rpsF, rpsH, rplM, rpsQ, rplI, rpsE, rplR, rplP, priB, rpsC, rplC, rpsJ, rplB, rplW, rplD, rpsS, rplV*
	BluR	*rcnA, yegR, yhjX, yjbJ, ycgZ, ymgA, ariR, ymgC*
Metal Homeostasis	Fur-1	*sodB, ftnA, grcA, tonB, ybiI, fecC, exbB, exbD, fhuF, efeU_2, pqqL, ydiE, fepE, fecA, yddA, yqjH, fhuA, efeB, fhuD, yddB, fecI, fhuC, efeO, sufA, fhuB, sufC, sufB, sufS, sufE, fecR, fepD, nrdH, sufD, fepB, fepG, yncE, fepC, nrdI, ybdZ, entS, nrdF, nrdE, entD, ybiX, fhuE, fes, entA, entH, entF, entC, entB, entE, fepA, cirA, fiu*
	Uncharacterized-5	*sodB, yoeG_1, dppB, dppF, dppD, dppC, fumA, nuoH, nuoI, grxD, iscS, nuoG, iscA, nuoC, nuoL, nuoM, sdhB, iscU, fur, fdx, narU, nuoN, nuoF, iscX, nuoE, pepB, fhuA, nuoJ, sodA, bfr, fecR, fecI, gpmA, ybiX, fecA, entD, ybdZ, mntH, yjjZ*
	CusR	*yedV, yncJ, cpxP, yjjZ, yedW, cueO, copA, cusS, cusR, cusA, cusB, cusF, cusC*
Respiration	FnR/IscR	*iscR, iscS, hscB, iscU, nfuA, cysE, erpA, iscA, hypB, nikB, yehD, ydhY, abrB, ynfO, napF, ychH, ylcI, yfcC, yoeA_1, ynfE, nrdD, glpA, hybO, ydfZ, yhbV, fdnG, feoA, ynjE, dmsA, nirB, ydjY, yjjI, yhcC, ttdR, narK, yecH, focA, bssR, ydjX, grcA, ynfK, ompW, yhbU, nikA*
	ArcA	*feoB, cydA, lomR_1, rsd, bluF, ydgC, bssR, lldR, efeU_2, yjiJ, ugpA, yjiR, fadI, fumA, kgtP, actP, ylaC, ugpE, ndk, gltA, hcaR, msrB, sucB, prpR, ugpB, sucD, sucC, aldA, sucA, lldP, astA, sthA, puuA, acs, astC, sdhB, fadB, yigI, mhpR, sdhA, phoH, puuD, sdhD, ydcI, glcC, sdhC, yejG*
	Pyruvate	*aceA, pta, yohJ, aceF, yjiA, ykgE, ackA, mqo, ldhA, yhjX, yjiY*
Protein Folding	RpoH	*rhsA, ybeX, htpX, rlmE, ldhA, yeaD, yhjX, topA, gapA, bssS, miaA, tusB, mfd, fliK, ybeY, zntR, yhdN, ybeZ, ybeD, hspQ, rsmJ, prlC, lon, ybbN, hslO, hslR, mutM, grpE, hslU, hslV, ibpB, ycjF, fxsA, ycjX, ibpA, dnaJ, groL, groS, clpB, htpG, dnaK*
Nucleotide and Amino Acid Biosynthesis	PurR-2	*pyrC, pyrD, upp, ridA, codA, carA, codB, carB, uraA, pyrI, pyrB*
	PurR-1	*ydhC, add, cspG, ydiJ, purN, purF, cvpA, purD, ghxP, purH, purC, purL, purT, purK, purE, purM, xanP*
	ArgR	*yagI, carB, argE, dtpD, carA, argH, argD, argG, argB, argA, artJ, argC, argF, argI*
	CysB	*dcyD, gsiD, yecC, yecS, gsiB, tauA, ygbE, tauB, gsiA, iaaA, nlpA, cysM, fliY, cysK, ydjN, yeeD, cbl, yeeE, cysH, sbp, cysA, cysI, cysW, cysC, cysN, yciW, cysU, cysJ, cysD, cysP*
	MetJ	*fur, ilvC, yghB, folE, metJ, yiiX, metQ, metC, metI, metK, mmuM, metL, metN, ybdH, metR, metE, ybdL, metB, mmuP, metA, metF*
	His-tRNA	*hisI, hisF, hisA, hisH, hisB, hisD, hisC, hisG*
	Tryptophan	*tnaA, tnaB, aroL, tyrA, aroF, aroH, trpB, trpC, trpA, mtr, trpD, trpE*
	BCAA-2	*ilvE, ilvA, ilvD, ilvC, thrC, thrB, thrA, ilvG_1, ilvM, ilvG_2*
	BCAA-1	*hmp, ilvC, leuD, ilvB, leuA, leuC, leuB, ilvN*
	TyrR	*mtr, tyrP, aroL, tyrA, aroF*

3.2.4 Shifts in i-modulon activities during heterologous gene expression relative to WT

We calculated the percentage variance explained for the activities of each i-modulon across the 84 samples (Figure 3.3B). The eight i-modulons with the largest activity changes explain over 50% of the total variance. They are RpoS (13.07%), Fur-1 (12.28%), Uncharacterized-4 (5.74%), Uncharacterized-0 (4.65%), Uncharacterized-47 (4.62%), RpoH (4.50%), Fnr/IscR (3.96%), and CusR (3.34%). Each of these i-modulons represent major changes of gene expression between WT and protein expressing strains and they identify the processes that differentially respond to heterologous gene expression.

By evaluating the functions and activity levels of the i-modulons which account for a large percent variance of transcriptome, we categorized these i-modulons into four modes of host responses to heterologous protein expression (Figure 3.3A, Table 3.1). Below, we explore each of these four host responses and highlight the insights gained.

3.2.5 Fear vs. greed: The trade-off between expression and stress

The fear versus greed dimension of host cell response consists of three i-modulons: the RpoS i-modulon, the Translation i-modulon, and a previously uncharacterized i-modulon (96). This host response represents 16.55% of the total variance in the dataset and reflects one of the major balancing acts within the host: the expression of stress mitigation and "hedging" functions ("Fear"), as opposed to the expression of growth related functions ("Greed") [28, 29]. This tradeoff is made apparent by plotting the activities of the RpoS i-modulon against those of the Translational i-modulon [17].

The RpoS i-modulon is the i-modulon that exhibits the largest percentage variance across

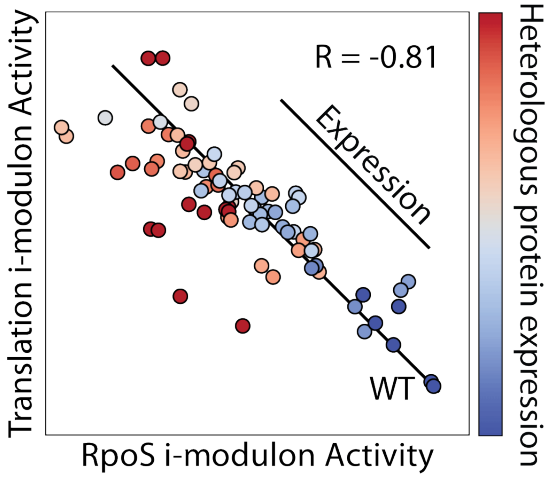

Figure 3.4: The "fear vs. greed" tradeoff is represented by the inverse correlations of the activity level of the RpoS and Translation i-modulons. The RpoS i-modulon is composed of various functions upregulated during stressful conditions - i.e., "fear", and is negatively correlated with the Translation i-modulon, consisting of various ribosomal proteins that are related to growth - i.e., "greed". Expression levels of heterologous proteins have been overlaid over these i-modulon activities demonstrating that increased expression has the effect of migrating upwards along the fear/greed tradeoff line.

the dataset (13.06%), and consists of genes which are controlled by the stress response sigma factor (RpoS). The Translation i-modulon consists of 18 ribosomal proteins, and represents the cellular translational capacity. We were interested in exploring how the expression of heterologous protein varied with the strain's position on this fear/greed tradeoff line. As previously shown [17], the activity of the RpoS i-modulon shows a clear negative correlation with the activity of the Translation i-modulon (R = -0.81, p-value = 4.02 x 10^{-21}). Interestingly, when we overlaid the expression levels of heterologous proteins, we find that strains which express higher levels of heterologous mRNA actually exhibit lower levels of stress compared to wild type (Figure 3.4).

Further, we find another previously uncharacterized i-modulon which is highly correlated with the RpoS i-modulon (R = 0.82, p-value = 4.03 x 10^{-21}) (Supp. Figure B.4). This i-modulon is made up of genes *ycgZ*, *ymgA*, *ariR*, and *ymgC*, all of which are predicted to have stress and biofilm related functions and are co-regulated by the transcriptional repressor BluR and RpoS

[30], hence its being labeled the BluR i-modulon. An increase in clumping and biofilm formation in *E. coli* during stress caused by overexpression of foreign protein has been noted by previous studies [31–33].

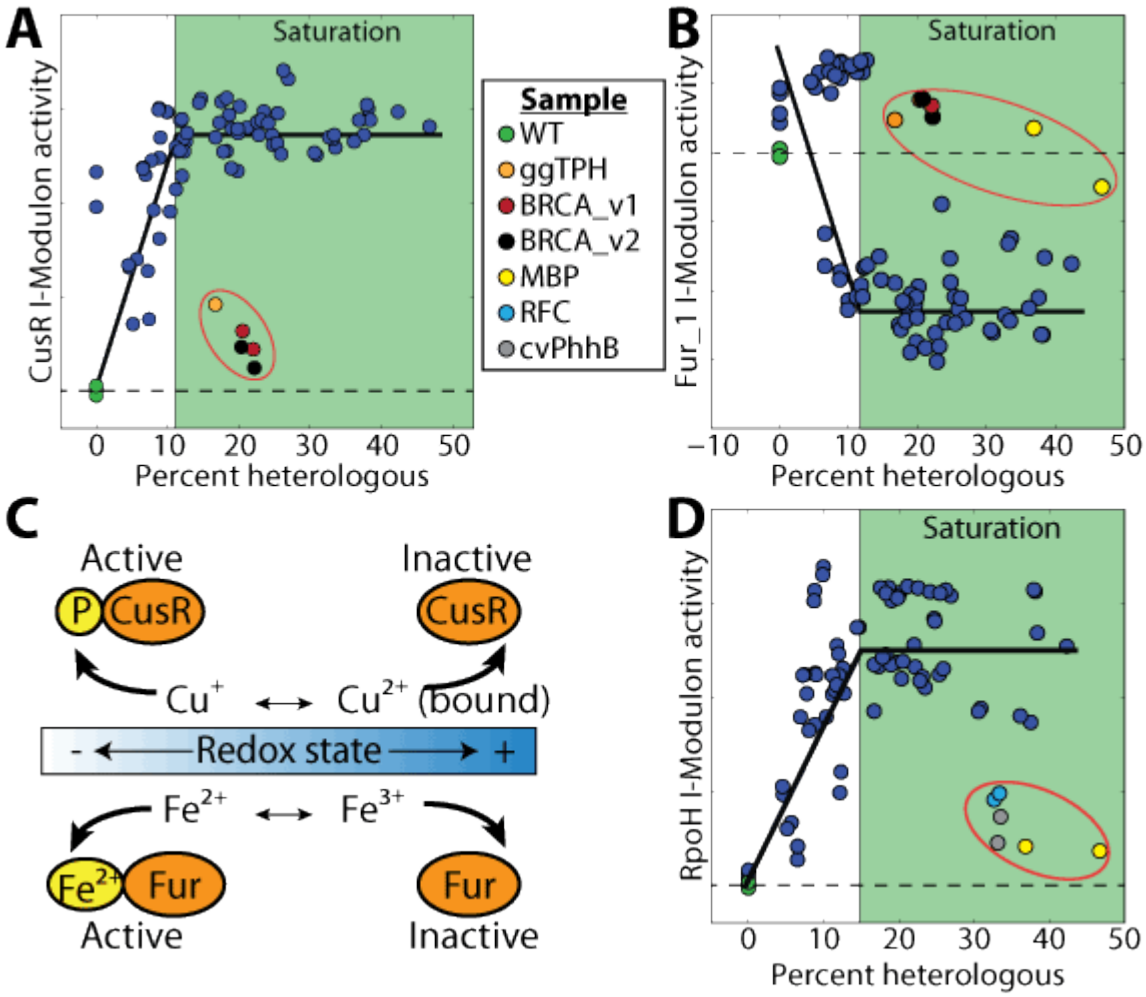

Figure 3.5: Activity levels of metal homeostasis and RpoH i-modulons are correlated with expression levels. A and B: Activity levels of CusR (A) and Fur-1 (B) i-modulons. CusR activity levels increase with heterologous protein expression levels, while Fur-1 decreases with heterologous protein expression. All activity levels are normalized to WT (in green) which is set to 0. Both display a saturation effect at high levels of heterologous protein expression. Both CusR and Fur-1 i-modulons have similar outlier samples ggTPH, BRCA_v2, BRCA_v1 (circled in red), suggesting a common underlying driver. C: Both CusR and Fur regulons have been shown to be affected by oxygen availability and redox state in the cell due to changes in oxidation states of the metals [34, 35]. D: RpoH i-modulon shows a positive correlation to heterologous protein expression and exhibits a similar saturation effect at high heterologous protein expression levels.

3.2.6 Metal homeostasis and respiration: CusR and Fur regulons are activated during heterologous protein expression

The Fur-1 and CusR i-modulons reflect a large percentage of the variance in the RNAseq dataset, representing 11.29% and 3.41%, respectively. These i-modulons contain genes controlled by Fur and CusR, respectively, most of which maintain metal homeostasis of iron and copper ions in *E. coli* [36, 37]. Interestingly, the CusR and Fur i-modulons exhibit a positive and negative correlation with the level of heterologous mRNA expression, respectively. Both also have similar outlier proteins (ggTPH, BRCA_v1, and BRCA_v2) with the addition of MBP for Fur, suggesting that their activity levels could share a common underlying cause (Figure 3.5AB).

The cellular response for respiration consists of the i-modulons for ArcA and Fnr/IscR and Pyruvate which are found to account for 1.17%, 3.96%, and 1.77% of the variance in the dataset, respectively. The activity levels of these i-modulons were generally poorly correlated with the level of heterologous gene expression (ArcA: R = 0.07, p-val = 0.49, Fnr/IscR: R = 0.16, p-value = 0.13, Pyruvate: R = -0.17, p-value = 0.12) indicating that oxygen availability does not have a direct effect on the expression levels. However, we find that the activity of the CusR and FNR/IscR i-modulon are correlated (R = 0.49, p-val = $1.5\text{x}10^{-6}$) (Supp. Fig. B.5).

In addition, we also find a previously uncharacterized i-modulon (uncharacterized-5) which is highly correlated with the Fur-1 i-modulon (R = 0.95, p-value = 1.29e^{-43}, Supp. Fig. B.6). This i-modulon contains several genes regulated by Fur such as *fecA*, *fecR*, and *fecI*, as well as *fur* itself, but also contains several genes regulated by ArcA and Fnr such as *sodA*, *sodB*, and those encoding for the chains of NADH:ubiquinone oxidoreductase. The main activity for this i-modulon is driven by genes co-regulated by Fur and Fnr or ArcA (Supp. Fig. B.7). Involvement of multiple transcription factors reflects coordination between metal homeostasis

and respiration.

3.2.7 Protein folding: Heterologous genes show a varied protein folding response

The overexpression of heterologous protein results in the heat shock response, characterized by the upregulation of genes such as *ibpA*, *ibpB*, *groL*, and *lon* [12, 14, 38–41], See Table 3.1. These genes are in the RpoH i-modulon, which represented approximately 4.77% of the variance in the dataset. The RpoH i-modulon represents the third highest variance among well-characterized i-modulons, after RpoS and Fur-1.

For most of the samples in our dataset, the activity level of the RpoH i-modulon increases with heterologous gene expression up to a saturation level. One of the samples, maltose-binding protein (MBP), has previously been found to fold extremely well when overexpressed, and has even been used as a solubility factor to help increase solubility when fused to problematic foreign proteins [42]. Predictably, MBP activated the RpoH i-modulon much less than would be expected based on its expression levels, suggesting that the host cells experience much lower protein misfolding stress than the rest of the other heterologous proteins. Interestingly, two other proteins, RFC and cvPhhB, show a similar low activation of RpoH activity, indicating that the protein misfolding response is target specific (Figure 3.5D).

3.2.8 Amino acid and nucleotide biosynthesis

One of the major effects of heterologous protein expression on the host is elevated metabolic burden and distortion in the use of biosynthetic pathways, specifically through the elevated need for nucleotides and certain amino acids [27]. Maintenance of a plasmid requires

the synthesis of additional nucleotides above the WT strain's needs for replication and DNA repair. This problem is exacerbated when using high copy number plasmids. Expression of both the protein of interest as well as additional plasmid associated proteins such as selection markers and the promoter proteins require additional usage of amino acids. Collectively, the eight amino acid and two nucleotide biosynthesis i-modulons which were uncovered during ICA decomposition accounted for 9.02% of the total variance in the dataset.

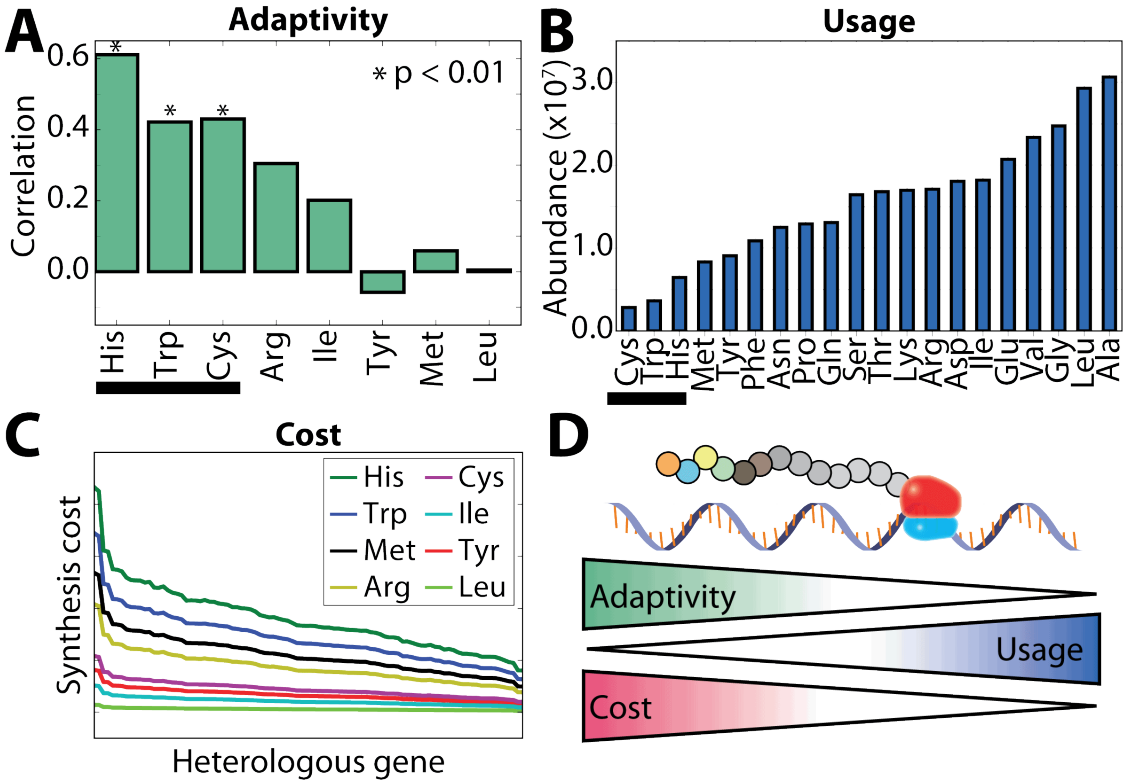

Figure 3.6: Adaptivity, usage, and costs of amino acids are linked during heterologous gene expression. A: Correlation of amino acid biosynthesis i-modulons with their global usage in the transcriptome across the dataset. Histidine, tryptophan, cysteine, isoleucine, and arginine were positively correlated while tyrosine, methionine, and leucine were not, showing varied sensitivity of amino acids to their usage. B: The three most correlated amino acids match with the three least commonly used amino acids in wild type under induction conditions. C: Computationally predicted synthesis cost of the top eight highest cost amino acids. Amino acids which are correlated well with their usage have a higher predicted synthesis cost. D: Cells have to balance the biosynthetic costs of each amino acid against their usage, and therefore exert a tighter transcriptional control over the biosynthetic pathways of costly amino acids (adaptivity).

3.2.9 Histidine, tryptophan, cysteine, isoleucine, and arginine i-modulons show sensitivity to global amino acid usage

To further explore the changes in metabolic burden, we examined the amino acid biosynthesis related i-modulons. Eight i-modulons were found that were related to amino acid biosynthesis pathways (Met, Arg, Trp, His, Cys, Tyr, Leu, Ile/Thr). Of these eight i-modulons, the activity levels of the i-modulons for His-tRNA, Tryptophan, CysB, Isoleucine, and ArgR were correlated with the usage of their respective amino acid residues in the transcriptome, while the activities for the i-modulons related to Tyr, Met, and Leu were not (Figure 3.6A). This leads to the conclusion that E. coli is more adaptive to changes in the usage of some amino acids than others. The highest correlations were for histidine, tryptophan, and cysteine, which correspond to the least commonly used amino acids in the proteome in WT cells (Figure 3.6B).

In order to better understand the cost of biosynthesis of each amino acid in the context of heterologous protein production, we made use of a genome-scale model of metabolism and expression in *E. coli* (ME model) [43] to simulate expression of each heterologous protein. The marginal cost of a metabolite (known as the 'shadow price'), of each amino acid was calculated for each simulation [43] and, with the exception of methionine, the amino acids with the highest shadow price were found to match the amino acid biosynthesis i-modulons that were most highly correlated with their usage (Figure 3.6C). Together, these results contribute to a better understanding of amino acid usage in the cell during heterologous protein expression. Amino acids such as histidine, tryptophan, and cysteine tend to be the most rarely used during normal growth, and are also the most costly to biosynthesize. As such, cells exert the tightest controls over the transcription of their biosynthetic pathways, leading to the enhanced adaptivity we see in our dataset of the most expensive and rare amino acids (Figure 3.6D).

3.3 Discussion

Heterologous protein expression in *E. coli* continues to be a major issue in biotechnology and the rational design of protein sequences for successful expression remains a challenge. Here, we address the complementary issue, and categorize the major host responses when confronted with expression of a range of 40 heterologous proteins. Analysis of RNAseq data reveals four key host responses that vary in intensity for a particular heterologous protein.

The first, and most prominent, result is the fear versus greed tradeoff and its implications when it is overlaid with heterologous gene expression levels. Single mutations on the *rpoB* or *rpoC* gene have previously been seen to dramatically increase cell fitness and growth rate by migrating these mutants along the greed/fear tradeoff line [17], increasing production of ribosomal proteins at the expense of stress functions [28, 44]. Intuitively, increasing overexpression of a single heterologous protein should increases the stress response experienced by the cell. However, across various heterologous proteins, higher levels of heterologous gene expression is instead correlated with lower RpoS response compared to wild type. The inverse correlation with growth, as well as expression, suggests that this set of genes are a possible avenue for host cell genetic manipulation in order to increase expression.

Second, we found that two transcription factors associated with metal homeostasis, Fur and CusR, were well correlated with the expression of heterologous genes. The large variability of their i-modulons is surprising in this experiment because all the samples were exposed to identical external concentrations of these ions. One explanation is that this reflects a destabilization of the redox balance within the host. One of the major roles of Fur in the cell is to sense oxidative stress and protect essential iron-containing Fe-S enzymes from damage. Changes to the redox potential in the host has been shown to affect both iron and copper homeostasis by increasing

the availability of Fe^{2+} binding to Fur, changing its regulatory footprint [34] as well as increasing the toxicity of copper via reduction of Cu^{2+} to the more dangerous Cu^{+} [35](Figure 3.5C). It has been shown that recombinant protein synthesis resulted in excess NADPH being produced in the cell [45], as well as increased demand on ATP production resulting in increased NADH and O_2 consumption [46], while recombinant protein expression has been shown to disrupt redox balance in *S. cerevisiae* through consumption of glutathione and GABA [47]. The correlation of the Fur i-modulon activity with the uncharacterized-5 (Supp. Fig. B.6) i-modulon, as well as the correlation between the CusR i-modulon activity and the Fnr/IscR i-modulon (Supp. Fig. B.5), further serves to reinforce this link between the redox state of the cell and the metal homeostasis regulons.

Third, the RpoH host response serves to illustrate a previously known phenomenon from the perspective of the host. High levels of heterologous gene expression have been known to induce the heat shock response to the large amounts of foreign protein which require folding. However, proteins have differential folding requirements as is demonstrated in this study in the RpoH i-modulon response. MBP is a particularly interesting case because it is often used in recombinant protein expression as a solubility tag, able to help solubilize insoluble protein when appended [42]. Of interest are the saturation dynamics we see with some of the important i-modulons such as CusR, Fur, and RpoH. During overexpression, cells have to contend with limits to their transcriptional, translational, and proteomic capacity [29, 48], and maintain a balance between core cellular processes of replication, growth, and energy generation, and stress functions of protein misfolding and metal homeostasis. As the proportion of the transcriptome taken up by the heterologous gene grows, lower capacity is available for these secondary functions.

