UNIVERSITY OF CALIFORNIA, SAN DIEGO

# Model-driven discovery of adaptive mechanisms and underground metabolism in *Escherichia coli*

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

in

Bioengineering

by

Gabriela Guzmán Lopez Aguado

Committee in charge:

Marcos Intaglietta, Chair
Eric E. Allen
Adam M. Feist
Christian M. Metallo
Victor F. Nizet
Milton H. Saier

2018

The Dissertation of Gabriela Guzmán Lopez Aguado is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2018

DEDICATION

For my mom,

for her constant love and support,

for instilling within me a determination and passion for knowledge

# EPIGRAPH

*Science and everyday life cannot and should not be separated.*

—Rosalind Franklin

*If you're too sloppy, then you never get reproducible results, and then you never can draw any conclusions; but if you are just a little sloppy, then when you see something startling, you [can] nail it down ... I called it the "Principle of Limited Sloppiness".*

—Max Delbrück

*If you don't like bacteria, you're on the wrong planet.*

—Stewart Brand

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# ACKNOWLEDGEMENTS

No PhD is an island; every PhD is a piece of an academic community, the result of hours of collaboration and mentorship. Over the course of the past six years, I have been lucky enough to learn from members of the Systems Biology Research Group (SBRG). They have helped shape me as a scientist and I am very grateful for their knowledge and patience. I would first like to thank the many graduate students, post-docs, and staff members who overlapped with my stay here at the SBRG. To Zachary King, thank you for always being so engaging with me about my projects and helping answer all of my coding questions. You helped me see the elegance of coding and helped me to appreciate the challenge even when I was getting ready to pull out my hair and give up. Thank you to Jose Utrilla Carreri, for being such a wonderful and patient experimental mentor, for teaching me how to do my first knockout experiments, and for always being available to answer questions and discuss science. Thank you Elizabeth Brunk for introducing me to protein structures, for always giving me such critical and helpful feedback on my writing, and for introducing me to the folks up at JBEI (Joint BioEnergy Institute). To the ALE team members, Ryan LaCroix, Troy Sandberg, and Joon Ho Park, thanks so much for helping me run so many of my experiments on the ALE machines and for being good friends. To my first experimental mentor, Haythem Latif, thank you for teaching me the ropes of the wet lab. To Aarash Bordbar, thanks for trusting a first year PhD student to help run experiments during my rotation project. Thanks to Ali Ebrahim, for always being so kind and optimistic about science and available to answer questions about *cobrapy* and sequencing processing pipelines. Thanks to Patrick Phaneuf for being so helpful with DNAseq processing; ale-analytics has been such an incredibly useful tool. Thank

you to Joanne Liu, Donghyuk Kim, Ye Gao, and Amitesh Anand, for being the best officemates! Another very special thank you to Richard Szubin and Ying Hefner for your positive attitudes and for being so helpful in constructing sequencing libraries (this PhD would have taken twice as long if it weren't for you!). Thanks to Marc Abrams and Helder Balelo for always being available to help and answer administrative questions. Thank you to Jan Lenington for being such an awesome administrator, resource, and advocate for so many graduate students. Finally thanks to Jon Monk, Teddy Obrien, Josh Lerman, Ana Moreno, Ke Chen, Jenni Levering, Daniel Zielinski, Nathan Mih, Connor Olson, Andreas Dräger, Nate Lewis, and everyone else who has been part of the group over the past six years. It has been such a pleasure to work with you!

I would like to thank the many collaborators outside of the SBRG and UC San Diego who helped to enrich my work and experience. I would first like to thank Balázs Papp at the Biological Research Centre of the Hungarian Academy of Sciences for being so helpful in providing extensive feedback and contributing to the our work on enzyme promiscuity and adaptation to non-native carbon sources presented in Chapter 4. Furthermore, thanks to Richard Notebaart at the Laboratory of Food Microbiology at the Wageningen University and Research as well as Csaba Pál, and Ákos Nyerges for their contributions to our research on enzyme promiscuity and evolution described in Chapter 4. I would also like to thank Sergey Nurk at the St. Petersburg Academic University for his contribution to analysis of genome amplifications in our samples related to Chapter 2. Thank you to Markus de Raad and Trent Northen at the Environmental Genomics and Systems Biology Division at Lawrence Berkeley National Laboratory for assisting with enzyme activity characterizations used in the work

presented in Chapter 4. I would also like to thank Anna Lechner and Jay Keasling for mentoring me in the purification of enzymes and *in vitro* characterizations and hosting me for 8 weeks at the Joint BioEnergy Institute (JBEI). Finally I would like to thank Rebecca Lennen and Elsayed Mohamed at the Novo Nordisk Foundation Center for Biosustainability.

To Dr. Adam Feist, thank you for giving me the opportunity to join the SBRG and for mentoring me over the course of my PhD. It is difficult to list the various ways in which you have helped me, because they are so many. In a lab as large as the SBRG, graduate students often look to post-docs and project scientists for the day-by-day questions, and you have been an incredibly helpful resource. First I'd like to thank you for introducing me to adaptive laboratory evolution and getting me excited about model-driven discovery. Thank you for being patient with me as I worked on my first, first-author publication. Thank you for encouraging me to share our research at various conferences. It has been an honor and a pleasure to work with you and I hope to work with you again in the future. On a personal note, thank you for being a friend and a great guy to share a beer with.

My dissertation work was made possible by the Novo Nordisk Foundation (Grant Number NNF10CC1016517). I would also like to thank the Siebel Foundation and the ARCS (Achievement Rewards for College Students) Foundation for providing me with extra support this last year of my PhD. It has been a wonderful networking opportunity to receive these recognitions as well as a great source of financial support for myself and my family.

Last, but not least, I need to save the biggest thank you for my amazing family. I could not have done this without your love and support! To my talented husband,

Zak, thank you for being there for me every day and bringing so much joy to my life. To my beautiful daughter, Elise, you always light up my day and I have been delighted to watch you discover the world over the past two years. To my mom, there are no words to describe how grateful I am for your unconditional love and support. You have always been there for me and in these past two years you have also been there for Elise when I needed to really bunker down and finish this PhD. I am so lucky to have had such a strong and intelligent role model in my mom, who from a very young age taught me to be an honest, hard worker, and to always stand up for what I think is important. Thanks also of course to my dad for teaching me to value a good education and to strive for success and to be an 'alpha woman' as he likes to say. Another big thank you to my siblings for their continual love and encouragement. This was all possible thanks to this amazing support system!

Chapter 2 is a reprint of a published manuscript: Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O., and Feist, A. M. (2015). "Model-driven discovery of underground metabolic functions in Escherichia coli". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.3, pp. 929–934. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 3 is a version of a manuscript under review at *BMC Systems Biology*: Guzmán, G. I., Olson, C. A., Hefner, Y., Phaneuf, P., Catoiu, E., Crepaldi, L. B., Goldschmidt Micas, L., Palsson, B. O., Feist, A. M. (2017) "Reframing essentiality in terms of adaptive flexibility". The dissertation author was the primary author of the manuscript and was responsible for the research.

Chapter 4 is a version of a manuscript in preparation for submission: Guzmán, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, A., Papp, H., de Raad, M., King, Z.

A., Northen, T. R., Notebaart, R. A., Pál, C., Palsson, B. O., Papp, B., Feist, A. M. (2017) "Enzyme promiscuity shapes evolutionary innovation and optimization". The dissertation author was the primary author of the paper and was responsible for the research.

## VITA

| | |
|---|---|
| 2010 | B. S. in Bioengineering, University of California, Berkeley |
| 2018 | Ph. D. in Bioengineering, University of California, San Diego |

## PUBLICATIONS

LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O., and Feist, A. M. (2015). "Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium". en. In: *Appl. Environ. Microbiol.* 81.1, pp. 17–30.

Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O., and Feist, A. M. (2015). "Model-driven discovery of underground metabolic functions in Escherichia coli". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.3, pp. 929–934.

Kim, D., Seo, S. W., Gao, Y., Nam, H., Guzman, G. I., Cho, B.-K., and Palsson, B. O. (2018). "Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP". en. In: *Nucleic Acids Res.*.

Guzmán, G. I., Olson, C. A., Hefner, Y., Phaneuf, P., Catoiu, E., Crepaldi, L. B., Goldschmidt Micas, L., Palsson, B. O., Feist, A. M. (2017) "Reframing essentiality in terms of adaptive flexibility". Manuscript under review at *BMC Systems Biology*.

Guzmán, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, A., Papp, H., de Raad, M., King, Z. A., Northen, T. R., Notebaart, R. A., Pál, C., Palsson, B. O., Papp, B., Feist, A. M. (2017) "Enzyme promiscuity shapes evolutionary innovation and optimization". Manuscript in preparation for submission.

ABSTRACT OF THE DISSERTATION

**Model-driven discovery of adaptive mechanisms and underground metabolism in *Escherichia coli***

by

Gabriela Guzmán Lopez Aguado

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2018

Marcos Intaglietta, Chair

Evidence suggests that novel enzyme functions evolved from low-level promiscuous activities in ancestral enzymes. Yet, the evolutionary dynamics and physiological mechanisms of how such side activities contribute to systems-level adaptations are poorly understood. Furthermore, it remains untested whether knowledge of an organism's promiscuous reaction set ('underground metabolism') can aid in forecasting the genetic basis of metabolic adaptations. In this dissertation, novel approaches toward exploring promiscuity in the space of a metabolic network are described. The work leverages genome-scale models, which have been widely used for predicting

growth phenotypes in various nutrient environments and following genetic perturbation in *Escherichia coli*. Failure modes of model predictions in relation to gene essentiality are explored as opportunities for targeting biological discovery, suggesting the presence of unknown underground pathways stemming from enzymatic cross-reactivity or suggesting limitations of experimental conditions stemming from short growth tests. Workflows are presented that couple constraint-based modeling and bioinformatic tools with knockout strain analysis and long-term growth experiments for the purpose of enhancing knowledge and predictability of enzyme promiscuity at the genome scale. Furthermore, a computational model of underground metabolism and laboratory evolution experiments are employed to examine the role of enzyme promiscuity in the acquisition and optimization of growth on predicted non-native substrates. Promiscuous enzyme activities played key roles in multiple phases of adaptation. Genes underlying the phenotypic innovations were accurately predicted by genome-scale model simulations of metabolism with enzyme promiscuity. Thus, it is shown that computational approaches will be essential to synthesize the complex role of promiscuous activities in models of evolutionary adaptation.

# Chapter 1

# Introduction

Organisms have evolved to take advantage of their environments. Our underlying biochemical building blocks are believed to drive this adaptability. What are the detailed biological mechanisms that lead to observable phenotypic changes? How do incremental adaptations to novel surroundings occur? In the age of whole genome sequencing, we can begin to answer these questions at the genetic and molecular level. The bacterium *Escherichia coli* is an ideal organism for studying questions of evolution and adaptation as it has a small genome and a short lifespan allowing us to track evolution in the laboratory for thousands of generations. By probing this organism's ability to respond to environmental and genetic perturbations in the laboratory, we can come closer to fully annotating its genetic components and better understand the genotype - phenotype relationship of life.

Beyond understanding the biological components of life, a major scientific goal is to be able to predict metabolic phenotypes and evolutionary trajectories. The relatively new field of systems biology provides the tools necessary to integrate our knowledge of the genetic and molecular components of life. Genome-scale models

synthesize our knowledge of genes, proteins, and reactions into whole-cell pictures of metabolism (Bordbar et al. 2014; McCloskey, Palsson, and Feist 2013). Upon these metabolic network representations of a cell, we can apply constraints based on growth conditions and we are then able to make predictions about the physiological states observed in real life. Although much progress has been made in the field of genome-scale modeling, there remain gaps in our knowledge of metabolism which lead to errors in predictions (Orth and Palsson 2012).

The work presented in this dissertation aims to resolve the inconsistencies found between model predictions and experimental observations and thereby improve the biological knowledge base for the model organism of *E. coli*. Beyond resolving modeling errors, current knowledge is utilized to make predictions about attainable phenotypes via adaptive laboratory evolution. I will start by introducing the key topics and tools related to adaptive evolution and systems biology that are fundamental to the work presented in this dissertation.

## 1.1 A Systems Biology Approach to Discovery

The following dissertation is composed of three main chapters that seek to expand our knowledge of metabolism in *E. coli*. The advent of genome-scale biology with next-generation sequencing technologies has resulted in more complete genome annotations and genome-scale reconstructions of metabolism. Genome-scale modeling relies upon well curated knowledge of enzyme reactivity; however, even for model organisms like *E. coli* genome-protein-reaction associations are incomplete. There are gaps in our knowledge of metabolism and this can lead to modeling errors in which *in silico* computations and *in vivo* observations do not coincide (Orth and Palsson

2012).

Gene essentiality is often utilized as a clear test of the predictive power of models (Figure 1.1). Experimental data sets related to gene essentiality have become widely available for various organisms offering genome-wide metabolic phenotypes to be predicted and tested (Baba et al. 2006; Christen et al. 2011). By placing specific growth constraints on a genome-scale model, including the systematic removal of individual genes to simulate a knockout strain, one can perform flux balance analysis to make growth or no growth predictions (Orth, Thiele, and Palsson 2010). These predictions may then be compared to experimental screens of gene essentiality. Those instances where a model predicts a gene to be essential that has been shown to be non-essential experimentally are termed **false negative** predictions. These inconsistencies can occur due to missing information in the genome-scale model. For instance, there may exist alternate isozymes or pathways capable of supporting growth despite the gene knockout. Conversely, instances where a model predicts a gene to be non-essential that has shown to be essential experimentally are termed **false positive** predictions. These inconsistencies can be the result of incorrect knowledge being included in the genome-scale model. It is possible that un-realistic reactions may have been added to a metabolic network reconstruction based on low confidence evidence. On the other hand, false positive predictions could also occur due to errors or limitations of essentiality screens. High-throughput essentiality screens have previously been conducted in microtiter plate formats and conducted over the period of 24 or 48 hours. These conditions may lead to false calls of essentiality for strains requiring longer incubation periods. Thus, false negative and false positive predictions of essentiality may be used to identify areas for further study and enable the systematic filling of

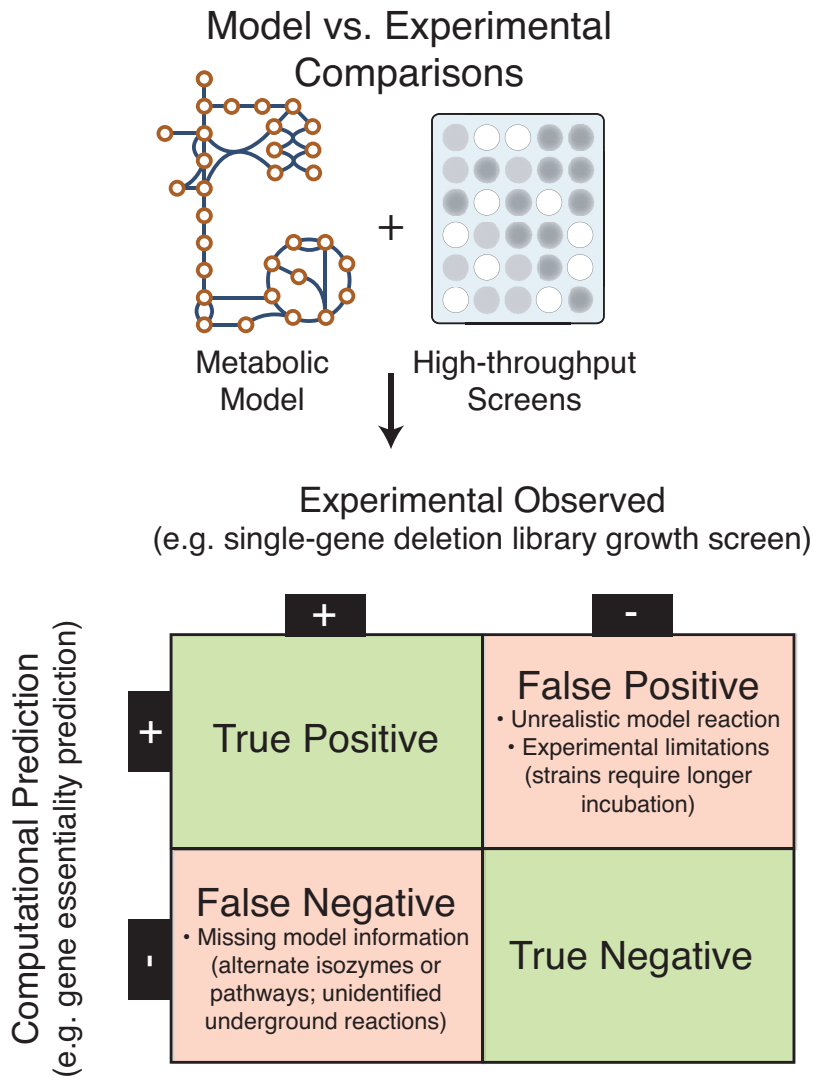


**Figure 1.1**: **Computational and experimental comparisons identify knowledge gaps.** Two modeling inconsistencies may be identified by comparing computational predictions of growth/no growth to experimental observations of growth/no growth, namely false negative and false positive predictions. False negative predictions occur when the model predicts no growth, but experiments show growth. Conversely, false positive predictions occur when the model predicts growth, but experiments show no growth.

knowledge gaps.

Furthermore, constraint-based modeling methods make predictions about the optimized state of a cell, implying an evolved phenotype. Provided a short adaptation period, the phenotype of a cell could change to match the predicted state. In this sense, it is possible that models may be used to predict evolutionary outcomes (Papp, Notebaart, and Pál 2011; Notebaart et al. 2014; Lässig, Mustonen, and Walczak 2017). However, not much work has been done to test whether constraint-based approaches can accurately predict the adaptive mutations that may occur *in vivo*. Much may be learned about metabolic capabilities by conducting adaptive evolution experiments, which directly feed into the knowledge contained in genome-scale reconstructions of metabolism and allows for iterative model-driven discovery processes (Figure 1.2). In the following section, we shall explore aspects of adaptive evolution in *E. coli*, which played a dominant role in the discovery efforts of this dissertation.

## 1.2 Adaptive Evolution in *Escherichia coli*

One underlying, fundamental gap in our knowledge of metabolism, which is a common theme found in the chapters of this dissertation, is an underestimation of the role of enzyme promiscuity in cell growth and adaptation. All enzymes possess the ability to use alternate substrates to a varying degree. The evolution of new gene functions is believed to be driven by the molecular infidelity of enzymes. Enzymes display flexibility in terms of substrate specificity and catalytic promiscuity. This enzymatic feature of promiscuity has been observed in a limited number of laboratory experiments (Jensen 1976; Khersonsky and Tawfik 2010; Loo et al. 2010; Notebaart et al. 2014); however, a larger 'underground' network of reactions may occur within

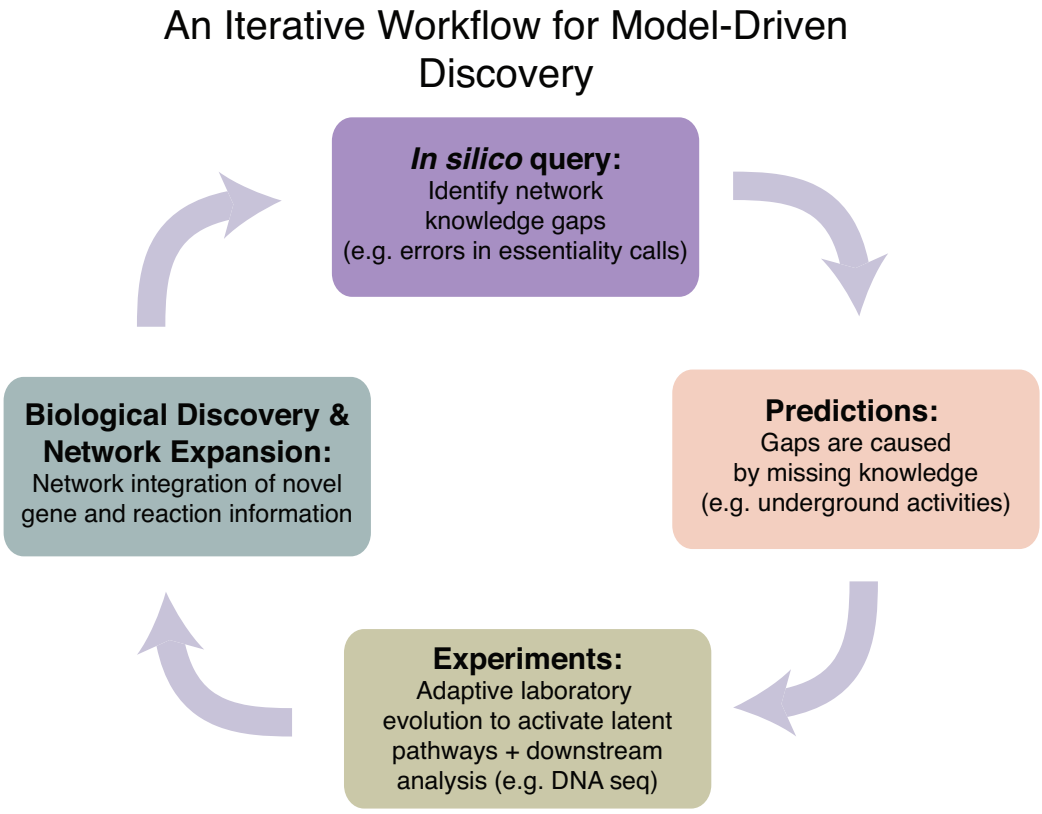


**Figure 1.2**: **An iterative workflow for model-driven discovery.** Metabolic model network knowledge gaps are identified by comparing *in silico* queries to experimental observations. This leads to predictions about the source of the knowledge gap, possibly missing isozyme or alternate pathway information or knowledge of an underground activity. These predictions are tested experimentally, possibly utilizing adaptive laboratory evolution experiments and genome resequencing and mutation analysis to test the hypothesis. Finally the new biology discovered from experiments is incorporated into the metabolic network and the loop is repeated.

a cell below the level of detection. It is not until a cell's metabolic capabilities are probed that these novel functions may come to light.

The term *underground metabolism* has been defined in the context of evolutionary biology as "reactions catalyzed by normal (unmutated) enzymes acting on substrate analogues which are themselves endogenous metabolites" (D'Ari and Casadesús 1998). It is believed that enzyme promiscuity has allowed for the adaptation of organisms to novel environments and it is those secondary, underground reactions that are normally not detected that drives the evolution of new gene functions. Studies of enzyme families and superfamilies are reflective of the role of enzyme promiscuity in the evolution of new, more specialized functions (Khersonsky and Tawfik 2010). The specific mechanistic route to enzyme specialization from a more broad-specificity ancestor, however, has not been as well characterized *in vivo*. It is believed, however, that new gene functions stemming from secondary activities of a pre-cursor enzyme must be selected for by providing an immediate advantage in a selecting growth environment, and specialization may occur by a small number of mutation events (Khersonsky and Tawfik 2010). By evolving the model organism *E. coli* in the laboratory, we can begin to examine these mechanisms at the genetic level and put some of the previously suggested theories to test. Current whole-genome sequencing technologies have enabled the rapid generation of large mutational datasets, that provide a picture of the evolutionary trajectories to an optimized endpoint. The information gleaned from these experiments allow us to make progress towards elucidating the extent of underground activities and metabolic plasticity of an organism.

Beyond single nucleotide polymorphisms (SNPs), and small insertion, and deletion mutation events that occur during adaptive evolutions, a mechanism of

adaptation that commonly occurs in growing bacteria is large regions of genome amplification. Genome duplication amplifications (GDAs) are not commonly called by sequencing analysis pipelines; however, they appear to commonly play an important role in selecting for and magnifying advantageous underground activities. GDA is believed to occur to some extent in approximately 10% of cells in a non-selective growth medium (Andersson and Hughes 2009). When a population of bacteria is placed in a selective growth environment, beneficial over-expression of a gene with advantageous underground reactivity may be selected for, and the population will reflect this by having a higher frequency of GDAs in the region conferring the benefit. Upon removal of the selection pressure or following acquirement of beneficial mutations and novel gene functionalities, it is believed that genome duplication amplifications are lost (Andersson and Hughes 2009; Bergthorsson, Andersson, and Roth 2007). Thus, GDAs and underground reactivity can play integral roles in evolutionary mechanisms. The results presented in this dissertation are reflective of the importance and prevalence of GDAs in the adaptation of *E. coli* to genetic perturbations and non-native growth environments.