Fourth, we examined the metabolic burden of plasmid maintenance and heterologous

protein expression that has been described in terms of bottlenecks in amino acid and nucleotide biosynthesis [27]. We find that certain amino acid biosynthesis pathways show more sensitivity to their usage, namely histidine, tryptophan and cysteine, while other amino acids act as free variables. One explanation could be that the biosynthesis of more energetically expensive amino acids could be more tightly coupled to their usage to reduce wastage. An example of this is histidine, which is one of the most expensive amino acids to synthesize, requiring between 31-41 ATP per molecule [49, 50]. The His-tRNA i-modulon consists of genes in the histidine biosynthetic pathway in the histidine operon (hisA, hisB, hisC, hisD, hisF, hisG, and hisI). Transcription of the histidine regulon is regulated by an attenuation mechanism whereby low levels of charged his-tRNA result in translational stalling along the histidine leader peptide, preventing formation of a rho-independent terminator and allowing the transcription of the rest of the operon [51, 52]. This mechanism effectively couples synthesis of the histidine biosynthesis genes directly to available levels of histidine in the host, reducing overproduction and waste of cellular resources.

Taken together, expressing a set of heterologous protein under the same conditions in the same host, coupled with ICA analysis of the corresponding RNAseq data set, reveals the host processes that are differentially activated during heterologous protein expression. With these host responses identified and the concomitant fundamental understanding of the underlying mechanisms, further efforts should be focused on the identification of approaches that mitigate these responses. If successful, predictive host engineering or controlling conditions for improved heterologous protein expression will have been achieved.

Acknowledgements

J.T., B.G.V., S.W.S. and B.O.P. designed the study, J.T., K.S.F, S.P.B and A.H. performed the experiment. J.T. and A.V.S. analyzed the data. J.T. and B.O.P. wrote the manuscript, with contributions from all the other co-authors.

This work was funded by the Novo Nordisk Foundation Grant Number NNF10CC1016517. We would like to thank Rebecca Lennen for her contribution of the gene templates and help in designing the plasmid backbone. We would also like to Stefan Kol, Laurence Yang and Daniel Zielinski for many valuable discussions. We would like to thank Marc Abrams for assistance with manuscript editing.

Chapter 3 in part is a reprint of material published in: **J Tan**, AV Sastry, KS Fremming, SP Bjørn, A Hoffmeyer, SW Seo, BG Voldborg, BO Palsson. "Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression" *Submitted*. The dissertation author is the primary author.

3.4 References

1. Selas Castiñeiras, T., Williams, S. G., Hitchcock, A. G. & Smith, D. C. *E. coli* strain engineering for the production of advanced biopharmaceutical products. en. *FEMS microbiology letters* **365.** ISSN: 0378-1097, 1574-6968. doi:10.1093/femsle/fny162 (Aug. 2018).

2. Gileadi, O., Burgess-Brown, N. A., Colebrook, S. M., Berridge, G., Savitsky, P., Smee, C. E. A., Loppnau, P., Johansson, C., Salah, E. & Pantic, N. H. High throughput production of recombinant human proteins for crystallography. en. *Methods in molecular biology* **426,** 221–246. ISSN: 1064-3745 (2008).

3. Savitsky, P., Bray, J., Cooper, C. D. O., Marsden, B. D., Mahajan, P., Burgess-Brown, N. A. & Gileadi, O. High-throughput production of human proteins for crystallization: the SGC experience. en. *Journal of structural biology* **172,** 3–13. ISSN: 1047-8477, 1095-8657 (Oct. 2010).

4. Agashe, D., Martinez-Gomez, N. C., Drummond, D. A. & Marx, C. J. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous muta-

tions in a key enzyme. en. *Molecular biology and evolution* **30,** 549–560. ISSN: 0737-4038, 1537-1719 (Mar. 2013).

5. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. en. *Science* **324,** 255–258. ISSN: 0036-8075, 1095-9203 (Apr. 2009).

6. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. en. *Nature reviews. Genetics* **12,** 32–42. ISSN: 1471-0056, 1471-0064 (Jan. 2011).

7. Frumkin, I., Schirman, D., Rotman, A., Li, F., Zahavi, L., Mordret, E., Asraf, O., Wu, S., Levy, S. F. & Pilpel, Y. Gene Architectures that Minimize Cost of Gene Expression. en. *Molecular cell* **65,** 142–153. ISSN: 1097-2765, 1097-4164 (Jan. 2017).

8. Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K.-H., Su, M., Luff, J. D., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P. & Hunt, J. F. Codon influence on protein expression in *E. coli* correlates with mRNA levels. en. *Nature* **529,** 358–363. ISSN: 0028-0836, 1476-4687 (Jan. 2016).

9. Cambray, G., Guimaraes, J. C. & Arkin, A. P. *Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli* 2018. doi:10.1038/nbt.4238.

10. Sastry, A., Monk, J., Tegel, H., Uhlén, M., Palsson, B. O., Rockberg, J. & Brunk, E. Machine Learning in Computational Biology to Accelerate High-Throughput Protein Expression. en. *Bioinformatics*. ISSN: 1367-4803, 1367-4811. doi:10.1093/bioinformatics/btx207 (Apr. 2017).

11. Farkas, Z., Kalapis, D., Bódi, Z., Szamecz, B., Daraba, A., Almási, K., Kovács, K., Boross, G., Pál, F., Horváth, P., Balassa, T., Molnár, C., Pettkó-Szandtner, A., Klement, É., Rutkai, E., Szvetnik, A., Papp, B. & Pál, C. Hsp70-associated chaperones have a critical role in buffering protein production costs. en. *eLife* **7.** ISSN: 2050-084X. doi:10.7554/eLife.29845 (Jan. 2018).

12. Lesley, S. A., Graziano, J., Cho, C. Y., Knuth, M. W. & Klock, H. E. Gene expression response to misfolded protein as a screen for soluble recombinant protein. en. *Protein engineering* **15,** 153–160. ISSN: 0269-2139 (Feb. 2002).

13. Dong, H., Nilsson, L. & Kurland, C. G. Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. en. *Journal of bacteriology* **177,** 1497–1504. ISSN: 0021-9193 (Mar. 1995).

14. Sharma, A. K., Mahalik, S., Ghosh, C., Singh, A. B. & Mukherjee, K. J. Comparative transcriptomic profile analysis of fed-batch cultures expressing different recombinant proteins in *Escherichia coli*. en. *AMB Express* **1,** 33. ISSN: 2191-0855 (Oct. 2011).

15. Arís, A., Corchero, J. L., Benito, A., Carbonell, X., Viaplana, E. & Villaverde, A. The expression of recombinant genes from bacteriophage lambda strong promoters triggers the

SOS response in *Escherichia coli*. en. *Biotechnology and bioengineering* **60,** 551–559. ISSN: 0006-3592 (Dec. 1998).

16. Dürrschmid, K., Reischer, H., Schmidt-Heck, W., Hrebicek, T., Guthke, R., Rizzi, A. & Bayer, K. Monitoring of transcriptome and proteome profiles to investigate the cellular response of *E. coli* towards recombinant protein expression under defined chemostat conditions. en. *Journal of biotechnology* **135,** 34–44. ISSN: 0168-1656 (May 2008).

17. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. *The Escherichia coli Transcriptome Consists of Independently Regulated Modules* en. Apr. 2019.

18. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. en. *Nature communications* **9,** 1090. ISSN: 2041-1723 (Mar. 2018).

19. Karczewski, K. J., Snyder, M., Altman, R. B. & Tatonetti, N. P. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. en. *PLoS genetics* **10,** e1004122. ISSN: 1553-7390, 1553-7404 (Feb. 2014).

20. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Applied and environmental microbiology* **83.** ISSN: 0099-2240, 1098-5336. doi:10.1128/AEM.00410-17 (July 2017).

21. Goltermann, L., Borch Jensen, M. & Bentin, T. Tuning protein expression using synonymous codon libraries targeted to the 5' mRNA coding region. en. *Protein engineering, design & selection: PEDS* **24,** 123–129. ISSN: 1741-0126, 1741-0134 (Jan. 2011).

22. Tsukuda, M. & Miyazaki, K. Directed evolution study unveiling key sequence factors that affect translation efficiency in Escherichia coli. en. *Journal of bioscience and bioengineering* **116,** 540–545. ISSN: 1389-1723, 1347-4421 (Nov. 2013).

23. Evfratov, S. A., Osterman, I. A., Komarova, E. S., Pogorelskaya, A. M., Rubtsova, M. P., Zatsepin, T. S., Semashko, T. A., Kostryukova, E. S., Mironov, A. A., Burnaev, E., Krymova, E., Gelfand, M. S., Govorun, V. M., Bogdanov, A. A., Sergiev, P. V. & Dontsova, O. A. Application of sorting and next generation sequencing to study 5'-UTR influence on translation efficiency in *Escherichia coli*. en. *Nucleic acids research* **45,** 3487–3502. ISSN: 0305-1048, 1362-4962 (Apr. 2017).

24. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q.-A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P. & Endy, D. Precise and reliable gene expression via standard transcription and translation initiation elements. en. *Nature methods* **10,** 354–360. ISSN: 1548-7091, 1548-7105 (Apr. 2013).

25. Nieuwkoop, T., Claassens, N. J. & van der Oost, J. Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. en. *Microbial biotechnology* **12,** 173–179. ISSN: 1751-7915 (Jan. 2019).

26. Sharp, P. M. & Li, W. H. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. en. *Nucleic acids research* **15,** 1281–1295. ISSN: 0305-1048 (Feb. 1987).

27. Wu, G., Yan, Q., Jones, J. A., Tang, Y. J., Fong, S. S. & Koffas, M. A. G. Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. en. *Trends in biotechnology* **34,** 652–664. ISSN: 0167-7799, 1879-3096 (Aug. 2016).

28. Utrilla, J., O'Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M. & Palsson, B. O. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. en. *Cell systems* **2,** 260–271. ISSN: 2405-4712 (Apr. 2016).

29. Scott, M., Klumpp, S., Mateescu, E. M. & Hwa, T. Emergence of robust growth laws from optimal regulation of ribosome synthesis. en. *Molecular systems biology* **10,** 747. ISSN: 1744-4292 (Aug. 2014).

30. Tschowri, N., Lindenberg, S. & Hengge, R. Molecular function and potential evolution of the biofilm-modulating blue light-signalling pathway of *Escherichia coli*. *Molecular microbiology* **85,** 893–906. ISSN: 0950-382X (2012).

31. Gomes, L. C. & Mergulhão, F. J. Effect of heterologous protein expression on *Escherichia coli* biofilm formation and biocide susceptibility (2016).

32. Glick, B. R., Brooks, H. E. & Pasternak, J. J. Physiological effects of plasmid DNA transformation on *Azotobacter vinelandii*. *Canadian journal of microbiology* **32,** 145–148. ISSN: 0008-4166 (Feb. 1986).

33. Castonguay, M.-H., van der Schaaf, S., Koester, W., Krooneman, J., van der Meer, W., Harmsen, H. & Landini, P. Biofilm formation by *Escherichia coli* is stimulated by synergistic interactions and co-adhesion mechanisms with adherence-proficient bacteria. en. *Research in microbiology* **157,** 471–478. ISSN: 0923-2508 (June 2006).

34. Beauchene, N. A., Mettert, E. L., Moore, L. J., Keleş, S., Willey, E. R. & Kiley, P. J. O2 availability impacts iron homeostasis in *Escherichia coli*. en. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 12261–12266. ISSN: 0027-8424, 1091-6490 (Nov. 2017).

35. Fung, D. K. C., Lau, W. Y., Chan, W. T. & Yan, A. Copper efflux is induced during anaerobic amino acid limitation in *Escherichia coli* to protect iron-sulfur cluster enzymes and biogenesis. en. *Journal of bacteriology* **195,** 4556–4568. ISSN: 0021-9193, 1098-5530 (Oct. 2013).

36. Seo, S. W., Kim, D., Latif, H., O'Brien, E. J., Szubin, R. & Palsson, B. O. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. en. *Nature communications* **5,** 4910. ISSN: 2041-1723 (Sept. 2014).

37. Munson, G. P., Lam, D. L., Outten, F. W. & O'Halloran, T. V. Identification of a copper-responsive two-component system on the chromosome of *Escherichia coli* K-12. en. *Journal of bacteriology* **182,** 5864–5871. ISSN: 0021-9193 (Oct. 2000).

38. Gill, R. T., Valdes, J. J. & Bentley, W. E. A comparative study of global stress gene regulation in response to overexpression of recombinant proteins in *Escherichia coli*. en. *Metabolic engineering* **2,** 178–189. ISSN: 1096-7176 (July 2000).

39. Rinas, U. Synthesis rates of cellular proteins involved in translation and protein folding are strongly altered in response to overproduction of basic fibroblast growth factor by recombinant *Escherichia coli*. en. *Biotechnology progress* **12,** 196–200. ISSN: 8756-7938, 1520-6033 (Mar. 1996).

40. Schweder, T. & Jürgen, B. in *Recombinant Protein Production with Prokaryotic and Eukaryotic Cells. A Comparative View on Host Physiology: Selected articles from the Meeting of the EFB Section on Microbial Physiology, Semmering, Austria, 5th–8th October 2000* (eds Merten, O.-W., Mattanovich, D., Lang, C., Larsson, G., Neubauer, P., Porro, D., Postma, P., de Mattos, J. T. & Cole, J. A.) 359–369 (Springer Netherlands, Dordrecht, 2001). ISBN: 9789401597494. doi:10.1007/978-94-015-9749-4_27.

41. Sørensen, H. P. & Mortensen, K. K. Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. en. *Journal of biotechnology* **115,** 113–128. ISSN: 0168-1656 (Jan. 2005).

42. Sun, P., Tropea, J. E. & Waugh, D. S. Enhancing the solubility of recombinant proteins in *Escherichia coli* by using hexahistidine-tagged maltose-binding protein as a fusion partner. en. *Methods in molecular biology* **705,** 259–274. ISSN: 1064-3745, 1940-6029 (2011).

43. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology* **9,** 693. ISSN: 1744-4292 (Oct. 2013).

44. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Applied and environmental microbiology* **81,** 17–30. ISSN: 0099-2240, 1098-5336 (Jan. 2015).

45. Heyland, J., Blank, L. M. & Schmid, A. Quantification of metabolic limitations during recombinant protein production in *Escherichia coli*. en. *Journal of biotechnology* **155,** 178–184. ISSN: 0168-1656, 1873-4863 (Sept. 2011).

46. Weber, J., Hoffmann, F. & Rinas, U. Metabolic adaptation of *Escherichia coli* during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. *Biotechnology and Bioengineering* **80,** 320–330 (2002).

47. De Ruijter, J. C., Koskela, E. V., Nandania, J., Frey, A. D., *et al.* Understanding the metabolic burden of recombinant antibody production in *Saccharomyces cerevisiae* using a quantitative metabolomics approach. *Yeast.* ISSN: 0749-503X (2018).

48. O'Brien, E. J., Utrilla, J. & Palsson, B. O. Quantification and Classification of *E. coli* Proteome Utilization and Unused Protein Costs across Environments. en. *PLoS computational biology* **12,** e1004998. ISSN: 1553-734X, 1553-7358 (June 2016).

49. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. en. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 3695–3700. ISSN: 0027-8424 (Mar. 2002).

50. Ingle, R. A. Histidine biosynthesis. en. *The Arabidopsis book / American Society of Plant Biologists* **9,** e0141. ISSN: 1543-8120 (Feb. 2011).

51. Kulis-Horn, R. K., Persicke, M. & Kalinowski, J. Histidine biosynthesis, its regulation and biotechnological application in *Corynebacterium glutamicum*. *Microbial biotechnology* **7,** 5–25 (2014).

52. Carlomagno, M. S., Chiariotti, L., Alifano, P., Nappo, A. G. & Bruni, C. B. Structure and function of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons. en. *Journal of molecular biology* **203,** 585–606. ISSN: 0022-2836 (Oct. 1988).

Chapter 4

Adaptation to oxidative stress

4.1 Abstract

Bacterial response to oxidative stress is of fundamental importance. Oxidative stresses are endogenous, such as reactive oxidative species (ROS) production during respiration, or exogenous in industrial biotechnology, due to culture conditions, or product toxicity. The immune system inflicts strong ROS stress on invading pathogens. In this study we make use of Adaptive Laboratory Evolution (ALE) to generate two independent lineages of *Escherichia coli* with increased tolerance to superoxide stress by up to 500% compared to wild type. We found: 1) that the use of ALE reveals the genetic basis for and systems biology of ROS tolerance, 2) that there are only 6 and 7 mutations, respectively, in each lineage, five of which reproducibly occurred in the same genes (iron-sulfur cluster regulator *iscR*, putative iron-sulfur repair protein *ygfZ*, pyruvate dehydrogenase subunit E *aceE*, succinate dehydrogenase *sucA*, and glutamine tRNA *glnX*), and 3) that the transcriptome of the strain lineages exhibits two different routes of tolerance: the direct mitigation and repair of ROS damage and the up-regulation of cell motility

and swarming genes mediated through phosphate starvation, which has been linked to biofilm formation and aggregation. These two transcriptomic responses can be interpreted as 'flight' and 'fight' phenotypes.

4.2 Introduction

During aerobic respiration, leakage of high energy electrons from respiratory quinones and electron transport chain components creates reactive oxidative species (ROS), such as superoxide and peroxide radicals. Studies have placed the production rate of hydrogen peroxide during normal aerobic growth in *E. coli* to be as high as 10-15μM/s [1, 2]. In addition to the endogenous production of ROS, bacteria also experience oxidative stress from external sources such as the host immune response during pathogenic infections [3]. Production strains of *E. coli* also tend to experience redox stress due to high oxygen concentrations used in fermentation vessels [4], or after genetic modifications to the host cell such as the knockout of thioredoxin and glutathione biosynthesis genes [5].

Due to their reactivity, ROS cause damage to macromolecules in the cell. DNA damage is caused by direct oxidation of individual bases and cross-linking between strands [6]. Lipids are peroxidized, changing membrane permeability. ROS damage to proteins occurs via direct oxidation of vulnerable amino acids such as cysteine and methionine, and loss of metal cofactors [7]. A major target of ROS damage are iron-sulfur clusters, found in important catalytic proteins and vital for central carbon metabolism and amino acid biosynthesis [8, 9]. Iron-sulfur clusters readily change redox states [10], making them particularly vulnerable to damage by ROS: an initial loss of an electron causes destabilization and a subsequent loss of an iron ion [11], with further oxidation causing a complete degradation and loss of the iron-sulfur cluster. *E. coli* relies

on two systems, the Isc and Suf cluster assembly systems, to synthesize and repair damage to iron-sulfur clusters [11].

Free cytoplasmic iron is a double-edged sword during oxidative stress. On the one hand, it is used as the metal cofactor in superoxide dismutase sodB, which scavenges superoxide species; on the other, it participates in the Fenton reaction, which converts superoxide and peroxide molecules into the even more reactive hydroxyl radical [12]. Damage to iron-sulfur clusters caused by oxidative stress not only inactivates the iron-sulfur containing protein, but also releases iron into the cytoplasm. Fine control of iron concentrations in the cytoplasm is thus vital to adequately deal with oxidative stress. *E. coli* does this via the transcriptional regulator Fur which represses iron uptake systems when bound to free Fe^{2+}. Oxidative stress has been shown to upregulate Fur, resulting in the repression of iron-uptake systems and a reduction of iron levels in wild type cells [13].

Various methods have been explored in order to increase tolerance of E. coli to exogenous oxidative stress. Those that were successful made use of mutations to CRP [14], or the overexpression of damage mitigation genes such as sodC and katG [15, 16]. *E. coli* also exhibits a phenomenon known as cross-tolerance where various stressors such as cold shock and osmotic shock have also been found to induce the oxidative stress response [17, 18]. When previously exposed to sublethal levels of one type of stress, *E. coli* has shown improved tolerance to a subsequent exposure to lethal levels of another type of stress [19].

Due to its role in pathogenesis and importance in industrial biotechnology, understanding tolerance to ROS remains a significant area of research. In this paper we make use of tolerization adaptive laboratory evolution (TALE) [20] to increase the tolerance of *E. coli* to the redox-cycling compound paraquat. We then make use of genome resequencing, transcriptomics, and ribosome

profiling to gain insight into these adaptations and the mechanisms through which they increase tolerance.

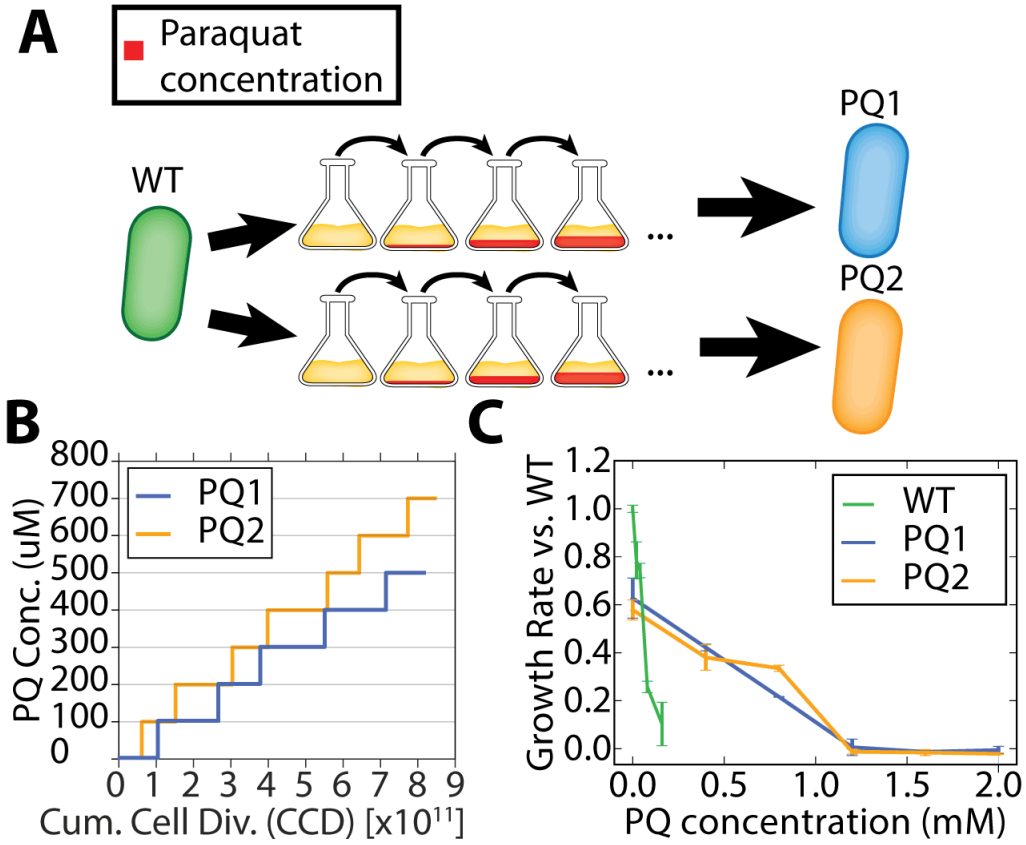

Figure 4.1: Description of the ALE experiment run and physiological characterization of the evolved strains. A: TALE was used to increase tolerance to paraquat. Paraquat concentration was raised in small increments over the course of the evolution, resulting in two end point strains. B: Graph showing paraquat concentration used in TALE against the number of cumulative cell divisions. C: Growth characterization of end point strains. WT shows loss of viability above 0.16mM, whilst end point strains show loss of viability at 0.8mM.

Table 4.1: Mutations found in evolved strains following TALE to increase tolerance to paraquat

Category	Gene	Strain	Mutation
Carbon flux	*aceE*	PQ1	Q791*
		PQ2	Q409*
	sucA	PQ1	R182L
		PQ2	G594S
	gltA	PQ2	L8P
Iron-sulfur cluster	*ygfZ*	PQ1	V107E
		PQ2	T108P
	iscR	PQ1	V55L
		PQ2	C104S
tRNA	*glnX*	PQ1	(C→ T 35/75nt)
		PQ2	(C→ T 35/75nt)
Phosphate uptake	*pitA*	PQ1	INS (+T 5/1500nt)
Polymyxin resistance	*arnA*	PQ2	A21A

4.3 Results

4.3.1 TALE increased tolerance to oxidative stress

Paraquat was chosen as the ROS stressor in these experiments since it is a known generator of internal superoxide stress through redox cycling. In order to differentiate mutations arising from adaptation to the culture media from those arising from oxidative stress, we make use of an *E. coli* K-12 MG1655 strain previously evolved on glucose minimal media as the starting strain [21]. Exponentially growing cultures were passed into incrementally higher paraquat concentrations to increase the tolerance to superoxide stress (Figure 4.1A).Two replicate end points (PQ1 and PQ2) were generated from the TALE which took place over 30 days, representing over 8 x 10^{11} cumulative cell divisions (CCD) [22] (Figure 4.1B).

To quantify the increase to paraquat tolerance, we performed growth experiments of the evolved strains across a range of paraquat concentrations between 0mM and 2.0mM. Wild type showed loss of viability at 0.16mM, while PQ1 and PQ2 only lost viability at paraquat concentrations above 0.8mM, a tolerance increase of 500% (Figure 4.1C).

4.3.2 Whole genome resequencing and mutation analysis reveal the genetic basis for increased ROS tolerance

Clonal isolates of the end point populations (PQ1, PQ2) were subject to whole genome resequencing in order to identify the genetic basis of tolerance to paraquat stress. Genetic mutations were called using the breseq computational pipeline (Materials and Methods). Key mutations related to tolerance to oxidative stress were determined by identifying genes or genetic regions that were mutated across multiple isolates from independent samples (Table 4.1). The full list of mutations is deposited in ALEdb [23]. Only six and seven mutations for PQ1 and PQ2, respectively, were required to confer tolerance to five times the maximum concentration of paraquat compared to WT.

Five genes were mutated in common: *aceE*, *ygfZ*, *iscR*, *sucA*, and *glnX* (Table 1). Two of these five genes (*aceE* and *sucA*) are related to carbon metabolism and the TCA cycle. *aceE* was one of the first genes to be mutated across both replicates, with both independent mutations resulting in truncation of the gene, at residues 791 and 409 for PQ1 and PQ2, respectively. *aceE* encodes the E1 subunit of the pyruvate dehydrogenase complex which catalyzes the reaction that converts pyruvate to acetyl-CoA, a key step for the entry of carbon flux into the TCA cycle during aerobic growth on glucose as the sole carbon source. *sucA*, on the other hand, encodes the E1 subunit of the 2-oxoglutarate dehydrogenase complex that catalyzes the conversion of 2-oxoglutarate into succinyl-CoA and CO2 along with the production of NADH. The TCA cycle is the main source of high energy electrons during aerobic respiration, and both these mutations would result in a reduction in carbon flux through the TCA cycle, reducing redox load. Interestingly, both *sucA* and *aceE* are the thiamine-binding E1 components of their respective dehydrogenase complexes, suggesting a significant vulnerability in thiamine production under

oxidative stress. Indeed, 2-iminoacetate synthase, coded for by the gene *thiH*, is a key enzyme in the thiamine-biosynthetic pathway and was found to contain a redox sensitive iron-sulfur cluster [24].

Another two commonly mutated genes are related to iron-sulfur cluster synthesis and repair (*iscR* and *ygfZ*). *iscR* is the transcriptional regulator for iron-sulfur cluster and biosynthesis and is involved in the regulation of the isc and suf operons [25]. The DNA binding affinity of *iscR* is dependent on the presence of iron-sulfur clusters bound to the protein. The *icsR* mutation site C104S in PQ2 is known to be one of the three conserved cysteine residues involved in iron-sulfur cluster binding [26], and mutations at this location have been shown to have an effect on the regulation of the *iscRSUA* and *sufABCDSE* operons [27]. *ygfZ*, was also found to be mutated in both PQ1 and PQ2 at amino acid residue positions 107 and 108 respectively. Though the function of *ygfZ* in *E. coli* is still unclear, it has been shown to be a folate-binding enzyme potentially involved in either the synthesis or repair of iron-sulfur cluster proteins [28, 29]. *ygfZ* has been hypothesized to also have a direct role in the degradation of plumbagin (a redox stress causing compound)[30], whilst inactivation of *ygfZ* has been found to result in increased sensitivity to oxidative stress in *E. coli* [29]. Positions 107 and 108 have previously determined to be conserved across *E. coli*, *M. tuberculosis* and *K. pneumoniae* [30] suggesting that this region might be critical to improving or modifying one or both of *ygfZ*'s hypothesized functions of Fe-S repair or plumbagin degradation.

The last common mutation was a mutation in *glnX*. *glnX* is one of the 4 glutamine tRNAs in *E. coli* which decodes the CAG codon in wild type. This mutation occurred at nucleotide 35 in the gene, changing the sequence of the anticodon from CTG to CTA, allowing the suppression of the amber stop codon TAG.

The end points also contained one and two strain specific mutations respectively. PQ1 contains a frameshift mutation in the *pitA* gene, encoding the low affinity phosphate transporter [31]. *pitA* is the major route for phosphate uptake in the cell under phosphate replete conditions, and disruption of its would severely limit phosphate availability in PQ1. PQ2 contains an additional mutation in *gltA*, the gene encoding citrate synthase, another member of the TCA cycle, and a silent mutation in arnA, a protein involved in polymyxin resistance. These mutations occur close to the 5' end of the gene at residue 8 and 21 respectively, which has been shown to affect gene expression levels through changes to mRNA secondary structure [32].

4.3.3 Knockout of *aceE* improves fitness at low levels of ROS stress

One of the first mutations to show up in both replicates during the first phase of tolerization was a truncation mutation at residue 409 and 791 for PQ1 and PQ2 respectively, suggesting that the disruption of *aceE* activity was important for the resistance to ROS stress. To better understand the role of the *aceE* truncation in improving strain fitness, we knocked out *aceE* in the wild type strain. The loss of *aceE* had a negative impact on strain viability in M9 minimal media with glucose as the sole carbon source, necessitating the addition of 10% LB to the media which greatly improved tolerance to paraquat, so paraquattolerance is not directly comparable to the evolved strains. However, we see that compared to WT grown in the same media, $\Delta aceE$ shows improved fitness when concentration of paraquat is low, but loses this fitness advantage at higher levels of paraquat (Supplementary Figure C.1).

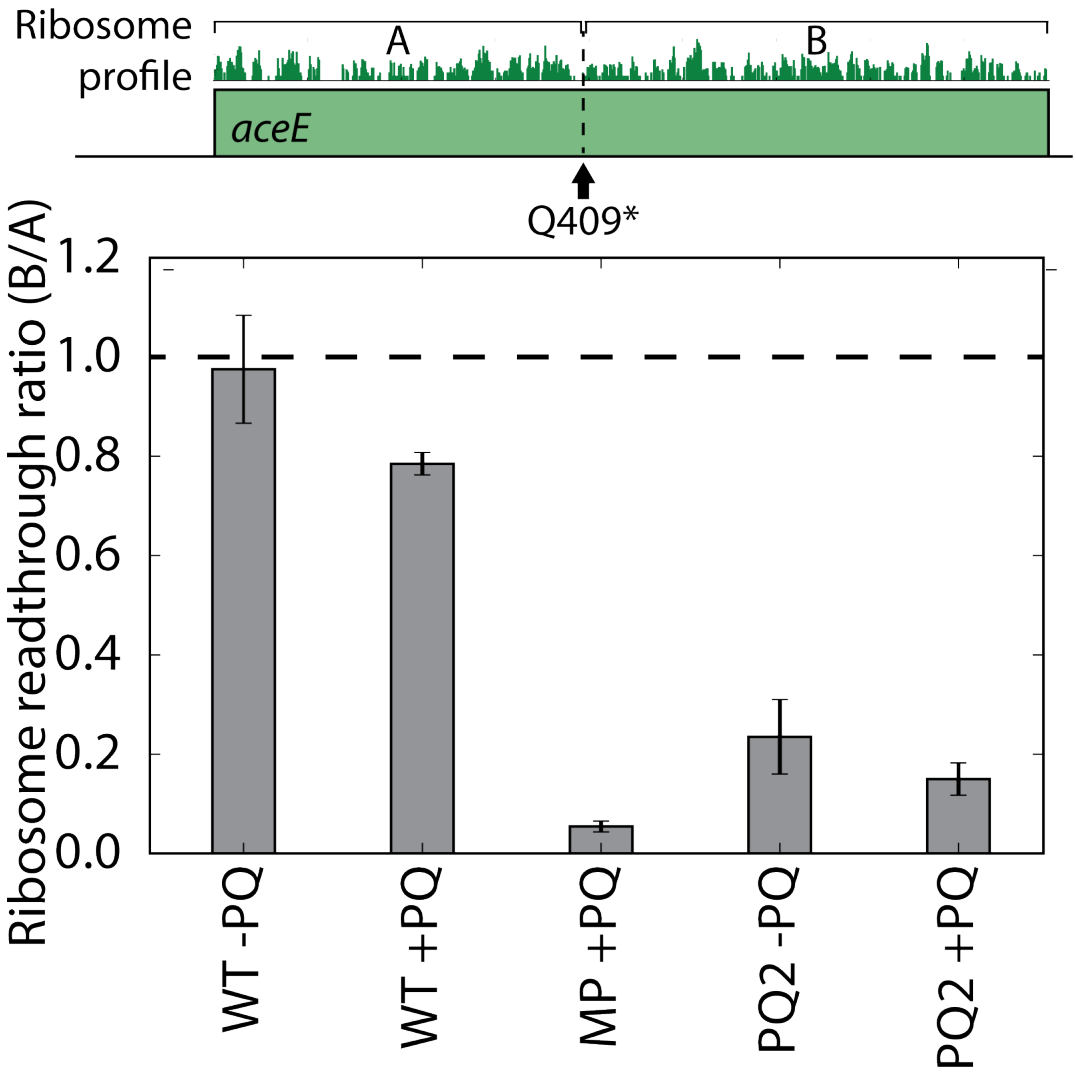

Figure 4.2: The *glnX* suppressor mutation allows limited readthrough of the non-sense mutation. We calculate the ratio of ribosomal density downstream and upstream of the non-sense mutation. WT shows readthrough ratios close to 1.0, indicating that there is no truncation along the aceE gene. Conversely, a midpoint clonal isolate with the aceE non-sense mutation but without the glnX suppressor mutation shows a very low ratio, suggesting that translation terminates at the non-sense mutation. PQ2 end point strains with both the *aceE* non-sense mutation and the *glnX* suppressor mutation show increased readthrough of the non-sense mutation, suggesting the combination of these mutations result in a tuning of *aceE* levels in the cell.

4.3.4 *glnX* mutation affects *aceE* expression

The truncation mutation which occurred in both replicates make use of TAG stop codon, and both strains then independently developed a mutation in *glnX* which allow suppression of the TAG stop codon. Stop codon suppression is known to be incomplete owing to competition between the native ribosome release factor and the suppression tRNA [33]. In order to see the effect of the *glnX* suppressor mutation on the expression level of *aceE*, we make use of ribosome profiling to determine the ribosomal density on the *aceE* gene. Due to the great selective disadvantage of the loss of *aceE*, we were unable to revive the PQ1 midpoint, and hence are looking only at the PQ2 midpoint. We calculated the ratio of the ribosomal density before the truncation and after the truncation to determine the percentage of stop codon readthrough. In WT ribosome density downstream of the truncation location (Q409*) is almost equal to the ribosome density upstream (0.97 ± 0.11). With the non-sense mutation but without the *glnX* suppressor mutation, this ratio drops to 0.05 ± 0.01 due to premature truncation of translation at the non-sense mutation. In PQ2 with the *glnX* suppressor mutation, we see that the ratio increases to 0.23 ± 0.08, indicating that the *glnX* mutation serves to tune the translation level of *aceE* (Figure 4.2B).

4.3.5 Dysregulation of iron-uptake genes under stress

We calculated differential gene expression of the evolved strains PQ1 and PQ2 when subject to 0.25mM of paraquat for 20 minutes relative to no paraquat exposure. The response of WT E. coli to superoxide stress has previously been well-characterized, involving the up-regulation of the SoxRS and OxyR regulons [12, 34, 35], thus we subtracted the ROS stress response in WT from the set of differentially expressed genes in the evolved strains in order to

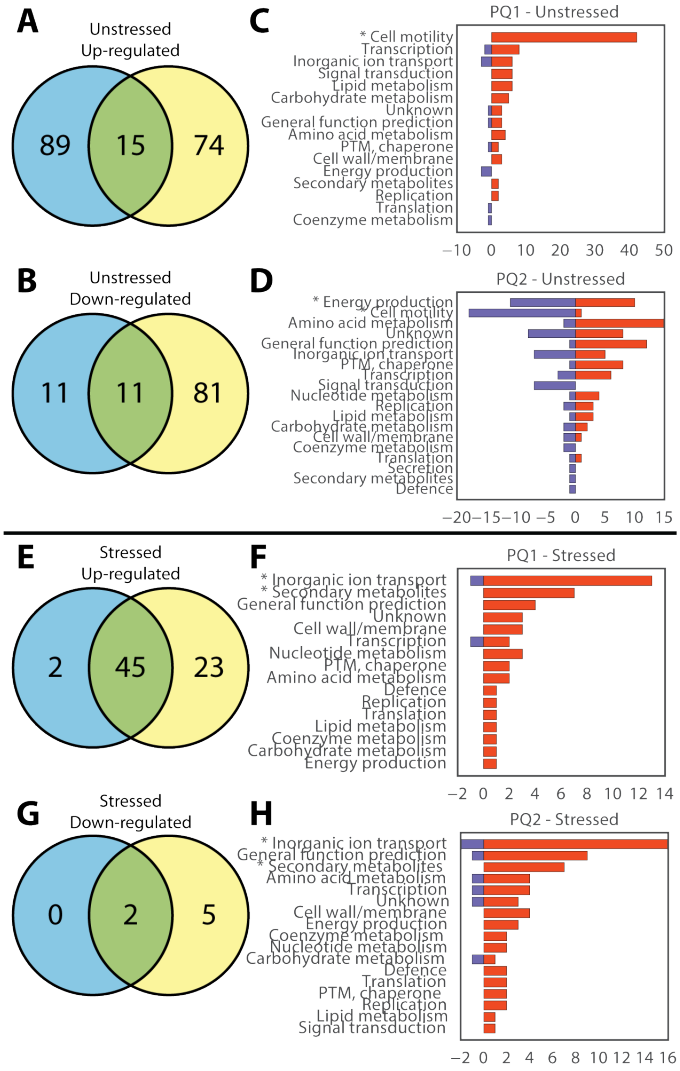

Figure 4.3: Differential expression of the evolved strains with and without the addition of paraquat. A,B: Differential expression of genes in evolved strains relative to WT in the absence of paraquat. PQ1 and PQ2 up-regulate 15 genes and down-regulate 11 genes in common vs WT. C,D: Differentially expressed genes in PQ1 are highly enriched for the COG category cell motility (p-value $=4.71\text{x}10^{-94}$), whilst in PQ2 they are enriched for both cell motility (p-value$=1.43\text{x}10^{-17}$) and energy production (p-value $= 1.02\text{x}10^{-5}$). E,F: We also calculated the differential expression between each evolved strain with and without the addition of paraquat to determine how each strains reaction to stress. We subtracted WT reaction to stress in order to isolate evolved strain specific responses. PQ1 and PQ2 up-regulate 45 genes and down-regulate 2 genes in common. G, H: These up-regulated genes are highly enriched for the COG categories of Inorganic ion transport (PQ1 p-value $= 2.04\text{x}10^{-16}$, PQ2 p-value $= 1.71\text{x}10^{-18}$) and Secondary metabolites (PQ1 p-value $= 3.17\text{x}10^{-12}$, PQ2 p-value $= 8.64\text{x}10^{-10}$).

Table 4.2: Important differentially regulated genes with and without paraquat stress. In the unstressed condition, differential expression of evolved strains was calculated relative to WT strain. In stressed conditions, differential expression of evolved strains under paraquat stress was calculated relative to the same strain without paraquat stress. Genes differentially regulated in WT under paraquat stress were subtracted to isolate evolved strain specific adaptations.

Condition	Strain	Function	Genes	Regulation
Unstressed	Common	Not well characterized/ non-functional	*ygfZ*, *ybcKLM*, *bcsAB*	Up-regulated
		Osmotic Stress	*proVWX*	Up-regulated
		Fatty acid degradation	*fadAB*	Up-regulated
		Sulfur uptake	*tauABC*	Down-regulated
	PQ1	Phosphate homeostasis	*pstABC, phoUBR*	Up-regulated
		Cell motility	*fliHIJKMOPR, flhABE, motAB, flgABEFGHIKLMN, fliACDEFGNSTZ, trg, cheAZYBRW, dgcZ, tar, tsr, tap*	Up-regulated
	PQ2	Iron-sulfur cluster	*iscASRU, sufABCDES, fdx, hscAB*	Up-regulated
		DNA synthesis	*nrdHIEF*	Up-regulated
		Cell motility	*fimE*	Down-regulated
			cheAWR, motAB, tap, tar, fimABCDFGI	Up-regulated
		Fermentative respiration	*ldhA, pta, ackA*	Up-regulated
		TCA cycle	*sdhBCD*	Down-regulated
Stressed	Common	Iron uptake and storage	*entSABCDEFH, fepABCDGRI, yncD, fes, fiu, fhuE*	Up-regulated
		DNA synthesis	*nrdEFH*	Up-regulated
		DNA repair	*ligA*	Up-regulated
	PQ1	Iron-sulfur cluster	*iscR*	Up-regulated
	PQ2	Formate dehydrogenase	*fdnH, fdoG*	Up-regulated
		Motility	*ybjN*	Up-regulated
		marR operon	*marA, mdaB*	Up-regulated
		ROS scavenging	*sodB*	Down-regulated

isolate adaptation specific transcriptional responses. This process left us with a total of 49 DEGs in PQ1 (47 up-regulated, and 2 down-regulated) and 75 DEGs in PQ2 (68 up-regulated, and 7 down-regulated), that were not part of the canonical superoxide stress response in WT. We find that both PQ1 and PQ2 have a convergent transcriptional response to ROS stress. Also in both, 45 genes were up-regulated and 2 genes were down-regulated. The most differentially regulated COG categories during stress were secondary metabolites (PQ1 p-value = $3.17\text{x}10^{-12}$, PQ2 p-value = $8.64\text{x}10^{-10}$) and Inorganic ion transport (PQ1 p-value = $2.04\text{x}10^{-16}$, PQ2 p-value = $1.71\text{x}10^{-18}$) (Figure 4.3EFGH, Table 4.2).

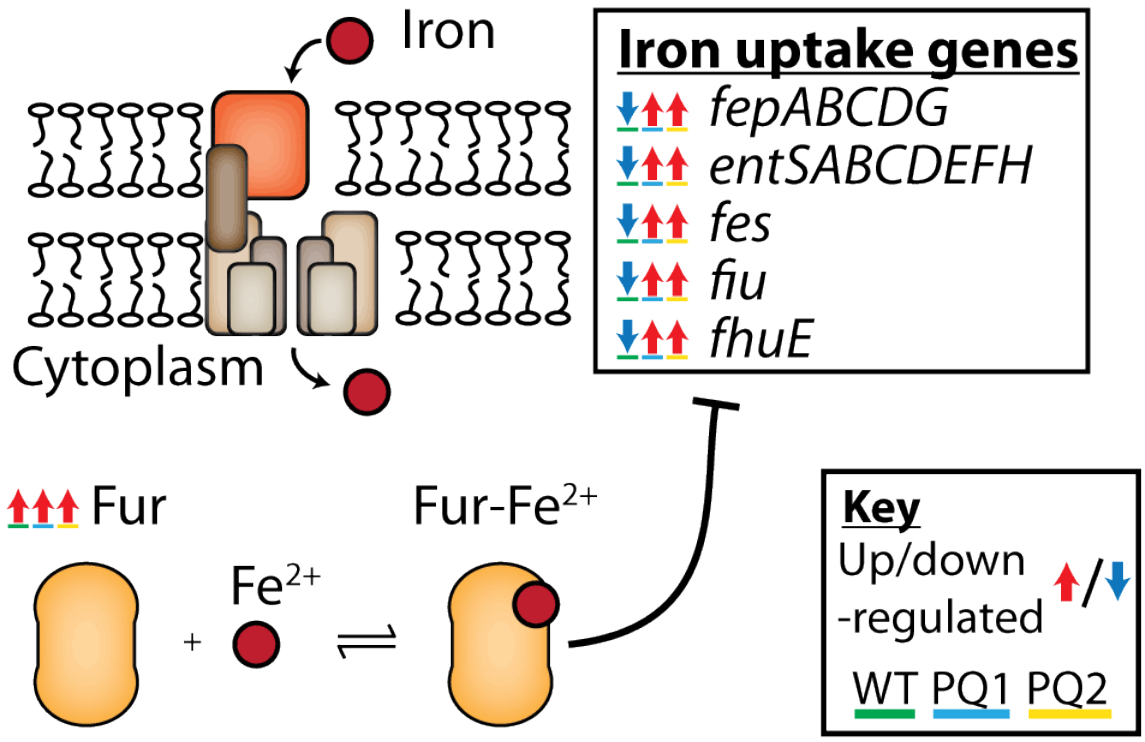

Figure 4.4: Evolved strains show dysregulation of iron-uptake genes regulated by Fur-Fe2+ while under paraquat stress. Differential gene expression was calculated for each strain under paraquat stress relative to the same strain under normal growth conditions, significantly differentially expressed genes are circled in bold. We see that both WT and the evolved strains up-regulate Fur, which in WT proceeds to repress the iron-uptake genes below. In contrast, the evolved strains instead see an up-regulation of these iron-uptake genes.

These 45 genes are highly enriched for regulation by Fur (p-value = 8.91 x 10^{-27}), and almost all of them are directly related to the uptake and storage of iron. In contrast, we see that WT actually down-regulates these genes, despite all three strains directly upregulating Fur (Figure 4.4). During superoxide stress, free iron in the cell aggravates damage to cellular components through formation of hydroxyl radicals via the Fenton reaction [36]. As the regulatory activity of Fur is mediated by availability of Fe^{2+}, this could indicate differences in the availability of free iron in the cell. Other commonly upregulated genes during stress are related to DNA synthesis and repair such as ribonucleotide reductase *nrdEFH* and DNA ligase *ligA* which would allow

cells to repair DNA damaged by ROS at an increased rate compared to WT.

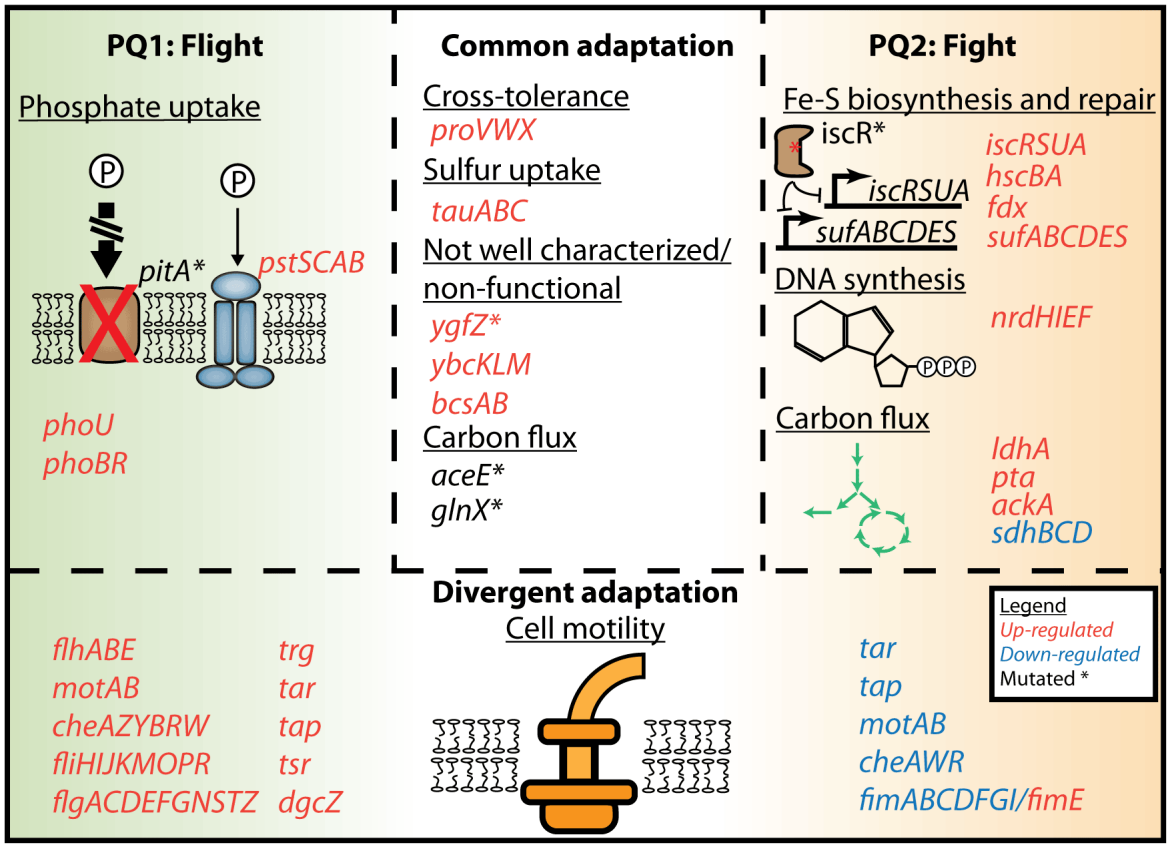

Figure 4.5: PQ1 and PQ2 show two different phenotypes growing under normal conditions. Differential regulation is calculated for the evolved strains growing without paraquat stress relative to wild type. Up-regulated genes are highlighted in red, down-regulated genes are highlighted in blue, mutated genes are denoted by an asterisk *. The "Flight" phenotype used by PQ1 makes use of a pitA disruption to induce phosphate starvation, providing a regulatory impetus for increase cell motility and aggregation, increasing tolerance to oxidative stress. On the other hand, PQ2 exhibits the "Fight" phenotype where it up-regulates genes involved in the repair moieties damaged by ROS such as iron-sulfur clusters and DNA, as well as redirects carbon flux from the TCA cycle towards fermentative pathways.