## 1.3 Introducing the Thesis

In the chapters that follow, methods for probing promiscuous activities at the genome-scale are presented. Genome-scale reconstructions of metabolic networks are combined with gene knockout analysis, adaptive laboratory evolution, and next-generation sequencing techniques such as DNAseq and RNAseq to probe the largely uncharacterized space of underground metabolism. In doing so, cellular mechanisms of regulation and adaptation are explored. The methods presented, which can be

extended to other organisms, become increasingly important when designing drugs targeting pathogenic bacteria or engineering enzymes and bacteria for biotechnology applications.

In Chapter 2, a model-driven workflow for discovering underground metabolic functions in *E. coli* is proposed. The method aims to resolve inconsistencies between modeling predictions of gene essentiality and experimental observations. The workflow identifies potential targets for analysis with flux balance analysis (FBA) gene essentiality simulations in *E. coli* utilizing the *i*JO1366 metabolic reconstruction. Putative isozyme targets are identified by using sequence homology. It is predicted that isozymes are up-regulated in the strain where the primary enzyme is knocked out and this is explored by performing qPCR. If the putative isozyme is up-regulated, multi-knockout strains are constructed in the hopes of finding a synthetic lethal interaction and concluding the workflow. In Chapter 3, previously reported false positive predictions of gene essentiality on defined minimal medium for *E. coli* are explored. Of the twenty false positive strains available in the Keio gene knockout collection, eleven strains are shown to grow with longer incubation periods. The strains that grew reproducibly showed lag phases ranging from less than one day to more than 7 days. Whole genome sequencing of the populations reveal that more than half of the strains that grew acquired mutations. Comparison of mutations and model predictions of alternate pathways/isozymes demonstrated agreement for many of the cases analyzed. It is demonstrated that longer-term growth experiments followed by whole genome sequencing can provide a better understanding of gene essentiality as well as elucidate adaptative mechanisms that occur during these growth screens. In Chapter 4, a computational model of underground metabolism and laboratory

evolution experiments are employed to examine the role of enzyme promiscuity in the acquisition and optimization of growth on predicted non-native substrates in *E. coli* K-12 MG1655. Promiscuous enzyme activities are shown to play key roles in multiple phases of adaptation. Altered promiscuous activities not only established novel high-efficiency pathways, but also suppressed undesirable metabolic routes. Genes underlying the phenotypic innovations were accurately predicted by genome-scale model simulations of metabolism with enzyme promiscuity. Model-driven methods like those presented in this dissertation have the advantage of being driven by a top-down, systems analysis in the context of whole cell metabolism. Such work holds promise for advancement of fields of metabolic engineering and pharmacology that are continuously adapting enzymes and metabolic pathways for desired phenotypes.

# Chapter 2

# Model-driven discovery of underground metabolic functions in *Escherichia coli*

## 2.1 Introduction

The notion that enzymes are highly specialized to carry out a single function is often untrue. It has been demonstrated that many enzymes exhibit flexibility, or promiscuity, in regards to what substrates their catalytic pockets recognize. This lack of substrate specificity can lead to accuracy-rate tradeoffs that may affect evolutionary trajectories (Tawfik 2014). How has enzyme promiscuity shaped the evolution and divergence of organisms? The 'patchwork' model theorizes that primitive enzymes possessed a high degree of substrate promiscuity because it conferred a greater degree of catalytic versatility when the pool of available enzymes was limited (Jensen 1976; Lazcano and Miller 1996; Rison and Thornton 2002; Khersonsky and Tawfik 2010).

The existence of promiscuous proteins further serves as a starting point for evolving new functions, allowing for novel adaptations. Thus, organisms may exhibit latent, underground metabolic pathways that form the basis of their capacity to adapt to changing environments (D'Ari and Casadesús 1998; Nam et al. 2012; Notebaart et al. 2014). Substrate promiscuity, also referred to as 'moonlighting activity' and 'cross-reactivity', has thus been studied in terms of evolution and ties have been made between enzymes and their superfamilies (Furnham, Beer, and Thornton 2012). How novel enzyme functions arise within superfamilies is thus examined, and provides a basis for predicting promiscuous behavior among these protein families. However, defining targets for studies of promiscuity outside of these families and on a larger scale can become quite challenging.

Enzyme promiscuity has become widely accepted and examined on the enzyme level from a biochemical standpoint (Loo et al. 2010). These detailed biochemical studies provide an *in vitro* view of enzyme promiscuity and may be extended to reflect the promiscuity of other proteins based on sequence homology or enzyme familial relationships. In the present study, this task is approached from a different perspective by taking advantage of *in vivo* experimental techniques in order to gain insight into activities that are more physiologically relevant. In this way, as has been demonstrated in other *in vivo* studies, many of the challenges associated with removing enzymes from their native environment are circumvented (Notebaart et al. 2014). Specifically, the present study focuses on the examination of the regulatory and evolutionary capacity of a cell *in vivo*. Theories regarding genome duplications have suggested that an enzyme with a side activity that is selected for may be enhanced via gene duplication followed by mutation accumulation (Andersson and Hughes 2009). Thus, laboratory

evolutions may provide insight into these evolutionary mechanisms involving enzyme promiscuity. Furthermore, exploration of an underground metabolic network that takes advantage of enzyme cross-reactivity through native regulatory adaptations is best examined in the context of a whole cell (D'Ari and Casadesús 1998; Notebaart et al. 2014).

A top-down, model-driven approach coupled with *in vivo* experimentation to explore enzyme promiscuity could provide new insights into the physiological role of underground metabolism and complement the current approaches to enzyme research. Computational predictions of gsne essentiality are a commonly utilized application of genome-scale models and constraint-based modeling (McCloskey and Palsson 2013; Bordbar et al. 2014). When these models fail to predict gene essentiality, it signifies a missing link in our knowledge of metabolism and provides targets for further exploration (Orth and Palsson 2012). Various computational algorithms – SMILEY, GrowMatch – have been published with the intent of reconciling such knowledge gaps (Reed et al. 2006; Kumar and Maranas 2009). The following is a proof-of-principle study that demonstrates the advantages of a workflow for examining promiscuity at the genome-scale that also encompasses an adaptive laboratory evolution (ALE) framework. Three cases are explored to illustrate the capabilities of such a targeted, top-down approach to uncover the underground, latent activities of enzymes that reconcile gaps in our knowledge of metabolism.

## 2.2 Results and Discussion

### 2.2.1 Developing a model-driven workflow for isozyme discovery

The results from this study demonstrated that a top-down, systems approach could be used to drive the discovery of enzyme substrate promiscuity by using three genes, *aspC*, *argD*, and *gltA*, that were incorrectly identified to be essential as inputs. The isozyme discovery workflow presented in this study is a prime example of targeted analysis based on systems-level insights, in this case: the inconsistencies between modeling predictions and experimental observations (Figure 2.1).

The first step in the isozyme discovery workflow was to identify the targets for exploration. These targets come from performing flux balance analysis (FBA) gene essentiality simulations in *E. coli* utilizing the *i*JO1366 metabolic reconstruction (Orth et al. 2011; Kauffman, Prakash, and Edwards 2003). When discussing computational gene essentiality predictions, the term *false negative prediction* refers to a situation in which a gene is predicted to be essential but experimentally observed to be non-essential. This type of prediction failure can stem from lack of knowledge of an alternate pathway or isozyme (Orth and Palsson 2012). All genes associated with false negative predictions in *i*JO1366 were identified, and those genes with high-confidence candidate isozymes, based on sequence homology, were used as examples for this study. To identify potential isozymes based on sequence homology, NCBI's BLASTp algorithm (Altschul et al. 1997) was run for each protein sequence (results summarized in Table S1). An expect value of <E-40 and high sequence identity percentage were utilized as a cut-off for candidates.

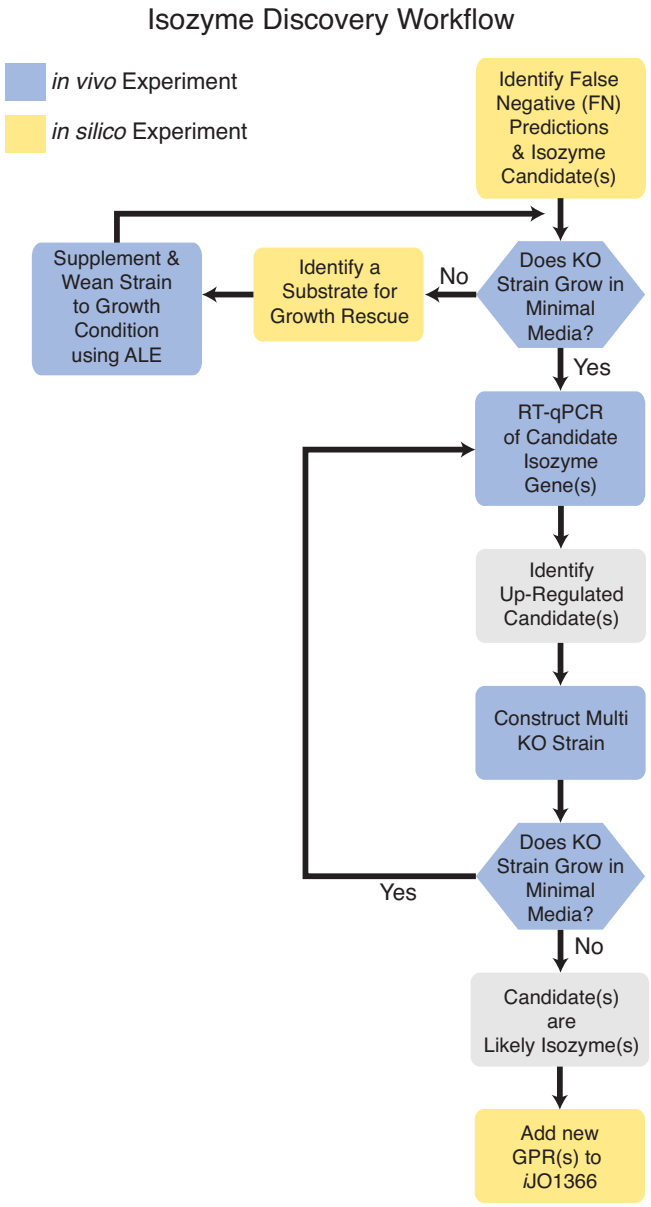


**Figure 2.1**: **A schematic of the general workflow utilized for isozyme discovery involving both *in vivo* and *in silico* experiments.** Starting from the top-most box, false negative model predictions and isozyme candidates were identified utilizing FBA and BLASTp. The workflow was then followed vertically downward examining KO strain growth, expression levels of candidate isozyme genes, and multi-KO strain phenotypes. Deviations from the schematic occurred when growth discrepancies were encountered. The workflow was terminated once a synthetic lethal interaction of false negative gene and isozyme candidate(s) were identified. The output was new enzymatic activities characterized and added to the current genome scale model reconstruction of *E. coli*.

Following identification of false negative targets and potential isozymes, experiments were conducted to determine which isozyme candidate might explain the modeling failure. The knockout strain corresponding to the false negative target was examined. Growth on glucose minimal medium of the knockout strain was confirmed. It was then hypothesized that an isozyme was compensating for the lost function of the primary gene that was knocked out. This hypothesis was tested by exploring expression of the putative isozymes in the primary knockout strain. RT-qPCR analysis was performed and if an isozyme candidate was up-regulated, the next step in the workflow was followed (Figure 2.1).

Following confirmation of up-regulation of the candidate isozyme, a double knockout (DKO) strain was constructed. It was thus hypothesized that removal of the up-regulated isozyme candidate would lead to a synthetic lethal interaction if there remained no other isozymes. The next step in the workflow was to test the growth of the DKO strain and confirm a synthetic lethal interaction. If a synthetic lethal interaction was verified following at least one week of incubation, then the isozyme was deemed to be correctly identified based on genetic and transcriptional evidence. A possible deviation from the above steps is also taken into consideration in this study. For example, if a DKO strain was not lethal, the possibility of an alternate isozyme was explored. The following sections describe the specific workflows followed for the three false negative cases examined in this study (Figure S1).

Implementation of this workflow resulted in three main findings: 1) discovery of seven new isozyme associations and adaptive regulatory mechanisms for partially-characterized enzymes in *E. coli*; 2) development of a enzyme substrate promiscuity discovery tool which can easily be extended to fill other knowledge gaps in *E. coli*

as well as other organisms; and 3) the establishment of a more rigorous assessment of lethality with longer growth incubations in order to prevent false statements of lethality, particularly for high throughput screens.

### 2.2.2 Case 1: *aspC*–aspartate aminotransferase

The aspartate aminotransferase in *E. coli*, encoded by the gene *aspC*, has been characterized as a broad-substrate, multi-functional enzyme that catalyzes the formation of aspartate, phenylalanine, and tyrosine (Fotheringham et al. 1986). Early studies have drawn links to the overlapping functions of the aminotransferases encoded by *aspC*, *tyrB*, and *ilvE* (Gelfand and Steinberg 1977). The aromatic aminotransferase encoded by *tyrB* has shown activity in the synthesis of phenylalanine, tyrosine, and leucine (Powell and Morrison 1978), whereas the branched-chain aminotransferase, encoded by *ilvE*, has been associated with the synthesis of isoleucine, leucine, and valine (Lee-Peng, Hermodson, and Kohlhaw 1979). Previous studies reported that *aspC* knockout strains are viable in both rich media and glucose minimal medium (Baba et al. 2006); however, *i*JO1366 model simulations predicted no growth on minimal medium. Given this false negative prediction, BLASTp was then utilized and results pointed to *tyrB* as an isozyme candidate (see Table S1).

Initial growth tests were performed to verify reports of non-essentiality. The growth data for the Δ*aspC* strain is illustrated in Figure 2.2B. Growth of the Δ*tyrB* strain was also validated in this study. Following completion of initial growth characterizations, RT-qPCR analysis of the *tyrB* isozyme target was performed in the Δ*aspC* and wild type strains. qPCR analysis showed up-regulation of *tyrB* in the Δ*aspC* strain with a fold change of 4.7 compared to the wild type strain (Figure 2.2A). Therefore,

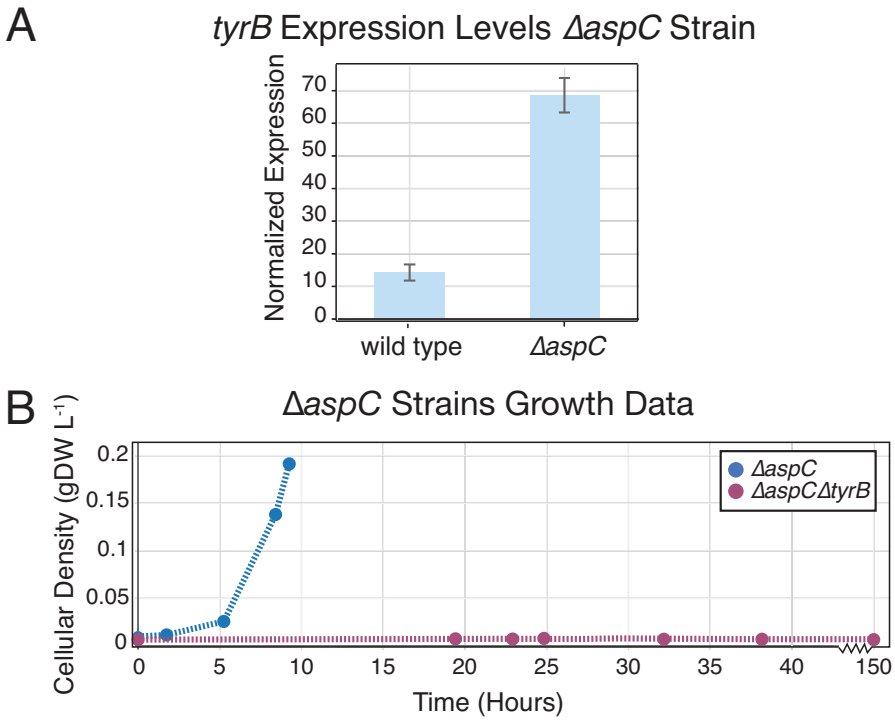


**Figure 2.2**: **The workflow-guided results utilized to discover isozymes of *aspC*.** A) A bar chart of the qPCR results in terms of normalized expression of the *tyrB* isozyme candidate in the Δ*aspC* and wild type strains (standard error ratio was calculated, p-value<0.05, N=1, 2 biological duplicates, 6 technical replicates). A fold increase of 4.8 is observed in the Δ*aspC* strain compared to wild type. B) Growth data on glucose minimal medium in terms of cellular density is reported for Δ*aspC* and Δ*aspC*Δ*tyrB* strains. The Δ*aspC*Δ*tyrB* shows no growth for >150 hours.

the next step in the workflow was followed and construction of the Δ*aspC*Δ*tyrB* DKO strain was performed. Growth of the Δ*aspC*Δ*tyrB* strain was monitored (Figure 2.2B) and the Δ*aspC*Δ*tyrB* KO pair was deemed synthetically lethal based on this genetic evidence. Therefore, successful execution of the workflow identified the isozyme link between *aspC* and *tyrB*.

As a secondary result of executing this method was the discovery of an association between *ilvE* and L-tyrosine biosynthesis. Efforts were placed on finding amino acid supplements that would enable growth of the Δ*aspC*Δ*tyrB* strain to further validate the functions of this interrelated trio of genes. Growth characterizations were

performed utilizing various combinations of amino acid supplementation including L-aspartate, L-tyrosine, L-phenylalanine, and L-leucine. Gene knockout simulations of growth on glucose plus supplementation of all combinations of the *aspC* and *tyrB* associated amino acids resulted in an expected requirement of aspartate and tyrosine for growth rescue of the Δ*aspC*Δ*tyrB* strain (Table S2). Experimental observations, however, showed that only aspartate was required for growth rescue. It was therefore speculated that the aminotransferase encoded by *ilvE* was fulfilling the role of tyrosine synthesis. The enzyme encoded by *ilvE* has shown some, though minimal, specific activity with phenylalanine and tyrosine in an *in vitro* assay (Lee-Peng, Hermodson, and Kohlhaw 1979). Thus, the overlapping *in vivo* functionality of these aminotransferases, *aspC*, *tyrB*, and *ilvE*, appeared to be greater than previously expected, thus these functionalities are assumed to be enabled by the native gene.

Finally, in order to examine the possibility of acquired mutations in the strains constructed in this study, genome resequencing was performed for the Δ*aspC* and Δ*tyrB* strains (Barrick et al. 2009). A summary of the output for these strains is shown in Table S3. The Δ*aspC* and Δ*tyrB* strains showed no apparent mutations in the coding regions of the related isozymes examined.

From the study of the *aspC* false negative, it was proposed that two new functions should be considered to occur in *E. coli* K-12: 1) the annotated tyrosine aminotransferase, *tyrB*, apart from being the 'aromatic amino acid' aminotransferase, can also perform the role of aspartate aminotransferase and 2) the isoleucine aminotransferase, *ilvE*, apart from being the 'branched-chain amino acid' aminotransferase, can also perform the role of tyrosine aminotransferase.

### 2.2.3 Case 2: *argD*–acetylornithine aminotransferase/ N-succinyldiaminopimelate aminotransferase

Another member of a generalist enzymatic class (Nam et al. 2012) explored in this study was the enzyme encoded by the gene *argD*. This aminotransferase was previously identified as having dual functionality, involved in both lysine and arginine biosynthesis (Ledwidge and Blanchard 1999). The *argD* gene is predicted to be an essential gene because of its role in amino acid synthesis; however, knockout studies have repeatedly shown the non-essentiality of this gene on glucose minimal medium (Baba et al. 2006). The putative isozyme targets explored for acetylornithine/N-succinyldiaminopimelate aminotransferase based on sequence homology were *astC*, *gabT*, *patA*, and *puuE*.

Initial examination of the growth of a Δ*argD* strain was performed. Following this confirmation of growth, RT-qPCR analysis was performed to examine the expression of isozyme candidate genes in the Δ*argD* strain compared to a wild type strain (Figure 2.3A). The candidate genes *astC* and *gabT* showed the greatest fold difference from wild type in expression: 3.97-fold and 4.06-fold, respectively. The up-regulation of these two genes prompted the construction of two DKO strains, Δ*argD*Δ*astC* and Δ*argD*Δ*gabT*. Of the two, Δ*argD*Δ*astC* was the DKO strain initially chosen for examination due to a previously drawn relationship (Newman et al. 2013). The growth exhibited by this strain is displayed in Figure 2.3B. The Δ*argD*Δ*astC* strain demonstrated only a mild difference in growth fitness compared to the Δ*argD* strain; therefore, further analysis of the remaining candidates was performed.

A second round of RT-qPCR was performed on the Δ*argD*Δ*astC* DKO strain to further identify isozyme candidates. The gene *gabT* continued to be up-regulated in

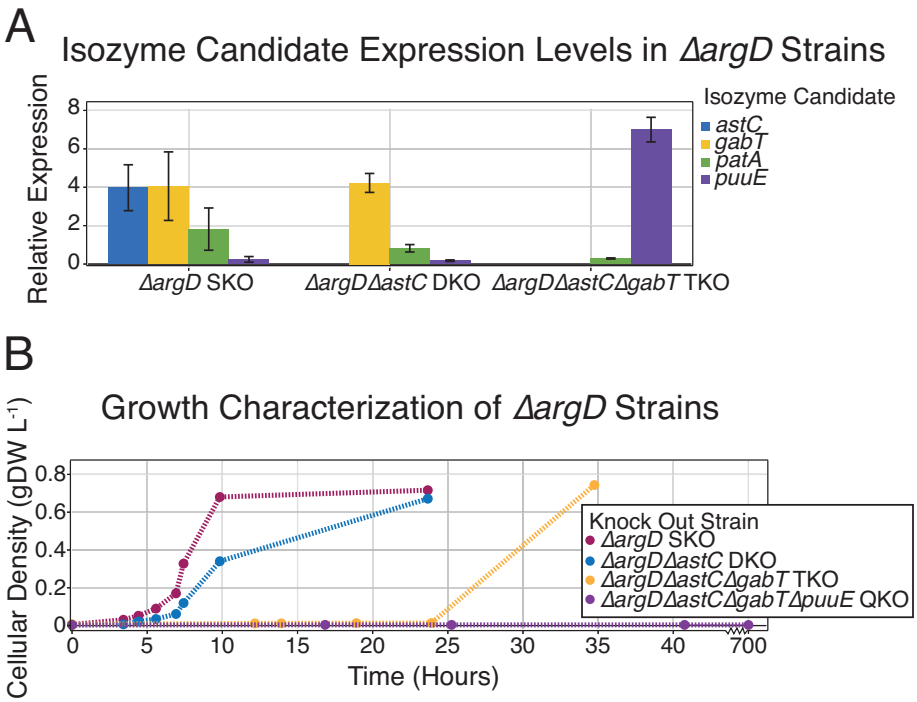


**Figure 2.3**: **The workflow-guided results utilized to discover isozymes of *argD*.** A) A bar chart of the relative expression, compared to wild type, of candidate isozyme genes in the SKO, DKO, and TKO strains shows up-regulation that guided the multi-KO strain construction (standard error ratio was calculated, p-value<0.05, N=1, 2 biological duplicates, 6 technical replicates). Note that *puuE* was not up-regulated until the construction of the $\Delta argD\Delta astC\Delta gabT$ TKO strain. B) Growth data on glucose minimal medium in terms of cellular density is reported for the four strains iteratively constructed as guided by the workflow. The last strain constructed $\Delta argD\Delta astC\Delta gabT\Delta puuE$ continued to show no growth after 700 hours of incubation.

this DKO strain with a relative expression ratio of 4.22 (Figure 2.3A). This result led to the construction of a triple knockout (TKO) strain, $\Delta argD\Delta astC\Delta gabT$. A significant reduction in growth fitness was observed in the TKO strain, with an exhibited lag phase of approximately 24 hours (Figure 2.3B). The eventual growth of the strain suggested the need for further examination of the remaining two candidates, *puuE* and *patA*.