4.3.6 Transcriptomic characterization of end point strains under normal growth conditions reveals two ROS tolerant phenotypes

When grown without paraquat stress, we find that PQ1 and PQ2 differentially express 126 and 182 genes, respectively, relative to wild type (PQ1: 104 up-regulated, 22 down-regulated,

PQ2: 89 up-regulated, 92 down-regulated). Of these, only 15 genes are commonly up-regulated and 11 genes are commonly down-regulated (Table 4.2, Figure 4.3AB). PQ1 significantly up-regulates while PQ2 significantly down-regulates genes with the COG annotation of Cell motility (PQ1 p-value = $4.71\text{x}10^{-94}$, PQ2 p-value = $1.43\text{x}10^{-17}$), whilst PQ2 also differentially regulates genes in the COG category of Energy production (p-values = $1.02\text{x}10^{-5}$) (Figure 4.3CD).

Several of the commonly upregulated genes, *ygfZ*, *ybcKLM*, and *bcsAB*, are not well characterized or thought to be non-functional in K-12 strains. *ygfZ*, discussed above, is annotated as a putative iron-sulfur cluster repair protein and is both mutated and up-regulated in PQ1 and PQ2 [29]. The combination of the mutation and up-regulation of *ygfZ* indicates that it plays an important role in tolerance to paraquat stress and warrants further investigation. *ybcL*, which is in an operon with *ybcM* directly downstream of *ybcK*, has previously been found to inhibit neutrophil migration during bladder infections by the pathogenic strain of *E. coli* UTI89, although several mutations in the K-12 variant have caused a loss of its original pathogenic function [37]. Regulation of the *ybcLM* operon is not well understood, and up-regulation of these genes in the evolved strains suggests a vestigial virulent response. *bcsA* and *bcsB* synthesize cellulose in non-K12 *E. coli* strains, but in K-12 strains a nonsense SNP in upstream gene *bcsQ* results in decreased expression of *bcsA* [38].

The evolved strains also commonly up-regulate the genes *proVWX* which are related to tolerance to osmotic stress. Previous studies have shown that previous exposure to one type of stress, such as pH, osmotic, oxidative, or thermal stress, has a positive effect on culture tolerance to other stresses, possibly due to the induction of general stress response pathways through RpoS [19]. Evolution of these strains in the presence of oxidative stress might cause a constitutive upregulation of stress response pathways, resulting in upregulation of *proVWX* even

in the absence of paraquat.

Two genes utilized in fatty acid degradation, *fadAB*, are commonly upregulated. The truncation of i would result in a decrease of conversion from pyruvate to acetyl-CoA, lowering the available pool of acetyl-CoA which is important for synthesis of various amino acids through the glyoxylate shunt and TCA cycle intermediates. Upregulation of the fatty acid degradation pathway would increase levels of acetyl-CoA in the cell through oxidation of fatty acids.

PQ1 - Exhibits a "Flight" Phenotype

As mentioned above, the frame shift mutation in *pitA* disrupts its function, severely limiting phosphate uptake in PQ1 [39]. Expectedly, PQ1 up-regulates the high affinity phosphate transport complex genes *pstABCS* and phosphate regulators *phoBR* and *phoU* even when not under paraquat stress, suggesting that the cell is constitutively experiencing phosphate starvation. Phosphate limitation is observed to activate a swarming phenotype in *P. aeruginosa* [40]. We observe the similar up-regulation of cell motility related genes in PQ1. 42 of the 126 differentially regulated genes in PQ1 are related to cell motility. These genes include flagellar biosynthesis genes *fliOIHKJPRM* and *flhABE*, flagellar components *motAB*, *flgABEFGHIKLMN*, and fliACDEFGNSTZ, chemotaxis related proteins trg and cheAZYBRW, as well as motility regulators *dgcZ*, *tar*, *tsr*, and *tap*. The up-regulation of these genes has also been associated with auto-aggregation and biofilm formation, which has been shown to increase the tolerance of *E. coli* to various stresses including oxidative stress [41].

Surprisingly, this regulatory adaptation in PQ1 does not intuitively counter ROS stress. Instead, it increased cellular motility (Supplementary Figure C.2), which might lead to a secondary effect of oxidative stress tolerance. We therefore term this the "Flight" phenotype (Fig-

ure 4.5).

PQ2 - Exhibits a "Fight" Phenotype

In contrast to PQ1, PQ2 downregulates cell motility related genes *cheAWR*, *motAB*, *tap*, and *tar*, and fimbriae genes *fimABCDFGI*, while up-regulating fimE. These changes to the FimB/FimE ratio are consistent with the overall downregulation of type 1 fimbriae expression [42, 43].

We find that the mutation in the iron-sulfur cluster binding region of the protein IscR causes a dysregulation of its regulon. Genes regulated by *iscR*, such as *iscASRU*, *hscSB*, *fdx*, *sufABCDES*, and *nrdHIEF* are all up-regulated in PQ2 even in the absence of paraquat stress. These genes synthesize and repair iron-sulfur clusters and DNA, which might increase the rate of cellular repair of ROS damage. Interestingly, *nrdEF* has been found to only be induced under iron starvation [44, 45], providing further support that iron levels in the evolved strains might differ from wild type, leading to the differences found in regulation of iron uptake genes.

In addition to mutations found in *aceE*, *gltA*, and *sucA* genes, PQ2 also upregulates *ldhA*, *pta*, and *ackA*, indicating a switch from aerobic respiration to fermentation for the production of lactate or acetate. Succinate dehydrogenase subunits *sdhBCD* are also downregulated, further decreasing flux through the oxidative arm of the TCA cycle.

Overall, PQ2 seems to adopt a different tolerance strategy from PQ1: it up-regulates genes that allow it to repair the damage caused by ROS to iron-sulfur clusters and DNA, and changes the flow of carbon flux towards fermentative pathways to mitigate endogenous and paraquat cycling ROS production. We have accordingly termed it the "Fight" phenotype (Figure 4.5).

4.4 Discussion

Bacteria encounter and respond to oxidative stress in many environments, both while fighting an immune response and responding to culture conditions in industrial biotechnology [4, 5]. Our results reveal a relatively simple genetic basis for adaptation of *E. coli* to extremely high levels of superoxide stress. They also uncover several novel genetic adaptations and regulatory strategies used by *E. coli* for adaptation to oxidative stress, including the use of a non-sense/suppressor mutation pair to control translation of pyruvate dehydrogenase in order to reduce carbon flux into the TCA cycle; dysregulation of the Fur regulon; phosphate starvation as a regulatory impetus for oxidative stress tolerance; and direct mitigation of ROS damage through. These findings pave the way to a better understanding of host-pathogen interactions as well as suggesting avenues for designing host strains to better withstand oxidative stress.

Over the course of this study, we have generated several significant findings. First, the use of experimental evolution was successful in developing strains that were tolerant to increased levels of paraquat. *E. coli* can develop up to a 500% increase in tolerance to paraquat relative to wild type in the span of 30 days, with an average of only 6.5 mutations, demonstrating the ease with which *E. coli* adapts to elevated oxidative stress.

Second, the low number of mutations required for adaptation makes the genetic basis of adaptation to oxidative stress simple and interpretable. These mutations fall into two main categories: modulation of metabolic flux through the TCA cycle through the *aceE*, *sucA*, and *glnX* mutations, and the increased synthesis and repair of iron-sulfur clusters through mutations in *ygfZ* and *iscR*. Modulating flux through the TCA cycle is a preventative measure, limiting the availability of high energy electrons which reduces endogenous ROS production and redox cycling by paraquat. On the other hand, mutations that increase synthesis and repair rates of

iron-sulfur clusters work to mitigate damage to iron-sulfur clusters caused by ROS. One of the more surprising findings from this study was the occurrence of an amber suppressor mutation in *glnX*, which allowed control of pyruvate dehydrogenase levels in the cell through inefficient readthrough of a non-sense mutation, an evolutionary strategy that has not been previously seen in other experiments. While their role is not well understood, suppressor mutations are commonly found in bacteria that interact directly with hosts, such as those in the gut microbiome [46], as well as in laboratory strains of *E. coli* that have been subject to long periods of mutagenesis [47]. These occurrences, as well as the emergence of the suppressor mutation while under oxidative stress, point to the role of suppressor mutations as stress-tolerance mechanisms.

Third, transcriptomics provides useful insight into the tolerization mechanisms of the evolved strains. One interesting finding in this study is the dysregulation of iron-uptake genes. It has long been known that iron plays an important role in oxidative stress in bacteria [48]. In the presence of oxidative stress, damage to iron-sulfur clusters as a result of the production of ROS results in the release of free iron into the cytoplasm [11]. In WT, Fur binds to free Fe^{2+} repressing the expression of iron-uptake genes in order to control levels of cytoplasmic iron. In the evolved strains, up-regulation of iron-uptake genes instead, in spite of the up-regulation of Fur, might indicate a decreased Fe^{2+} level. This suggests that the evolved strains reduce Fe^{2+} concentration under oxidative stress through a yet to be uncovered mechanism.

Finally, analysis of the evolved strains from a systems biology perspective leads to the identification of two phenotypic states. The first of these makes use of cross-tolerization to phosphate starvation as a means to gain resistance to oxidative stress. The mutation in the *pitA* gene in PQ1 results in a frameshift mutation, disrupting activity of the major phosphate transporter in *E. coli*. The ensuing phosphate starvation phenotype could augment tolerance to oxidative

stress in three ways. The first is via the reduction of the availability of high energy electrons for redox cycling and endogenous ROS production [49, 50]. The second is as a regulatory impetus for up-regulation of general stress response through the RpoS regulon [51]. The appearance of the amber suppressor mutation that has previously been observed to relax the stringent response [52] might further optimize the response towards oxidative stress. Third, there is evidence that phosphate starvation causes a virulence phenotype which might impart tolerance to oxidative stress as a secondary effect. The *pstABC* complex, upregulated during phosphate starvation, has been found to play an important role in virulence and tolerance to oxidative stress in Avian Pathogenic *E. coli* [53], while phosphate starvation has been linked to the induction of cellular motility, auto-aggregation, and the formation of biofilms [54], a physical method of increasing resistance to oxidative stress [55]. The exposure to low levels of one type of stress resulting in a gain in tolerance to various other seemingly unrelated stressors has previously been observed in several experiments [56], but this study reveals a potential application of this phenomenon. This result could find a use in metabolic engineering applications where oxidative stress to the cell factory is a concern through simple modifications of phosphate availability in the media.

The second of these phenotypic states increases tolerance to oxidative stress by directly combating the damaging effects of ROS. The mutation in *iscR* in PQ2 causes a constitutive up-regulation of genes in its regulon, such as *iscSUA*, *sufABC*, and *nrdEFH*. The high level of IscSUA and SufABC could confer tolerance to oxidative stress by both increasing the iron-sulfur synthesis rates as well as increasing the rate of repair for damaged iron-sulfur clusters, while NrdEFH might increase DNA synthesis rates in order to mitigate damage done to DNA. In conjunction, a shift from aerobic respiration towards a fermentative mode through coordinated mutations and regulation to reduce metabolic flux through the TCA cycle and up-regulation of

genes such as *ldhA*, *ackA*, and *pta* reduces the availability of high energy electrons, reducing the overall production of both endogenous and redox-cycling ROS.

In this study we found that laboratory evolution of E. *coli* leads to adapted strains that can withstand up to five times increased paraquat concentrations compared to wild type. Analysis of these strains revealed insights into the genomic basis for and systems biology of ROS tolerance. Laboratory evolution resulted in adapted strains whose properties were consistent with known targets of ROS damage, yet achieved tolerance through non-intuitive mechanisms. Taken together, the properties of the adapted strains encourage continued work to build a more complete understanding of adaptation to ROS stress.

4.5 Acknowledgements

J.T., J.H.P., A.M.F. and B.O.P. designed the study, J.H.P., C.A.O. performed the ALE experiment, R.S. and Y.H. generated genomic sequencing libraries of the strains, J.T. generated transcriptomic and ribosome profiling libraries. J.T., P.V.P. and A.V.S. analyzed the data. J.T. and B.O.P. wrote the manuscript, with contributions from all the other co-authors.

This work was funded by the Novo Nordisk Foundation Grant Number NNF10CC1016517 and by National Institute of General Medical Science grant number GM057089. We would like to thank Marc Abrams for his assistance with manuscript editing and Amitesh Anand for helpful discussions and assistance with growth screens.

Chapter 4 in part is a reprint of material published in: **J Tan**, CA Olson, JH Park, AV Sastry, PV Phaneuf, L Yang, R Szubin, Y Hefner, AM Feist, BO Palsson "Experimental evolution reveals the genetic basis and systems biology of superoxide stress tolerance" *Submitted*. The dissertation author is the primary author.

4.6 References

1. Imlay, J. A. Diagnosing oxidative stress in bacteria: not as easy as you might think. *Current opinion in microbiology* **24,** 124–131. ISSN: 1369-5274, 1879-0364 (Apr. 2015).

2. Seaver, L. C. & Imlay, J. A. Are respiratory enzymes the primary sources of intracellular hydrogen peroxide? *The Journal of biological chemistry* **279,** 48742–48750. ISSN: 0021-9258 (Nov. 2004).

3. Kohanski, M. A., Dwyer, D. J., Hayete, B., Lawrence, C. A. & Collins, J. J. A common mechanism of cellular death induced by bactericidal antibiotics. *Cell* **130,** 797–810. ISSN: 0092-8674 (Sept. 2007).

4. Baez, A. & Shiloach, J. *Escherichia coli* avoids high dissolved oxygen stress by activation of SoxRS and manganese-superoxide dismutase. *Microbial cell factories* **12,** 23. ISSN: 1475-2859 (Mar. 2013).

5. Bessette, P. H., Aslund, F., Beckwith, J. & Georgiou, G. Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 13703–13708. ISSN: 0027-8424 (Nov. 1999).

6. Cadet, J. & Wagner, J. R. DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *Cold Spring Harbor perspectives in biology* **5.** ISSN: 1943-0264. doi:10.1101/cshperspect.a012559 (Feb. 2013).

7. Birben, E., Sahiner, U. M., Sackesen, C., Erzurum, S. & Kalayci, O. Oxidative stress and antioxidant defense. *The World Allergy Organization journal* **5,** 9–19. ISSN: 1939-4551 (Jan. 2012).

8. Imlay, J. A. Iron-sulphur clusters and the problem with oxygen. *Molecular microbiology* **59,** 1073–1082. ISSN: 0950-382X (2006).

9. Roche, B., Aussel, L., Ezraty, B., Mandin, P., Py, B. & Barras, F. Reprint of: Iron/sulfur proteins biogenesis in prokaryotes: formation, regulation and diversity. *Biochimica et biophysica acta* **1827,** 923–937. ISSN: 0006-3002 (Aug. 2013).

10. Broderick, J. B. Iron–Sulfur Clusters in Enzyme Catalysis. *Comprehensive Coordination Chemistry II,* 739–757 (2003).

11. Djaman, O., Outten, F. W. & Imlay, J. A. Repair of oxidized iron-sulfur clusters in *Escherichia coli*. *The Journal of biological chemistry* **279,** 44590–44599. ISSN: 0021-9258 (Oct. 2004).

12. Farr, S. B. & Kogoma, T. Oxidative stress responses in *Escherichia coli* and Salmonella typhimurium. *Microbiological reviews* **55,** 561–585. ISSN: 0146-0749 (Dec. 1991).

13. Zheng, M., Doan, B., Schneider, T. D. & Storz, G. OxyR and SoxRS Regulation offur. en. *Journal of bacteriology* **181,** 4639–4643. ISSN: 0021-9193, 1098-5530 (Aug. 1999).

14. Basak, S. & Jiang, R. Enhancing E. coli tolerance towards oxidative stress via engineering its global regulator cAMP receptor protein (CRP). *PloS one* **7,** e51179. ISSN: 1932-6203 (Dec. 2012).

15. Battistoni, A., Pacello, F., Folcarelli, S., Ajello, M., Donnarumma, G., Greco, R., Grazia Ammendolia, M., Touati, D., Rotilio, G. & Valenti, P. Increased Expression of Periplasmic Cu,Zn Superoxide Dismutase Enhances Survival of *Escherichia coli* Invasive Strains within Nonphagocytic Cells. *Infection and Immunity* **68,** 30–37 (2000).

16. Smith, A. H., Imlay, J. A. & Mackie, R. I. Increasing the oxidative stress response allows *Escherichia coli* to overcome inhibitory effects of condensed tannins. *Applied and environmental microbiology* **69,** 3406–3411. ISSN: 0099-2240 (June 2003).

17. Smirnova, G. V., Zakirova, O. N. & Oktiabr'skiui. Role of the antioxidant system in response of *Escherichia coli* bacteria to cold stress. *Mikrobiologiia* **70,** 55–60. ISSN: 0026-3656 (2001).

18. Smirnova, G. V., Muzyka, N. G. & Oktyabrsky, O. N. The role of antioxidant enzymes in response of *Escherichia coli* to osmotic upshift. *FEMS microbiology letters* **186,** 209–213. ISSN: 0378-1097 (May 2000).

19. Rodriguez-Rojas, A., Kim, J., Johnston, P., Makarova, O., *et al.* Non-lethal oxidative stress boosts bacterial survival and evolvability under lethal exposure. *BioRxiv* (2019).

20. Mohamed, E. T., Wang, S., Lennen, R. M., Herrgaard, M. J., Simmons, B. A., Singer, S. W. & Feist, A. M. Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution. en. *Microbial cell factories* **16,** 204. ISSN: 1475-2859 (Nov. 2017).

21. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Applied and environmental microbiology* **81,** 17–30. ISSN: 0099-2240, 1098-5336 (Jan. 2015).

22. Lee, D.-H., Feist, A. M., Barrett, C. L. & Palsson, B. O. Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of *Escherichia coli*. *PloS one* **6,** e26172. ISSN: 1932-6203 (Oct. 2011).

23. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. *Nucleic acids research* **47,** D1164–D1171. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

24. Ayala-Castro, C., Saini, A. & Outten, F. W. Fe-S cluster assembly pathways in bacteria. *Microbiology and molecular biology reviews: MMBR* **72,** 110–25, table of contents. ISSN: 1092-2172, 1098-5557 (Mar. 2008).

25. Schwartz, C. J., Giel, J. L., Patschkowski, T., Luther, C., Ruzicka, F. J., Beinert, H. & Kiley, P. J. IscR, an Fe-S cluster-containing transcription factor, represses expression of *Escherichia coli* genes encoding Fe-S cluster assembly proteins. en. *Proceedings of the National*

Academy of Sciences of the United States of America **98,** 14895–14900. ISSN: 0027-8424 (Dec. 2001).

26. Fleischhacker, A. S., Stubna, A., Hsueh, K.-L., Guo, Y., Teter, S. J., Rose, J. C., Brunold, T. C., Markley, J. L., Munck, E. & Kiley, P. J. Characterization of the [2Fe-2S] cluster of *Escherichia coli* transcription factor IscR. *Biochemistry* **51,** 4453–4462. ISSN: 0006-2960, 1520-4995 (June 2012).

27. Rajagopalan, S., Teter, S. J., Zwart, P. H., Brennan, R. G., Phillips, K. J. & Kiley, P. J. Studies of IscR reveal a unique mechanism for metal-dependent regulation of DNA binding specificity. *Nature structural & molecular biology* **20,** 740–747. ISSN: 1545-9993, 1545-9985 (June 2013).

28. Waller, J. C., Ellens, K. W., Hasnain, G., Alvarez, S., Rocca, J. R. & Hanson, A. D. Evidence that the Folate-Dependent Proteins YgfZ and MnmEG Have Opposing Effects on Growth and on Activity of the Iron-Sulfur Enzyme MiaB. *Journal of Bacteriology* **194,** 362–367 (2012).

29. Waller, J. C., Alvarez, S., Naponelli, V., Lara-Nunez, A., Blaby, I. K., Da Silva, V., Ziemak, M. J., Vickers, T. J., Beverley, S. M., Edison, A. S., Rocca, J. R., Gregory 3rd, J. F., de Crecy-Lagard, V. & Hanson, A. D. A role for tetrahydrofolates in the metabolism of iron-sulfur clusters in all domains of life. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 10412–10417. ISSN: 0027-8424, 1091-6490 (June 2010).

30. Lin, C.-N., Syu, W.-J., Sun, W.-S. W., Chen, J.-W., Chen, T.-H., Don, M.-J. & Wang, S.-H. A role of ygfZ in the *Escherichia coli* response to plumbagin challenge. *Journal of biomedical science* **17,** 84. ISSN: 1021-7770, 1423-0127 (Nov. 2010).

31. Harris, R. M., Webb, D. C., Howitt, S. M. & Cox, G. B. Characterization of PitA and PitB from*Escherichia coli*. en. *Journal of bacteriology* **183,** 5008–5014. ISSN: 0021-9193, 1098-5530 (Sept. 2001).

32. Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J. & Gustafsson, C. Design parameters to control synthetic gene expression in *Escherichia coli*. *PloS one* **4,** e7002. ISSN: 1932-6203 (Sept. 2009).

33. Eggertsson, G. & Soll, D. Transfer ribonucleic acid-mediated suppression of termination codons in *Escherichia coli*. *Microbiological reviews* **52,** 354. ISSN: 0146-0749 (1988).

34. Seo, S. W., Kim, D., Szubin, R. & Palsson, B. O. Genome-wide Reconstruction of OxyR and SoxRS Transcriptional Regulatory Networks under Oxidative Stress in *Escherichia coli* K-12 MG1655. *Cell reports* **12,** 1289–1299. ISSN: 2211-1247 (Aug. 2015).

35. Rui, B., Shen, T., Zhou, H., Liu, J., Chen, J., Pan, X., Liu, H., Wu, J., Zheng, H. & Shi, Y. A systematic investigation of Escherichia coli central carbon metabolism in response to superoxide stress. *BMC systems biology* **4,** 122. ISSN: 1752-0509 (Sept. 2010).

36. Winterbourn, C. C. Toxicity of iron and hydrogen peroxide: the Fenton reaction. *Toxicology letters* **82-83,** 969–974. ISSN: 0378-4274 (Dec. 1995).

37. Lau, M. E., Loughman, J. A. & Hunstad, D. A. YbcL of uropathogenic *Escherichia coli* suppresses transepithelial neutrophil migration. *Infection and immunity* **80,** 4123–4132. ISSN: 0019-9567, 1098-5522 (Dec. 2012).

38. Serra, D. O., Richter, A. M. & Hengge, R. Cellulose as an architectural element in spatially structured *Escherichia coli* biofilms. *Journal of bacteriology* **195,** 5540–5554. ISSN: 0021-9193, 1098-5530 (Dec. 2013).

39. Rosenberg, H., Gerdes, R. G. & Chegwidden, K. Two systems for the uptake of phosphate in *Escherichia coli*. *Journal of bacteriology* **131,** 505–511. ISSN: 0021-9193 (Aug. 1977).

40. Bains, M., Fernandez, L. & Hancock, R. E. W. Phosphate starvation promotes swarming motility and cytotoxicity of Pseudomonas aeruginosa. *Applied and environmental microbiology* **78,** 6762–6768. ISSN: 0099-2240, 1098-5336 (Sept. 2012).

41. Laganenka, L., Colin, R. & Sourjik, V. Chemotaxis towards autoinducer 2 mediates autoaggregation in *Escherichia coli*. *Nature communications* **7,** 12984. ISSN: 2041-1723 (Sept. 2016).

42. Holden, N., Blomfield, I. C., Uhlin, B.-E., Totsika, M., Kulasekara, D. H. & Gally, D. L. Comparative analysis of FimB and FimE recombinase activity. *Microbiology* **153,** 4138–4149. ISSN: 0026-2617, 1350-0872 (Dec. 2007).

43. Klemm, P. Two regulatory fim genes, fimB and fimE, control the phase variation of type 1 fimbriae in *Escherichia coli*. *The EMBO journal* **5,** 1389–1393. ISSN: 0261-4189 (June 1986).

44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12. ISSN: 2226-6089, 2226-6089 (May 2011).

45. Andrews, S. C. Making DNA without iron–induction of a manganese-dependent ribonucleotide reductase in response to iron starvation. *Molecular microbiology* **80,** 286–289. ISSN: 0950-382X (2011).

46. Marshall, B. & Levy, S. B. Prevalence of amber suppressor-containing coliforms in the natural environment. *Nature* **286,** 524–525. ISSN: 0028-0836 (July 1980).

47. Belin, D. Why are suppressors of amber mutations so frequent among *Escherichia coli* K12 strains?. A plausible explanation for a long-lasting puzzle. *Genetics* **165,** 455–456. ISSN: 0016-6731 (Oct. 2003).

48. Touati, D. Iron and oxidative stress in bacteria. *Archives of biochemistry and biophysics* **373,** 1–6. ISSN: 0003-9861 (Jan. 2000).

49. Marzan, L. & Shimizu, K. Metabolic regulation of *Escherichia coli* and its phoB and phoR genes knockout mutants under phosphate and nitrogen limitations as well as at acidic condition. *Microbial Cell Factories* **10,** 39 (2011).

50. Moreau, P. L. Diversion of the metabolic flux from pyruvate dehydrogenase to pyruvate oxidase decreases oxidative stress during glucose metabolism in nongrowing *Escherichia*

coli cells incubated under aerobic, phosphate starvation conditions. *Journal of bacteriology* **186,** 7364–7368. ISSN: 0021-9193 (2004).

51. Mandel, M. J. & Silhavy, T. J. Starvation for different nutrients in *Escherichia coli*— results in differential modulation of RpoS levels and stability. *Journal of bacteriology* **187,** 434–442. ISSN: 0021-9193 (Jan. 2005).