A third round of RT-qPCR was performed on the constructed TKO strain. Although the *puuE* gene had been down-regulated in the SKO and DKO strain compared to the wild type strain (relative expression ratios of 0.24 and 0.18, respectively), qPCR showed its up-regulation in the TKO strain (Figure 2.3A). A relative expression ratio of 7.00 was found for the *puuE* gene, thereby prompting the construction of a quadruple KO (QKO) strain. The strain $\Delta argD\Delta astC\Delta gabT\Delta puuE$ was screened for growth for more than four weeks in multiple attempts and a conclusion of lethality was made. This result closed the experimental loop in the workflow. As a final validation, all remaining DKO and TKO combinations were constructed and their growth validated to ensure the synthetic lethal interaction was as expected (Figure S2).

The results from examining isozymes of *argD* suggested the presence of a regulatory hierarchy regarding isozyme activation that emerged following serial knockout and expression analysis of the multi-KO strains. The mechanisms influencing this regulatory response is suggested as an avenue for further study beyond the scope of work presented here.

In order to examine the possibility of acquired mutations in the strains constructed and utilized in this study, genome resequencing and analysis was performed

for $\Delta argD$, $\Delta puuE$, and their descendent TKO strains. A summary of the results for these strains is provided in Table S4. The only obvious mutation of interest was that of the new junction call at *puuR*/ *puuC* in the $\Delta argD\Delta astC\Delta gabT$ TKO strain. Further read-depth coverage analysis of this region revealed a 962 bp deletion between 1,360,264 bp 1,361,226 bp, deleting a large section of the *puuR* gene (Figure S3). The genes *puuR* and *puuC* are both in the same operon as the isozyme *puuE* explored in this study, with *puuR* acting as a repressor for this operon under conditions of low putrescine concentration (Nemoto et al. 2012). These results thus point to a rarer mechanism of up-regulation via promoter mutation (Andersson and Hughes 2009).

A potential mechanism for the large up-regulation of the *puuE* isozyme in the TKO strain is elucidated from a structural analysis of the repressor protein, *puuR*. The *puuR* DNA-binding transcriptional repressor consists of two domains, namely a helix-turn-helix DNA-binding domain and a Cupin-family domain (Nemoto et al. 2012). Using the high number of homologous templates available, a homology model was constructed from the amino acid sequence of the *puuR* gene via the available toolkits (Roy, Kucukural, and Zhang 2010; Hildebrand et al. 2009; Kelley and Sternberg 2009) (with an average confidence of 97% across templates, see Figure S4). The resulting structure demonstrates that the observed deletion in the Cupin domain from the read depth coverage analysis is in direct contact with the helix-turn-helix motif in the DNA binding domain. Thus it was concluded that removing this part of the protein would drastically compromise protein integrity and prevent DNA binding, and, consequently, the ability of the *puuR* protein to repress the *puuE* gene.

Finally, examination of this case demonstrated the potential importance of extending incubation times in essentiality screens. Often in high throughput data

sets, growth cut-off times are made for the sake of analysis (Nichols et al. 2011; Baba et al. 2006; Gerdes et al. 2003), which could lead to misleading reports of essentiality. This study provided initial data on a range for incubation times required to make essentiality calls with higher accuracy. For the $\Delta argD\Delta astC\Delta gabT$ TKO strain, longer lag phases than those typically observed in *E. coli* were measured. Interestingly, mutations were observed that were implicated in rescuing the growth and loss of the primary enzyme(s) under examination. As strain resequencing becomes more accessible, it is possible that similar mutations acquired during extended lag phases will be observed (Finkel 2006). As demonstrated here, strains exhibiting delayed or slow growth may present an interesting opportunity for discovery.

### 2.2.4 Case 3: *gltA*–citrate synthase

The last false negative case examined in this study was the growth of a citrate synthase, *gltA*, knockout strain. Previous studies demonstrated the ability of a 2-methylcitrate synthase to perform the same catalytic function as citrate synthase and suggested that mutagenesis is required for this transition to take place (Patton et al. 1993). Independently, utilizing the approach presented here, a BLASTp analysis also pointed to *prpC* as a putative isozyme based on sequence homology. This case was thereby examined as a final validating case, demonstrating the up-regulation of an isozyme in the absence of the primary catalyzing enzyme.

Upon initial characterization of a $\Delta gltA$ strain, the strain did not grow on minimal medium despite the fact that it was listed as a positive growth phenotype in the initial Keio screen (Baba et al. 2006). Thus, the case could be considered a true negative prediction; however, given the strong evidence for a possible homolog

(Table S1) as well as previous literature reports (Patton et al. 1993), this case was further explored. Adaptive evolution was utilized to see if the putative isozyme, *prpC*, could indeed rescue a *gltA* deficient strain when a selective pressure was applied. To promote growth, a 'weaning' ALE was performed (Lee and Palsson 2010). FBA was conducted to determine which metabolite could be added to the medium in order to rescue growth (Table S5). It was predicted that supplementation with citrate would allow for utilization of glucose and support growth. Although other supplements were predicted to improve growth, citrate was selected due to its close relation to the citrate synthase reaction as well as the inability of *E. coli* to utilize citrate as its sole carbon source, thereby forcing metabolism of glucose (Koser 1923). $\Delta gltA$ was therefore grown with citrate supplementation in two parallel ALE experiments. Growth was observed in $\Delta gltA$ ALE 1 and $\Delta gltA$ ALE 2 with supplementation; and following two passages robust growth was observed without supplementation (Figure 2.4A). The final apparent cell densities for each ALE experiment showed an approximate 8 and 6 fold increase (ALE experiment 1 and 2 respectively) from the initial supplemented state.

Expression analysis of the evolved endpoints showed significant up-regulation (more than 500 fold difference) of the isozyme target, *prpC*, thereby providing evidence of its isozyme function and allowing for a linear progression through the workflow (Figure 2.4B). Furthermore, the growth rate of each isolate clone correlated well with the expression level of *prpC* (Figure S5). Following observation of these results, an attempt was made to knock out *prpC* from the four endpoint clones. Although knockout confirmation primers suggested successful removal of the *prpC* gene, the strains continued to demonstrate growth. This suggested the possibility of duplication

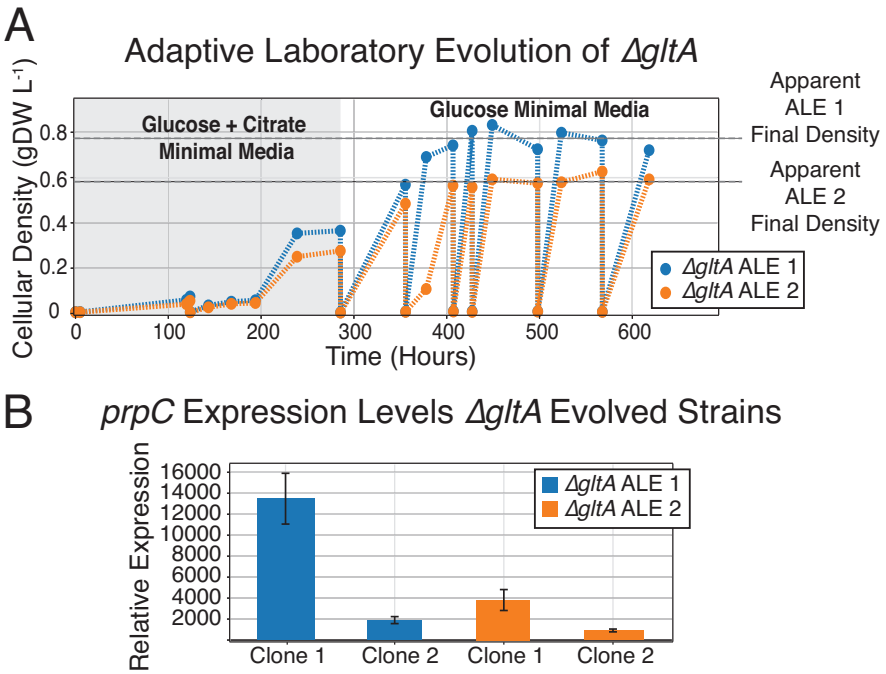


**Figure 2.4**: **The workflow-guided results utilized to discover isozymes of *gltA*.** A) Cellular density results from the ALE of Δ*gltA* on glucose minimal medium are illustrated. A vertical drop in cellular density corresponds to manual passaging of a fraction of the cell culture for a fresh batch of medium. The independent ALE experiments reached a different apparent final density. B) A bar chart showing qPCR results as a fold increase in expression of the *prpC* isozyme candidate in four ALE endpoint clones in relation to a wild type strain (standard error ratio was calculated, p-value<0.05, N=1, 2 biological duplicates, 6 technical replicates).

of the *prpC* gene elsewhere in the genome. Therefore, an in-depth whole-genome resequencing and analysis was performed.

Two clones from each experiment, Δ*gltA* ALE1 and Δ*gltA* ALE2, were isolated and resequenced along with the parent Δ*gltA* Keio strain. Resequencing results revealed the possibility of new junction evidence and mutations within the *prp* operon (Table S6) for three out of the four endpoint isolates. Further analysis showed elevated read coverage of the region between 279 and 371 kb in the ALE endpoint clones but not in the Δ*gltA* parent strain (Figure S6). The same coverage abnormality was also confirmed in ALE clones isolated from the third passage cultures (Figure S6). A single relevant novel junction was then identified by applying a custom pipeline based on *de novo* fragment assembly, providing significant evidence for tandem duplication. The presence of the junction was further verified in all endpoint isolates by performing read mapping onto the corresponding sequence. The duplicated region was flanked by a 182 bp repeat which is a part of the IS30 mobile element. The rightmost flanking copy is not present in the GenBank reference K-12 strain (Benson et al. 2005) and is unique to the Keio K-12 parent strain. Furthermore, plots from Figure S6 clearly suggest high multiplicity of the 100 kb duplicated fragment. To perform the copy number analysis, per base coverage data was normalized to the average coverage of the particular strain and to the position specific biases, inferred from the Keio parent strain coverage distribution (Figure S7). Then the multiplicity was estimated by the median of the normalized coverage values across the duplicated region. Predicted multiplicities of the duplicated region are 7, 11, 16, and 8 for the different clones, respectively. Finally, analysis of novel junctions also predicted a smaller-scale, 1 kb, duplication event in ALE 1 clone 1. The genome coordinates of this duplication are

348,810 - 349,895, which spans part of the *prpC* gene and is encompassed by the larger-scale 100 kb duplication as well.

The results from the coverage and whole genome resequencing analysis helped to explain the inability to construct the DKO construction. Based on the suspected high copy number of duplication, it was deemed unreasonable to knockout all copies of the *prpC* gene and the experimental workflow was concluded. Thus, given the transcriptional and mutational evidence, it is likely that the *prpC* gene is indeed an isozyme for *gltA*, as has previously been reported. Furthermore, these results expand upon the theories surrounding previous reports of genome duplication amplifications as an evolutionary mechanism (Andersson and Hughes 2009; Finkel 2006).

Insight into biological adaptability requiring evolution was thus gained from exploring the presence of large-scale duplications. The mutation event was required to rescue growth and activate the known isozyme, *prpC*, similar to a previous study (Patton et al. 1993). This mutation event occurred after two ALE passages in this study (Figure S6). Interestingly, all four individual endpoint clones, as well as the two clones from Pass 3 that were isolated and sequenced, exhibited the same large-scale duplication of the 100 kb region, thereby implying a clear evolutionary pressure to up-regulate this particular region. Published theories regarding genome duplication amplifications have remarked on the instability of large duplications and their subsequent loss in the absence of selection pressures (Andersson and Hughes 2009). Thus, although the duplications were detected after the third passage and again after the ninth, they could be lost with further adaptation.

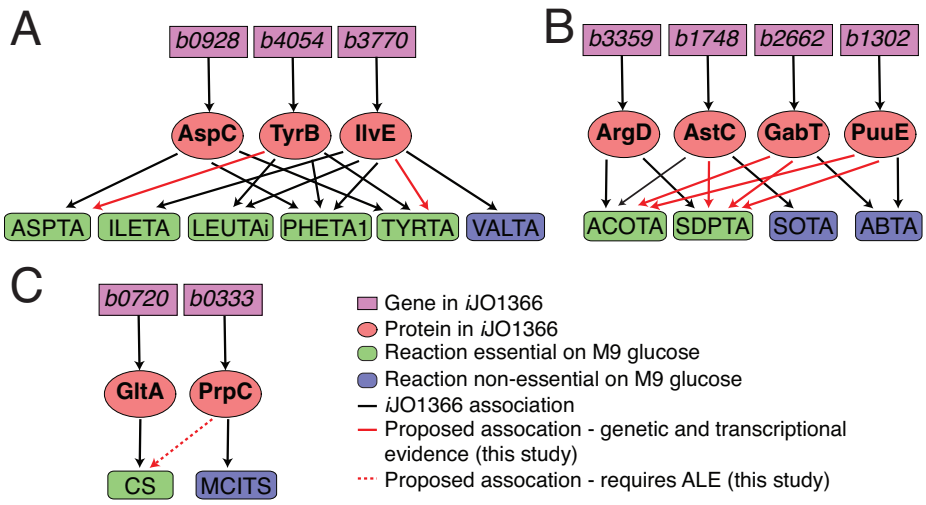


**Figure 2.5**: **A summary of gene-protein-reaction (GPR) associations to be added to the *E. coli* metabolic network reconstruction *i*JO1366 based on findings from three cases, *gltA*, *aspC*, and *argD*.** Novel associations are highlighted in red. A) GPR additions for *tyrB* and *ilvE*. B) GPR additions for the quadruple synthetic lethal interaction set. C) GPR for *gltA* and *prpC* highlights the requirement of evolution or mutagenesis for the suggested association.

### 2.2.5 Gene-protein-reaction (GPR) analysis and conservation of isozymes across 55 related strains

Inconsistencies between *in silico* predictions and *in vivo* data guided this study and resulted in the discovery of seven new links between known, partially-characterized enzymes and reactions that are conditionally essential to the metabolic network in *E. coli*. In this study, we moved to complete the missing links that propose a solution to the computational and experimental inconsistencies observed through *in vivo* studies. Suggested changes to the GPR association in *i*JO1366 based on genetic and transcriptional evidence are presented in Figure 2.5 (abbreviations are defined in Table S7). The expanded reaction associations for the cases examined in this study support the hypothesis that functional overlap occurs for enzymes across metabolism, forming the basis of an underground metabolic network, and this concept has been supported by other recent works (D'Ari and Casadesús 1998; Notebaart et al. 2014).

As a preliminary expansion of this study, we investigated the conservation of the newly discovered isozymes in 55 closely related strains of *E. coli* and *Shigella* that have existing metabolic models (Monk et al. 2013). It was determined that the same GPR changes should be made in the majority of these models. However, some of the putative isozymes discovered in this study have no corresponding gene in the related strains (Figure S8). For example, eight Shigella strains examined are lacking *prpC*, the newly discovered *gltA* isozyme. Also, 14 *E. coli* strains from different clades lack *puuE*, one of newly discovered *argD* isozymes. Finally, five of the *Shigella* strains lack *astC*, another one of the *argD* isozymes. Therefore new GPR associations are available for each of the 55 models, but they must be adjusted in a strain-specific manner. Furthermore, analysis of isozyme and regulatory region sequence conservation between different strains of *E. coli* could illuminate divergent evolutionary strategies in the *E. coli* species.

## 2.3 Conclusions

Enzymatic promiscuity and a cell's ability to adapt to genetic perturbations were explored in the execution of the model-driven workflow developed in this study. The results suggest that a hierarchy of latent metabolic solutions exist, as highlighted by the analysis of the false negative *argD*. Furthermore, this study emphasized the possibility of discovering novel regulatory responses following long-term culturing studies and genome resequencing when determining gene essentiality. Lastly, the developed methods can be readily extended to other organisms and gene targets where gap-filling is required. For example, a gap-filling study utilizing the *E. coli i*JO1366 metabolic reconstruction has identified a total of 265 false negative predictions (corresponding

to 59 unique genes) which could be explored under various environmental and genetic conditions (Orth and Palsson 2012). The extendability of this work to pathogenic organisms could be particularly advantageous in searching for anti-microbial targets. As genome-scale models and organism-specific knowledge bases expand, their ability to predict biological behavior for both basic science and biotechnology applications will increase. This likely expansion is evident in the appearance of complementary studies (Notebaart et al. 2014) which use computational modeling to isolate specific predicted functionalities through gene knockout or over-expression and through determining media conditions which focus pressure on the predicted function(s). A comparison of the present study to the aforementioned study (Notebaart et al. 2014) shows some overlap between genes explored and thought to exhibit underground activity (*ilvE* and *tyrB*), although the suggested activities reported in each study for these genes are distinct. Workflows for discovering promiscuous and latent activities such as the one presented here will be critical for advancement of model-driven science.

Although the strengths of the presented method were demonstrated with the cases explored in this study, there is room for improvement to broaden the applicability of the workflow. For false negative model gaps, there is the possibility that an alternate pathway is rescuing the growth of the cell. For such alternate pathway solutions, isozyme analysis, could result in fruitless effort as those solutions are not captured by the workflow. Another area of improvement proposed for the workflow is in selecting bioinformatic algorithms. There are many enzymes for which BLASTp, a purely sequence homology driven algorithm, will not result in the identification of candidate isozymes. In order to expand this list of putative isozymes, the use of protein structure similarity or substrate structure similarity identifying algorithms is suggested (Zhang

et al. 2012). Finally, the utilization of RNA-seq or other larger-scale omics methods could capture a more complete picture of transcriptional changes in response to KO perturbations rather than qPCR analysis. These modifications to the workflow presented may result in a more robust method for filling model gaps, which can be applied to other organisms as well.

## 2.4 Materials and Methods

### 2.4.1 Identifying isozyme candidates with BLASTp

Enzyme protein sequences corresponding to the false negative gene targets were utilized as input for NCBI's BLASTp algorithm. These sequences were compared with all other protein sequences in the organism *E. coli* K-12 MG1655. Only those with high alignment scores (>150) and Expect values (E-values) <1E-40 were considered for this study (Table S1).

### 2.4.2 Bacterial strains and plasmids

All bacterial strains utilized in this study were descendants of *E. coli* K-12 strain. Strains included a control, characteristically wild-type strain, *E. coli* K-12 MG1655 (bop27), as well as several strains taken from the single-gene knockout Keio collection (Baba et al. 2006) derived from the parent strain, *E. coli* K-12 strain BW25113. Keio strains examined were Δ*aspC*, Δ*gltA*, Δ*aldA* and Δ*argD*. Knockout strains constructed utilizing the Keio parent as a starting strain are summarized in Table 2.1 of Materials and Methods.

Plasmids utilized in this study were pKD3, pKD13, pKD46, pCP20 (Datsenko

and Wanner 2000).

### 2.4.3 Culture conditions and growth characterization

All strains were grown in either M9 minimal medium or Luria-Bertani (LB) broth in 250 mL Erlenmyer flasks containing magnetic stir bars for aeration. The M9 minimal medium was composed of 2 g $L^{-1}$ D-glucose, 100 $\mu$M $CaCl_2$, 2 mM $MgSO_4$, 6.8 g $L^{-1}$ $Na_2HPO_4$, 3 g $L^{-1}$ $KH_2PO_4$, 0.5 g $L^{-1}$ NaCl, 1 g $L^{-1}$ $NH_4Cl$, and 250 $\mu$L $L^{-1}$ trace element solution. The trace element solution was composed of (per liter) 1 g of $FeCl_3 \cdot 6H_2O$, 0.18 g of $ZnSO_4 \cdot 7H_2O$, 0.12 g of $CuCl_2 \cdot 2H_2O$, 0.12 g of $MnSO_4 \cdot H_2O$ and 0.18 g of $CoCl_2 \cdot 6H_2O$. For strains with antibiotic resistances and during knockout procedures, LB and M9 minimal mediums were supplemented with antibiotics at concentrations of 50 $\mu$g $mL^{-1}$ kanamycin, 50 $\mu$g $mL^{-1}$ chloramphenicol or 100 $\mu$g $mL^{-1}$ ampicilin.

Growth screens on 0.2% D-glucose M9 minimal medium were conducted in liquid culture. Cell cultures were first grown to an optical density at 600 nm ($OD_{600}$) of 0.6-1.0 in LB broth containing selective antibiotics. The cells were then washed twice with M9 minimal medium (no carbon source) prior to inoculation of the screening growth media. The target initial $OD_{600}$ for inoculations was 0.01-0.02.

### 2.4.4 *In silico* modeling

The latest genome scale metabolic model for *Escherichia coli* K-12 MG1655 (Orth et al. 2011), *i*JO1366, was utilized in this study for all growth and knockout simulations using the constrain-based modeling package, COBRApy (Ebrahim et al. 2013). Growth simulations were performed by optimizing for the default core biomass

**Table 2.1**: **Knockout Strains** *KO not fully constructed

| Keio Parent Strain | Evolved Clone | Double KO | Triple KO | Quadruple KO |
|---|---|---|---|---|
| $\Delta gltA$ | $\Delta gltA$ ALE 1 Clone 1 | *$\Delta gltA$ ALE 1 Clone 1 $\Delta prpC$ | | |
| | $\Delta gltA$ ALE 1 Clone 2 | *$\Delta gltA$ ALE 1 Clone 2 $\Delta prpC$ | | |
| | $\Delta gltA$ ALE 2 Clone 1 | *$\Delta gltA$ ALE 2 Clone 1 $\Delta prpC$ | | |
| | $\Delta gltA$ ALE 2 Clone 2 | *$\Delta gltA$ ALE 2 Clone 2 $\Delta prpC$ | | |
| $\Delta aspC$ | | $\Delta aspC\Delta tyrB$ | | |
| $\Delta argD$, $\Delta puuE$ | | $\Delta argD\Delta astC$ | $\Delta argD\Delta astC\Delta gabT$ | $\Delta argD\Delta astC\Delta gabT\Delta puuE$ |
| | | $\Delta argD\Delta gabT$ | $\Delta argD\Delta puuE\Delta astC$ | |
| | | $\Delta argD\Delta puuE$ | $\Delta argD\Delta puuE\Delta gabT$ | |
| | | $\Delta puuE\Delta astC$ | $\Delta puuE\Delta astC\Delta gabT$ | |
| | | $\Delta puuE\Delta gabT$ | | |

objective function, which is a representation of essential biomass compounds, from cellular components in stoichiometric amounts (Feist and Palsson 2010). Determining gene or reaction essentiality on simulated glucose minimal medium was achieved by removing the desired gene or reaction from the model and then running a flux balance analysis (FBA) simulation under conditions mimicking the *in vivo* screen (aerobic growth with glucose as a substrate) as previously described in detail (Orth and Palsson 2012) with the glucose exchange reaction lower bound set to -10 mmol $gDW^{-1}hr^{-1}$ and the oxygen exchange reaction lower bound set to -1000 mmol $gDW^{-1}hr^{-1}$. A gene or reaction was deemed essential if the predicted flux through the biomass objective function was less than zero (a threshold of 0.001 or less). LB medium was simulated by opening all exchange reactions (setting their lower bound to -10 mmol $gDW^{-1}hr^{-1}$).

### 2.4.5 Adaptive laboratory evolution by weaning off of supplementation

Adaptive laboratory evolution was started isolating a single colony of the confirmed $\Delta gltA$ Keio strain grown on a solid LB plate with 50 $\mu$g $mL^{-1}$ kanamycin. This single colony was then grown to mid-log in LB, washed twice with M9 minimal medium with no carbon source and then used to inoculate two 50 mL flasks of M9 minimal medium with 0.2% glucose and 0.02% citric acid at 37 °C and utilizing a magnetic stir bar for mixing and aeration. The citric acid supplementation was completely removed after two passages once a notable increase in growth fitness was measured. For ALE cultures, $OD_{600}$ was measured once a day or before passages. ALE cultures were allowed to grow to stationary phase before each passage. Passages were conducted every other day with a targeted inoculation $OD_{600}$ of 0.05. Before

each passage, a glycerol stock of each culture was prepared and stored at -80 °C. ALE was conducted over a period of approximately 25 days once the final apparent cellular density of each flask had appeared to stabilize for over 5 days. Samples taken from the endpoint ALE culture flasks were plated on LB rich medium and two colonies from each ALE were selected for further characterization (qPCR, resequencing and growth characterization). The endpoint clones were named Δ*gltA* ALE 1 clone 1, Δ*gltA* ALE 1 clone 2, Δ*gltA* ALE 2 clone 1 and Δ*gltA* ALE 2 clone 2.

### 2.4.6 Generation of knockout strains

Knockout strains were generated using the λ-Red recombination system described by Datsenko and Wanner(Datsenko and Wanner 2000). When generating DKO strains, the pKD3 plasmid was utilized to amplify the FRT-flanked chloramphenicol resistance cassette, allowing for the selection of DKO strains with a dual antibiotic resistance. For the generation of TKO strains, pCP20 was utilized to remove antibiotic resistance cassettes and standard knockout procedures were followed utilizing a FRT-flanked kanamycin resistance cassette generated by PCR from the pKD13 plasmid (Datsenko and Wanner 2000).

### 2.4.7 qPCR

Cells were harvested for RNA extraction during exponential phase (OD600 .4-.7). Cells were collected in two volumes of Qiagen RNA-protect Bacteria Reagent, pelleted and stored at -80 °C. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K and 20% SDS for 20 minutes at 37 °C. Total RNA was was isolation using Qiagen RNeasy Mini Kit columns and following vendor

procedures. A 30 minute on-column DNase-treatment was performed prior to elution. RNA was quantified on a NanoDrop. Total RNA quality was assessed by running an RNA-nano chip on a Agilent Bioanalyzer.

Reverse-transcription was performed on 10 $\mu$g of total RNA. The reaction mixture (60 $\mu$L) contained total RNA, 75 $\mu$g random primers, 1X first strand buffer, 10 mM dithiothreitol, 0.5 mM deoxyribonucleotide triphosphates, 20 U of SUPERase-In, and 600 U of SuperScript II reverse-transcriptase. The mixture was incubated in a thermocycler at 25 °C for 10 minutes, 37 °C for one hour, 42 °C for one hour and 70 °C for 10 minutes to inactivate SuperScript II. Remaining RNA was removed by adding 20 $\mu$L of 1 N NaOH to the reaction mixture and incubating at 65 °C for 30 minutes. The reaction was neutralized by the addition of 20 $\mu$L of 1 N HCl. cDNA was then purified utilizing a QIAquick PCR Purification column, following vendor procedures. cDNA quantification was performed using a NanoDrop.

Real-time quantitative PCR was performed on the synthesized cDNA using the QuantiTect SYBR Green PCR Kit. The 25 $\mu$L qPCR mixtures contained 12.5 $\mu$L 2x QuantiTect SYBR Green PCR Master Mix, 0.2 $\mu$M forward primer, 0.2 $\mu$M reverse primer and cDNA template. Each qPCR was performed in triplicate in the BioRad iCycler under the following conditions: 95 °C for 15 minutes, followed by 40 cycles of denaturation at 94 °C for 15 seconds, annealing at 52 °C for 30 seconds, and extending at 72 °C for 30 seconds at which point the SYBR fluorescence was measured for the qPCR curve generation. The biological experiments were performed in duplicate and compared to wild type (MG1655) under the same growth conditions. Binding affinity of each primer set was assessed by constructing a standard curve for each primer. This allowed for calculation of a reaction efficiency. Relative quantities of

cDNA were calculated using the standard curve and normalizing to the quantity of an housekeeping gene, *hcaT*, in the same sample. hcaT was chosen as the housekeeping gene as it was most stably expressed across experiments and as it has been identified as an ideal internal control gene in previous experiments (Zhou et al. 2011). Results were reported in a bar chart showing relative normalized enrichment ratios. The standard error ratios are reported after having performed a right-tailed t-test analysis, assuming normal variables and distribution, with a p-value less than or equal to 0.05.

### 2.4.8 Genome resequencing

Following supplementation of growth media and culture passage of 10 days for *gltA* mutant strains, genomic DNA assessment was prompted to examine mutation accumulation over the course of the growth experiments. Genomic DNA was isolated using NucleoSpin Tissue XS Purification Kit. The quality of DNA was assessed with UV absorbance ratios 260/280 and 260/230. DNA was quantified utilizing the Qubit dsDNA High Sensitivity Assay. Paired-end resequencing libraries were generated following Illumina's Nextera XT standard protocol with and input of 1 ng genomic DNA. Libraries were run on a MiSeq platform using a 250 cycle kit.

### 2.4.9 Mutation analysis

The output library sequences were aligned to a reference genome utilizing the computational pipeline tool, *breseq* (Barrick et al. 2009). The $\Delta gltA$ Keio parent strain was sequenced and mutations were analyzed comparing those mutations accumulated in the passaged strains to those that were present in the parent strain.

Additionally, custom approach based on *de novo* fragment assembly was used

for complementary analysis of novel junctions in *gltA* strains.

Chapter 2 is a reprint of a published manuscript: Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O., and Feist, A. M. (2015). "Model-driven discovery of underground metabolic functions in Escherichia coli". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.3, pp. 929–934. The dissertation author was the primary author of the paper and was responsible for the research.

# Chapter 3

# Reframing essentiality in terms of adaptive flexibility

## 3.1 Introduction

The term essential has been used to define those components of the cell that are required to sustain cell growth. Defining the essential components of life in organisms large and small has been a topic of great scientific interest, and, on one extreme, there is a growing effort to understand the basic principles of life by studying and even synthesizing minimal organisms (Mobegi et al. 2017; Hutchison et al. 2016). Beyond understanding the basic genotype-phenotype connection of life, studies of gene essentiality provide knowledge for medical and industrial applications. Essential genes provide targets for antibacterial drug discovery. For example, by carefully targeting an essential cell component in a bacterium as virulent as *Mycobacterium tuberculosis*, we can strive to treat pathogenic bacterial infections in a targeted and rational manner while at the same time avoiding harmful side-effects to the host organism (Chung,

Dick, and Lee 2013). Strategic, logical drug design has become increasingly important in the face of rising numbers of antibiotic resistant pathogens (Fischbach and Walsh 2009; Hughes and Andersson 2016; Ventola 2015). Thus, research studies of gene essentiality have become important knowledge sources for the advancement of science and medicine.

How do we go about defining the set of essential genes for an organism? Increased availability of genome sequences has led to detailed experimental and computational examination of gene functions at a genome scale in several model organisms including *E. coli*. A clear method for studying gene essentiality is the systematic experimental disruption of genes. One such study resulted in a collection of 3985 single-gene deletion strains for the BW25113 strain of *E. coli* (Baba et al. 2006). Other studies have utilized high-throughput transposon mutagenesis as a tool for gene disruption and identification of all essential genome elements beyond protein-coding sequences (Christen et al. 2011). Beyond experimentally defining essential genes, computational tools such as constraint-based modeling have been used for predicting the essential metabolic components of cells (Feist and Palsson 2008).

Expanding knowledge of the cellular components contributing to metabolism has allowed for the construction of genome-scale models of metabolism. The comprehensive metabolic model for *E. coli* *i*JO1366 contains information related to 1366 metabolic genes and their associated 2251 reactions. Such models can be used to study bacteria from a whole-cell, systems biology perspective (Orth et al. 2011; Bordbar et al. 2014; Monk et al. 2017). By removing genes from the model and performing flux-balance analysis, predictions about gene essentiality on defined growth media can be made. These predictions can then be compared to experimental data and provide insight into

existing knowledge gaps when inconsistencies are encountered (Orth and Palsson 2012). False positive predictions are inconsistencies that occur when the model predicts a gene to be non-essential, but experiments show the gene to be essential. Such instances can be attributed to the inclusion of unrealistic reactions in the model. They can also, however, be attributed to flaws in the experimental data. For example, high-throughput growth screens conducted in plate format are often stopped after 24 or 48 hours of growth (Baba et al. 2006; Nichols et al. 2011; Gerdes et al. 2003). These screens might not capture those strains that are slower to grow. Furthermore, it is also possible that the models predict growth that is not feasible without some form of genetic change or adaptation. High-throughput screens are rarely followed by whole genome sequencing because of the assumption that mutations are not accrued in such a short period of time. Thus, growth that is accompanied by genetic change is not captured by such growth screens.

Essentiality is widely accepted to be conditional (D′Elia, Pereira, and Brown 2009; Ish-Am, Kristensen, and Ruppin 2015). Genes essential for growth in one environment might not be essential in another, given the right nutrient composition. However, essentiality may also be discussed in evolutionary terms. Upon the removal of an essential gene, it is possible that a short period of adaptation is sufficient to activate a redundant pathway or isozyme and enable growth. On the other hand, some genes may be essential regardless of whether or not an adaptive period is provided. Thus, we can also consider a spectrum of essentiality that is related to adaptability. This has been discussed and demonstrated in studies of multi-copy suppression and adaptive laboratory evolution (Patrick et al. 2007a; Guzmán et al. 2015). The extent of redundant pathways in *E. coli* is yet to be fully elucidated;

however, underground metabolism and enzyme promiscuity have been shown to play critical roles in adaptation to new growth environments or in response to genetic perturbation (D'Ari and Casadesús 1998; Notebaart et al. 2014; Guzmán et al. 2015).

In this study, we utilize previously reported false positive predictions of essentiality (Orth and Palsson 2012) to identify gene-deletion strains that may be considered 'non-essential' given a longer incubation period. This study examines gene-deletion strains and categorizes them into three categories (expanding on definitions previously used (D′Elia, Pereira, and Brown 2009; Ish-Am, Kristensen, and Ruppin 2015)). First, if a knockout strain cannot grow on a defined medium where the wild type strain can grow, 'conditional essentiality' is established. Second, if a knockout strain is able to grow on the defined medium where the wild type strain can also grow, 'non-essentiality' is established. Third, if a knockout strain does not initially grow on a defined medium, but is able to grow given an adaptive period and the acquisition of mutations, 'non-essentiality with mutations' is established. Longer growth tests are followed by whole genome sequencing and interpretation of any resulting mutations to determine the adaptive mechanisms required for the rescue of the knockout strains analyzed. The results presented demonstrate a striking agreement between model-predicted alternate isozymes/pathways and observed mutations and shed light on the dynamics of growth observed for various non-essential genes.

## 3.2 Results

### 3.2.1 Identifying Gene Targets

The genes explored in this study were chosen because of essentiality discrepancies observed between *in silico* predictions and *in vivo* observations (Orth and Palsson 2012). Such discrepancies indicate areas for discovery or better understanding as they point out differences between *in vivo* screens and computations based on the collected wealth of knowledge for a given organism (Feist and Palsson 2008). These discrepancies were previously identified as false positive (FP) model predictions (instances where the model predicts that a gene is non-essential, but experimental studies have identified the gene as essential in the particular growth environment). FP model predictions are believed to occur due to the inclusion of a 'non-physiological' model reaction such as an unrealistic alternate isozyme or pathway. On the other hand, FP predictions may also occur if there are errors in experimental calls of essentiality. It is possible that short-term high-throughput growth screens could result in genes being identified as essential when in actuality they require more incubation time to display growth. In these cases, it is possible that the metabolic model annotation of an alternate pathway or isozyme is correct. This study examined the possibility that FP predictions are caused by experimental limitations. Furthermore, the dynamics and mechanisms behind a cell's ability to still display reproducible and robust growth during time frames longer than normal wild-type growth were also examined. This was accomplished by performing more extensive growth analysis of FP genes identified for the *E. coli* metabolic model, *i*JO1366. 41 genes previously identified as FP predictions (Orth and Palsson 2012) were utilized as a starting point for this study (Figure 3.1A).

These were genes associated with FP predictions in defined minimal media conditions.

### 3.2.2 Growth screens considering longer time scales

Extended growth tests were performed on FP associated gene-deletion strains with the hypothesis that model-predicted alternate pathways would rescue growth given longer incubation. Of the 41 FP associated gene knockout (KO) strains identified as potential targets for long growth incubations, only 20 were available in the Keio collection and confirmed by PCR. Unavailable strains were not considered for the longer growth test presented in this study (Baba et al. 2006) (Figure 3.1A). It is possible and perhaps likely that these strains are essential on nutrient-rich media; however, a thorough study using alternate gene disruption analyses (Christen et al. 2011) is required to further elucidate essentiality for this set. The 20 confirmed gene-deletion strains (Table S1) were grown in a rich nutrient undefined medium (Luria-Bertani LB broth) and then used to inoculate minimal medium for the long growth test. Growth of these 20 gene-deletion strains was monitored periodically over the course of approximately two weeks or until growth was observed. If growth was observed, the culture was passed to fresh media to ensure that the growth would persist and was not a by-product of residual LB media. Nine strains did not grow during the extended growth test and were classified as conditionally essential (not essential in LB, but essential for growth in the minimal medium tested) (Figure 3.1A). For the KO strains that did display growth, such long incubation growth tests showed that eleven of the twenty previously identified essential genes were actually non-essential (Figure 3.1, Table 3.1). To confirm the reproducibility of these results, these eleven gene-deletion strains were grown again in triplicate (with some strains tested more

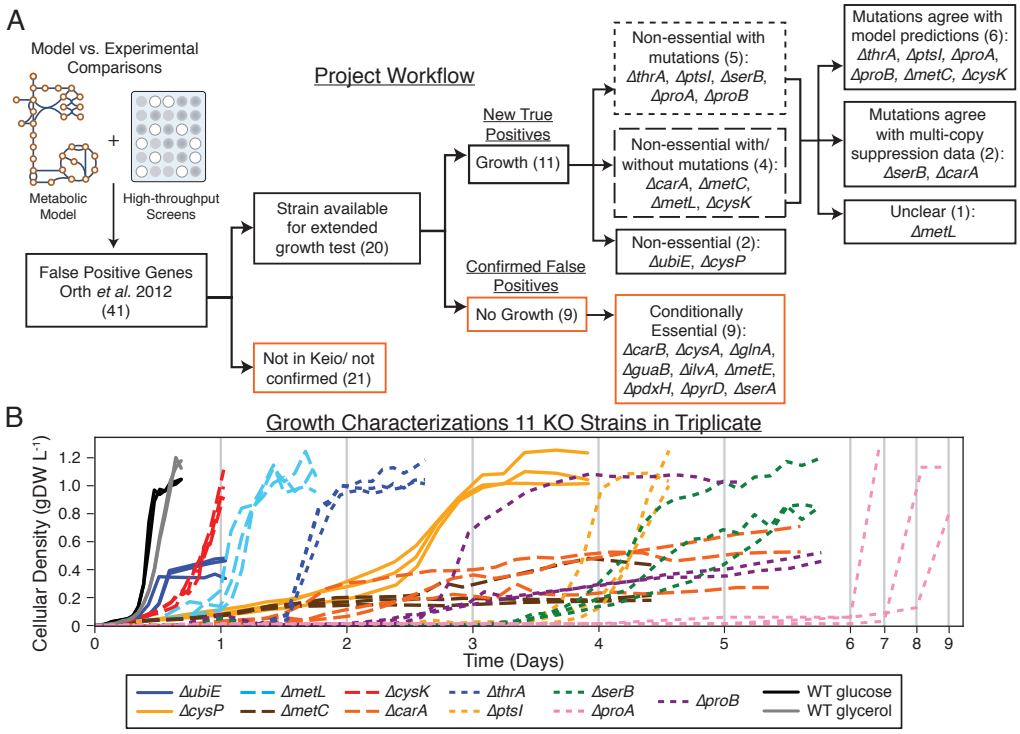


**Figure 3.1**: **Project workflow and growth characterizations of false positive strains.** A. A workflow summarizing the sequence of analyses followed and results from this study is shown. A Keio gene knockout collection strain was not available for approximately half of the corresponding false positive genes listed in (Orth and Palsson 2012), possibly due to essentiality on rich growth media. Of those strains that were available for a longer growth test, approximately half showed growth. Those strains that grew were analyzed for mutations. Five false positive strains showed mutations in all replicate experiments sequenced. Four strains showed mixed results, meaning that only some populations accrued mutations. Two strains showed no mutations in any replicate samples. The nine strains that showed mutations in at least some populations were further analyzed in the context of model-predicted alternate pathways and historical data. Six of these cases showed agreement with model-predictions, two showed agreement with previous reports of multi-copy suppression (Patrick et al. 2007b), and mutation analysis for one case was not clearly linked to either. B. Growth curves of eleven Keio collection strains associated with false positive predictions is displayed. Growth data in terms of cellular density in grams of dry weight per Liter (gDW/L) is reported for the FP gene KO strains. Those strains that accrued mutations in all replicate populations during this growth test are noted with small dashed lines. Those strains that showed mixed results, showing mutations in only some populations, are noted with larger dashed lines. All Keio strains were grown in M9 minimal medium with glucose as the carbon source with the exception of $\Delta cysK$ and $\Delta cysP$ which utilized a glycerol carbon source. Growth of the wild type strain in glucose and glycerol is also provided as a point of reference (black and grey growth curves).

than triplicates, up to nine biological replicates) and their cell density monitored more closely to acquire a more detailed view of their growth trajectories (Figure 3.1B). These eleven gene KO strains were the main focus of this work.

Growth experiments showed a great deal of fitness diversity among the eleven gene-deletion strains (Figure 3.1B, Table 3.1). While four of the eleven KO strains ($\Delta cysK$, $\Delta metL$, $\Delta thrA$, $\Delta ubiE$) showed reproducible growth to their respective final cellular densities within the first 48 hours of incubation, the remaining strains displayed more variability in the time necessary to display growth. Furthermore, some replicates showed a range in growth dynamics between replicate experiments, which is reflected in the standard deviation of the mean time required to reach at least half of the final density observed during these growth tests (Table 3.1). For example, $\Delta proB$ experiments showed a high degree of variability between the three replicates tested. One replicate reached its final density, near that of wild-type, around Day 4 whereas the other two replicates had reached half this level around Day 6 (Figure 3.1B). Several gene-deletion strains also showed variability in the final density achieved. While several strains such as $\Delta thrA$, $\Delta cysK$, and $\Delta metL$ showed typical growth trajectories similar to the wild type strain (with longer lag phases), other strains such as $\Delta carA$, $\Delta metC$, and $\Delta ubiE$ displayed significantly slower growth rates and reached approximately half of the cell density observed for the wild type during the testing period. We hypothesized that the diverse range of growth phenotypes observed could be attributed to differences in adaptive mechanisms required for growth. This was further studied by examining mutations acquired during the growth experiments.

**Table 3.1**: **Strain details from growth characterizations.** [1] Evidence for this described in (Wissenbach, Ternes, and Unden 1992). [2]*In silico* prediction, *i*JO1366 (Orth and Palsson 2012). [3]Experimental multicopy suppression evidence (Patrick et al. 2007a). *The data used from triplicate experiments represented in Figure 1B was used to calculate means, standard deviations (St. Dev.), and percent relative standard deviations (%RSD).

| Keio Strain | Predicted Alternate Genes | *Mean, Std. Dev., %RSD final cell density (gDW/L) | *Mean, Std. Dev., %RSD time to > half final density (Hrs) | Mutations Flask1 Population? |
|---|---|---|---|---|
| WT glucose | - | 1.0, <0.01, 0.2% | 10, <1, <0.01% | No |
| WT glycerol | - | 1.1, 0.04, 3% | 12, <1, 0.04% | No |
| Δ*ubiE* | alternate growth using demethylmenaquinone[1] | 0.43, 0.08, 20% | 14, 1, 9% | No |
| Δ*cysK* | *cysM*[2] | 1.0, 0.1, 10% | 21, <1, <0.01% | Variable |
| Δ*metL* | (*thrL* or *malY*)[2] | 1.0, 0.1, 9% | 28, 2, 7% | Variable |
| Δ*metC* | (*tnaA* or *malY*)[2], (*malY*, *alr*, *fimE*)[3] | 0.28, 0.1, 50% | 36, 9, 20% | Variable |
| Δ*thrA* | (*metL* or *lysC*)[2] | 1.1, 0.1, 8% | 43, <1, 0.04% | Yes |
| Δ*carA* | (*yahI* or *arcC* or *yqeA*)[2], (*carB*, *ygiT*, *cho*, *yncK*)[3] | 0.50, 0.2, 40% | 58, 17, 30% | Variable |
| Δ*cysP* | (*modA* + *modB* + *modC*)[2] | 1.1, 0.1, 10% | 65, <1, 0.2% | No |
| Δ*proB* | *argE*[2] | 0.66, 0.3, 50% | 83, 10, 10% | Yes |
| Δ*ptsI* | *galP*[2], (*fucP*, *xylE*, *galE*)[3] | 1.1, 0.1, 10% | 100, 5, 5% | Yes |
| Δ*serB* | *glyA*[2], (*gph*, *hisB*, *ytjC*)[3] | 0.95, 0.2, 20% | 110, 8, 7% | Yes |
| Δ*proA* | *argE*[2] | 1.1, 0.3, 20% | 190, 26, 10% | Yes |

**Table 3.2**: **Flask 1 Population Mutations.** GDA abbreviation stands for genome duplication amplification event. [1]Binding site recognition sites predicted in (Liu, Blackwell, and States 2001). [2]Evidence to support binding site in (El Qaidi et al. 2009). [3]Attenuator-model of regulation for histidine operon described in (Frunzio, Bruni, and Blasi 1981; Johnston et al. 1980; Di Nocera et al. 1978; Artz and Broach 1975). [4]Information related to the MalI transcriptional repressor in (Reidl and Boos 1991; Reidl et al. 1989).

| Keio Strain | Exp. # | Fraction Population | Gene | Protein Change | Perceived Impact |
|---|---|---|---|---|---|
| Δ*thrA* | 2 | 0.23 | *metJ* | V46E | Reduce MetJ repression[1] |
| | 1 | 0.10 | *metJ/metB* | Intergenic (-200/-75) | Reduce MetJ repression[1] |
| | 1 | 0.79 | *metJ/metB* | Intergenic (-211/-66) | Reduce MetJ repression[1] |
| Δ*ptsI* | (1, 4, 5, 6) | (1.0, 1.0, 0.36, 0.39) | *metK/galP* | Intergenic (+328/-96) | Reduce GalR repression[2] |
| | 5 | 0.25 | *metK/galP* | Intergenic (+333/-91) | Reduce GalR repression[2] |
| | 3 | 1.0 | *metK/galP* | Intergenic (+334/-90) | Reduce GalR repression[2] |
| | 6 | 0.57 | *metK/galP* | Intergenic (+339/-85) | Reduce GalR repression[2] |
| | 3 | 1.0 | *crp* | T141P | - |
| | 5 | 0.70 | *crp* | G142S | - |
| | 6 | 0.59 | *crp* | G142D | - |
| | 4 | 1.0 | *crp* | R143H | - |
| | 1 | 1.0 | *crp* | A145V | - |
| | 6 | 0.43 | *crp* | I187T | - |
| | 7 | 1.0 | 96 genes [*rrsC-rrlA*] | 99 kbp, 2X GDA | Increased Expression *cyaA* |
| Δ*serB* | 1 | 0.67 | *hisL/hisG* | Intergenic (+41/-105) | Increase *his* operon expression[3] |
| | 6 | 0.88 | *hisR* | His tRNA (5/77 bp) | Increase *his* operon expression[3] |
| | 2 | 0.82 | *hisR* | His tRNA (48/77 bp) | Increase *his* operon expression[3] |
| | 4 | 0.92 | *hisR* | His tRNA (67/77 bp) | Increase *his* operon expression[3] |
| | 7 | 0.71 | *hisR* | His tRNA (72/77 bp) | Increase *his* operon expression[3] |
| Δ*proA* | 2 | 1.0 | *proB* | 1 bp Del | - |
| | 1 | 0.77 | *argD* | G282D | Reduce ArgD activity |
| | 2 | 0.56 | *argD* | Del (772-774/1221 bp) | Reduce ArgD activity |
| | 4 | 0.43 | *argD* | Q154* | Reduce ArgD activity |
| | 3 | 0.34 | *argD* | G49R | Reduce ArgD activity |
| Δ*proB* | 6, 7 | 0.83, 0.72 | *glnA* | F463L | Reduce GlnA activity |
| | 5 | 0.86 | *glnA* | D187E | Reduce GlnA activity |
| | 4 | 0.45 | *glnA* | G179C | Reduce GlnA activity |
| | 4 | 0.23 | *glnA* | H172R | Reduce GlnA activity |
| | 1 | 1.0 | *glnA* | G171S | Reduce GlnA activity |
| | 2 | 0.86 | *glnA* | E156D | Reduce GlnA activity |
| | 3 | 1.0 | *glnA* | S148F | Reduce GlnA activity |
| Δ*carA* | 7 | 0.27 | *carA/ carB* | Intergenic (+2/-16) | Increase Expression *carB* |
| | 5 | 0.86 | *carB* | L11L | - |
| | 3, 7 | 0.31, 0.81 | *rpoS* | E96* | - |
| | 3, 7 | 1.0 | 508 genes [*insD6-insD1*] | 520 kbp, 2X GDA | Increase Expression *carB* |
| Δ*metC* | 4 | 0.55 | *malI* | 10bp Dup (227/1029 bp) | Reduce MalI activity[4] |
| | 2 | 0.32 | *malI* | Q55* | Reduce MalI activity[4] |
| | 3 | 0.55 | *malX* | Q529Q | - |
| Δ*metL* | 2, 4 | 0.72, 0.38 | *rpoS* | L317R | - |
| | 4 | 1.0 | *metB* | 1 bp Del (1144/1161 bp) | - |
| Δ*cysK* | 1 | 1.0 | 2,062 genes [*insL3-insL1*] | 2.1 Mbp, 2X GDA | Increase Expression *cysM* |

### 3.2.3 Mutation analysis driven by parallel evolution

The guiding principle for mutation analysis was to identify evidence of parallelism between replicate experiments at the level of genes mutated to determine likely mechanisms of adaptation. Parallel evolution at the gene-level has been demonstrated to provide compelling evidence specific to applied selection pressures (Woods et al. 2006; Bailey, Rodrigue, and Kassen 2015) or, for this study, in response to genetic perturbations. Parallelism was examined by first identifying key mutation events across replicate experiments, such as multiple unique mutations occurring within the same gene or multiple unique mutations in linked metabolic genes and their regulatory elements (Table 3.2). The identification of even a single mutation shared between two samples at the gene level is highly unlikely (Fisher's exact test p-value $< 0.005$). Secondly, these key mutation events were interpreted in the context of model-predicted alternate isozymes and pathways or other experimental studies to further frame potential adaptive evolution events. Model-associated isozymes were identified by examining model gene-protein-reaction associations and model-associated alternate pathways were identified by examining those reactions associated with alternate growth solutions (Table 3.1).