52. Breeden, L. & Yarus, M. Amber suppression relaxes stringent control by elongating stringent factor. en. *Molecular & general genetics: MGG* **187,** 254–264. ISSN: 0026-8925, 1432-1874 (Oct. 1982).

53. Crepin, S., Lamarche, M. G., Garneau, P., Seguin, J., Proulx, J., Dozois, C. M. & Harel, J. Genome-wide transcriptional response of an avian pathogenic *Escherichia coli* (APEC) pst mutant. *BMC genomics* **9,** 568. ISSN: 1471-2164 (Nov. 2008).

54. Vogeleer, P., Vincent, A. T., Chekabab, S. M., Charette, S. J., Novikov, A., Caroff, M., Beaudry, F., Jacques, M. & Harel, J. *Escherichia coli* O157:H7 responds to phosphate starvation by modifying LPS involved in biofilm formation. doi:10.1101/536201.

55. Schembri, M. A., Hjerrild, L., Gjermansen, M. & Klemm, P. Differential expression of the *Escherichia coli* autoaggregation factor antigen 43. *Journal of bacteriology* **185,** 2236–2242. ISSN: 0021-9193 (Apr. 2003).

56. Gunasekera, T. S., Csonka, L. N. & Paliy, O. Genome-Wide Transcriptional Responses of *Escherichia coli* K-12 to Continuous Osmotic and Heat Stresses. en. *Journal of bacteriology* **190,** 3712–3720. ISSN: 0021-9193, 1098-5530 (May 2008).

Chapter 5

Conclusions

The development of microbial cell factories through metabolic engineering provides many advantages over traditional chemical synthesis such as less severe reaction conditions, increased steroselectivity, simpler raw materials, and the ability to produce complex biological molecules such as proteins. Creation of these microbial cell factories involves genetic manipulation of the host cell and the expression of heterologous genes to expand its native genetic capabilities. Concurrently the advent of high throughput next-generation sequencing has ushered in a new era in biology, allowing quantitative understanding of processes within the cell on an unprecedented scale. Methods such as genome resequencing, ChiP-Exo, RNASeq, and Ribosome profiling allow direct interrogation of biological processes at multiple levels from genotype to translation at a whole-cell scale. This new fullness of information has brought about a paradigm shift where we can begin to put together a mechanistic genotype-phenotype relationship from a systems perspective. In this dissertation, we apply several of these omics data types to improve our understanding of the model organism and common biotechnological workhorse *E. coli* for the purposes of microbial protein production.

In the first chapter of this dissertation "Multi-omic data integration reveals hidden biological regularities", we examined *E. coli* during expression of its native proteins in order to gain insight into successful translation. Unlike heterologous proteins, almost all native proteins would have evolved for optimal translation, reducing wasteful ribosomal stalling and drop off except where necessary. We describe how translation rates along a protein are encoded through various means including Shine-Dalgarno-like codons to allow time for proper folding of protein secondary structure as they exit the ribosome. Additionally, we see that whilst translation efficiency is variable across genes, it is consistent across conditions, and is thus largely an intrinsic property of each gene, suggesting that it is influenced by sequence properties which we can manipulate.

The second chapter of this dissertation "Independent component analysis of *E. coli* 's transcriptome reveals the cellular processes that respond to heterologous gene expression" describes a large scale transcriptomic study of heterologous protein expression in E. coli to uncover the dimensions of host cell response. Independent component analysis decomposes the transcriptomic response across 40 proteins into 4 major host cell responses: Fear vs Greed, Metal Homeostasis and Respiration, Protein Folding and Amino Acid and Nucleotide Biosynthesis. These represent the major perturbations to the cell during expression of a variety of heterologous proteins, and give us insight into the interplay between growth rates, expression levels, stress and the metabolic burden. It also identifies clear avenues in which we should focus our future research efforts in order to improve heterologous protein expression.

In the last chapter of this dissertation "Experimental evolution reveals the genetic basis and systems biology of superoxide stress tolerance" we examine adaptations which increase *E. coli* 's tolerance to reactive oxygen species, a stressor which is common in industrial biotechnology due to high level heterologous protein expression, culture conditions or product toxicity.

We generate two strains which display enhanced tolerance to superoxide stress, as well as elucidate several novel adaptive mechanisms. These mechanisms, attenuating flux through central carbon metabolism, phosphate starvation and increased damage repair rates provide pathways to increasing host cell fitness in an industrial setting.

The use of microbes as cell factories promises to be the next industrial revolution if we can improve yields and efficiency of production. In this dissertation we cover three key aspects of improving protein expression in *E. coli* : identifying translation dynamics along the length of a transcript for proper folding of protein secondary structure, major host cell responses during expression of a variety of heterologous proteins, and lastly genetic adaptations to increase oxidative stress tolerance. These findings improve our understanding of *E. coli* and propel its development as a platform production strain for industrial biotechnology.

Appendix A

Multi-omic data integration enables discovery of hidden biological regularities - Supplementary Information

A.1 Supplementary Notes

A.1.1 Supplementary Note 1: Ribosome profiling pause site analysis

It has been established that translation rate is not constant along a gene. This has been demonstrated in-vitro using proteins such as firefly luciferase [1, 2] and epoxide hydrolases12. Various phenomena such as codon usage [1, 3], mRNA secondary structure[4], anti-Shine-Dalgarno-like sequences [5], poly-proline stretches16 as well as protein domains [6–8] have

been implicated in determining the locations of these pause sites. Slower translation rate have been shown anecdotally to improve solubility of heterologously expressed proteins [1, 5] as well as improve yield and function [2, 9], and has been implied to be necessary for proper co-translational folding [10–13], yet at the same time impose a fitness cost on the cell [14]. Here we show a link between secondary structure motifs, anti-Shine-Dalgarno-like sequences, and pausing locations along the length of a transcript.

Hypergeometric enrichment testing of the pause sites downstream from the end of each structure points towards the enrichment of pausing at certain codon locations (Supplementary Figure A.2A, C, E). At the same time, sequence analysis near the ends of the secondary structures which have pause sites at these locations show increased occurrences of anti-Shine-Dalgarno-like sequence, approximately 6 codons upstream from the enriched pause site for each secondary structure motif (Supplementary Figure A.2B, D, F). This indicates that secondary structures tend to have pause sites downstream from their ends, and anti-SD-like sequences might be used to induce the pausing. It is noted here that in contrast to other studies which determine the pausing propensity with relation to the assumed A site, assigning ribosome density to the 3' end of the read results in a pause site which is around 2-3 codons further downstream. For example, Li *et. al.*[5] found that anti-SD-like sequences were linked to pausing 8-11 nucleotides (3rd-4th codon) downstream, which in our analysis occurs on the 6th codon downstream instead because of the different positional assignment of the read. Datasets from Mohammad *et. al.* 2016 which make use of an altered protocol in order to reduce biases also showed pause site enrichments at similar location, and show the corresponding increase in SD-like sequences near the ends of secondary structures (Supplementary Figure A.9).

Overall, across both MOPS minimal and MOPS rich growth conditions, around 50%

of ribosomal density and 60% of pause sites can be attributed to either anti-SD-like sequences or secondary structure motifs (Supplementary Figure A.4). While this indicates that these two factors play a major role in ribosome pausing locations, the overlap between them is only between around 15% of all pause sites. This suggests that there are other factors unaccounted for in this study, which either require, or induce pausing, and should be the subject of further investigation.

A.1.2 Supplementary Note 2: Structure of ME models

Here we briefly introduce key formulations of the ME model, and refer the interested reader to a more complete supplementary information in O'Brien *et. al.* 2013[15]. The ME model is a steady state growth model which accounts for metabolism (M) and gene expression (E). Based on an input of available nutrients to the cell such as carbon and nitrogen sources, it predicts: a) the cell's maximum growth rate in a specific condition, and at the maximum growth rate, b) metabolite uptake and excretion rates, c) metabolic reaction fluxes, and d) gene expression fluxes such as translation and transcription rates. This is achieved through formulating transcription, translation, transport and metabolic fluxes into a quasi-linear problem and solving for maximum growth rate, taking into account compartmentalization of proteins and metabolites into the cytoplasm, periplasm, and extracellular region. This is done as follows:

1. The metabolic network is described by a stoichiometric matrix, similar to those used in M-models4 where rows represent metabolites and columns represent reactions. Coefficients represent the metabolites consumed (negative value) or produced (positive value) in each reaction. At steady state, there is no change in metabolite concentrations, hence we get: $Sv=0$ where S is the stoichiometric matrix, and v is the flux vector, allowing us to solve for v.

2. In the ME model, translation is accounted for by enforcing that proteins have to be

produced for all enzyme-catalyzed reactions, proportional to the flux for that particular reaction. Protein synthesis is balanced by protein dilution (due to cell division), which is proportional to the amount of protein required to hold the predicted flux. Note that in this case we disregard direct protein degradation as it has been shown to negligible in a growing *E. coli* cell for most metabolic genes [16, 17]. While active degradation is important for many signaling proteins such as transcription factors, these proteins are a very small proportion of the whole proteome in rapidly growing cell. The depletion of most of the modeled proteins is therefore mostly through dilution caused by growth, and the depletion rate is simply the growth rate. Mathematically, this is represented below:

$$v_{\text{translation},i} = v_{\text{diluation},i} = \mu[E_i]$$

$$v_{\text{reaction},i} = k_{\text{eff},i}[E_i]$$

$$v_{\text{translation},i} = \frac{\mu}{k_{\text{eff},i}} v_{\text{reaction},i}$$

The above relationships link the required translation rate $v_{\text{reaction},i}$ to the flux through the reaction $v_{\text{translation},i}$ as well as the predicted growth rate μ and the proteins effective catalytic rate $k_{\text{eff},i}$. Unlike traditional M models where biomass has to be explicitly modeled, this inherently takes protein and macromolecule dilution into account during growth, and allows prediction of optimal protein production, and hence gene expression.

It is important to note that, in previous ME models [15], the effective catalytic rate was set to be proportional to the respective enzyme's solvent accessible surface area (SASA). Here, we make use of condition-specific proteomics data, which vastly improves this parameter and the predictive scope of the model itself. More details of this parameterization procedure is found in the following section.

Finally, translation of proteins are catalyzed by ribosomes, requiring tRNA and its synthethases and the cost of production of these molecules are also explicitly accounted for in the ME model. Analogous to metabolic enzymes, ribosome efficiency k_{ribo} is a necessary parameter for coupling the ribosome production rate to the overall proteome translation rates, and is shown below:

$$v_{\text{synthesis, ribosome}} = v_{\text{dilution, ribosome}} = \sum_i \text{length}\left(\text{peptide}_i\right) \cdot \frac{\mu}{k_{\text{ribo}}} \cdot v_{\text{translation},i}$$

Because ribosomes are themselves partially made up of peptides which need to be synthesized, this coupling creates an asymptotic relationship between ribosome production and growth rate. Placing a limitation on cell size replicates the natural phenomenon where growth rate is constrained by the balance between enzyme and ribosome production even in the overabundance of nutrients.

3. In the ME model presented in this contribution, transcription is now achieved through the production of mRNA from nucleotides, catalyzed by RNA polymerase. This is also handled in a similar way to ribosomes and metabolic enzymes through dilution of RNA polymerase, proportional to the total flux through all transcription reactions.

A.1.3 Supplementary Note 3: ME model coupling parameters

In addition to metabolic reactions, the ME model describes various biological processes as biochemical reactions. For example, translation of a protein is described by a reaction assembling the component amino acids into peptides. An enzyme complex is then formed by a reactions which assemble together the various peptides and cofactors, and more reactions which apply the necessary post-translational modifications. Dependent reactions are linked through what are termed "coupling constraints" because they force a certain amount of flux through one of the

reactions based on the level of flux through the other, effectively coupling the processes. For example, flux through a metabolic flux is coupled to production of the catalytic enzyme. A translation reaction is coupled to both reactions which produce ribosomes and reactions which transcribe the mRNA. In iOL1650-ME, these coupling constraints are expressed as inequalities as functions of growth rate to reflect how the nature of many of these constraints are nonlinear in growth. However, the constants in these functions are set for each individual coupling constraint. A detailed description of all the processes and coupling parameters in the iOL1650-ME model is available in the supplementary information of that manuscript [15].

One of the most critical coupling constraints to the function of the ME model is the k_{eff}, which describes the amount of enzyme required on average to sustain a unit of flux under in vivo conditions. An example coupling constraint for a metabolic reaction has the form:

$$v_{\text{metabolic}} \leq \frac{k_{\text{eff}}}{\mu} v_{\text{enzyme}}$$

This coupling constraint expresses both how as growth rate increases, more enzyme must be made to pass on to each daughter cell, and how the amount of enzyme production relates to its efficiency through the k_{eff}parameter. While the k_{eff}parameter has units of 1/time like a k_{cat}, it is strictly less than the k_{cat}, as it is describing the amount of enzyme which must be made to catalyze a unit of flux instead of the maximal flux a particular enzyme can sustain. The reason this parameter is critical to the function of the ME model is because it describes the relative cost of each enzyme. An appropriate analogy is an econometric model. Where the reactions themselves express operational costs, and the cost of the enzyme is the capital cost. A good example of this the difference between the "expensive" but efficient pyruvate dehydrogenase complex compared to pyruvate formate lyase. Pyruvate dehydrogenase (PDH) and pyruvate formate lyase (PFL) both convert pyruvate to acetyl coA to allow carbon to flow through the

TCA cycle. Flux through PFL will result in secretion of formate, whereas flux through PDH will result in production of CO2, which is the in vivo behavior. In *E. coli*, PDH is one of the most expensive metabolic enzymes because it consists of 60 subunits. However, an optimally efficient *E. coli* cell might still produce this enzyme over PFL because its catalytic rate is more than commensurately higher, as it exploits phenomena such as substrate tunneling and multiple catalytic sites, and will therefore have a lower protein cost per unit of flux catalyzed. Therefore, in order for an ME model to correctly predict flux through PDH, it must have keff parameters for PDH and PFL which represents this tradeoff accurately. In a genome-scale ME model, these parameters affect the cost between all the various alternate pathways and isozymes. Therefore, the accuracy of the model predictions of protein expression and the physiological state of the cell will depend on reasonable relative values of these parameters.

A.1.4 Supplementary Note 4: Simulation of Batch Growth with ME

One of the advantages of the ME model over traditional M models is its ability to account for proteome limitation. This gives the model the ability to correctly simulate batch growth, where a cell has a surplus of nutrients around it, but is limited by its ability to produce the proteins which are required to process these nutrients. Unlike M models, which can predict growth rates given a specific substrate uptake rate, a ME model can predict the maximal growth rate given an unbounded substrate uptake (allowing the model to take up as much substrate as it wants). O'Brien *et. al.* investigated how proteome limitation begins to take effect as the substrate uptake rate approaches the optimal value, eventually causing a maximal growth rate3. Beyond this optimal growth rate, the model is infeasible with any substrate uptake rate because of the proteome limitation constraints. This optimal growth rate can be computationally identified

by doing a binary search. Because the experimental data used in this study were all generated under batch growth conditions, the simulations of growth used this batch growth procedure, which computes a proteome-limited state. Supplementary Note 5: Simulating ME with estimated parameters In the absence of in vivo experimental measurements, the original iOL1650-ME model estimated k_{eff}values based on the solvent-accessible surface area (SASA), which is a function of protein size [15, 18]. We sought to evaluate the effect of our new set of estimated parameters values on predictions of differential gene expression. We simulated a switch of the primary carbon substrate from fumarate to acetate, which could be validated by the previously generated mRNA sequencing data sets (GSE59759). Using the original k_{eff}values obtained from protein size based estimation we observed a significant number of false positive predictions due to incorrect pathway usage, with 17 false positives found out of 37 total predictions. Using the consensus k_{eff}parameters derived from sets A + B yielded only 6 false positives with the same number of correct predictions. Interestingly, almost all of the improvement in prediction came from using parameter set A alone (Supplementary Figure A.6), which constrained parameters in central carbon metabolism. This result suggests that the accuracy of a ME model is most sensitive to these key 28 k_{eff}parameters which lie in its high flux backbone [19].

After showing that the set of estimated k_{eff}values gives better predictions for previously examined nutrient shift, we generated new experimental data for growth on dual substrates. We computed predicted differential gene expression after media supplementation with four key nutrients: Adenine, Glycine, L-Threonine, or L-Tryptophan, using 1) k_{eff}found in set A, and 2) k_{eff}found in both sets A and B (TableA.1). Genes that were computed to change in expression by more than a factor of 16 after supplementation were considered to be predictions of differential expression (Supplementary A.8). Prediction validation was performed using mRNA sequenc-

ing data [19], some of which was taken from a previous study [19], to experimentally determine differentially expressed genes. Predictions were made with a slightly modified iOL1650-ME (Supplementary Figure A.5). For all four supplementations, the accuracy of ME predictions (Table 1) was higher than those resulting from sampling M model flux states (Supplementary FigureA.8). Moreover, the accuracy increased when comparing parameter set A + B to parameter set A alone. Using parameter set A + B, the accuracy of predictions of ranged from 55 to 100%, and all predicted sets were significantly enriched for differentially expressed genes ($p < 0.05$ using a hypergeometric distribution).

These results suggest that a parameterized genome-scale ME model, due to its incorporation of enzyme biosynthetic costs, gives a significant improvement in prediction of gene expression over existing methods using M models alone. For example, the ME model correctly predicts the often non-intuitive shift of amino acid precursors. Specifically, when supplementing with L-Threonine, the ME model will produce L-Serine from L-Threonine directly, as observed experimentally. On the other hand, the ME model predicts no significant up regulation of glyA to produce L-Serine from Glycine supplementation, as seen in the expression profiling data. The iJO1366 M-model will incorrectly predict this change (Supplementary Figure A.6c), demonstrating how the ability to account for protein expression costs improves the accuracy of the predicted flux state. Moreover, as a result of its quantitative numerical predictions of gene expression, ME models can also predict partial up and down regulation of genes, in addition to binary responses when a gene goes from active to inactive. For example, after supplementation with Adenine, the model correctly predicts downregulation of *purHD* and *purMN* (Supplementary Figure A.6c).

A.2 Supplementary Methods

A.2.1 mRNA seq

Cells were harvested at mid-log (OD600 $\approx$ 0.3) in biological duplicates for each condition. From each sample, 3mL of culture were mixed with 6mL RNAprotect Bacteria Reagent (Qiagen), incubated for 5 minutes, and then centrifuged at 5000g for 10 minutes at room temperature. Total RNA samples were then isolated from the pellet using the RNeasy Plus Mini kit (Qiagen). Samples were quantified using a NanoDrop 1000 (Thermo Scientific) and an Agilent RNA 6000 Nano Kit with an Agilent 2100 Bioanalyzer. Strand-specific mRNA libraries were created using the dUTP method, with ribosomal rRNA subtraction with the Ribo-Zero rRNA Removal Kit (Epicentre). Libraries were run on a Miseq (Illumina, CA) multiplexed with 2 duplicates per run as per the manufacturer's instructions. Expression values were computed using the bowtie [20] and cufflinks [21] packages. The processed data were uploaded to GEO under accession numbers GSE59759 and GSE59760.

A.2.2 Structural Data Retrieval and Manipulation

Incorporating protein-related information into a GEM involves four stages of semi-automated curation: (i) map the genes of the organism to available experimental protein structures, found in publicly available databases, such as the Protein Data Bank (PDB); (ii) determine genes with and without available protein structures and perform homology modeling using the I-TASSER suite of programs [22] to fill in gaps where crystallographic or NMR structures are not available; (iii) perform ranking and filtering of PDB structures for each gene based on a set selection criteria (e.g., resolution, number of mutations, completeness); (iv) map GEM genes to other databases (e.g., BRENDA [23, 24], SwissProt [25], Pfam [26], SCOP [27]) for complementary

protein-structure derived data. The quality of the reconstruction expansion process to include high confidence protein structures is considered by carrying out a series of QC/QA verification steps during the ranking and filtering stage. The GEM annotation of the organism of interest is stored in SBML and Matlab formats and many organisms can be found in the BiGG database [28]. Amino acid sequence of the proteins of interest are stored in FASTA format. To map protein structural data to a GEM, we make use of Python modules, ProDy [29, 30] and Biopython [31] to parse information in the PDB files. The molecular visualization software VMD37 was used for viewing the 3D structure of the modeled protein and the predicted functional sites and the creation of images. Installation of PfamScan and HMMER3 algorithms are required for generating protein fold families for certain proteins [32, 33]. Open source software for protein structural predictions are available and are used in conjunction with the IPython framework.

A.2.3 Predictions of mRNA expression in parametrized conditions

We sought to evaluate the effect of the estimated parameter values on predictions of differential gene expression under identical experimental conditions as the proteomics used to parameterize the model. We simulated a switch of the primary carbon substrate from fumarate to acetate, and obtained mRNA sequencing data (GSE59759) for *E. coli* K-12 BW25113 (obtained from the Coli Genetic Stock Center) cultivated under identical experimental conditions to those of the proteomic data. Bach growth simulations were then performed using 3 different sets of keff parameters: the initial solvent-accessible-surface area parameters, keff parameters derived from fluxomics (set A), and keff parameters derived from fluxomics and our proteomics-based algorithm (set A + B). Predicted differential expression was compared to the set of genes identified as differentially expressed by cufflinks [21] at a false discovery rate of 0.05. In order to account

for accuracy on a level field with respect to sensitivity, we varied the computational cutoff so that we would have a similar number of correct predictions (as close to 20 as possible), allowing us to directly compare the number of incorrect predictions. Using the the original iOL1650-ME model keff values obtained from protein size based estimation, we observed a significant number of false positive predictions due to incorrect pathway usage, with 17 false positives found out of 37 total predictions. Using a consensus keff parameter set derived from both experimentally measured flux values [34] (set A) and model-predicted values (set B) yielded only 6 false positives with the same number of correct predictions. Additionally, we were able to improve predictions greatly using the 28 keff parameter values from set A alone (Supplementary Figure A.8), which only constrained parameters in central carbon metabolism. This result suggests that the accuracy of a ME model is most sensitive to keff parameters for reactions which lie along its high flux backbone.

A.2.4 Sampling of M-model flux states in iJO1366

The optGpSampler[35] software was used to sample flux states in the iJO1366 M model using its python API. First, the model was reduced by removing blocked reactions as identified by flux variability analysis in cobrapy42. For each simulation, the lower bound on the biomass reaction was set to 90% of its optimal value. Additionally, the reactions HXAND, XAND, and URIC were blocked (Supplementary Figure A.7). The sampling algorithm was then run for 100 steps and generated 10000 points for each simulation. Afterwards, the fluxes were linearly scaled such that the mean flux of the biomass reaction was 1. Sampling was run on the model with only D-Glucose uptake, and also with 10 mmol/gDw/hr uptake allowed of each of the supplements. Reactions which were unbounded (as determined by an FVA maximum greater than 500 or an

FVA minimum of less than -500) were excluded from the subsequent analysis. For each of the supplements, predicted changed reaction fluxes between the supplemented and unsupplemented samples were determined by finding reaction fluxes where (1) the mean changed by more than a factor of 2 and (2) the mean changed by more than the sum of the standard deviations for the supplemented and unsupplemented fluxes. These reactions were converted to gene predictions by assuming all genes in the gene reaction rule for the changing reaction were up or down regulated. These gene predictions were validated against mRNA sequencing data in the same manner as the ME gene differential expression predictions. This method was used instead of the more traditional method comparing pairs of samples as done in some other studies [9] because it resulted in higher accuracy than those methods with this data.