Following growth characterization experiments, genomic DNA was sequenced. Samples were taken from the first flask of growth in minimal medium and prepped for whole genome sequencing (referred to as flask 1 populations, see Materials and Methods). In addition to flask 1 population sample sequencing, the starting inoculation strain grown in nutrient-rich media was sampled and sequenced as a reference for mutation analysis. It is of importance to note that four starting strains isolated from the Keio collection and grown in rich medium ($\Delta carA$, $\Delta cysK$, $\Delta metC$, and $\Delta ptsI$)

contained mutations prior to growth on the defined minimal medium. These base mutations, possibly acquired during the construction of these strains, might have been selected for during growth on the nutrient rich medium and will be addressed later on a case-by-case basis. Of the eleven gene-deletion strains that grew during the long growth tests, two strains ($\Delta cysP$ and $\Delta ubiE$) did not reveal any prevalent mutations in the flask 1 populations that were sequenced and were thus considered non-essential and actually 'True Positive' model predictions (Figure 3.1A). Five ($\Delta thrA$, $\Delta ptsI$, $\Delta serB$, $\Delta proA$, and $\Delta proB$) accrued prevalent mutations in all flask 1 populations sequenced (occurring at a fraction of the total population $> 0.2$ as determined by read-depth) that were not present in the inoculating cultures (Table 3.2). These strains that acquired mutations during growth were considered non-essential with mutations (Figure 3.1A). Four strains ($\Delta cysK$, $\Delta metC$, $\Delta metL$, and $\Delta carA$) showed mutations in some of the flask 1 populations sequenced and were considered non-essential with/without mutations since it appeared that it was possible to attain growth without mutations; however, it was possible that mutations were below the detection criteria (occurring at a fraction $< 0.2$) and the population is highly heterogeneous with many mutations or the mutational events are outside the scope of the computational mutation identification pipeline utilized (e.g., genome rearrangements). In summary, those strains that showed prevalent mutations in some or all population samples sequenced were considered non-essential with mutations, whereas those that showed no mutations were considered non-essential and actually True Positive predictions (Figure 3.1A).

The nature of the mutations that were observed varied in terms of structural or regulatory mutations. Regulatory mutations observed included mutations in intergenic

regions, transcription factors, tRNAs, as well as large regions of genome amplification. Mutations were considered structural if they occurred within the coding region of a metabolic gene. The following sections highlight the diversity and extent of parallel mutation events observed during these extended growth experiments.

### 3.2.4 Mutation enrichment in genetic elements linked to predicted alternate pathways/isozymes

In order to elucidate the mechanism of adaptation for the FP gene-deletion strains, key mutations were analyzed in the context of model-predicted alternate pathways or isozymes (Table 3.1 and Table 3.2). The first few cases highlighted were in excellent agreement with the model-predicted alternate functional pathway. For the $\Delta thrA$ and $\Delta ptsI$ strains, mutations were enriched in intergenic regions that could be linked to model-predicted alternate isozymes (*metL*) or pathways (*galP*) (Figure 3.2A,B). ThrA is annotated as a bifunctional aspartate kinase and homoserine dehydrogenase. The metabolic model for *E. coli*, *i*JO1366, lists MetL as an alternative bifunctional enzyme capable of catalyzing the same reactions (Orth et al. 2011), which is also supported by *in vitro* enzyme assays (Falcoz-Kelly, Rapenbusch, and Cohen 1969). It was thus speculated that the intergenic mutations between *metJ* and *metB* (Figure 3.2A) affect transcription of *metL* (Liu, Blackwell, and States 2001). Furthermore, the mutation within the coding region of *metJ*, the transcriptional repressor for various *met* operon genes, was also proposed to influence expression of *metL*. Lastly, in another independent replicate, a genome duplication amplification was also detected (Table 3.2) which included the metL gene and was thus hypothesized to increase *metL* expression.

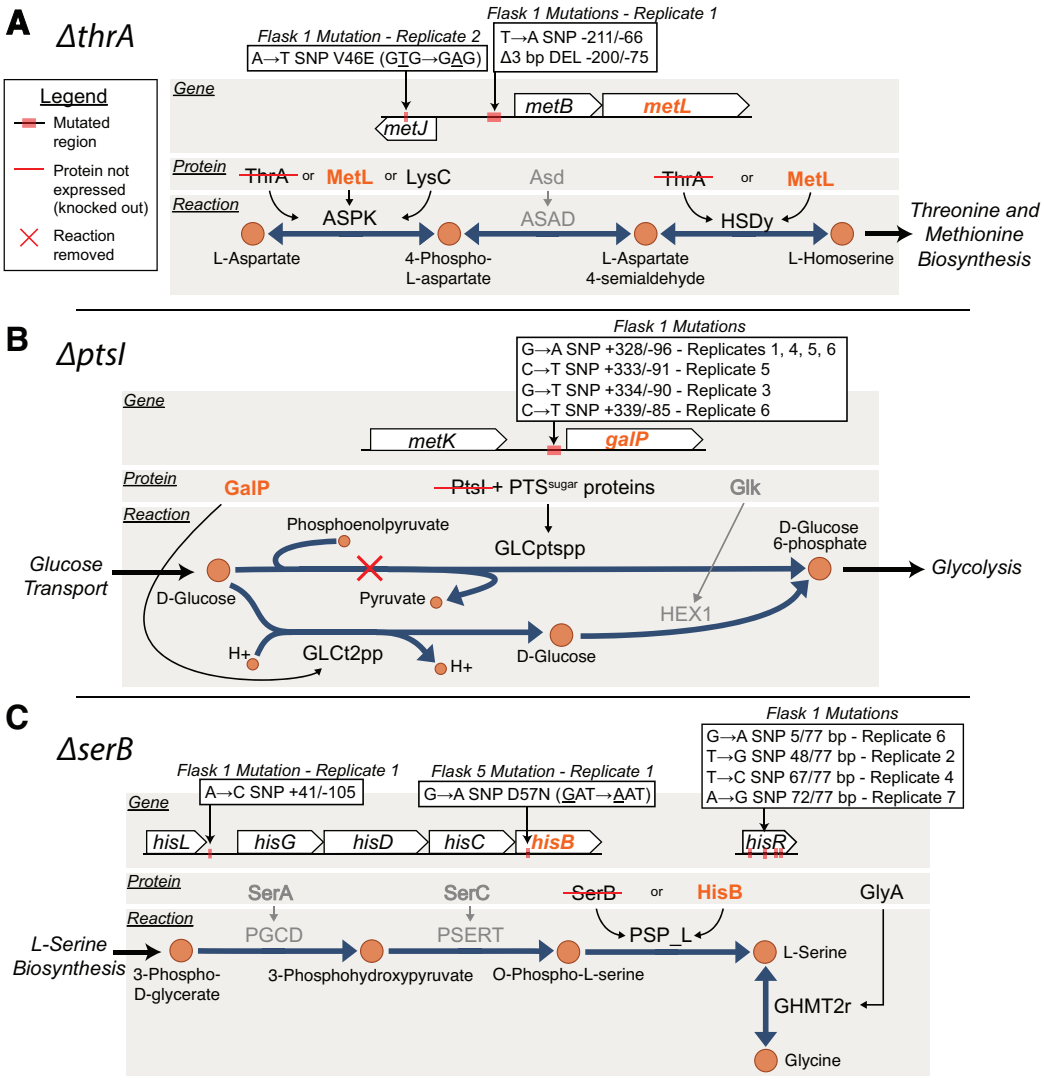


**Figure 3.2**: **Pathway maps related to Δ*thrA*, Δ*ptsI*, and Δ*serB* false positive cases.** Whole genome sequencing analysis revealed that, for two out of three of these cases the model prediction was in agreement with the observed utilized pathway as inferred from mutation analysis. The associated gene-protein-reaction information for each case is highlighted. In A., the mutation results for Δ*thrA* imply that MetL (highlighted in orange) is the enzyme responsible for the isozyme activity as predicted. B. Results for Δ*ptsI* suggest that the predicted alternate pathway related to GalP is utilized in the absence of PtsI. C. Results for Δ*serB* suggest that contrary to the predicted GlyA associated alternate pathway, HisB is responsible for rescuing growth in the absence of SerB.

Another false positive case for which mutations showed strong agreement with model-predictions was *ptsI* (Figure 3.2B). PtsI is part of the well-characterized phosphoenolpyruvate:sugar phosphotransferase system ($PTS^{sugar}$) (Postma, Lengeler, and Jacobson 1993; Chauvin, Brand, and Roseman 1996; Ginsburg and Peterkofsky 2002). This system is responsible for the phosphorylation and transport of various carbohydrate substrates including glucose; however, *E. coli* contains an alternative system for glucose transport and phosphorylation linked to *galP* in *i*JO1366 (Figure 3.2B). GalP is a proton symporter involved in galactose transport, but it has also been shown to transport glucose in *ptsG* and *ptsM* mutants (Henderson, Giddens, and Jones-Mortimer 1977). The mutation evidence observed in the $\Delta$*ptsI* experiments in this study was suggestive of D-glucose transport via the *galP* alternate pathway, as predicted by *i*JO1366 model simulations. Seven mutations were observed in the intergenic region upstream of *galP* in five independent replicate experiments (Table 3.2). Of these seven mutation events, four were identical at the nucleotide level of mutation (Table 3.2), thus demonstrating a high degree of parallel evolution for these replicate experiments and implicating that these intergenic mutations played an important role in adaptation to the *ptsI* perturbation. The likelihood of getting the same mutation at the nucleotide level in two independent samples is even less likely than at the gene level (Fisher's exact test $P < 5e\text{-}06$). The mutation event was suggested to be associated with increased expression of *galP* via reduced repression by the transcriptional repressor *GalP* based on binding site analysis (El Qaidi et al. 2009). Other mutations observed across the replicate $\Delta$*ptsI* populations were in CRP (cyclic-AMP regulatory protein). Six unique *crp* mutations were observed in five replicate experiments (Table 3.2). CRP is known to regulate the transcription of

approximately 100 genes, including *galP*, and it is activated by binding cyclic-AMP (cAMP) (Fic et al. 2009; Latif et al. 2016; Kim et al. 2018). Thus, the mutations observed could be linked to influencing the expression of *galP*. It is of interest to note, however, that a deleterious *cyaA* mutation was observed in the Keio parent strain used to inoculate all experiments. The mutation observed was a seven base-pair deletion leading to the truncation of the CyaA (cyclic-AMP synthase) protein, reducing it from 848 amino acids to 485 amino acids. CyaA activity is important for the activation of the regulator CRP (Franchini, Ihssen, and Egli 2015; Peterkofsky, Svenson, and Amin 1989) and it is thus likely that this deletion event influenced these growth study results.

While $\Delta ptsI$ strains showed widespread agreement in the locations that accrued mutations, the $\Delta metC$ strains showed mutations in only some of the populations sequenced (Table 3.2, no mutations were detected in Experiment #1). For $\Delta metC$, the model predicted that *malY* could compensate for the gene-deletion. The $\Delta metC$ populations showed mutations in *malI*, a regulatory protein that represses expression of *malY* (Reidl and Boos 1991; Reidl et al. 1989). Thus, we hypothesize that the two mutations observed in independent replicate experiments are likely responsible for increasing *malY* expression.

Mutation analysis for the remaining cases did not show as clear of an agreement with model predictions. The mutations observed in the $\Delta carA$ strains did not agree with the model-predicted alternate pathway; however, it did suggest agreement with previous multi-copy suppression results associated with over-expression of *carB* (Patrick et al. 2007a) (Table 3.1). One mutation observed in a $\Delta carA$ population was in the *carA*/*carB* intergenic region, suggesting a regulatory effect on the expression of *carB*.

Another was a synonymous mutation in the coding region of *carB*, indicating possible selective pressure for the use of *carB*, even though the protein sequence did not change. Lastly, there was also a genome duplication event observed in some of the replicate experiments which included *carB*.

For the Δ*serB* experiments, distinct protein-coding, intergenic, and tRNA non-coding mutations were observed that could be linked to increasing the expression and possibly the activity of an isozyme, HisB. This isozyme relationship was not included in the *i*JO1366 reconstruction; however, an alternate pathway for L-serine biosynthesis linked to *glyA* was predicted to suppress the *serB* deletion (Table 3.1, Figure 3.2C). Mutations observed in the replicate experiments, however, did not appear to be associated with this alternate pathway (Table 3.1, Figure 3.2C). Previous work has shown that plasmid over-expression of *hisB*, *gph*, and *ytjC* individually could rescue a *serB* knockout strain (Patrick et al. 2007a). Furthermore, directed evolution experiments have identified mutations in the corresponding enzymes (HisB, Gph, and YtjC) that could improve the isozyme activities that rescue a *serB* deletion (Yip and Matsumura 2013). One such mutation, a D57N HisB protein change (Yip and Matsumura 2013), was also observed in a flask 5 clonal sample (i.e., a clone taken after several passages of the starting strain) in this study (Figure 3.2C).

Parallel mutations linked to the regulation of the histidine operon were also observed in Δ*serB* flask 1 populations (Figure 3.2C). Previous work has supported an attenuator model of regulation for the histidine operon (Frunzio, Bruni, and Blasi 1981; Johnston et al. 1980; Di Nocera et al. 1978; Artz and Broach 1975). Transcription of the histidine operon is believed to be dependent on the secondary structure of a lead mRNA (intergenic region between *hisL* and *hisG*), which is affected by the translation

of a histidine-rich lead peptide (*hisL*). One key mutation observed was found in the *hisL*/*hisG* intergenic region (Figure 3.2C), likely increasing transcription of histidine operon genes (including HisB) by directly affecting the attenuator region. Four other replicate experiments, however, accrued four distinct mutations in *hisR*, a non-coding histidine tRNA (Table 3.2). Specifically, three (out of four) of these mutations were found in the acceptor-stem of $tRNA^{His}$–a region of tRNA important for recognition by aminoacyl-tRNA synthetases (aaRS) (Tian et al. 2015; Jahn, Rogers, and Söll 1991) and for proper cleavage of the pre-tRNA transcripts (Holm and Krupp 1992; Kirsebom and Svärd 1992). Moreover, $tRNA^{His}$ position A71 interacts with multiple residues of Histidyl-tRNA synthetase (HisRS) (Tian et al. 2015) and the A–>G (72/77nt) mutation found in replicate 7 has been shown to decrease the cleavage precision of pre-tRNAs by *E. coli* ribonuclease P (Holm and Krupp 1992). Previous studies have demonstrated that mature $tRNA^{His}$ can attenuate the transcription of the his operon genes (Frunzio, Bruni, and Blasi 1981; Johnston et al. 1980; Di Nocera et al. 1978; Artz and Broach 1975). Thus, the *hisR* mutations observed in the replicate experiments in this work are speculated to reduce the amount of mature $tRNA^{His}$ and its attenuator behavior upon the his operon by decreasing the efficacy of pre-$tRNA^{His}$ cleavage and amino-acylation, allowing for increased HisB expression. Overall, the highly reproducible mutations observed in this study appear to be linked to increasing expression and possibly the side-activity of HisB, a histidinol phosphatase which can also perform the phosphoserine phosphatase function of SerB (Yip and Matsumura 2013).

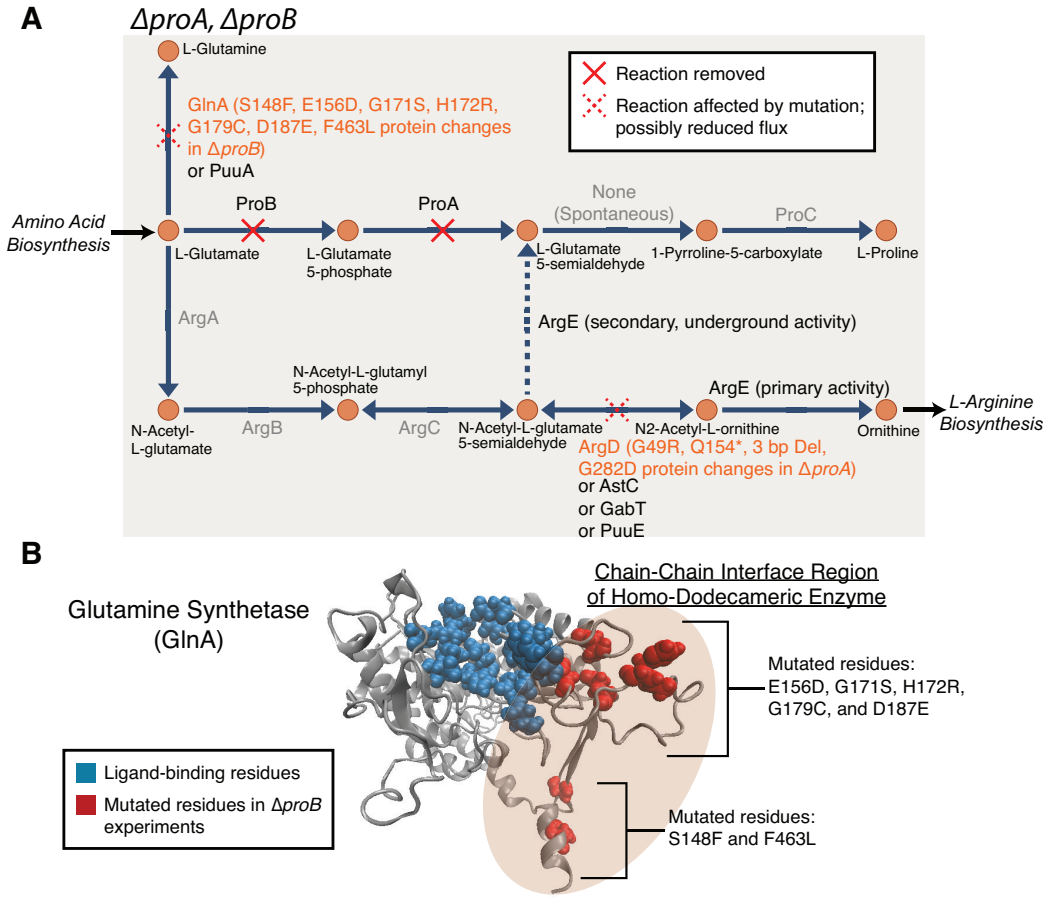


**Figure 3.3**: **Structural mutations observed in Δ*proA* and Δ*proB* experiments analyzed in relation to ArgE underground activity.** A. Metabolic pathway maps related to Δ*proA* and Δ*proB* false positive cases. Both are involved in L-proline synthesis. Model simulations predict using an alternate pathway related to arginine and ornithine synthesis to rescue a *proA*/*proB* deficient *E. coli* strain. Mutations were observed in the coding regions of the metabolic genes *argD* and *glnA*. It is suggested that reduced flux through these enzymes, increases flux through the ArgE associated underground activity, thus increasing production of L-proline and allowing for cell growth. B. Mutation analysis in relation to the glutamine synthetase (GlnA) protein structure. An I-Tasser-predicted protein structure is provided (Yang et al. 2015) and the amino acid residue associated with observed *glnA* mutations in the Δ*proB* populations are highlighted in red. Those residues associated with ligand binding based on the crystal structure of the *Salmonella typhimurium* GlnA enzyme (Gill and Eisenberg 2001) are highlighted in blue. The mutations appear to be in buried regions of the homo-dodecameric enzyme at the interface of chain-chain interactions.

### 3.2.5 Structural mutations are indirectly linked to an underground activity

The mutations observed in the Δ*proA* and Δ*proB* growth screen experiments could not be directly linked to the predicted alternate gene *argE* as in those FP cases previously discussed; however, analysis suggested that the mutations are indirectly related to the suppression of a *proA* or *proB* deletion phenotype. ProA and ProB are enzymes involved in the first two steps of proline biosynthesis in *E. coli* K-12 (Figure 3.3A). Previous work in *Salmonella typhimurium* and *E. coli* strains (Berg and Rossi 1974; Kuo and Stocker 1969; Itikawa, Baumberg, and Vogel 1968) have suggested that an underground activity of the ArgE enzyme in *E. coli*, typically involved in the arginine biosynthesis pathway, can catalyze the conversion of N-acetyl-L-glutamate-5-semialdehyde to L-glutamate 5-semialdehyde (Figure 3.3A) (D'Ari and Casadesús 1998). This side activity of ArgE does not typically occur at a significant enough level to rescue a *proA*/*proB* KO strain unless a mutation in *argD* occurs. The proposed mechanism of suppression is that a mutation inactivates ArgD activity leading to sufficient build up of the N-acetyl-L-glutamate-5-semialdehyde metabolite such that the underground activity of ArgE becomes significant (Itikawa, Baumberg, and Vogel 1968; D'Ari and Casadesús 1998). Thus, the four parallel mutation events in *argD* observed in the Δ*proA* replicate experiments are in agreement with these prior reports (Table 3.2). One observed mutation in this study was predicted to significantly affect ArgD activity by interfering with substrate binding since residues 283 and 284 have been identified as ligand binding residues and the mutation observed was of glycine 282 changing to aspartate. Other mutations observed in replicate experiments included a three base-pair deletion and introduction of an early stop codon (Figure 3.3A), also

likely to reduce ArgD activity.

While the mutations observed in the Δ*proA* experiments seemed to agree with previous reports, mutations observed in Δ*proB* appeared to be novel, but still related to the underground activity of ArgE. Eight mutations in seven independent Δ*proB* replicate experiments occured in the coding region of *glnA*, a glutamine synthetase encoding gene (Figure 3.3, Table 3.2). When the mutated amino acid residues are highlighted on a predicted GlnA protein structure (I-Tasser structure prediction (Yang et al. 2015)), they appear to be clustered in two distinct regions of the GlnA chain (Figure 3.3B). The mutations do not appear to be directly changing ligand-binding residues based on analysis of corresponding ligand-binding residues of a *Salmonella typhimurium* GlnA enzyme (Gill and Eisenberg 2001). The mutations appear to be in regions that are highly buried and involved in chain-chain interactions of the homo-dodecameric enzyme (Figure 3.3B). These mutations are likely to have some effect on GlnA enzyme activity. If GlnA activity were reduced, it is suggested that a larger L-glutamate pool could increase flux through the ArgABCE pathway (Figure 3.3A) and suppress the *proB* deletion. In summary, mutation analysis for the Δ*proA* and Δ*proB* experiments suggested distinct adaptive mechanisms indirectly related to the low-level, underground activity of the ArgE enzyme. This alternate pathway is in agreement with model predictions made with *i*JO1366.