A.3 Supplementary Figures

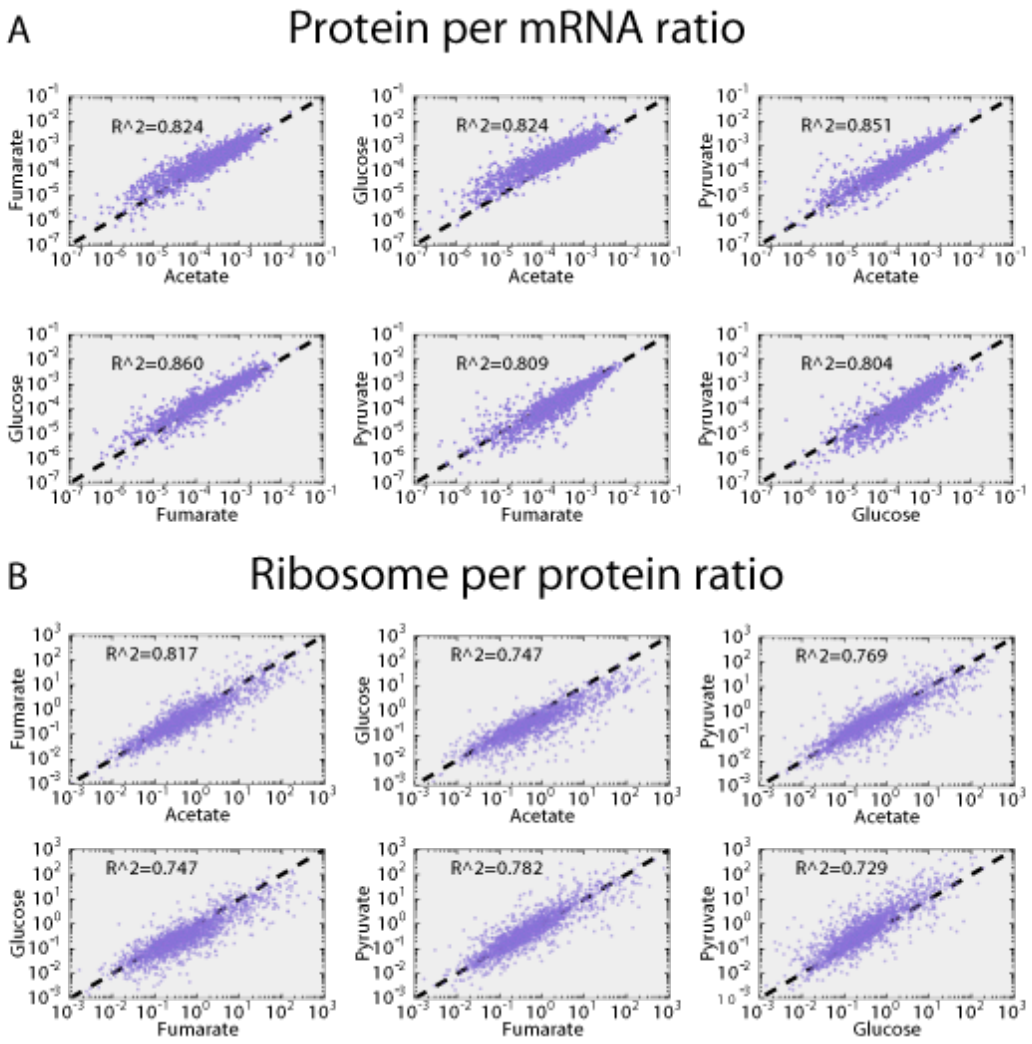

Figure A.1: Protein per mRNA ratios (A) and ribosome per protein (B) ratios across environments are highly conserved.
Ratios span several orders of magnitude across genes, but are highly conserved across different experimental conditions

Table A.1: Predicted expression changes confirmed experimentally.
The ME model predicted differential expression for the following nutrient supplementations to growth of E. coli on M9 Minimal Media with D-Glucose. Two separate sets of k_{eff} were used, one using only the 28 parameters in set A, and the other also adding in the 284 modeling-derived parameters in set B. The predictions were evaluated using the set of differentially expressed genes determined from mRNA sequencing for each nutrient supplementation. Two small modifications were made to the model for Adenine, which initially had an accuracy of only 26.2% due to the genes used for Adenine degradation in the model which are not expressed and may not be functional (for details see Figure A.7). The numbers provided in the correct column are the number of genes which are predicted to be differentially expressed and also are differentially expressed in the same direction in the RNA-seq data. The number in parenthesis refers to the number of correctly-predicted differentially expressed genes which are still expressed under both conditions but varied in their predicted quantitative values. The incorrect column contains the number of genes predicted to be differentially expressed which were not in the model. These numbers are used to predict the percent accuracy, and the p-value for a hypergeometric enrichment of differentially expressed genes in the predicted set.

	k_{eff} parameters from set A				**k_{eff} parameters from set A+B**			
	Correct	Incorrect	Accuracy	p	Correct	Incorrect	Accuracy	p
L-Tryptophan	6 (0)	3	66.7	0.031	7 (0)	0	100	0.011
L-Threonine	30 (5)	15	66.7	0.119	15 (5)	5	75	0.044
Adenine	12 (4)	35	25.5	0.663	11 (4)	7	61.1	4.00E-06
Glycine	15 (0)	20	42.9	0.391	5 (0)	4	55.6	0.024

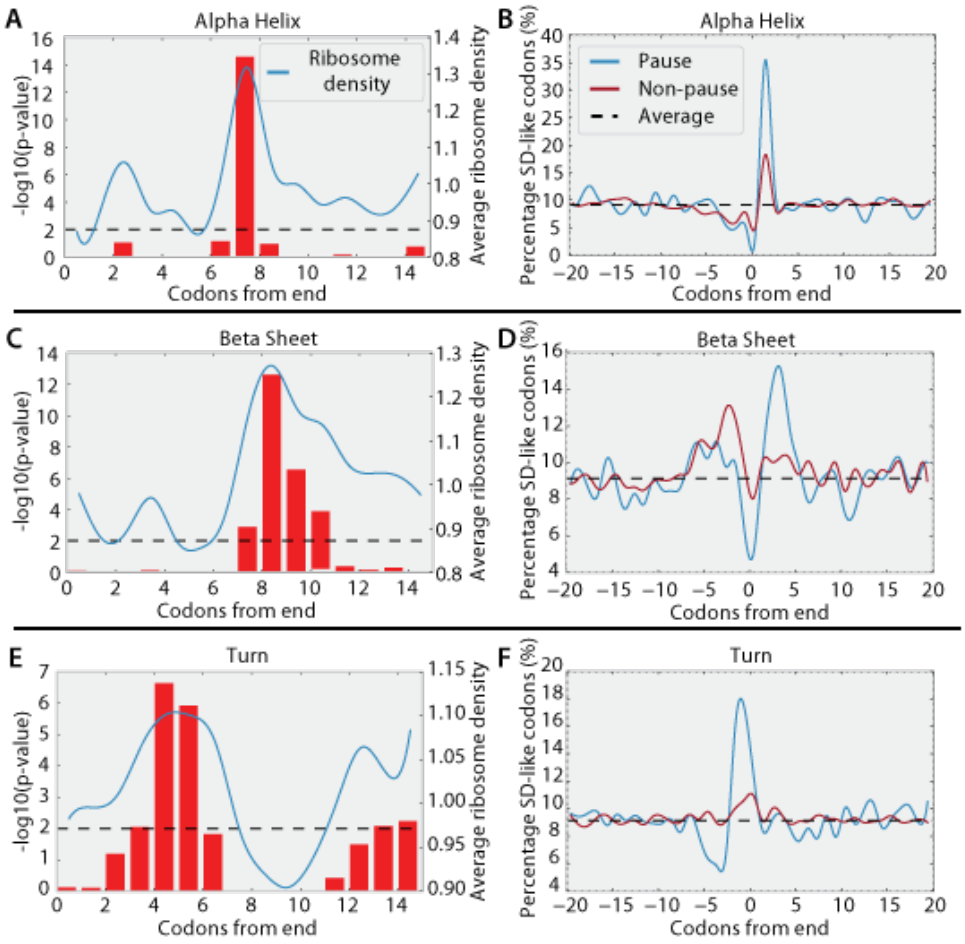

Figure A.2: Pause site enrichment and percent SD-like sequence near ends of secondary structures.

Hypergeometric enrichment testing was used to determine locations downstream of secondary structures which were enriched for pausing (A, C, E). Red bars represent p-value of enrichment. Blue line indicates the per gene normalized ribosome density at each codon position averaged across all occurrences of the secondary structure, showing a matching increase in average ribosome density. Based on the codon locations indicated by the enrichment test, all occurrences of each secondary structure were divided into those with corresponding pause sites (Pause) and those without (Non-Pause). The percentage occurrence of Shine-Dalgarno-like sequences appearing near the ends of these secondary structures is shown (B, D, E). The global average occurrence of an SD-like sequence is 9.13%, but this increases greatly at certain codon locations near the ends of secondary structures with corresponding pause sites (Blue line).

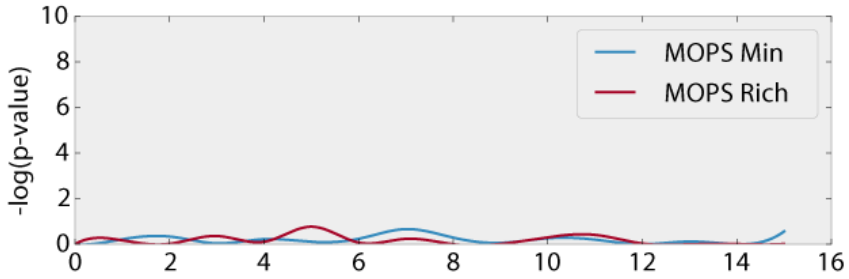

Figure A.3: Hypergeometric enrichment test of codons downstream from annotated SCOP domains.
No enrichment was found when codons downstream from the ends of domains were tested for pause site enrichment. Previous studies have found evidence that show slow translating regions between domains might be important for proper folding. However the location and even presence of pause sites might only occur on a case by case basis specific to particular domains, and fail to show enrichment when tested on the genome scale.

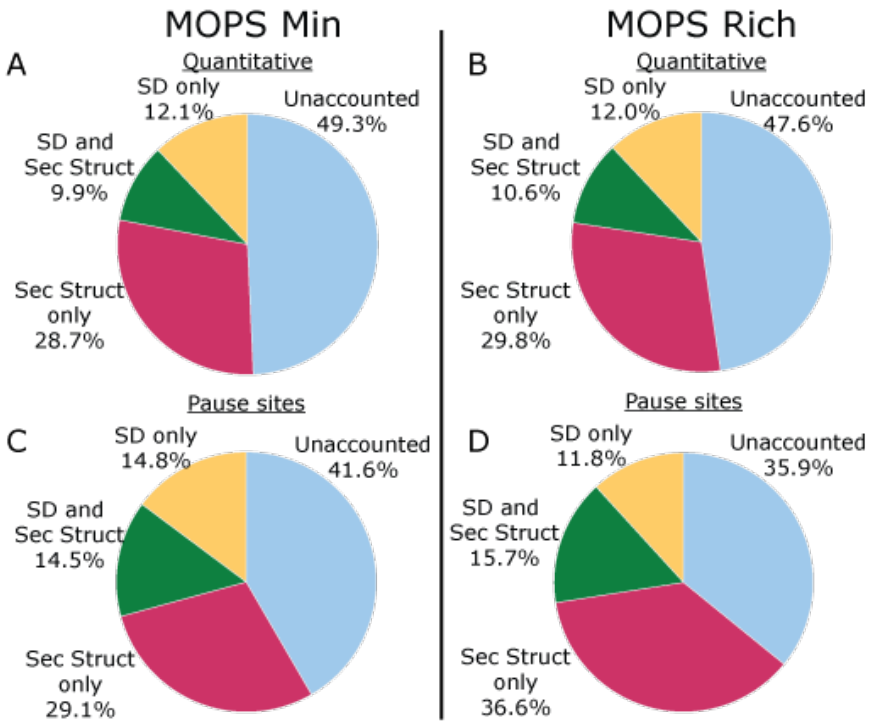

Figure A.4: Percentage of ribosome density and pause sites linked to SD-like sequences and/or Secondary Structure

Pie charts showing the distribution of pause sites linked to secondary structures and/or Shine-Dalgarno-like codons. In both MOPS minimal and MOPS Rich media, codons indicated to be pause-enriched for SD-like sequences accounted for around 20% of ribosomal density (A, B) and 30% of pause sites (C, D), while secondary structures accounted for 40% and 35% respectively. Of these, around 20–25% of these codons are indicated by both SD-like sequences and secondary structures. Around 50% of ribosomal density and 40% of all pause sites were still unaccounted for, indicating that there are other factors linked to pausing which were not included.

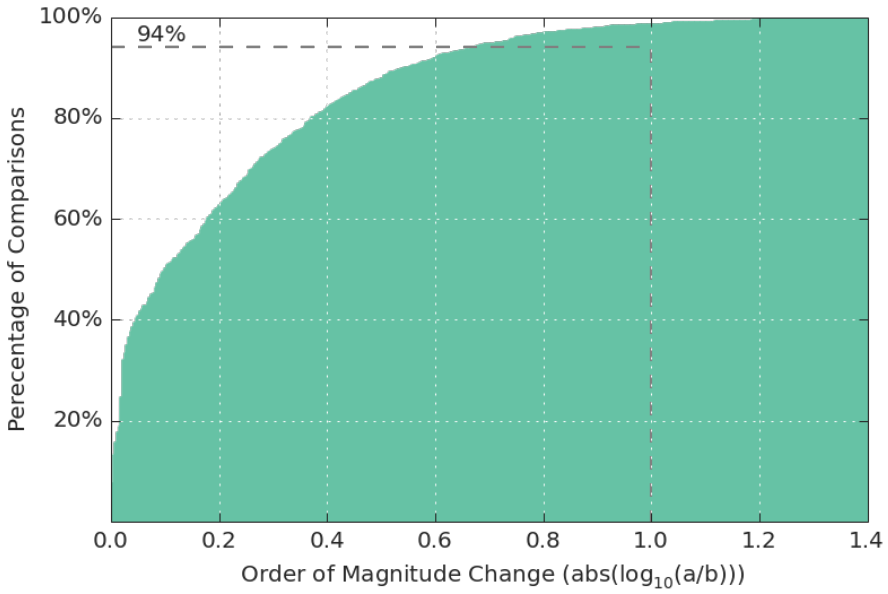

Figure A.5: Distribution of pairwise comparisons between computed k_{eff}.
Each of the 284 k_{eff} was compared between conditions (6 comparisons between each of the 4 conditions). Plotted is the cumulative distribution of all these pairwise comparisons in terms of the change in order of magnitude. We observe that 94% of these comparisons remain within an order of magnitude. For the comparisons which were not within an order of magnitude, proteins associated with those complexes were more likely to catalyze multiple reactions (42.5% catalyzing more than one reaction v.s. 27.8% catalyzing more than one reaction).

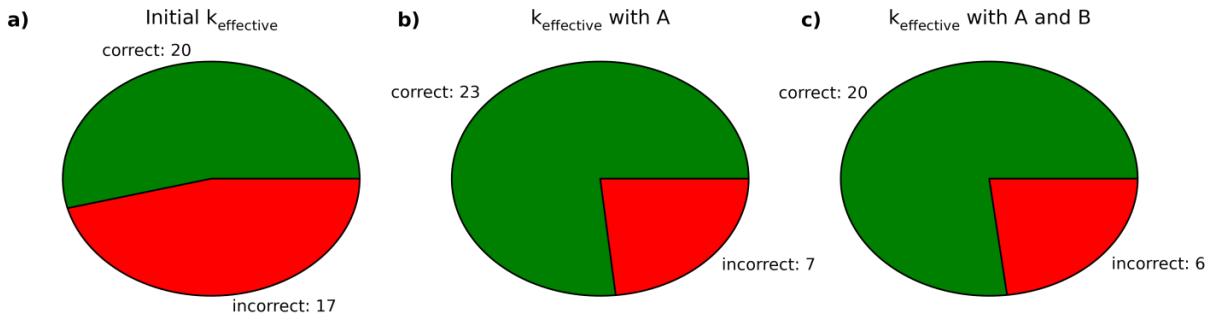

Figure A.6: Predicted Differential Expression between Fumarate and Acetate
The ME model was used to predict differential expression between growth on fumarate and growth on acetate as the main carbon substrates. These predictions were run with three different sets of k_{eff} and then validated using mRNA sequencing. This is described in more detail in the "Predictions of mRNA expression under identical conditions to proteomic data" section.

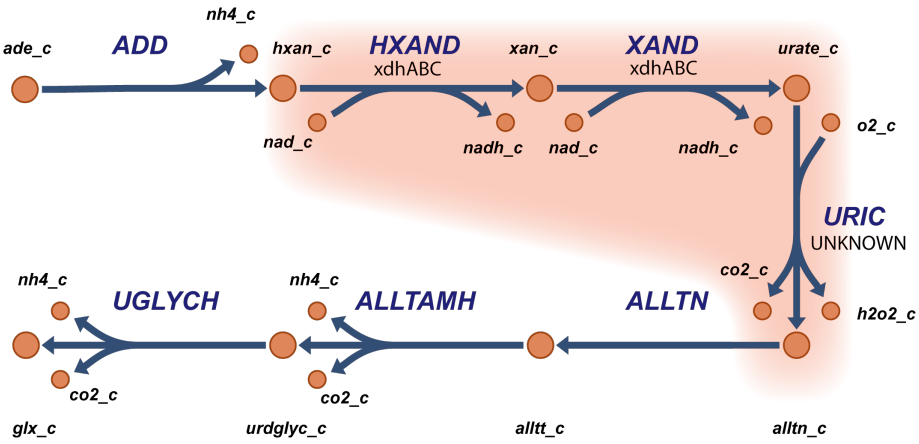

Figure A.7: Updates to the Adenine Degradation Pathway in the *E. coli* Metabolic Reconstruction.

As reconstructed in *i*JO1366 and *i*OL1650-ME, this pathway breaks down intracellular adenine to glyoxylate. However, our expression-profiling data with adenine supplementation suggests that the highlighted reactions do not occur to break down adenine, even though the upstream reaction adenine deaminase does occur. While these reactions were included in the model based on their homology to adenine degradation pathways in other organisms, it is likely that the pathway is latent or inactive in E. coli K-12. Therefore, the reactions HXAND and XAND were disabled during simulations on Adenine. Additionally, use of the unexpressed gene focB was also penalized.

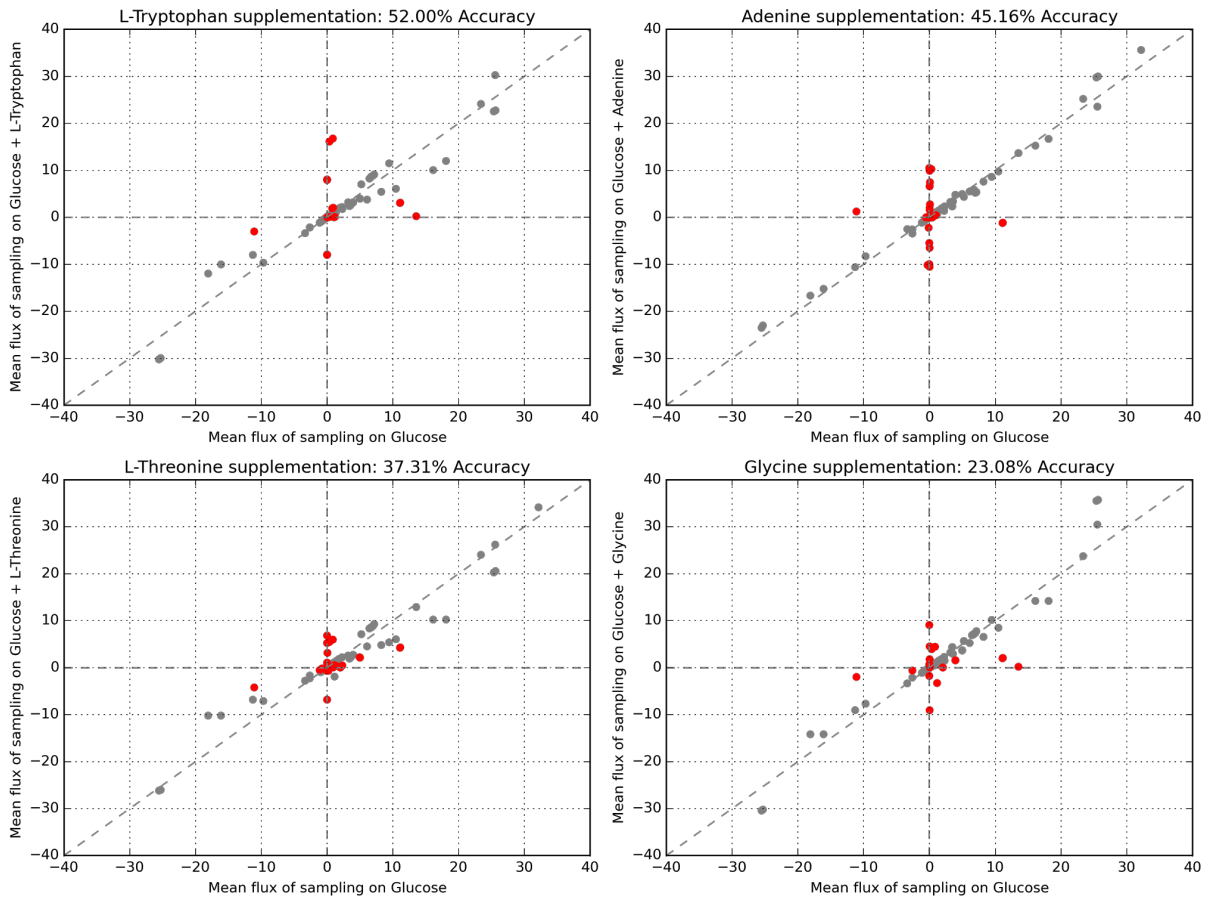

Figure A.8: Sampling *i*JO1366 Flux States Following Nutrient Supplementation.
To compare ME predictions of gene differential expression to that of the M model, sampling was run on *i*JO1366 to identify reactions which changed significantly in flux (colored red) after supplementation. For each of these four conditions, the mean flux from sampling on glucose is plotted on the x axis, and the mean flux from sampling on glucose with the nutrient supplementation is plotted on the y axis. The model gene reaction rules were used to convert these to gene expression predictions, which were compared to mRNA sequencing to give the reported accuracies.

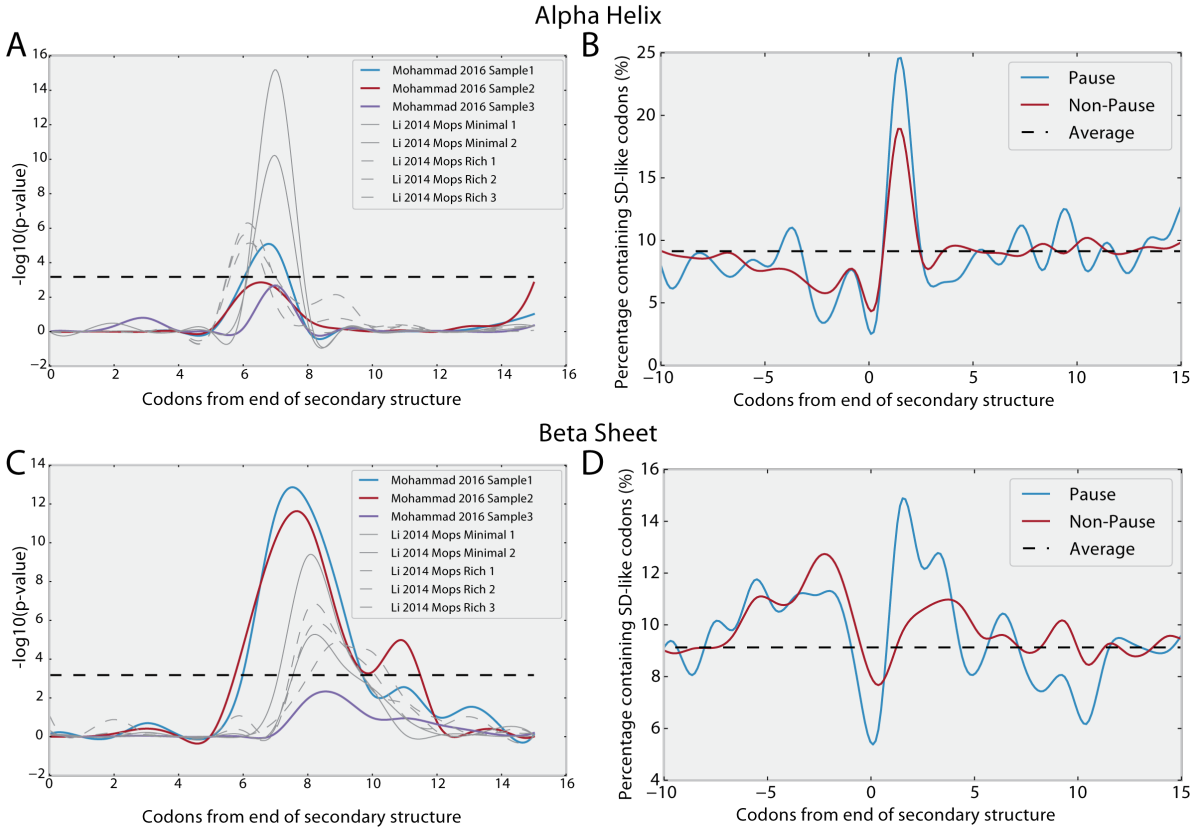

Figure A.9: Pause site enrichment downstream of secondary structures across datasets. The location of pause site enrichments downstream of alpha helix (A) and beta sheet (C) secondary structure motifs between several datasets from Mohammad *et. al.* 2016 [36] and Li *et. al.* 2014 [37], which make use of different protocols. The location of pause site enrichment is consistent across both protocols. Similar to the analysis done in Supplementary figure 2B, D, F, the datasets from Mohammad *et. al.* 2015 also show increased prevalence of SD-like codons at the ends of secondary structures with pause sites downstream as compared to those without (B, D).

A.4 References

1. Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S. & Liu, Y. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular cell* **59,** 744–754. ISSN: 1097-2765, 1097-4164 (Sept. 2015).

2. Siller, E., DeZwaan, D. C., Anderson, J. F., Freeman, B. C. & Barral, J. M. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *Journal of molecular biology* **396,** 1310–1318. ISSN: 0022-2836, 1089-8638 (Mar. 2010).

3. Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S. & Futcher, B. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* **3.** ISSN: 2050-084X. doi:10.7554/eLife.03735 (Oct. 2014).

4. Sørensen, M. A., Kurland, C. G. & Pedersen, S. Codon usage determines translation rate in Escherichia coli. *Journal of molecular biology* **207,** 365–377. ISSN: 0022-2836 (May 1989).

5. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484,** 538–541. ISSN: 0028-0836, 1476-4687 (Apr. 2012).

6. Hess, A.-K., Saffert, P., Liebeton, K. & Ignatova, Z. Optimization of translation profiles enhances protein expression and solubility. *PloS one* **10,** e0127039. ISSN: 1932-6203 (May 2015).

7. Zhang, G. & Ignatova, Z. Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PloS one* **4,** e5036. ISSN: 1932-6203 (Apr. 2009).

8. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology* **16,** 274–280. ISSN: 1545-9993, 1545-9985 (Mar. 2009).

9. Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J. M., Sachs, M. S. & Liu, Y. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495,** 111–115. ISSN: 0028-0836, 1476-4687 (Mar. 2013).

10. Saunders, R. & Deane, C. M. Synonymous codon usage influences the local protein structure observed. *Nucleic acids research* **38,** 6719–6728. ISSN: 0305-1048, 1362-4962 (Oct. 2010).

11. Purvis, I. J., Bettany, A. J., Santiago, T. C., Coggins, J. R., Duncan, K., Eason, R. & Brown, A. J. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. *Journal of molecular biology* **193,** 413–417. ISSN: 0022-2836 (Jan. 1987).

12. Komar, A. A., Lesnik, T. & Reiss, C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS letters* **462,** 387–391. ISSN: 0014-5793 (Dec. 1999).

13. Subramaniam, A. R., Zid, B. M. & O'Shea, E. K. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* **159,** 1200–1211. ISSN: 0092-8674, 1097-4172 (Nov. 2014).

14. Hingorani, K. S. & Gierasch, L. M. Comparing protein folding in vitro and in vivo: foldability meets the fitness challenge. *Current opinion in structural biology* **24,** 81–90. ISSN: 0959-440X, 1879-033X (Feb. 2014).

15. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. eng. *Molecular systems biology* **9,** 693. ISSN: 1744-4292 (2013).

16. Maurizi, M. R. Proteases and protein degradation in Escherichia coli. *Experientia* **48,** 178–201. ISSN: 0014-4754, 1420-9071 (1992).

17. Nath, K. & Koch, A. L. Protein degradation in Escherichia coli. II. Strain differences in the degradation of protein and nucleic acid resulting from starvation. *The Journal of biological chemistry* **246,** 6956–6967. ISSN: 0021-9258 (Nov. 1971).

18. Miller, S., Lesk, A. M., Janin, J. & Chothia, C. The accessible surface area and stability of oligomeric proteins. en. *Nature* **328,** 834–836. ISSN: 0028-0836 (Aug. 1987).

19. Bordbar, A., Nagarajan, H., Lewis, N. E., Latif, H., Ebrahim, A., Federowicz, S., Schellenberger, J. & Palsson, B. O. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Molecular systems biology* **10** (2014).

20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature methods* **9,** 357–359. ISSN: 1548-7091 (Apr. 2012).

21. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. en. *Nature biotechnology* **28,** 511–515. ISSN: 1087-0156 (May 2010).

22. Zhang, Y. I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins: Structure, Function, and Bioinformatics* **77,** 100–113 (2009).

23. Chang, A., Scheer, M., Grote, A., Schomburg, I. & Schomburg, D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic acids research* **37,** D588–D592. ISSN: 0305-1048 (2009).

24. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. & Schomburg, D. BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research* **32,** D431–D433. ISSN: 0305-1048 (2004).

25. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31,** 365–370. ISSN: 0305-1048 (2003).

26. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., *et al.* The Pfam protein families database. *Nucleic acids research* **32,** D138–D141. ISSN: 0305-1048 (2004).

27. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247,** 536–540. ISSN: 0022-2836 (1995).

28. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics* **11,** 213. ISSN: 1471-2105 (Jan. 2010).

29. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* **27,** 1575–1577. ISSN: 1367-4803 (2011).

30. McKinney, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython* (" O'Reilly Media, Inc.", 2012).

31. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423. ISSN: 1367-4803 (2009).

32. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* Pfam: the protein families database. *Nucleic acids research,* gkt1223. ISSN: 0305-1048 (2013).

33. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., *et al.* The Pfam protein families database. *Nucleic acids research,* gkp985. ISSN: 0305-1048 (2009).

34. Nanchen, A., Schicker, A. & Sauer, U. Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. *Applied and environmental microbiology* **72,** 1164–1172. ISSN: 0099-2240 (Feb. 2006).

35. Megchelenbrink, W., Huynen, M. & Marchiori, E. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PloS one* **9,** e86587. ISSN: 1932-6203 (Feb. 2014).

36. Mohammad, F., Woolstenhulme, C. J., Green, R. & Buskirk, A. R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell reports* **14,** 686–694. ISSN: 2211-1247 (Feb. 2016).

37. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635. ISSN: 0092-8674, 1097-4172 (Apr. 2014).

Appendix B

Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression - Supplementary Information

B.1 Methods

B.1.1 Bacterial strains and growth conditions

Media used was M9 minimal media with 0.2% w/v glycerol. Bacterial strains used were Gly2 glycerol evolved strains from [1] which grows optimally with glycerol as a carbon source.

B.1.2 Generation of library of gratuitous proteins

We generated a library of 46 strains each expressing a protein on a plasmid controlled by a rhamnose promoter (Supplementary Table 1). Gene blocks for each protein were generated by PCR from existing plasmids. These were inserted into a vector backbone based on a modified pNIC28-Bsa4 plasmid containing the rhamnose transcription regulators rhaS and rhaR through 3' LIC cloning[2] under the control of a rhamnose rhaBAD promoter. All proteins were preceded by a His-TEV tag at the N terminus as well as a bi-cistronic design (BCD) sequence to control translation initiation [3]. Sequences were verified by Sanger sequencing. Verified plasmids were transformed into Gly2 cells [1]. A negative control plasmid was included in the dataset (CC1) which included all plasmid elements except the heterologous protein sequence.

B.1.3 Culture conditions

Each heterologous protein expressing strain was grown up overnight in M9 minimal media with 0.2% w/v glycerol as a carbon source. 2mL of overnight culture was inoculated into 60mL of fresh M9/glycerol. Cultures were incubated shaking at 37C until they reached an A600 OD of 0.3. Rhamnose was added to a final concentration of 1mM to induce protein expression. Cultures were further incubated for an additional 2hrs after induction and harvested.

B.1.4 Transcriptomics

Cells were pelleted and lysed with a modified version of the RNAProtect Bacteria Reagent protocol, ribosomal RNA was depleted using Ribo-Zero rRNA Removal Kit for Gram-Negative bacteria (Illumina), libraries were created using KAPA RNA Library Preparation kit. Deviations from the kit protocols are mentioned below. 3mL of induced culture was added to 6mL of

RNAProtect Bacteria Reagent (Qiagen) and vortexed, then left at room temperature to incubate for 5 minutes. Cells were pelleted and then resuspended in 400uL elution buffer and then split into two tubes, with one kept as a spare. One pellet was then lysed enzymatically with the addition of lysozyme, proteinase-K and 20% SDS. SUPERase-In was added to maintain the integrity of the RNA. RNA isolation was then performed according to the rest of the kit protocol. rRNA was the depleted using the Ribo-Zero rRNA Removal Kit for Gram-Negative Bacteria according to the protocol, and libraries were constructed for paired-end sequencing using the KAPA RNA-Seq Library Preparation kit protocol. Libraries were sequenced on an Illumina NextSeq platform. Transcriptomic reads were mapped using Bowtie2 [4], and reads were counted using HTSeq [5]. Libraries were normalized using TPM including the heterologous gene to determine the percentage of heterologous transcripts to native genes transcripts. TPM was then calculated separately for only the native genes excluding the heterologous gene for calculating translation efficiencies and as input in the ICA algorithm. RNA-seq data is available in the Gene Expression Omnibus (GEO) database with accession number.

B.1.5 Independent Component Analysis

We combined the expression profiles generated in this study with a collection of 278 expression profiles previously generated in our research group. ICA was performed as described in [6]. Briefly, the expression compendium was centered, using the WT MG1655 Gly2 expression profile reported in this manuscript as the baseline condition. We executed FastICA 100 times with random seeds and a convergence tolerance of 10-7. We constrained the number of components in each iteration to the number of components that reconstruct 99% of the variance as calculated by principal component analysis. The resulting components were clustered using DB-

SCAN to identify robust independent components. I-modulons were extracted from independent components by iteratively removing genes with the largest absolute value and computing the kurtosis test statistic of the resulting distribution. Once the test statistic fell below a cutoff of 4 (identified through a sensitivity analysis), we designated the removed genes an i-modulon.

B.1.6 Model simulations and calculating simulated amino acid costs

We used a model for metabolism and gene expression to simulate the expression of each heterologous gene in *E. coli* [7]. Briefly, transcription and translation and dilution reactions were added for each heterologous gene to allow its synthesis. The model was then solved setting μ equal to the measured growth rate of each protein, optimizing for the flux through the dilution reaction which would symbolize the maximum dilution rate of the heterologous protein during replication of the cell. The solution to the optimization results in two vectors, the flux vector and the shadow price vector. Shadow prices represent the energetic cost of making each metabolite.

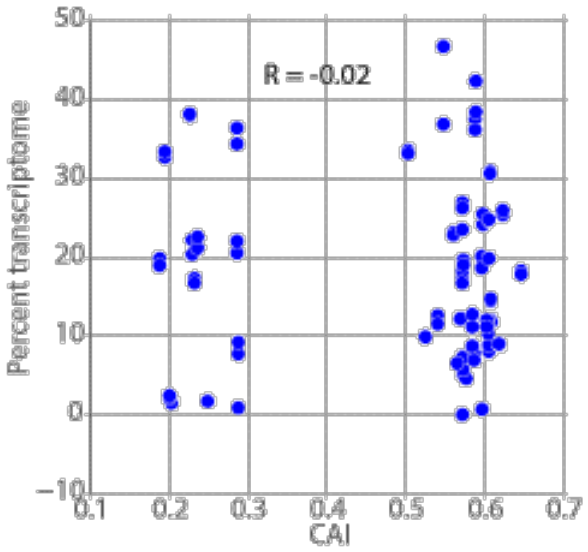

Figure B.1: Codon adaptation index plot against the expression level shows that expression level is not dependent on codon usage

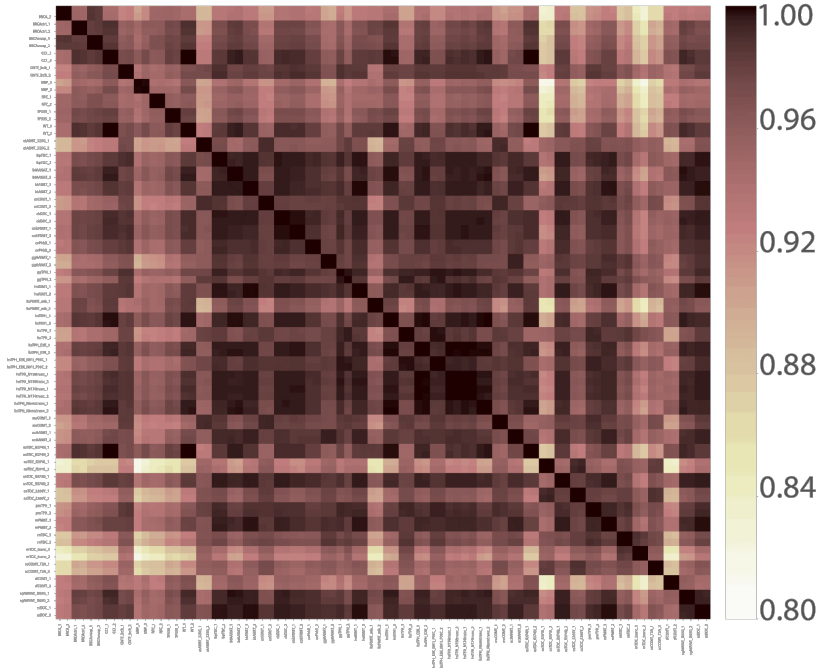

Figure B.2: Codon usage in the entire dataset is flexible across samples. Heatmap shows the correlation matrix of the normalized codon usage across datasets, showing that the transcriptome is not modulated to maintain codon preference.

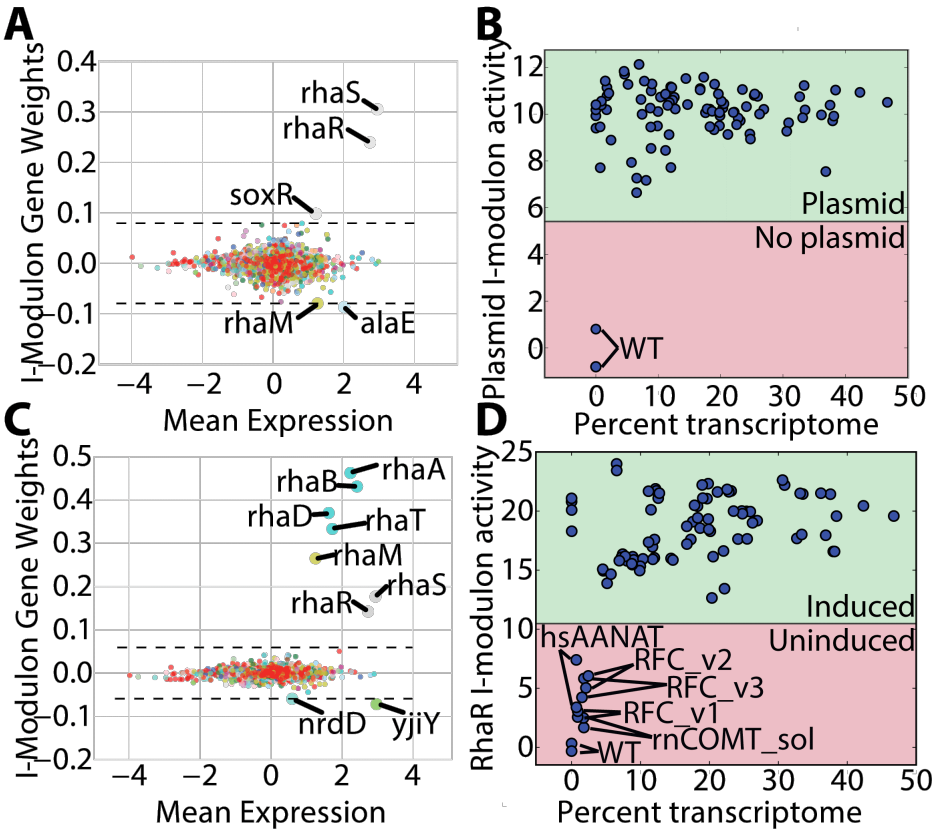

Figure B.3: ICA distinguishes between two closely related signals. A: An i-modulon showed high weighting in rhaS and rhaR, but not the rest of the rhamnose metabolism genes. The rhamnose inducible plasmid coded for both the rhaS and rhaR genes, hence this i-modulon probably represented the plasmid copy number. B: All samples show a high level of this i-modulon except wild type, which did not contain the plasmid. C: Genes which show significant weights in the RhaR i-modulon are members of the rhaSR-rhaBAD operon: rhaA, rhaB, rhaD, rhaT, rhaM, rhaS, rhaR, with nrdD and yjiY showing marginal significance. D: While most samples show a high activity in the RhaR i-modulon implying a successful induction (green), several samples show failure of induction (red).

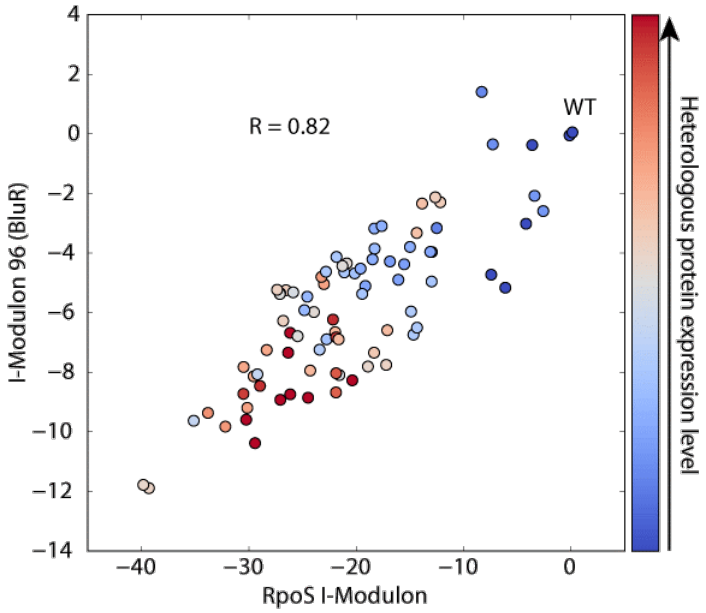

Figure B.4: RpoS and i-modulon 96 are highly correlated. Many genes in i-modulon 96 are regulated by both BluR and RpoS, suggesting a co-regulation under the conditions

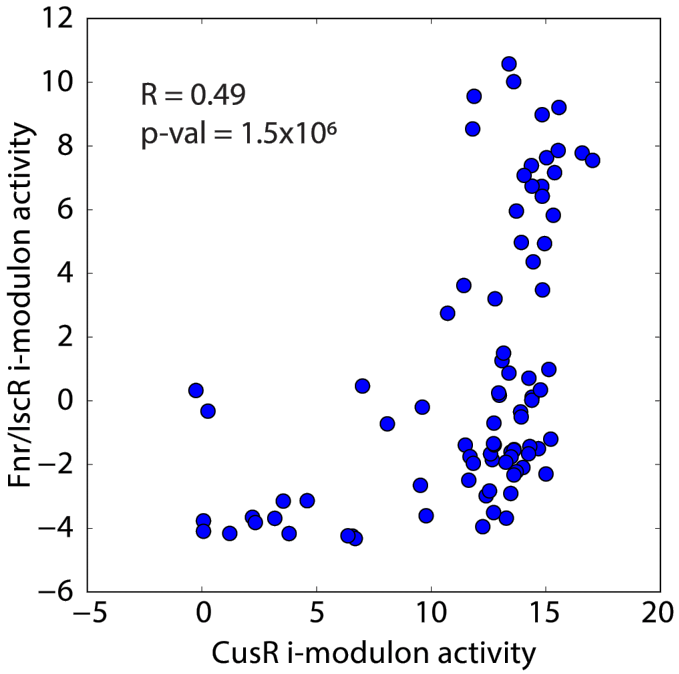

Figure B.5: CusR and Fnr/IscR I-modulon activities are correlated with each other, indicating a link between metal homeostasis and the aerobicity of the cell.

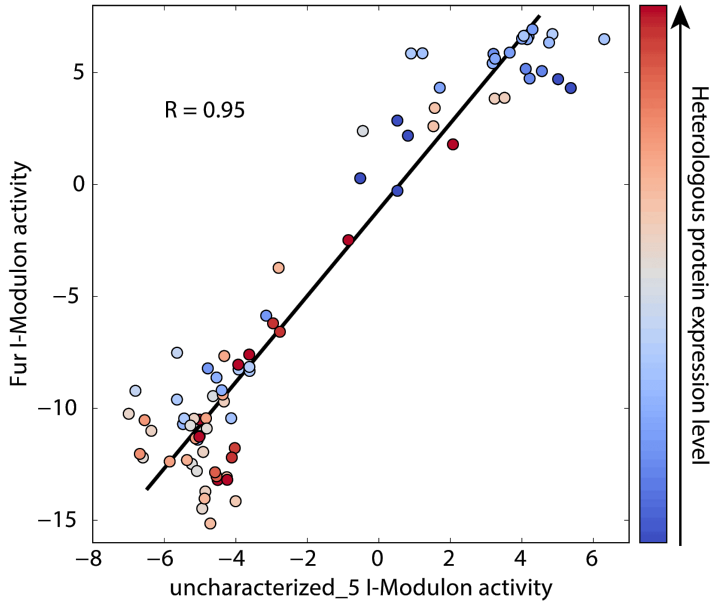

Figure B.6: Fur-1 i-modulon activity is strongly correlated with the activity of uncharacterized-5.

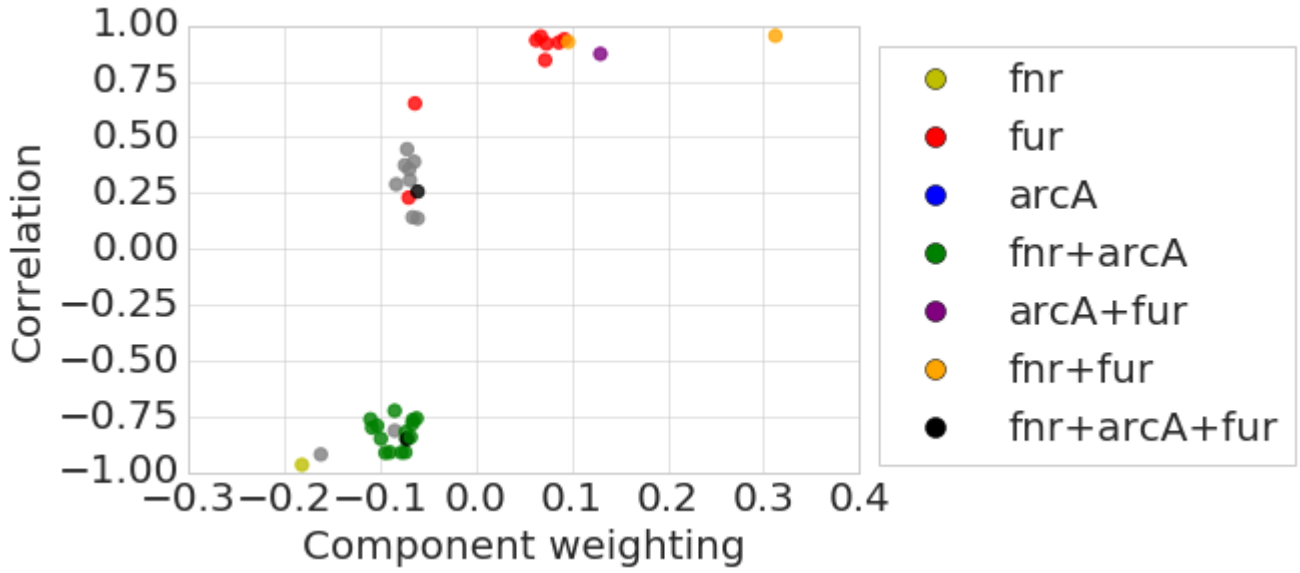

Figure B.7: Uncharacterized-5 is comprised of genes many of which are regulated by a combination of fnr, fur and arcA. We plot the correlation of each gene's expression with the activity of the uncharacterized-5 i-modulon to determine the main drivers of its activity within the dataset. The most highly correlated genes are those co-regulated by fur and either fnr or arcA, demonstrating a link between metal homeostasis and aerobicity within the dataset.

B.2 References

1. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Applied and environmental microbiology* **83.** ISSN: 0099-2240, 1098-5336. doi:10.1128/AEM.00410-17 (July 2017).

2. Savitsky, P., Bray, J., Cooper, C. D. O., Marsden, B. D., Mahajan, P., Burgess-Brown, N. A. & Gileadi, O. High-throughput production of human proteins for crystallization: the SGC experience. en. *Journal of structural biology* **172,** 3–13. ISSN: 1047-8477, 1095-8657 (Oct. 2010).

3. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q.-A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P. & Endy, D. Precise and reliable gene expression via standard transcription and translation initiation elements. en. *Nature methods* **10,** 354–360. ISSN: 1548-7091, 1548-7105 (Apr. 2013).

4. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature methods* **9,** 357–359. ISSN: 1548-7091, 1548-7105 (Mar. 2012).

5. Anders, S., Pyl, P. T. & Huber, W. *HTSeq: Analysing high-throughput sequencing data with Python* 2010.

6. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. *The Escherichia coli Transcriptome Consists of Independently Regulated Modules* en. Apr. 2019.

7. Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O'Brien, E. J., Liu, J. K. & Palsson, B. O. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. en. *PLoS computational biology* **14,** e1006302. ISSN: 1553-734X, 1553-7358 (July 2018).

Appendix C

Experimental evolution reveals the genetic basis and systems biology of superoxide stress tolerance - Supplementary Information

C.1 Methods

C.1.1 Strains

The initial strain used for the first phase of evolution was an MG1655 K-12 *E. coli* strain which had been evolved for optimal growth on glucose as a carbon source in M9 minimal media [1].

C.1.2 TALE

TALE was performed using a similar protocol to that in Mohamad et. al. 2017 [2]. Parallel cultures were started in M9 minimal medium by inoculation from isolated colonies. Evolution was performed in an automated platform with 15 mL working volume aerobic cultures maintained at 37°C and magnetically stirred at 1100 rpm. Growth was monitored by periodic measurement of the 600 nanometer optical density (OD600) on a Tecan Sunrise microplate reader, and cultures were passaged to fresh medium during exponential cell growth at an OD600 of approximately 0.3. Growth rates were determined for each batch of medium by linear regression of ln(OD600) versus time. At the time of passage, PQ concentration in the fresh medium batch was automatically increased if a specified growth rate had been met for several flasks. Samples were saved throughout the experiment by mixing equal parts culture and 50% v/v glycerol and storing at -80°C in glycerol. Mutation calling The breseq pipeline version 0.33.1 [3] was used to map the DNA-seq reads to an E. coli K12 MG1655 reference genome (NCBI accession NC_000913 version 3). DNA-seq quality control was accomplished using the software AfterQC version 0.9.7 [4].

C.1.3 Generation of aceE knockout

We used P1 phage transduction to transfer the *aceE* knockout from the Δ*aceE* strain in the Keio collection to WT [5]. Briefly, Δ*aceE* strain from the Keio collection was grown up and lysed using P1 phages. The lysate was filtered to remove cell debris leaving behind only P1 phages with packaged Δ*aceE* strain DNA. This was used to infect WT to transfect the DNA. Because the Δ*aceE* strain in the Keio collection contains a selective kanamycin marker at the gene lesion we were able to select for successful transfections.