### 3.2.6 Genome amplification events

A mechanism of adaptation observed in several gene KO growth experiments was genome amplification. This type of adaptation was observed most clearly in some of the Δ*cysK*, Δ*ptsI*, and Δ*carA* replicate experiments. Examination of the functional

significance of these large duplication events is statistically less compelling compared to small mutation events (Fisher's exact test resulting in larger p-values depending on size of duplication). However, although these large regions of amplification often contain hundreds of genes, their impact can sometimes be linked to a key gene of interest (Andersson and Hughes 2009; Guzmán et al. 2015; Mundhada et al. 2017). The largest duplication observed was in a $\Delta cysK$ flask 1 population (Table 3.2). CysK is a PLP-dependent enzyme involved in L-cysteine biosynthesis and has been annotated as a cysteine synthase and L-cysteine desulfhydrase (Kredich and Tomkins 1966; Boronat et al. 1984). There are multiple cysteine-desulfhydrases suspected for *E. coli* K-12 including *cysM*, *metC*, *tnaA*, and *malY* (Awano et al. 2005). CysM was the isozyme listed in *i*JO1366 and predicted to rescue a *cysK* knockout strain. Mutations in the flask 1 population sample for the *cysK* knockout strain did not reveal any clear key small mutations; however, a large region of genome amplification of 2X multiplicity was observed based on read-depth analysis (Figure 3.4A). The region spanning approximately 2 million basepairs, or slightly less than half of the genome, does contain the gene encoding the model-predicted alternate isozyme, *cysM*. The region of amplification was flanked by IS186 insertion elements (Figure 3.4A). We propose that these repetitive IS element sequences were instrumental in the mechanism of duplication by recombination as has been previously described (Andersson and Hughes 2009). Although *cysM* was included within the large region of amplification, we propose further follow-up studies conducting expression and/or knockout analysis in order to make more definitive claims of the alternate pathway used to compensate for the *cysK* gene deletion.

Two instances of genome duplication occurred in replicate experiments of $\Delta carA$

and Δ*ptsI* (Figure 3.4B, C). For the Δ*carA* populations, a region of approximately 520 kilo base pairs was duplicated in two replicate experiments (Table 3.2). This region of amplification was flanked by IS2 insertion elements. The region that was amplified does include the *carB* gene that has been shown to suppress a *carA* gene deletion in previous work (Patrick et al. 2007a) (Figure 3.4B). Thus, similar to the *cysK* case, the repetitive IS2 element seems to mediate the amplification and increased dosage of the enzyme encoded by *carB* (Figure 3.4B). Unlike the *cysK* and *carA* genome amplification events, the *ptsI* amplification was significantly smaller (99 kilo base pairs) and flanked by genes encoding ribosomal RNA (*rrlC* and *rrsC* on one side and *rrlA* and *rrsA* on the other) (Figure 3.4C). The gene pairs *rrlA* and *rrlC*, and *rrsA* and *rrsC* each share 99% sequence identity according to BLAST (basic local alignment search tool) alignment analysis (Altschul et al. 1990). Thus, these repetitive sequence regions are potential targets for duplication by recombination as is observed with IS elements (Andersson and Hughes 2009). Although this region did not contain the model-predicted gene of interest, *galP*, it did contain the gene *cyaA*, which encodes an adenylate cyclase. Adenylate cyclase is responsible for the synthesis of cyclic AMP, which is an important signaling molecule, and as previously mentioned, important for activation of the regulator CRP (Kim et al. 2018; Franchini, Ihssen, and Egli 2015; Peterkofsky, Svenson, and Amin 1989). Thus, this amplification appeared to be indirectly related with affecting expression/regulation of *galP*.

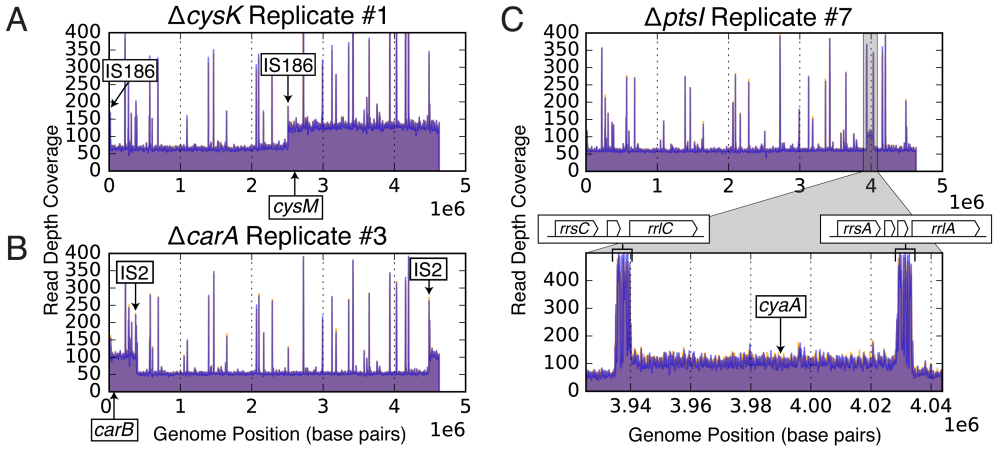


**Figure 3.4**: **Genome duplication amplification events observed in Δ*cysK*, Δ*carA*, and Δ*ptsI* experiments.** Read depth coverage (y-axis) is plotted against the genome position (x-axis) for flask 1 population samples for A. Δ*cysK*, B. Δ*carA*, C. and Δ*ptsI* experiments. For Δ*cysK* and Δ*carA* samples (on the left), the regions of amplification are flanked by IS elements, IS186 and IS2, respectively. These regions of amplification contain the genes (*cysM* and *carB*) associated with a model-predicted alternate isozyme (*cysM*) or a previously reported multi-copy suppressor (*carB*). For the Δ*ptsI* experiment (on the right), a smaller region is amplified and the this region is zoomed in on in the bottom plot. This region of amplification is flanked by ribosomal RNA genes and the identified metabolic gene of interest within this region is *cyaA*.

### 3.2.7 False Positive strains requiring no mutations for growth, or, True Positives

For those strains that did not acquire detectable mutations during the long growth experiments, Δ*ubiE* and Δ*cysP*, it is assumed that only regulatory responses were required to shift expression of alternate metabolic pathways and enable growth. For the case of Δ*ubiE*, however, the drastic reduction in final cell density observed in replicate experiments suggests that the associated reactions involved in ubiquinone and menaquinone biosynthesis are important for cellular energetics (Lee et al. 1997). Previous work has shown that *ubiE* mutant strains can grow using demethylmenaquinone as the sole respiratory quinone (Wissenbach, Ternes, and Unden 1992). Although reported to be important during anaerobic growth, demethylmenaquinone

was observed to have a small but significant capacity to function during aerobic growth as well (Sharma et al. 2012). Furthermore, previous high throughput growth screens show inconsistencies in labeling $\Delta ubiE$ as essential, probably due to cell density cut-offs utilized to label growth/ no growth (Baba et al. 2006; Feist et al. 2007; Monk et al. 2017). Overall, the results of this study (Figure 3.1B, Table 3.1) are consistent with prior reports and show that *ubiE* is non-essential for growth on glucose minimal medium. For the case of *cysP*, it is possible that the predicted alternate pathway (Table 3.1) could be used to enable growth and only regulatory shifts already wired in the wild-type strain are required. Detailed analysis of these regulatory shifts were not pursued in this study; however, future work could examine expression (RT-qPCR, RNAseq) of model-predicted alternate pathways, following workflows similar to those previously reported (Guzmán et al. 2015). These cases are no longer considered false positive model predictions, but instead were true positive predictions in agreement with the model.

## 3.3 Discussion

This study utilized a systematic model-driven approach to identify genes that were mistakenly labeled as essential in minimal media, as well as interpret and suggest mechanisms of adaptation to such genetic perturbations when combined with growth experiments and whole genome sequencing. Three key findings were supported by the results. Firstly, extended growth tests of gene KO strains were shown to result in the reversal of several calls of essentiality, in agreement with model predictions. This finding has direct implications to high-throughput screens of essentiality. Secondly, it was demonstrated that mutation events are likely even after relatively short incubation

times in response to genetic perturbations. Finally, results showed that analysis of parallel mutation events among replicate experiments have implications for expanding gene-protein-reaction associations in both knowledge bases and models.

Growth/ no growth calls made by large-scale growth screens of gene-deletion strain collections such as the Keio collection (Baba et al. 2006) serve as a comprehensive guide for strain are key to testing the predictive power of genome-scale metabolic models (Bordbar et al. 2014; Orth and Palsson 2012). Cellular acclimation to such genetic and metabolic disruptions may require greater time to make such growth/no growth calls as mechanisms of adaptation or regulatory responses might be required for detectable growth. Extended growth incubation of false positive KO strains in this study revealed that 55% (11 out of 20) of the false positive strains available and confirmed could be considered true positives (both experiments and predictions in agreement with calls of non-essentiality). Growth of the examined false positive KO strains was highly reproducible given growth conditions that were well-aerated and provided sufficient time to allow for extended lag-phases. Thus, this study outlines a quantitative time window in which high-throughput growth screens can be designed to call growth/no growth phenotypes going forward (Baba et al. 2006; Joyce et al. 2006; Feist et al. 2007). There was a great deal of phenotypic diversity observed for the different KO strains that grew sub-optimally (as compared to wild type) and this diversity is manifested in the different mechanistic responses of the cells as revealed through mutations.

Coupling population sequencing with extended growth tests in this study revealed that mutation events of interest were likely, even within a period of incubation as short as 48 hours. The false positive strains that were considered for this study

were ultimately placed in one of three categories of essentiality: conditionally essential, non-essential, and non-essential with mutations. Those strains that were repeatedly able to grow given longer periods of incubation were considered non-essential (with or without mutations) and thus in agreement with model predictions of growth (i.e., reassigned as true positives). Of these strains, mechanisms of adaptation to genetic perturbation were categorized broadly as either requiring mutations or not requiring mutations for growth. Population sequencing and mutation analysis of the false positive KO strains revealed that 82% (9 out of 11) of the strains that grew accrued mutations in at least some of the replicate population samples sequenced (Figure 3.1). This result is of general interest as short-term growth screens are commonly practiced with the assumption that mutations are not acquired during such short periods of growth. For those populations that did not accrue mutations, it is suggested that the annotated alternate pathways or isozymes listed in the genome-scale reconstruction and model of metabolism utilized in this study were likely correct. However, such confirmation was not the focus of this study and future work could examine this more comprehensively by performing additional cellular measures such as expression analysis of the predicted isozymes, as has been previously demonstrated (Guzmán et al. 2015), or a complementary approach such as ribosomal sequencing (Ingolia et al. 2012). Furthermore, it is also of general interest to note that some starting Keio strains grown and isolated on a nutrient rich medium possessed mutations that may have influenced growth on the minimal medium. There is strong evidence for this in the *ptsI* KO strain in the Keio collection. Given the wide usage of such gene-deletion libraries, it is important to understand baseline mutations and how they may influence downstream applications.

Examination of mutational parallelism at the gene level proved to be informative and provided compelling contextual evidence for correlation to modeling predictions in a number of the gene KO cases examined. For those strains that did require mutations, the mutations observed across replicate experiments (Table 3.2) allowed for the identification of proposed alternate pathways. Six of the nine cases examined ($\Delta thrA$, $\Delta ptsI$, $\Delta proA$, $\Delta proB$, $\Delta metC$, and $\Delta cysK$) showed key mutations that were interpreted to be in agreement with the model-predicted alternate pathways, thus allowing us to label them as newly assigned true positives. Mutation enrichment across replicate experiments has been shown to provide strong evidence that they were positively selected for (Woods et al. 2006; Bailey, Rodrigue, and Kassen 2015), and this coupled with previously reported data provided the basis for the proposed mechanisms of adaptation described in this study. The establishment of causality for each gene KO strain in detail, however, will require follow-up experiments isolating individual mutants and conducting more detailed experimental analysis as has been previously demonstrated (Utrilla et al. 2016). The results and mutations identified here are the starting reagents for such studies. Furthermore, there are additional key mutations which were identified to display parallel evolution (e.g., *metK* in the $\Delta metL$ strain) whose mechanism of adaptation was not immediately obvious and such cases provide additional targets for discovery (see Table 3.2).

## 3.4 Conclusions

Adaptive flexibility is critical for organisms evolving to novel ecological niches or responding to environmental stress. When examining gene essentiality for such applications as drug discovery or modifying industrial bioprocessing strains, one must

consider possible unanticipated adaptive mechanisms that may follow the intended genetic disruption. Underlying enzymatic side activities may rise to the surface after short adaptive periods leading to unwanted 'rogue' activities (Notebaart et al. 2017). This study shows that while high-throughput, short-term growth screens may capture a large-scale picture of gene essentiality, they may not reveal underlying metabolic capabilities attainable with slightly longer incubation or short adaptive periods. Furthermore, these findings suggest that many of the strains in large gene-deletion collections, such as the Keio collection, likely contain adaptive mechanisms to overcome the intended KO. Thus, sequencing is likely necessary prior to using such clones for the myriad of applications they enable. In conclusion, the results presented in this study highlight genetic and metabolic flexibility in response to gene disruption in the organism of *E. coli*. Furthermore, genome-scale reconstructions and metabolic models provide a promising avenue for the elucidation of adaptive mechanisms and for predicting observable *in vivo* phenotypes.

## 3.5 Materials and Methods

### 3.5.1 False Positives Selection and In silico Model Validation

The false positive strains identified for longer growth tests were taken from the previously published work (Orth and Palsson 2012). Those strains were the subset of genes considered in this study. They were described as false positive predictions on at least one substrate examined and had no experimental growth on any of 34 substrates experimentally tested (Orth and Palsson 2012). However, upon further examination, it was observed that $\Delta cysK$ and $\Delta cysP$ did have experimental evidence of growth on

glucose carbon source (Baba et al. 2006; Feist et al. 2007). These two cases were thus examined using a glycerol substrate on which they were still considered false positive predictions (Joyce et al. 2006).

The false positive predictions were verified as growth predictions *in silico* by utilizing the comprehensive metabolic reconstruction of *E. coli* K-12, *i*JO1366 (Orth et al. 2011). The flux balance analysis (FBA) simulations were conducted using the constraint-based modeling package COBRApy (Ebrahim et al. 2013). Simulations were conducted by optimizing the core biomass objective function, which is determined to be a stoichiometric representation of all core metabolic biomass components in the cell (Feist et al. 2007). To simulate a gene-deletion growth screen and thereby closely mimic experimental growth conditions, the desired gene was removed from the metabolic model and then a FBA simulation was run as previously described (Orth and Palsson 2012), setting the glucose (or glycerol) exchange reaction lower bound to -10 mmol $\cdot$ $gDW^{-1}h^{-1}$ (gDW is an abbreviation of gram of dry weight) and the oxygen exchange reaction lower bound to -1000 mmol $\cdot$ $gDW^{-1}h^{-1}$. All gene-deletion simulations were verified to result in a prediction of growth in agreement with previous reports (Orth and Palsson 2012).

Model-predicted alternate isozymes or alternate pathways listed in Table 1 were determined based on gene-protein-reaction associations listed in the *i*JO1366 model (for isozymes) or based on verification of alternate growth solutions using alternate reactions. Alternate reactions were identified by examining alternate pathways required for synthesis of the essential biomass component related to the gene knockout as described in an existing knowledge base (EcoCyc (Keseler et al. 2013)) and confirming that the model-predicted growth solution was associated with flux through a corresponding

model reaction (listed in Table 1). Alternate isozymes and pathways are also listed and described in a previous publication (Orth and Palsson 2012).

### 3.5.2 Strains Utilized and PCR verification

All strains utilized in this study were taken from the single-gene deletion Keio collection (Baba et al. 2006). These strains are all derived from the parent Keio strain *E. coli* K-12 BW25113. The reference strain utilized in growth screens and wherever 'wild type' is specified in this manuscript, the parent Keio strain without any deletions or Kanamycin resistance cassette was utilized.

The strains utilized in the growth screens were first verified by polymerase chain reaction (PCR) experiments utilizing the methods detailed in (Baba et al. 2006). For each strain that was used, they were verified by three PCR reactions utilizing 1)flanking primers, 2) internal K1 and forward flanking primer, and 3) internal K2 and reverse flanking primer as previously suggested in (Baba et al. 2006).

### 3.5.3 Culture Conditions and Growth Characterizations

Rich media utilized for pre-culture growth was Luria-Bertani Broth (LB). LB media consisted of an autoclaved 25 g/L LB Broth (EMD Millipore LB Broth, Miller - Novagen, catalog 71753) in Milli-Q water. The M9 minimal media utilized in the long term growth characterizations consisted of 0.1mM $CaCl_2$ , 2mM $MgSO_4$, 1x Trace elements Solution, 1x M9 salts solution, and either 2g/L glucose or 0.2% (by volume glycerol), in Milli-Q water. The 4000x trace elements solution consisted of 27 g/L $FeCl_3 \cdot 6H_2O$, 1.3 g/L $ZnCl_2$, 2 g/L $CoCl_2 \cdot 6H_2O$, 2 g/L $Na_2MoO_4$*$2H_2O$, 0.75 g/L $CaCl_2$, 0.91 g/L $CuCl_2 \cdot 2H_2O$, and 0.5 g/L $H_3BO_3$, in concentrated HCl. The 10x

M9 salt solution was composed of 68 g/L $Na_2HPO_2$, 30 g/L $KH_2PO_2$, 5 g/L NaCl, and 10 g/L NH4Cl, in Milli-Q water. The M9 media, trace elements solution, and M9 salt solutions were all sterile filtered. Except for the BW25113 wild-type strain, all LB and M9 cultures contained 25 mg/L Kanamycin A.

The twenty strains that were available in the Keio collection and PCR verified, were selected for an initial long-term growth test. Pre-cultures of these strains were grown overnight in 2-3 mLs of LB media in a 10 mL culture tube on a shaker plate. The following morning, 50 mL M9 minimal media cultures in 250-mL Erlenmeyer flasks containing magnetic stir bars for aeration were inoculated at a target OD600 of 0.01-0.02. The OD600 was monitored at least once a day for two weeks or until growth was observed, at which point the cells were passed to a new flask of M9 minimal media to ensure that the growth observed persisted. The cells were passed consecutively to 5 flasks to ensure the growth observed was consistent. At the end of the experiment, glycerol stock samples were frozen at -80 °C for future use and the flask 5 population was PCR validated as described above to ensure there was no contamination.

Following the initial growth screen, a more detailed growth characterization was conducted on an automated platform. Initial 15 mL LB pre-cultures were inoculated from glycerol frozen stocks of the Keio Knockout Collection strains and the Keio Knockout parent strain BW25113, and were grown overnight. Growth test cultures were then started in triplicate by pipetting 50 $\mu$L from a preculture into three 17 mL tubes containing M9 minimal media. Both the pre-cultures and growth tests were grown at 37 °C in magnetically stirred tubes, at a rate of 1,100 rpm to ensure full aeration. Optical density at 600 nm (OD600) sampling was performed using an automated system with a Tecan Sunrise Microplate Reader, using 100 $\mu$L of culture

for each measurement. Sampling frequency was initially between 6-12 hours, and was increased to 2-4 hours once growth was observed. The OD600 data was then converted to units of grams of dry weight per liter (gDW/L) using the conversion factor for the plate reader and sample volume (1.663 gDW/L/OD600). The growth curves depicted in Figure 3.1B were constructed by importing cell density data from the experiment into a Jupyter notebook (http://jupyter.org/) and utilizing the scientific computing library suite SciPy (http://www.scipy.org/).

### 3.5.4 Whole Genome Sequencing and Mutation Analysis

Genomic DNA was isolated using a Macherey-Nagel NucleoSpin Tissue Kit. DNA concentrations were determined using a Thermo Fischer Qubit dsDNA HS Assay Kit. Paired-end whole genome DNA sequencing libraries were prepared using a Kapa Biosystems KAPA HyperPlus Kit. Manufacturer protocols were followed for all kits. DNA sequencing libraries were then run on a Illumina HiSeq4000 platform with a 100/100 HiSeq 3000/4000 PE cycle kit (PE-410-1001).

The *breseq* pipeline (Deatherage and Barrick 2014) version 0.30.0 with bowtie2 version 2.2.6 was used to map sequencing reads and identify mutations relative to the *E. coli* BW25113 genome (NCBI accession CP009273.1). Mutations considered for analysis in this study were present in a population at a fraction $> 0.2$. Those mutations listed in Table 3.2, were further filtered so the key mutations discussed in the results are presented. Additionally, analysis of large regions of genome amplification (GDAs) was performed by analyzing read depth coverage utilizing a custom python script.

Chapter 3 is a version of a manuscript under review at *BMC Systems Biology*: Guzmán, G. I., Olson, C. A., Hefner, Y., Phaneuf, P., Catoiu, E., Crepaldi, L. B.,

Goldschmidt Micas, L., Palsson, B. O., Feist, A. M. (2017) "Reframing essentiality in terms of adaptive flexibility". The dissertation author was the primary author of the manuscript and was responsible for the research.

# Chapter 4

# Enzyme promiscuity shapes evolutionary innovation and optimization

## 4.1 Introduction

Understanding how novel metabolic pathways arise during adaptation to environmental changes remains a central issue in evolutionary biology. The prevailing view is that enzymes often display promiscuous (i.e., side or secondary) activities and evolution takes advantage of such pre-existing weak activities to generate metabolic novelties(Jensen 1976; Khersonsky and Tawfik 2010; Nam et al. 2012; Notebaart et al. 2014; Huang et al. 2012; Voordeckers et al. 2012; Näsvall et al. 2012; Schmidt et al. 2003; Copley 2000). However, it remains poorly explored how and at what evolutionary stages enzyme side activities contribute to environmental adaptations. Do genetic elements associated with promiscuous activities mutate mostly in the initial

'innovation' stage of adaptation when the population acquires the ability to grow on a new nutrient source(Copley 2000; Mortlock 2013) (i.e., innovation) or do they also contribute to improving fitness in subsequent stages (i.e., optimization)(Barrick and Lenski 2013)? Innovations have been linked to beneficial mutations that endow an organism with novel capabilities such as the ability to use a new carbon source and expand into a new ecological niche(Barrick and Lenski 2013; Wagner 2011). This is distinct from optimizations associated with mutations that improve upon the initial innovation. It is often observed that the mutations accrued within this optimization phase produce gradual benefits in fitness(Barrick and Lenski 2013). Typically, enzyme promiscuity has been linked to the innovation phase, for which mutations enhancing secondary activities may result in dramatic phenotypic improvements(Khersonsky and Tawfik 2010; Barrick and Lenski 2013). In this work, we demonstrate that enzyme promiscuity can be linked to fitness benefits in both the innovation and optimization stages of adaptive evolution.

A second open question concerns our ability to predict the genetic basis of adaptive evolution(Papp, Notebaart, and Pál 2011). There has been an increasing interest in studying empirical fitness landscapes to assess the predictability of evolutionary routes(Visser and Krug 2014). However, these approaches assess predictability only in retrospect and there is a need for computational frameworks that forecast the specific genes that accumulate mutations based on mechanistic knowledge of the evolving trait. A recent study suggested that a detailed knowledge of an organism's promiscuous reaction set (the so-called 'underground metabolism'(D'Ari and Casadesús 1998)) enables the computational prediction of genes that confer new metabolic capabilities when overexpressed(Notebaart et al. 2014). However, it remains to be tested whether

this approach can also accurately predict the adaptive mutations that occur in an evolving population of cells where many alternative adaptive routes may lead to the same phenotype. In this study, we address these issues by performing controlled laboratory experiments to adapt *E. coli* to novel carbon sources and by monitoring the temporal dynamics of adaptive mutations.