C.1.4 Growth curves

End point strains were inoculated from overnight cultures into M9 minimal media with glucose as a carbon source (0.4% w/v) and allowed to grow to A600 OD 0.5. They were then diluted down to OD 0.01 with glucose minimal media containing different concentrations of paraquat. These were loaded onto a Bioscreen C set to measure OD every 30 minutes for 24 hours at 37C at high shaking. Comparison between WT and aceE knockout was performed in a similar fashion except with 10% LB added to the media to allow growth of the aceE knockout strain.

C.1.5 Culture conditions

WT, PQ1 and PQ2 were grown overnight in M9 minimal media with 0.4% w/v glucose as a carbon source. Fresh media was inoculated with the overnight culture to an initial OD of 0.025. Cultures were aerated with a stir bar in a water bath maintained at 37°C until OD reached 0.3. 50mM paraquat was added to a final concentration of 250uM in stressed condition flasks. After 20 minutes both stressed and unstressed conditions were harvested for ribosome profiling and transcriptomics.

C.1.6 Ribosome profiling

Ribosome profiling libraries were created using a modified version of the protocol outlined in Latif et. al. [6]. Differences from the published protocol are outlined below. In order to negate the possible confounding effects of addition of chloramphenicol to the media at harvest, cells were lysed by grinding in liquid nitrogen. 50mL of cells were harvested by centrifugation for 4 minutes at 37C in a 50mL conical tube containing 0.400g of sand, supernatant was aspirated quickly

and the cell pellet was flash frozen in liquid nitrogen. Pellets were then transferred into a liquid nitrogen cooled mortar and pestle, 500uL of lysis buffer was added and the pellet was pulverised to lyse the cells. Lysate was transferred to a falcon tube to thaw on ice and centrifuged and the supernatant whole cell lysate was isolated to continue with the published protocol [6]. Reads were sequenced on an Illumina HighSeq machine using a single end 50bp kit. Ribosome profiling reads had adapters removed using CutAdapt v1.8 (M. Martin 2011), then mapped to E. coli genome MG1655 using Bowtie v1.0.0 (Langmead 2010) and scored at the 3' end to generate ribosome density profiles for each gene.

C.1.7 Transcriptomics

Cells were pelleted and lysed with a modified version of the RNAProtect Bacteria Reagent protocol, ribosomal RNA was depleted using Ribo-Zero rRNA Removal Kit for Gram-Negative bacteria (Illumina), libraries were created using KAPA RNA Library Preparation kit. Deviations from the kit protocols are mentioned below. 3mL of induced culture was added to 6mL of RNAProtect Bacteria Reagent (Qiagen) and vortexed, then left at room temperature to incubate for 5 minutes. Cells were pelleted and then resuspended in 400uL elution buffer and then split into two tubes, with one kept as a spare. One pellet was then lysed enzymatically with the addition of lysozyme, proteinase-K and 20% SDS. SUPERase-In was added to maintain the integrity of the RNA. RNA isolation was then performed according to the rest of the kit protocol. rRNA was the depleted using the Ribo-Zero rRNA Removal Kit for Gram-Negative Bacteria according to the protocol, and libraries were constructed for paired-end sequencing using the KAPA RNA-Seq Library Preparation kit protocol. Reads were sequenced on the Illumina NextSeq platform.

Transcriptomic reads were mapped using Bowtie2 [7], and reads were counted using HT-

Seq [8]. Differential expression of genes was called using the DESeq2 [9] package in Bioconductor. Genes with a log2fold change greater than 1 and an FDR-adjusted p-value smaller than 0.1 were considered to be significantly differentially expressed between conditions. Raw read counts were normalized to transcripts per million (TPM) for further analysis. Sequencing data is available in the Gene Expression Omnibus (GEO) database with accession number.

C.1.8 Enrichment analysis for COG categories

Differentially expressed genes between pairs of conditions were annotated with their Clusters of Orthologous Genes (COG) categories. We then performed a hypergeometric test to test for enrichment of each COG category amongst the set of differentially regulated genes. The Bonferroni correction was used to adjust for the FDR, and an adjusted p-value below 0.01 was considered significantly enriched.

C.1.9 Cell motility assay

Overnight cultures of each strain were inoculated into M9 minimal media plates containing 0.4% w/v glucose and 0.25% w/v agar by inserting a pipette tip containing 1uL of culture about 3mm into the center of the plate and ejected as the tip was lifted up. Plates were incubated at 37C for 72 hours.

C.2 Supplementary Figures

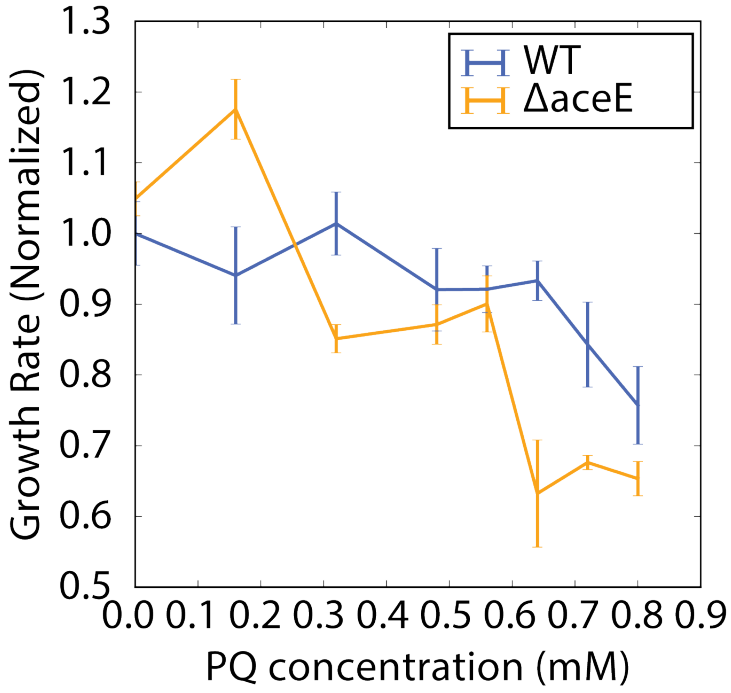

Figure C.1: The *aceE* knockout strain showed increased fitness over WT at low concentrations of paraquat but decreased fitness at higher concentrations. This growth curve was done with the addition of 10% w/v LB because the *aceE* knockout mutant is unable to grow in glucose minimal media. LB has been shown to greatly increase tolerance to oxidative stress due to the availability of amino acids in the media.

Figure C.2: Cell motility assay of WT and evolved strains shows increased cell motility in PQ1 and decreased cell motility in PQ2. PQ1 showed an up-regulation of cell motility related genes compared with wild type, whilst PQ2 showed a down-regulation of the same genes compared with wild type. We performed a cell motility assay on these strains and found that over 72 hours, PQ1 cells are indeed more motile than PQ2 and WT strains.

C.3 References

1. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Applied and environmental microbiology* **81,** 17–30. ISSN: 0099-2240, 1098-5336 (Jan. 2015).

2. Mohamed, E. T., Wang, S., Lennen, R. M., Herrgaard, M. J., Simmons, B. A., Singer, S. W. & Feist, A. M. Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution. *Microbial cell factories* **16,** 204. ISSN: 1475-2859 (Nov. 2017).

3. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in molecular biology* **1151,** 165–188. ISSN: 1064-3745, 1940-6029 (2014).

4. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M. & Gu, J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC bioinformatics* **18,** 80. ISSN: 1471-2105 (Mar. 2017).

5. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. en. *Molecular systems biology* **2,** 2006.0008. ISSN: 1744-4292, 1744-4292 (Jan. 2006).

6. Latif, H., Szubin, R., Tan, J., Brunk, E., Lechner, A., Zengler, K. & Palsson, B. O. A streamlined ribosome profiling protocol for the characterization of microorganisms. *BioTechniques* **58,** 329–332. ISSN: 0736-6205, 1940-9818 (2014).

7. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature methods* **9,** 357–359. ISSN: 1548-7091, 1548-7105 (Mar. 2012).

8. Anders, S., Pyl, P. T. & Huber, W. *HTSeq: Analysing high-throughput sequencing data with Python* 2010.

9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 550. ISSN: 1465-6906 (2014).

Appendix D

Datasets generated for the purpose of this dissertation

Table D.1: Datasets generated for the purpose of this dissertation

2	Glucose	Ribosome profiling	
2	Pyruvate	Ribosome profiling	
2	Fumarate	Ribosome profiling	
2	Acetate	Ribosome profiling	
3	atASMT_325G_1	RNASeq	GEO accession GSE133607
3	atASMT_325G_2	RNASeq	GEO accession GSE133607
3	bpTDC_1	RNASeq	GEO accession GSE133607
3	bpTDC_2	RNASeq	GEO accession GSE133607
3	BRCA_1	RNASeq	GEO accession GSE133607
3	BRCA_2	RNASeq	GEO accession GSE133607
3	BRCActrl_1	RNASeq	GEO accession GSE133607
3	BRCActrl_2	RNASeq	GEO accession GSE133607
3	BRCActrl_3	RNASeq	GEO accession GSE133607
3	BRCActrl_4	RNASeq	GEO accession GSE133607

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	BRCAswap_1	RNASeq	GEO accession GSE133607
3	BRCAswap_2	RNASeq	GEO accession GSE133607
3	btAANAT_1	RNASeq	GEO accession GSE133607
3	btAANAT_2	RNASeq	GEO accession GSE133607
3	btASMT_1	RNASeq	GEO accession GSE133607
3	btASMT_2	RNASeq	GEO accession GSE133607
3	CC1_3	RNASeq	GEO accession GSE133607
3	CC1_4	RNASeq	GEO accession GSE133607
3	CC1_5	RNASeq	GEO accession GSE133607
3	CC1_6	RNASeq	GEO accession GSE133607
3	CC1_1	RNASeq	GEO accession GSE133607
3	CC1_2	RNASeq	GEO accession GSE133607
3	ccCOMT_1	RNASeq	GEO accession GSE133607
3	ccCOMT_2	RNASeq	GEO accession GSE133607
3	ckDDC_1	RNASeq	GEO accession GSE133607
3	ckDDC_2	RNASeq	GEO accession GSE133607
3	CNTF_belt_1	RNASeq	GEO accession GSE133607
3	CNTF_belt_2	RNASeq	GEO accession GSE133607
3	cobPNMT_1	RNASeq	GEO accession GSE133607
3	cobPNMT_2	RNASeq	GEO accession GSE133607
3	cvPhhB_1	RNASeq	GEO accession GSE133607
3	cvPhhB_2	RNASeq	GEO accession GSE133607
3	ggAANAT_1	RNASeq	GEO accession GSE133607
3	ggAANAT_2	RNASeq	GEO accession GSE133607
3	ggTPH_1	RNASeq	GEO accession GSE133607
3	ggTPH_2	RNASeq	GEO accession GSE133607
3	hsAANAT_1	RNASeq	GEO accession GSE133607
3	hsAANAT_2	RNASeq	GEO accession GSE133607
3	hsASMT_1	RNASeq	GEO accession GSE133607

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	hsASMT_2	RNASeq	GEO accession GSE133607
3	hsPNMT_mb_1	RNASeq	GEO accession GSE133607
3	hsPNMT_mb_2	RNASeq	GEO accession GSE133607
3	hsTPH_1	RNASeq	GEO accession GSE133607
3	hsTPH_2	RNASeq	GEO accession GSE133607
3	hsTPH_E2K_1	RNASeq	GEO accession GSE133607
3	hsTPH_E2K_2	RNASeq	GEO accession GSE133607
3	hsTPH_E2K_N91I_P99C_1	RNASeq	GEO accession GSE133607
3	hsTPH_E2K_N91I_P99C_2	RNASeq	GEO accession GSE133607
3	hsTPH_N158trunc_1	RNASeq	GEO accession GSE133607
3	hsTPH_N158trunc_2	RNASeq	GEO accession GSE133607
3	hsTPH_N174trunc_1	RNASeq	GEO accession GSE133607
3	hsTPH_N174trunc_2	RNASeq	GEO accession GSE133607
3	hsTPH_Ntermtrunc_1	RNASeq	GEO accession GSE133607
3	hsTPH_Ntermtrunc_2	RNASeq	GEO accession GSE133607
3	hsTPH1_1	RNASeq	GEO accession GSE133607
3	hsTPH1_2	RNASeq	GEO accession GSE133607
3	MBP_1	RNASeq	GEO accession GSE133607
3	MBP_2	RNASeq	GEO accession GSE133607
3	msCOMT_1	RNASeq	GEO accession GSE133607
3	msCOMT_2	RNASeq	GEO accession GSE133607
3	ocAANAT_1	RNASeq	GEO accession GSE133607
3	ocAANAT_2	RNASeq	GEO accession GSE133607
3	osTDC_K374H_1	RNASeq	GEO accession GSE133607
3	osTDC_K374H_2	RNASeq	GEO accession GSE133607
3	osTDC_K374L_1	RNASeq	GEO accession GSE133607
3	osTDC_K374L_2	RNASeq	GEO accession GSE133607
3	osTDC_K374Q_1	RNASeq	GEO accession GSE133607
3	osTDC_K374Q_2	RNASeq	GEO accession GSE133607

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	osTDC_L360V_1	RNASeq	GEO accession GSE133607
3	osTDC_L360V_2	RNASeq	GEO accession GSE133607
3	pmTPH_1	RNASeq	GEO accession GSE133607
3	pmTPH_2	RNASeq	GEO accession GSE133607
3	RFC_1	RNASeq	GEO accession GSE133607
3	RFC_2	RNASeq	GEO accession GSE133607
3	RFCctrl_3	RNASeq	GEO accession GSE133607
3	RFCctrl_4	RNASeq	GEO accession GSE133607
3	RFCctrl_1	RNASeq	GEO accession GSE133607
3	RFCctrl_2	RNASeq	GEO accession GSE133607
3	RFCnopause_1	RNASeq	GEO accession GSE133607
3	RFCnopause_2	RNASeq	GEO accession GSE133607
3	RFCpluspause_1	RNASeq	GEO accession GSE133607
3	RFCpluspause_2	RNASeq	GEO accession GSE133607
3	rnCOMT_sol_1	RNASeq	GEO accession GSE133607
3	rnCOMT_sol_2	RNASeq	GEO accession GSE133607
3	rnPNMT_1	RNASeq	GEO accession GSE133607
3	rnPNMT_2	RNASeq	GEO accession GSE133607
3	rnTDC_2	RNASeq	GEO accession GSE133607
3	rnTDC_1	RNASeq	GEO accession GSE133607
3	rnTDC_trunc_1	RNASeq	GEO accession GSE133607
3	rnTDC_trunc_2	RNASeq	GEO accession GSE133607
3	saCOMT_T2A_1	RNASeq	GEO accession GSE133607
3	saCOMT_T2A_2	RNASeq	GEO accession GSE133607
3	sfCOMT_1	RNASeq	GEO accession GSE133607
3	sfCOMT_2	RNASeq	GEO accession GSE133607
3	sgAANAT_D63G_1	RNASeq	GEO accession GSE133607
3	sgAANAT_D63G_2	RNASeq	GEO accession GSE133607
3	ssDDC_1	RNASeq	GEO accession GSE133607

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	ssDDC_2	RNASeq	GEO accession GSE133607
3	TP53B_1	RNASeq	GEO accession GSE133607
3	TP53B_2	RNASeq	GEO accession GSE133607
3	WT_1	RNASeq	GEO accession GSE133607
3	WT_2	RNASeq	GEO accession GSE133607
3	atASMT_325G_1	Ribosome profiling	GEO accession GSE134324
3	atASMT_325G_2	Ribosome profiling	GEO accession GSE134324
3	bpTDC_1	Ribosome profiling	GEO accession GSE134324
3	bpTDC_2	Ribosome profiling	GEO accession GSE134324
3	BRCA_1	Ribosome profiling	GEO accession GSE134324
3	BRCA_2	Ribosome profiling	GEO accession GSE134324
3	BRCActrl_1	Ribosome profiling	GEO accession GSE134324
3	BRCActrl_2	Ribosome profiling	GEO accession GSE134324
3	BRCActrl_3	Ribosome profiling	GEO accession GSE134324
3	BRCActrl_4	Ribosome profiling	GEO accession GSE134324
3	BRCAswap_1	Ribosome profiling	GEO accession GSE134324
3	BRCAswap_2	Ribosome profiling	GEO accession GSE134324
3	btAANAT_1	Ribosome profiling	GEO accession GSE134324
3	btAANAT_2	Ribosome profiling	GEO accession GSE134324
3	btASMT_1	Ribosome profiling	GEO accession GSE134324
3	btASMT_2	Ribosome profiling	GEO accession GSE134324
3	CC1_3	Ribosome profiling	GEO accession GSE134324
3	CC1_4	Ribosome profiling	GEO accession GSE134324
3	CC1_5	Ribosome profiling	GEO accession GSE134324
3	CC1_6	Ribosome profiling	GEO accession GSE134324
3	CC1_1	Ribosome profiling	GEO accession GSE134324
3	CC1_2	Ribosome profiling	GEO accession GSE134324
3	ccCOMT_1	Ribosome profiling	GEO accession GSE134324
3	ccCOMT_2	Ribosome profiling	GEO accession GSE134324

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	ckDDC_1	Ribosome profiling	GEO accession GSE134324
3	ckDDC_2	Ribosome profiling	GEO accession GSE134324
3	CNTF_belt_1	Ribosome profiling	GEO accession GSE134324
3	CNTF_belt_2	Ribosome profiling	GEO accession GSE134324
3	cobPNMT_1	Ribosome profiling	GEO accession GSE134324
3	cobPNMT_2	Ribosome profiling	GEO accession GSE134324
3	cvPhhB_1	Ribosome profiling	GEO accession GSE134324
3	cvPhhB_2	Ribosome profiling	GEO accession GSE134324
3	ggAANAT_1	Ribosome profiling	GEO accession GSE134324
3	ggAANAT_2	Ribosome profiling	GEO accession GSE134324
3	ggTPH_1	Ribosome profiling	GEO accession GSE134324
3	ggTPH_2	Ribosome profiling	GEO accession GSE134324
3	hsAANAT_1	Ribosome profiling	GEO accession GSE134324
3	hsAANAT_2	Ribosome profiling	GEO accession GSE134324
3	hsASMT_1	Ribosome profiling	GEO accession GSE134324
3	hsASMT_2	Ribosome profiling	GEO accession GSE134324
3	hsPNMT_mb_1	Ribosome profiling	GEO accession GSE134324
3	hsPNMT_mb_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH_1	Ribosome profiling	GEO accession GSE134324
3	hsTPH_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH_E2K_1	Ribosome profiling	GEO accession GSE134324
3	hsTPH_E2K_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH_E2K_N91I_P99C_1	Ribosome profiling	GEO accession GSE134324
3	hsTPH_E2K_N91I_P99C_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH_N158trunc_1	Ribosome profiling	GEO accession GSE134324
3	hsTPH_N158trunc_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH_N174trunc_1	Ribosome profiling	GEO accession GSE134324
3	hsTPH_N174trunc_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH_Ntermtrunc_1	Ribosome profiling	GEO accession GSE134324

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	hsTPH_Ntermtrunc_2	Ribosome profiling	GEO accession GSE134324
3	hsTPH1_1	Ribosome profiling	GEO accession GSE134324
3	hsTPH1_2	Ribosome profiling	GEO accession GSE134324
3	MBP_1	Ribosome profiling	GEO accession GSE134324
3	MBP_2	Ribosome profiling	GEO accession GSE134324
3	msCOMT_1	Ribosome profiling	GEO accession GSE134324
3	msCOMT_2	Ribosome profiling	GEO accession GSE134324
3	ocAANAT_1	Ribosome profiling	GEO accession GSE134324
3	ocAANAT_2	Ribosome profiling	GEO accession GSE134324
3	osTDC_K374H_1	Ribosome profiling	GEO accession GSE134324
3	osTDC_K374H_2	Ribosome profiling	GEO accession GSE134324
3	osTDC_K374L_1	Ribosome profiling	GEO accession GSE134324
3	osTDC_K374L_2	Ribosome profiling	GEO accession GSE134324
3	osTDC_K374Q_1	Ribosome profiling	GEO accession GSE134324
3	osTDC_K374Q_2	Ribosome profiling	GEO accession GSE134324
3	osTDC_L360V_1	Ribosome profiling	GEO accession GSE134324
3	osTDC_L360V_2	Ribosome profiling	GEO accession GSE134324
3	pmTPH_1	Ribosome profiling	GEO accession GSE134324
3	pmTPH_2	Ribosome profiling	GEO accession GSE134324
3	RFC_1	Ribosome profiling	GEO accession GSE134324
3	RFC_2	Ribosome profiling	GEO accession GSE134324
3	RFCctrl_3	Ribosome profiling	GEO accession GSE134324
3	RFCctrl_4	Ribosome profiling	GEO accession GSE134324
3	RFCctrl_1	Ribosome profiling	GEO accession GSE134324
3	RFCctrl_2	Ribosome profiling	GEO accession GSE134324
3	RFCnopause_1	Ribosome profiling	GEO accession GSE134324
3	RFCnopause_2	Ribosome profiling	GEO accession GSE134324
3	RFCpluspause_1	Ribosome profiling	GEO accession GSE134324
3	RFCpluspause_2	Ribosome profiling	GEO accession GSE134324

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
3	rnCOMT_sol_1	Ribosome profiling	GEO accession GSE134324
3	rnCOMT_sol_2	Ribosome profiling	GEO accession GSE134324
3	rnPNMT_1	Ribosome profiling	GEO accession GSE134324
3	rnPNMT_2	Ribosome profiling	GEO accession GSE134324
3	rnTDC_2	Ribosome profiling	GEO accession GSE134324
3	rnTDC_1	Ribosome profiling	GEO accession GSE134324
3	rnTDC_trunc_1	Ribosome profiling	GEO accession GSE134324
3	rnTDC_trunc_2	Ribosome profiling	GEO accession GSE134324
3	saCOMT_T2A_1	Ribosome profiling	GEO accession GSE134324
3	saCOMT_T2A_2	Ribosome profiling	GEO accession GSE134324
3	sfCOMT_1	Ribosome profiling	GEO accession GSE134324
3	sfCOMT_2	Ribosome profiling	GEO accession GSE134324
3	sgAANAT_D63G_1	Ribosome profiling	GEO accession GSE134324
3	sgAANAT_D63G_2	Ribosome profiling	GEO accession GSE134324
3	ssDDC_1	Ribosome profiling	GEO accession GSE134324
3	ssDDC_2	Ribosome profiling	GEO accession GSE134324
3	TP53B_1	Ribosome profiling	GEO accession GSE134324
3	TP53B_2	Ribosome profiling	GEO accession GSE134324
3	WT_1	Ribosome profiling	GEO accession GSE134324
3	WT_2	Ribosome profiling	GEO accession GSE134324
4	ALE4A_0PQ	RNASeq	GEO accession GSE134256
4	ALE4B_0PQ	RNASeq	GEO accession GSE134256
4	ALE4A_PQ	RNASeq	GEO accession GSE134256
4	ALE4B_PQ	RNASeq	GEO accession GSE134256
4	18.24A_PQ	RNASeq	GEO accession GSE134256
4	18.24B_PQ	RNASeq	GEO accession GSE134256
4	16.28A_PQ	RNASeq	GEO accession GSE134256
4	16.28B_PQ	RNASeq	GEO accession GSE134256
4	18.36A_0PQ	RNASeq	GEO accession GSE134256

Table D.1: Datasets generated for the purpose of this dissertation, Continued

Chapter	Sample	Data Type	Data availability
4	18.36B_0PQ	RNASeq	GEO accession GSE134256
4	18.36A_PQ	RNASeq	GEO accession GSE134256
4	18.36B_PQ	RNASeq	GEO accession GSE134256
4	16.32A_0PQ	RNASeq	GEO accession GSE134256
4	16.32B_0PQ	RNASeq	GEO accession GSE134256
4	16.32A_PQ	RNASeq	GEO accession GSE134256
4	16.32B_PQ	RNASeq	GEO accession GSE134256
4	ALE4A_0PQ	Ribosome profiling	GEO accession GSE134256
4	ALE4B_0PQ	Ribosome profiling	GEO accession GSE134256
4	ALE4A_PQ	Ribosome profiling	GEO accession GSE134256
4	ALE4B_PQ	Ribosome profiling	GEO accession GSE134256
4	18.24A_PQ	Ribosome profiling	GEO accession GSE134256
4	18.24B_PQ	Ribosome profiling	GEO accession GSE134256
4	16.28A_PQ	Ribosome profiling	GEO accession GSE134256
4	16.28B_PQ	Ribosome profiling	GEO accession GSE134256
4	18.36A_0PQ	Ribosome profiling	GEO accession GSE134256
4	18.36B_0PQ	Ribosome profiling	GEO accession GSE134256
4	18.36A_PQ	Ribosome profiling	GEO accession GSE134256
4	18.36B_PQ	Ribosome profiling	GEO accession GSE134256
4	16.32A_0PQ	Ribosome profiling	GEO accession GSE134256
4	16.32B_0PQ	Ribosome profiling	GEO accession GSE134256
4	16.32A_PQ	Ribosome profiling	GEO accession GSE134256
4	16.32B_PQ	Ribosome profiling	GEO accession GSE134256