## 4.2 Results and Discussion

### 4.2.1 Experimental evolution of non-native carbon source utilizations

The non-native carbon sources explored in this study were selected based on enzymatic side activities computationally predicted to enable growth. This was accomplished by using a comprehensive network reconstruction of underground metabolism shown to predict novel functional states *in vivo* when the predicted enzyme side activity is overexpressed using plasmids(Notebaart et al. 2014). By adding a subset of underground reactions to the comprehensive metabolic reconstruction for *E. coli* K-12 MG1655, *i*JO1366(Orth et al. 2011), novel substrates were computationally identified to be tested experimentally (Table S1).

Adaptive evolution experiments were conducted in two distinct phases: first, an 'innovation'(Copley 2000; Mortlock 2013) stage during which cells acquired mutations to grow on the non-native carbon sources and, second, an 'optimization'(Barrick and Lenski 2013) stage during which a strong pressure was placed to select for the fastest growing cells on the novel carbon sources (Figure 4.1A).

During the initial innovation stage of laboratory evolution experiments (Fig-

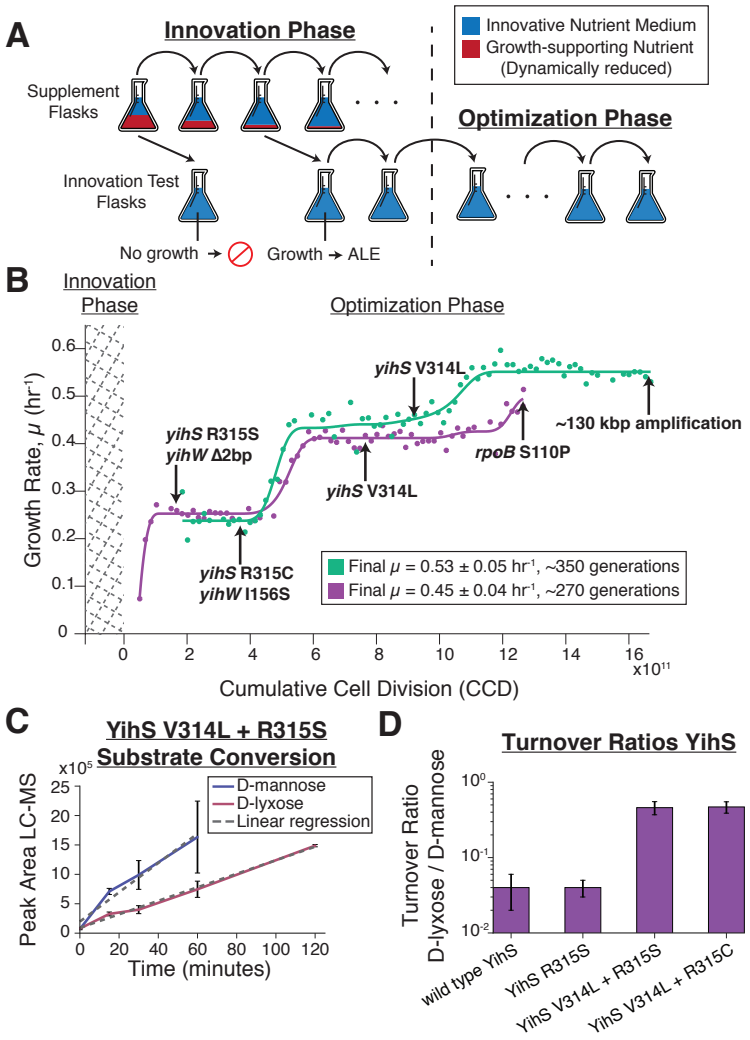


**Figure 4.1**: **Laboratory evolution method schematic and D-lyxose experiments.** A) A schematic of the two-part adaptive laboratory evolution (ALE) experiments. The innovation phase involved growing cells in supplemented flasks containing the innovative substrate (blue) and growth-promoting supplement (red). As cultures were serially passed, they were split into another supplemented flask as well as an 'innovation test flask' containing only the innovative nutrient to test for the desired evolved growth phenotype. The 'optimization' phase consisted of selecting for the fastest growing cells and passing in mid log phase. B) Growth rate trajectories for duplicate experiments (green and purple) for the example case of D-lyxose. Population growth rates are plotted against cumulative cell divisions. Clones were isolated for whole genome sequencing at notable growth-rate plateaus as indicated by the arrows. Mutations gained at each plateau are highlighted beside the arrows (mutations arising earlier along the trajectory persisted in later sequenced clones). C) YihS V314L + R315S mutant enzyme activity on D-mannose and D-lyxose. LC-MS was used to analyze YihS activity at saturating substrate concentrations to compare turnover rates on each substrate. Product formation was followed over time at a constant enzyme concentration. Turnover rates were calculated using linear regression. D) Turnover ratios of substrate conversion of D-lyxose / D-mannose are shown for the wild type YihS and mutant YihS enzymes. A ratio <1 indicates a higher turnover rate on D-mannose compared to D-lyxose. Error bars represent standard error calculated from the linear regression analysis.

ure 4.1A, see *SI Materials and Methods*), *E. coli* was successfully adapted to grow on five non-native substrates, specifically, D-lyxose, D-2-deoxyribose, D-arabinose, m-tartrate, and monomethyl succinate. Duplicate laboratory evolution experiments were conducted in batch growth conditions and in parallel on an automated laboratory evolution (ALE) platform using a protocol that uniquely selected for adaptation to conditions where the ancestor (i.e., wild-type) is unable to grow (Figure 4.1A)(LaCroix et al. 2015). In the innovation phase, *E. coli* was weaned off a growth-supporting nutrient (glycerol) onto novel substrates (Figure 4.1A, Table S2). Clones were isolated and sequenced shortly after the innovative growth phenotype was achieved and mutations analyzed for their associated causality (Figure 4.1B, Fig. S1, Dataset S1). Additional substrates were also chosen that did not result in successful laboratory evolution experiments (i.e., strains growing solely on the novel substrate). This could be attributed to various experimental and biological factors such as experimental duration limitations, the requirement of multiple mutation events, or stepwise adaptation events, as observed in an ethylene glycol adaptation study(Szappanos et al. 2016). There was no obvious pattern to these substrates which are listed in Table S1.

### 4.2.2 Underground metabolism accurately predicts the genes mutated during innovation

Strong signs of parallel evolution were observed at the level of mutated genes in replicate evolution experiments. Such parallelism provides evidence of the beneficial nature of the observed mutations and is a prerequisite for predicting the genetic basis of adaptation. Mutations detected in the evolved isolated clones for each experiment demonstrated a striking agreement with such predicted 'underground'

**Table 4.1**: **Key Innovative Mutations** [1]Substrate binding information (Itoh et al. 2008).[2]Protein family information in the Pfam database(Finn et al. 2016). [3]Binding regions found on EcoCyc(Keseler et al. 2013) based on (Huerta and Collado-Vides 2003).

| Gene Mutated | Substrate | Gene Prediction | Protein Change(s) (Experiment #) | Perceived Impact | Structural or Regulatory (S or R) |
|---|---|---|---|---|---|
| *yihS* | D-Lyxose | *yihS* | R315S(1) | Substrate binding[1] | S |
| | | | R315C(2) | Substrate binding[1] | S |
| | | | V314L(1 and 2) | Substrate binding[1] | S |
| *yihW* | D-Lyxose | *yihS* | Frameshift(1) | Loss of function, large truncation | R |
| | | | I156S(2) | - | R |
| *rbsK* | D-2-Deox. | *rbsK* | N20Y(1) | - | S |
| *rbsR* | D-2-Deox. | *rbsK* | Insertion Sequence(1) | Loss of function, increased *rbsK* expression | R |
| 181 kbp and 281 kbp Regions | D-2-Deox. | *rbsK* | Large Amplification(1) | Increased gene expression | R |
| *fucR* | D-Arabinose | *rbsK* | D82Y(1) | Pfam: DeoRC C terminal substrate sensor domain[2] | R |
| | | | S75R(1 and 2) | Pfam: DeoRC C terminal substrate sensor domain[2] | R |
| | | | *244C(2) | - | R |
| *dmlA* | m-Tartrate | *dmlA* | A242T(1) | - | S |
| *dmlR/dmlA* | m-Tartrate | *dmlA* | intergenic -50/-53(2) | sigma 70 binding: -10 of dmlRp3 promoter[3] | R |
| | | | intergenic -35/-68(2) | dmlRp3 promoter region[3] | R |
| *ybfF/seqA* | Mon. Succ. | *ybfF* | intergenic -73/-112(1) | sigma 24 binding: -35 of ybfFp1 promoter[3] | R |
| | | | intergenic -51/-123(2) | sigma 24 binding: -10 of ybfFp1 promoter[3] | R |

utilization pathways(Notebaart et al. 2014). Specifically, for four out of the five different substrate conditions, key mutations were linked to the predicted enzyme with promiscuous activity, which would be highly unlikely by chance ($P < 10^{-8}$, Fisher's exact test), (Table 4.1, Fig. S2). Not only were the specific genes (or their direct regulatory elements) mutated in 4 out of 5 cases, but few additional mutations (0-3 per strain, Dataset S1) were observed in the initial innovation phase, indicating that the innovations required only one or two mutational steps to activate the predicted growth phenotype and the method utilized was highly selective.

### 4.2.3 Mechanistic insights into metabolic innovations

In general, key innovative mutations could be categorized as regulatory (R) or structural (S) (Table 4.1). Of the sixteen mutation events outlined in Table 4.1, eleven were categorized as regulatory (observed in all five substrate conditions) and five were categorized as structural (three of five substrate conditions). For D-lyxose, D-2-deoxyribose, and m-tartrate evolution experiments, mutations were observed within the coding regions of the predicted genes, namely yihS, rbsK, and dmlA (Table 4.1, Figs. S3-S5). Regulatory mutations, occurring in transcriptional regulators or within intergenic regions–likely affecting sigma factor binding and transcription of the predicted gene target–were observed for D-lyxose, D-2-deoxyribose, m-tartrate, and monomethyl succinate (Table 4.1). Observing more regulatory mutations is broadly consistent with previous reports(Mortlock 2013; Toll-Riera et al. 2016). Regulatory mutations are believed to increase the expression of the target enzyme, thereby increasing the dose of the typically low-level side activity(Guzmán et al. 2015). This observation is consistent with 'gene sharing' models of promiscuity and adaptation

where diverging mutations that alter enzyme specificity are not necessary to acquire the growth innovation(Guzmán et al. 2015; Piatigorsky et al. 1988). Structural mutations are believed to improve the enzyme side activity to achieve the innovation, and this effect was experimentally verified.

The effects of structural mutations on enzyme activity were examined for the YihS isomerase enzyme that was mutated during the D-lyxose evolution (Figure 4.1B, Table 4.1). The activities of the wild-type YihS and three mutant YihS enzymes (YihS R315S, YihS V314L + R315C, and YihS V314L + R315S) were tested in vitro. A cell-free in vitro transcription and translation system(Shimizu et al. 2001; Raad et al. 2017) was used to express the enzymes and examine conversions of D-mannose to D-fructose (a primary activity(Itoh et al. 2008)) and D-lyxose to D-xylulose (side activity) (Figure 4.1C, Fig. S6). The ratios of the turnover rates of D-lyxose to the turnover rates of D-mannose were calculated and compared (Figure 4.1D). The double mutant YihS enzymes showed approximately a ten-fold increase in turnover ratio of D-lyxose to D-mannose compared to wild type ($P < 0.0003$, ANCOVA). These results suggest that the mutations indeed shifted the affinity towards the innovative substrate (enzyme side activity), while still retaining an overall preference for the primary substrate, D-mannose (ratio <1). This is in agreement with 'weak trade-off' theories of the evolvability of promiscuous functions(Khersonsky and Tawfik 2010) in that only a small number of mutations could result in significant improvements in the promiscuous activity of an enzyme without greatly affecting the primary activity.

The causality of the observed key innovative mutations was explored by re-introducing them into the ancestor wild-type strain using a recently developed genome engineering method (pORTMAGE)(Nyerges et al. 2016). Genome editing was per-

formed for screening mutation causality(Herring et al. 2006) on all novel substrate conditions, except for monomethyl succinate, which only contained a single mutation (Table 4.1). Individual mutants were isolated after pORTMAGE reconstruction, and their growth was monitored on the innovative growth medium over the course of one week. The growth test revealed that single mutations were sufficient for growth on D-lyxose, D-arabinose, and m-tartrate (Table S3). In the case of D-2-deoxyribose, an individual mutation was not sufficient for growth, thereby suggesting that the mechanism of adaptation to this substrate is more complex, requiring multiple mutation events (in this case, both regulatory and structural mutations). Overall, these causality assessments support the notion that underground activities open short adaptive paths towards novel phenotypes.

Are the mutations observed in our laboratory experiment relevant for environmental adaptations in the wild? Previous studies have found that predominantly intestinal and extraintestinal strains of *E. coli*, as well as some *Salmonella* species, can use D-2-deoxyribose as a sole carbon source as they possess a pathogenicity island containing the deoxyribokinase deoK(Bernier-Fébreau et al. 2004; Monk et al. 2013; Tourneux et al. 2000). Four such reported pathogenic strains (three *E. coli* and one *Salmonella*)(Bernier-Fébreau et al. 2004; Monk et al. 2013; Tourneux et al. 2000) can grow on D-2-deoxyribose and possesses a deoxyribokinase (DeoK) with a tyrosine residue at the equivalent N20Y position (Fig. S4). This information suggests that the N20Y mutation may have improved the ribokinase underground activity in the strains evolved here on D-2-deoxyribose. Therefore the genetic basis of adaptation observed in the laboratory is indeed relevant to evolution in the wild.

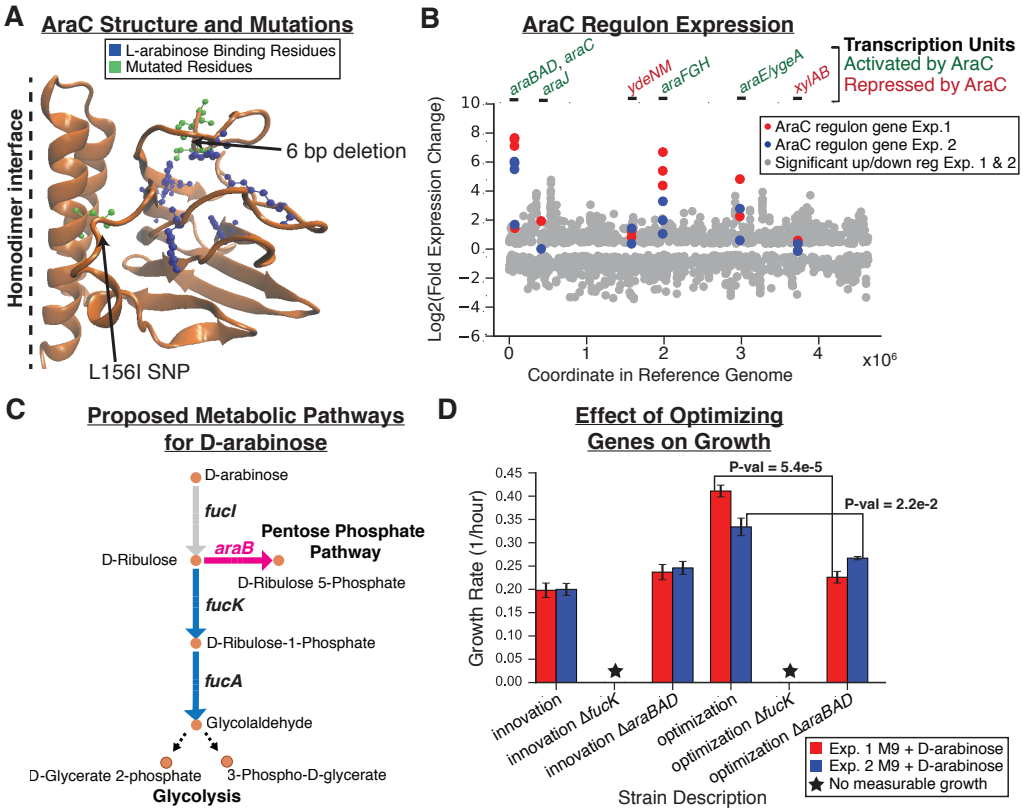


**Figure 4.2**: **Optimization mutation analysis for D-arabinose evolution experiments.** A) Structural mutations observed in sequencing data of Experiments (Exp.) 1 and 2 (green) as well as residues previously identified as important for binding L-arabinose (blue) are highlighted on one chain of the AraC homodimer protein structure. The six base pair deletion observed in Exp. 1 appears to be most clearly linked to affecting substrate binding. B) Expression data (RNAseq) for significantly differentially expressed genes (q-value <0.05). Scatter plot shows log2(fold change) of gene expression data comparing endpoint to initial populations for Exp. 1 and Exp. 2 (grey dots) with the location of the gene in the reference genome as the x-axis. Those genes that are associated with AraC transcription units are highlighted (red dots for Exp. 1 and blue dots for Exp. 2). Above the plot, the transcription units are labeled green if AraC activates expression (in the presence of arabinose) or red if AraC represses expression of those genes. C) The proposed two pathways for metabolizing D-arabinose. The pink pathway is enabled by the optimizing mutations observed in *araC*. D) Growth rate analysis of various innovation (starting point of optimization phase) and optimization (endpoint of optimization phase) strains with or without *fucK* or *araB* genes knocked out. Strains were grown on M9 minimal media with D-arabinose as the sole carbon source.

### 4.2.4 Contribution of enzyme side activities to the optimization phase of adaptation

Once the roles of mutations acquired during the innovation phase were established, adaptive mechanisms required for optimizing or fine-tuning growth on the novel carbon sources were explored. Specifically, of major interest for this study was the role of enzyme promiscuity during this second 'optimization'(Barrick and Lenski 2013) phase of the evolutions. Analysis of mutations in the optimization phase led to identification of additional promiscuous enzyme activities, above and beyond the innovative mechanisms, impacting the phenotypes of the evolved strains in three of the five nutrient conditions. Discovery of these optimizing activities was driven by a systems-level analysis consisting of mutation and transcriptome analyses coupled with computational modeling of optimized growth states on the novel carbon sources.

The 'optimization' phase of the evolution experiments consisted of serially passing cultures in the early exponential phase of growth in order to select for cells with the highest growth rates (Figure 4.1A). Marked and repeatable increases in growth rates on the non-native carbon sources was observed in as few as 180?420 generations (Table S1). Whole genome sequencing of clones was performed at each distinct growth-rate 'jump' or plateau during the optimization phase (Figure 4.1B, Fig. S1). Such plateaus represent regions where a causal mutation has fixed in a population(LaCroix et al. 2015). Out of the total set of 41 mutations identified in the growth optimization regimes (Datasets S1, S2), a subset (Table 4.2) was explored where the same gene was repeatedly mutated in replicate experiments or across all endpoint sequencing data on a given carbon source. To unveil the potential mechanisms for improving growth on the non-native substrates, the transcriptome of initial and endpoint populations (right

after the innovation phase and at the end of the optimization phase) was analyzed using RNAseq. Differentially expressed genes were compared to genes containing optimizing mutations (or their direct targets) and targeted gene deletion studies were performed.

Mutations acquired during the optimization phase leading to large gains in fitness were directly linked to the influence of enzyme promiscuity. The clearest example of an important optimizing mutation was found in the D-arabinose experiments occurring in the araC gene, a DNA-binding transcriptional regulator associated with L-arabinose metabolism(Bustos and Schleif 1993). Based on structural analysis of AraC (Figure 4.2A), the mutations observed in the parallel experiments likely affect substrate binding regions given their proximity to a bound L-arabinose molecule (RCSB Protein Data Bank entry 2ARC)(Soisson et al. 1997), possibly increasing its affinity for D-arabinose. Expression analysis revealed that the araBAD transcription unit associated with AraC regulation(Gama-Castro et al. 2016) was the most highly up-regulated set of genes (expression fold increase ranging from approximately 45-65 for Exp 1 and 140-200 for Exp 2, $P < 10^{-4}$) in both experiments (Figure 4.2B). Further examination of these up-regulated genes revealed that the ribulokinase (AraB) has a similar kcat on four 2-ketopentoses (D/L- ribulose and D/L- xylulose)(Lee, Gerratana, and Cleland 2001) despite the fact that araB is consistently annotated to only act on L-ribulose (EcoCyc)(Keseler et al. 2013) or L-ribulose and D/L-xylulose (BiGG Models)(King et al. 2016). It was thus reasoned that AraB was catalyzing the conversion of D-ribulose to D-ribulose 5-phosphate in an alternate pathway for metabolizing D-arabinose (Figure 4.2C).

The role of the proposed second pathway in optimizing growth on D-arabinose

**Table 4.2**: **Optimizing Mutations** **pyrE* is located in the large region of amplification (first entry of the table).

| Gene Mutated | Substrate | Mutation Type | Proposed Impact | Associated with Underground Activity? |
|---|---|---|---|---|
| 131 kbp Region | D-Lyxose | Amplification | Increased *xylB* expression | No |
| 183 kbp Region | D-2-Deoxyribose | Deletion | Decreased *aldA* expression | Yes |
| *araC* | D-Arabinose | 6 bp Deletion | Increased *araB* expression | Yes |
| *ygbI* | m-Tartrate | 20 bp Deletion, SNP | Increased *ygbJKLMN* expression | Maybe |
| *pyrE* | D-Lyxose*, m-Tartrate | Amplification*, Intergenic | Increased *pyrE* expression | No |

was analyzed both computationally and experimentally. Flux balance analysis of a model utilizing the AraB underground reaction, instead of the FucK associated ribulokinase reaction (pathway of D-arabinose metabolism associated with innovative mutations), resulted in an approximately 10% higher simulated growth rate (Fig. S7). Experimental growth rate measurements of clones carrying either the *fucK* or *araBAD* genes knockouts showed that the FucK enzyme activity was essential for growth on D-arabinose (Figure 4.2D, Table S4). However, removal of *araB* from optimized endpoint strains reduced the growth rate of the strain to the approximate growth rate of the initial innovative strain (Figure 4.2D), suggesting that the proposed *araB* encoded pathway was the primary optimizing adaption responsible for the jump in growth rate. A similar pathway has been described in mutant *Klebsiella aerogens* W70 strains(St Martin and Mortlock 1977), providing further support for the proposed alternate pathway. Overall, underground activities of both the *fuc* operon (innovative mutations) and *ara* operon (optimizing mutations) encoded enzymes appeared to be important for the adaptation to efficiently metabolize D-arabinose and the *ara* mutated operon did not solely support growth. A similar mechanism of amplification of growth enhancing promiscuous activities in the m-tartrate optimization regime of adaptation is also described (Supporting Text, Fig. S8).

### 4.2.5 Loss of an enzyme side activity improves fitness

Finally, analysis of D-2-deoxyribose adaptation revealed a conceptually novel way by which alterations in promiscuous enzyme activities contribute to growth optimization. It is suggested that suppression of a side reaction of aldehyde dehydrogenase A (AldA) enhanced growth on the novel carbon source. Several lines of observation

are consistent with this scenario. The optimizing mutation event was a large deletion event spanning 171 genes (Fig. S9). Of these, the metabolic gene that was most significantly expressed in the ancestor was *aldA* (Fig. S9). AldA has been described as a broad substrate specificity enzyme and has shown catalytic activity on acetaldehyde(Rodríguez-Zavala, Allali-Hassani, and Weiner 2006). Computational modeling showed that forcing increased flux through acetaldehyde to acetate conversion decreased the overall growth rate (Fig. S9, Dataset S3), suggesting a clear mechanism behind the growth-rate enhancing deletion event. This finding demonstrates that not only enhancement, but also suppression of side reactions plays pivotal roles in adaptation to novel environments. Two additional proposed mechanisms for growth optimization on m-tartrate and D-lyxose were related to the primary activities of *pyrE* and *xylB* and are discussed in the *Supporting Text* (Fig. S8 and Fig. S10).

## 4.3 Conclusions

Taken together, the results presented show that enzyme promiscuity is prevalent in metabolism and plays a major role in both phenotypic innovation and optimization. It was demonstrated that enzyme side activities can confer a fitness benefit in two distinct ways. First, side activities contributed to the establishment of novel metabolic routes that enabled or improved the utilization of a new nutrient source. Second, suppression of an undesirable underground activity that diverted flux from a newly established pathway conferred a fitness benefit.

The results of this study have direct relevance for understanding the role of promiscuous enzymatic activities in evolution and for utilizing computational models to predict the trajectory and outcome of molecular evolution(Papp, Notebaart, and

Pál 2011; Lässig, Mustonen, and Walczak 2017). Here, it was demonstrated that computational metabolic network models which include the repertoire of enzyme side activities made it possible to predict the genetic basis of adaptation to novel carbon sources. As such, systems models and analyses are likley to contribute signficanlty towards representing the complex implications of promiscuity in theoretical models of molecular evolution(Lässig, Mustonen, and Walczak 2017).

## 4.4 Materials and Methods

### 4.4.1 *In silico* modeling

The most current version of the genome scale model for *Escherichia coli* K-12 MG1655, *i*JO1366(Orth et al. 2011), was utilized in this study as the base model before adding underground reactions related to the five substrates analyzed as previously reported(Notebaart et al. 2014). The underground reactions previously reported were added to *i*JO1366 using the constraint-based modeling package, COBRApy(Ebrahim et al. 2013). All growth simulations using parsimonious flux balance analysis were conducted using COBRApy. Growth simulations were performed by optimizing the default core biomass objective function (a representation of essential biomass compounds in stoichiometric amounts)(Feist and Palsson 2010). To simulate aerobic growth on a given substrate, the exchange reaction lower bound for that substrate was adjusted to -10 mmol $gDW^{-1}hr^{-1}$.

Sampling was conducted to determine the most likely high flux metabolic pathways for growth on D-2-deoxyribse (Dataset S3). The Artificial Centering Hit-and-Run algorithm, optGpSampler(Megchelenbrink, Huynen, and Marchiori 2014),

was utilized to sample the steady-state solution space. The lower bound of the biomass objective function was set to 90% of the optimum in order to better simulate realistic growth conditions. The number of sample points used was two times the number of reactions in the *i*JO1366 model (5186 sample points) and the step count was set to 25000 in order to ensure a nearly uniformly sampled solution space. Thus for each reaction, a distribution of likely flux states was acquired. The high flux reaction set was pulled out from these distributions as those reactions which had a mean flux greater than 1.0, but less than 100 and whose standard deviation was less than 50. The results from this analysis for D-2-deoxyribose are summarized in Dataset S3 and Fig. S9.

### 4.4.2 Laboratory Evolution Experiments

The bacterial strain utilized in this study as the starting strain for all evolutions and MAGE manipulations was an *E. coli* K-12 MG1655 (ATCC 4706). Laboratory evolution experiments were conducted on an automated platform using a liquid handling robot as previously described(LaCroix et al. 2015; Sandberg et al. 2014). As described above, the experiments were conducted in two phases, an ?innovation? phase and an 'optimization' phase. At the start of the innovation phase, cultures were serially passaged after reaching stationary phase in a supplemented flask containing the non-native carbon source at a concentration of 2 g/liter and the growth-supporting supplement (glycerol) at a concentration of 0.2%. Cultures were passaged in stationary phase and split into another supplemented flask and a test flask containing only the non-native carbon source at a concentration of 2 g/liter. As the innovation phase progressed, the concentration of the growth-supporting nutrient was adjusted to

maintain a target max OD600 (optical density 600 nm) of 0.5 as measured on a Tecan Sunrise plate reader with 100 $\mu$L of sample. This ensured that glycerol was always the growth limiting nutrient. If growth was not observed in the test flask within three days, the culture was discarded; however, once growth was observed in the test flask, this culture was serially passaged to another test flask. Once growth was maintained for three test flasks, the second phase of the evolution experiments commenced - the optimization phase. The optimization phase was conducted as in previous studies(LaCroix et al. 2015; Sandberg et al. 2014). The culture was serially passaged during mid-exponential phase so as to select for the fastest growing cells on the innovative carbon source. The growth rate was monitored by periodically taking OD600 measurements. The evolution experiments were concluded once increases in the growth rate were no longer observed for several passages.

Growth data from the evolution experiments was analyzed with an in-house MATLAB package. Growth rates were calculated for each flask during the 'optimization' phase of the evolution experiments by using a least-squares linear regression. Calculated growth rates were rejected if fewer than three OD measurements were sampled, the range of OD measurements were less than 0.2 or greater than 0.4, or if the R2 correlation was <0.98. Growth-rate trajectory curves (Figure 4.1B, Fig. S1) curves were produced in MATLAB by fitting a monotonically increasing piecewise cubic spline to the data as reported previously(Sandberg et al. 2014; LaCroix et al. 2015). Evolution experiment parameters were also calculated with the MATLAB script (Table S2) including the cumulative number of cell divisions (CCDs), which were calculated as previously described(Lee et al. 2011).

### 4.4.3 Growth Media Composition

All strains were grown in M9 minimal medium. The M9 minimal medium was composed of the carbon source at a concentration of 2 g/L unless otherwise specified (for example, during the 'innovation' phase of the ALE experiments the total amount of carbon source varied as the growth supporting nutrient concentration was dynamically decreased). Carbon sources were purchased from Sigma Aldrich (D-(-)-Lyxose 99% catalog #220477, 2-Deoxy-D-Ribose 97% catalog #121649, D-(-)-Arabinose >=98% catalog #A3131, meso-Tartaric acid monohydrate >=97% catalog #95350, and mono-Methyl hydrogen succinate 95% catalog #M81101). The growth supporting nutrient used was glycerol. Other components of the M9 minimal medium were 0.1 mM $CaCl_2$, 2.0 mM $MgSO_4$, 6.8 g $L^{-1}$ $Na_2HPO_4$, 3.0 g $L^{-1}$ $KH_2PO4$, 0.5 g $L^{-1}$ NaCl, 1.0 g $L^{-1}$ NH4Cl, and trace elements solution. A 4,000x trace element solution consisted of 27 g $L^{-1}$ $FeCl_3 \cdot 6\ H_2O$, 2 g $L^{-1}$ $NaMoO_4 \cdot 2\ H_2O$, 1 g $L^{-1}$ $CaCl_2 \cdot H_2O$, 1.3 g $L^{-1}$ $CuCl_2 \cdot 6\ H_2O$, 0.5 g $L^{-1}$ $H_3BO_3$, and concentrated HCl dissolved in double-distilled $H_2O$ and sterile filtered. The final concentration in the media of the trace elements solution was 1x.

### 4.4.4 Whole Genome Sequencing and Mutation Analysis

Genomic DNA was isolated using the Machery-Nagel Nucleospin Tissue Kit using the support protocol for bacteria provided by the manufacturer user manual. The quality of genomic DNA isolated was assessed using Nanodrop UV absorbance ratios. DNA was quantified using Qubit dsDNA high-sensitivity assay. Paired-end whole genome DNA sequencing libraries were generated utilizing either a Nextera XT kit (Illumina) or KAPA HyperPlus kit (Kapa Biosystems). DNA sequencing libraries

were run on an Illumina Miseq platform with a paired-end 600 cycle v3 kit.

DNA sequencing fastq files were processed utilizing the computational pipeline tool, *breseq* (Deatherage and Barrick 2014) , aligning reads to the *E. coli* K-12 MG1655 genome (NC000913.3) (Datasets S1 and S2). Additionally, identification of large regions of genome amplification were identified using a custom python script that utilizes aligned files to identify regions with more than 2x (minus standard deviation) of mean read depth coverage.

### 4.4.5 Enzyme activity characterization

All enzymes used in this study were generated by cell-free in vitro transcription and translation using the PURExpress in vitro Protein Synthesis Kit (New England Biolabs). Linear DNA templates utilized in all cell-free in vitro transcription and translation reactions were generated by PCR from dsDNA blocks encoding the enzymes with transcription and translations elements synthesized by Integrated DNA Technologies. Linear DNA templates were purified and concentrated using phenol/chloroform extraction and ethanol precipitation. The encoded enzymes were produced using PURExpress according to manufacturer?s protocol with linear DNA templates concentrations of 25 ng/ 1 $\mu$L reaction.

The activities of the wild type YihS and three mutant YihS enzymes towards D-Mannose and D-Lyxose over time was determined using LC/MS. Substrate (10 mM) was added to 7.5 $\mu$L of PURExpress reaction in a buffered solution (50 mM Tris, 100 mM KCl, 10 mM MgCl2, pH 8) for a total volume of 250 $\mu$L and incubated at 37 °C. At different time points (0, 15, 30, 60, 120, 240 and 1320 minutes), 10 $\mu$L samples were taken and quenched with 90 $\mu$L of LC/MS grade ethanol. Next, samples were

dried under vacuum (Savant SpeedVac Plus SC110A) and resuspended in 50 $\mu$L of LC/MS grade methanol/water (50/50 v/v). The samples were filtered through 0.22 $\mu$m microcentrifugal filtration devices and transferred to 384-well plate for LC/MS analysis. An Agilent 1290 LC system equipped with a SeQuant® ZIC® -HILIC column (100 mm x 2.1 mm, 3.5 $\mu$m 200 Å, EMD Millipore) was used for separation with the following LC conditions: solvent A, H2O with 5 mM ammonium acetate; solvent B, 19:1 acetonitrile:H2O with 5 mM ammonium acetate; timetable: 0 min at 100% B, 1.5 min at 100% B, 6 min at 65% B, 8 min at 0% B, 11 min at 0% B, 12.5 min at 100% B and 15.5 min at 100% B; 0.25 mL/min; column compartment temperature of 40 °C. Mass spectrometry analyses were performed using an Agilent 6550 quadrupole time of flight mass spectrometer. Agilent software Mass Hunter Qualitative Analysis (Santa Clara, CA) was used for naïve peak finding and data alignment. Analysis of covariance (ANCOVA) was used to determine if the slopes of mutants for both xylose and mannose are significantly different from the wild type slopes. Detailed instrument information and data are provided in Table S6 and Dataset S6.

### 4.4.6 pORTMAGE Library Construction/Isolation of individual mutants

Mutations were introduced and their corresponding combinations accumulated during the laboratory evolution experiments into the ancestral *E. coli* strain using pORTMAGE recombineering technology (Nyerges et al. 2016). ssDNA oligonucleotides, carrying the mutation or mutations of interest, were designed using MODEST (Bonde et al. 2014) for *E. coli* K-12 MG1655 (ATCC 4706). To isolate individual mutants,

a single pORTMAGE cycle was performed separately with each of the 15 oligos in *E. coli* K-12 MG1655 (ATCC 4706) + pORTMAGE3 (Addgene ID: 72678) according to a previously described pORTMAGE protocol (Nyerges et al. 2016). Following transformation, cells were allowed to recover overnight at 30 °C and were plated to Luria Bertani (LB) agar plates to form single colonies. Presence of each mutation or mutation combinations was verified by High-Resolution Melting (HRM) colony-PCRs with Luminaris HRM Master Mix (Thermo Scientific) in a Bio-Rad CFX96 qPCR machine according to the manufacturer?s guidelines. Mutations were confirmed by capillary-sequencing. pORTMAGE oligonucleotides, HRM PCR and sequencing primers are listed in Dataset S4.

### 4.4.7 RbsK Comparison to DeoK/kinases in other Enterobacteriacae

Protein sequence alignment was conducted for the *E. coli* MG1655 RbsK N20Y mutant sequence from this study and DeoK sequences reported for *E. coli* strains (Bernier-Fébreau et al. 2004; Monk et al. 2013), three pathogenic (AL862, 55989, and CFT073) and one commensal (EC185), as well as the DeoK sequence reported for *S. enterica serovar Typhi* (Tourneux et al. 2000). The sequence alignments were performed using the multiple sequence alignment package, T-Coffee (Notredame, Higgins, and Heringa 2000) (Figure S4).

Enzyme protein sequences of RbsK, YihS, and DmlA modified with the protein changes observed in the ALE whole genome sequencing data were used as input sequences to NCBI?s BLASTp algorithm (Altschul et al. 1997). These sequences were compared to all other protein sequences in other Enterobacteriacae. The resulting

alignments were saved as and analyzed using biopython and the python data analysis library, pandas. Results were filtered to identify sequences which aligned to the observed mutated protein change. For RbsK (the only sequence for which alignments containing tyrosine in the mutated N20Y position were observed) the alignments were further filtered to only include results with an Expect value (e-value) <1E-2 and alignment lengths longer than 200 residues. These filtered results for RbsK are summarized in Dataset S5.

### 4.4.8 Individual mutant growth test

Isolated mutants were tested for growth over the course of one week (Table S3). Individual colonies were isolated on LB agar plates and used to inoculate pre-cultures grown overnight in 2 mL of glucose M9 minimal liquid media in 10 mL tubes. The following morning, pre-cultures were pelleted at 5000 rpm and gently resuspended in M9 minimal medium with no carbon source and this spinning and resuspension was repeated twice to wash the cells of residual glucose. The final resuspension was in 2 mL of M9 minimal medium with no carbon source. The growth test tubes consisting of 2 mL of M9 minimal medium plus the corresponding innovative carbon source were inoculated with the washed cells at a dilution factor of 1:200. Growth was monitored over the course of one week by visually inspecting for increased cellular density. Once growth was observed, colony PCR was conducted (Qiagen HotStarTaq Master Mix Kit) with the primer sequences listed in Table S5. DNA sequencing of PCR products was conducted by Eton Bioscience Inc using their SeqRegular services. DNA sequencing was utilized to confirm the designed mutations were as expected and to confirm that no other mutations had been acquired in the regions of interest during the growth

test.

### 4.4.9 RNA sequencing

RNA sequencing data were generated under conditions of aerobic, exponential growth on M9 minimal medium plus the corresponding innovative carbon source (D-lyxose, D-2-deoxyribose, D-arabinose, or m-tartrate). Cells were harvested using the Qiagen RNA-protect bacteria reagent according to the manufacturer's specifications. Prior to RNA extraction, pelleted cells were stored at -80 °C. Cell pellets were thawed and incubated with lysozyme, SuperaseIn, protease K, and 20% sodium dodecyl sulfate for 20 min at 37 °C. Total RNA was isolated and purified using Qiagen?s RNeasy minikit column according to the manufacturer's specifications. Ribosomal RNA (rRNA) was removed utilizing Ribo-Zero rRNA removal kit (Epicentre) for Gram-negative bacteria. The KAPA Stranded RNA-seq kit (Kapa Biosystems) was used for generation of paired-end, strand-specific RNA sequencing libraries. RNA sequencing libraries were then run on an Illumina HiSeq 2500 using the 'rapid-run mode' with 2 x 35 paired end reads.

Reads were mapped to the E. coli K-12 genome (NC_000913.2) using bowtie2 (Langmead and Salzberg 2012). Cufflinks (Trapnell et al. 2010) was utilized to calculate the expression level of each gene in units per kilobase per million fragments mapped (FPKM). This information was then utilized to run cuffdiff (Trapnell et al. 2013) to calculate gene expression fold change between endpoint and initial growth populations using a geometric normalization and setting a maximum false discovery rate of 0.05. Gene expression fold change was considered significant if the calculated q-value was smaller than 0.05.

### 4.4.10 Metabolic Map Generation and Data Superimposition

All metabolic pathway maps generated in Figure 4.2 and Figs. S8-S10 were generated using the pathway visualization tool Escher (King et al. 2015).

### 4.4.11 Bioscreen growth test of mutants

Individual sequenced clones from the D-arabinose evolution experiments (Exp. 1 and Exp. 2) along with the wild-type *E. coli* K-12 MG1655 strain were utilized for bioscreen growth tests and gene knockout manipulations. These were clones isolated from the initial innovation and optimized endpoint populations (Dataset S1). A P1-phage transduction mutagenesis protocol based on a previously reported method (Donath, Dominguez, and Withers 2011) was followed to replace the *fucK* gene in the evolution and wild-type strains with a Kanamycin resistance cassette from the *fucK Keio* strain (Baba et al. 2006). The BW25113 Keio collection strain is effectively missing the *araBAD* genes, so the *yabI* Keio strain was utilized for the P1-phage transduction of all strains to transfer this neighboring *araBAD* deletion along with the *yabI*-replaced Kanamycin resistance cassette. It was deemed that a *yabI* deletion would not significantly affect the results of the growth experiments since *yabI* is a non-essential inner membrane protein that is a member of the DedA family (Doerrler et al. 2013). *E. coli* K-12 contains seven other DedA proteins and it is only collectively that they are essential (Boughner and Doerrler 2012).

The growth screens were conducted in a Bioscreen-C system machine. Pre-cultures were started from frozen stocks of previously isolated clones and grown overnight in M9 minimal medium + 0.2% glycerol. These pre-cultures were used to inoculate the triplicate bioscreen culture wells at 1:100 dilution of M9 minimal medium

supplemented with either 2g/L D-arabinose or 0.2% glycerol. The final volume for each well was 200 $\mu$l. The growth screen was conducted under continuous shaking conditions at 37 °C. OD600 (optical density at 600 nm) readings were taken every 30 minutes over the course of 48 hours.

Chapter 4 is a version of a manuscript in preparation for submission: Guzmán, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, A., Papp, H., de Raad, M., King, Z. A., Northen, T. R., Notebaart, R. A., Pál, C., Palsson, B. O., Papp, B., Feist, A. M. (2017) "Enzyme promiscuity shapes evolutionary innovation and optimization". The dissertation author was the primary author of the paper and was responsible for the research.

# Chapter 5

# Conclusions and Outlook

## 5.1 Expanding Model-Driven Discovery

The results presented in this dissertation demonstrate the pervasive nature of enzyme promiscuity in metabolism as well as the power of systems biology methods to excavate underground activities in a systematic way. The case studies presented, however, merely scratch at the surface of underground metabolism as it is assumed to be vast. How do we go beyond these case studies? Is it a realistic pursuit to fully elucidate the underground metabolic network of reactions in *E. coli* and other organisms?

### 5.1.1 How deep is the underground?

A previous limitation in the study of enzyme promiscuity was the sensitivity of enzyme characterization. More rigorous and quantitative measures of enzyme side activities and substrate ambiguity could aid in expanding the database of feasible reactions within a cell. Progress has recently been demonstrated in high throughput

characterizations of enzyme activities (Greving et al. 2012; Sévin et al. 2017). Recent progress in mass-spectrometry-based enzyme assays have demonstrated the potential for highly sensitive, high-throughput enzyme characterization. Acoustic deposition methods with nanostructure initiator mass-spectrometry (NIMS) coupled with array imaging-based activity readouts have shown the potential for rapidly characterizing multiple reactions and reaction pathways (Greving et al. 2012). Furthermore, non-targeted *in vitro* metabolomics techniques have shown the potential for identifying previously uncharacterized enzymes in *E. coli* (Sévin et al. 2017). This method consisted of utilizing mass-spectrometry to monitor the accumulation or depletion of metabolites when overexpressed or purified proteins were incubated in a metabolome extract containing hundreds of biologically relevant substrates. The resulting knowledge gained from such techniques could directly aid in expanding the database of enzyme activities and allow for the inclusion of such data in reconstructions of underground metabolism. A version of this type of reconstruction has already proven to be fruitful in predicting evolutionary trajectories in non-native growth environments as was described in Chapter 4 of this dissertation. By including the enzyme activities listed in the BRENDA enzyme database (Placzek et al. 2017), the computational model of underground metabolism (Notebaart et al. 2014) expanded upon the genome scale metabolic model of metabolism to predict attainable phenotypes accessible via underground metabolic routes. Thus, it is my opinion that the current advances in high-throughput metabolomics and enzyme activity assays could be directly leveraged to more completely elucidate the network of underground metabolism.

### 5.1.2 Where might the underground take us?

Within the biotechnology industry, there is a strong desire to produce highly efficient bioprocessing strains (Manzer, Waal, and Imhof 2013). In strain design, it is typical to target specific pathways to be up-regulated or knocked out in order to improve the desired phenotype; however, the biological systems that we work with tend to push back on such manipulations. One can imagine the removal of one pathway resulting in the over-expression of a compensating underground activity, or what may colloquially be referred to as a 'rogue' enzyme activity. These undesired compensating mechanisms, which are related to the adaptive mechanisms of an organism, are a challenge in strain engineering. However, as our knowledge of underground metabolism expands, our ability to anticipate the 'rogue' activities will likely aid in making rapid strain design more successful.

Beyond engineering bioprocessing strains for industrial applications, the themes presented in this dissertation have far-reaching potential in the field of medicine as well. For example, recent work has shown that brain cancer-associated mutations in the cytosolic metabolic enzyme isocitrate dehydrogenase 1 result in its ability to catalyze a reaction other than its primary reaction (Dang et al. 2009). This new reaction leads to the increased production of a metabolite (R(-)-2-hydroxyglutarate) and these elevated levels have been implicated in an increased risk of malignant brain tumors. Thus, it is evident that expanding our understanding of the potential secondary activities enzymes could aid in better analyzing disease states as well as lead to possibly more targeted drugs and treatments.

### 5.1.3 Constraint-Based Modeling and Laboratory Evolution for Discovery

The work presented in this dissertation highlighted the potential for discovery resulting from the coupling of *in silico* computational methods and *in vivo* experimental methods. Constraint-based modeling methods allow us to utilize our knowledge of biological building blocks (genes, proteins, and reactions) to make predictions about observable, measurable phenotypes. Flux balance analysis computes an optimal growth solution; however, experimental conditions are often not optimized for cell growth. Thus, laboratory evolution methods help to achieve these optimal growth states *in vivo*. The comparison can then be made between the computed and adapted cell state to identify knowledge gaps and therefore potential avenues of discovery. The iterative process of model-driven discovery should result in the regular upkeep and improvement of metabolic models, which can then aid in understanding and predicting *in vivo* phenotypes and move towards more complete genome annotations.

Beyond studying the model organism of *E. coli*, the workflows applied in this dissertation could be extended to other organisms. For example, studying relevant adaptive mechanisms in pathogenic strains of bacteria could aid in the understanding of antibiotic resistance. Future avenues of study could also examine the evolution of populations of mixed species and look at the commonly used adaptive mechanism of horizontal gene transfer. I hope that the work presented in this dissertation may serve as an impetus for future discovery efforts.

## 5.2 Conclusion

As we continue to unearth the biological foundations that support the evolution of elegant biological structures, the coupling of computational efforts and experimental studies will greatly benefit our understanding and expand our ability to engineer solutions to industrial and medical challenges. The iterative nature of model-driven discovery provides an avenue for continual improvement of our knowledge base. Current improvements to the sensitivity in the field of mass spectrometry and enzyme characterization, as well as next-generation sequencing technologies, and laboratory evolution experiments coupled with constraint-based modeling methods make the elucidation of underground metabolism a tangible goal. Being able to predict adaptive mechanisms will greatly benefit our ability to engineer bioprocessing strains and anticipate medical and environmental responses to human interventions.

# Bibliography


Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). "Basic local alignment search tool". en. In: *J. Mol. Biol.* 215.3, pp. 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". en. In: *Nucleic Acids Res.* 25.17, pp. 3389–3402.

Andersson, D. I. and Hughes, D. (2009). "Gene amplification and adaptive evolution in bacteria". en. In: *Annu. Rev. Genet.* 43, pp. 167–195.

Artz, S. W. and Broach, J. R. (1975). "Histidine regulation in Salmonella typhimurium: an activator attenuator model of gene regulation". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 72.9, pp. 3453–3457.

Awano, N., Wada, M., Mori, H., Nakamori, S., and Takagi, H. (2005). "Identification and functional analysis of Escherichia coli cysteine desulfhydrases". en. In: *Appl. Environ. Microbiol.* 71.7, pp. 4149–4152.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006). "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection". en. In: *Mol. Syst. Biol.* 2, p. 2006.0008.

Bailey, S. F., Rodrigue, N., and Kassen, R. (2015). "The effect of selection environment on the probability of parallel evolution". en. In: *Mol. Biol. Evol.* 32.6, pp. 1436–1448.

Barrick, J. E. and Lenski, R. E. (2013). "Genome dynamics during experimental evolution". en. In: *Nat. Rev. Genet.* 14.12, pp. 827–839.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., and Kim, J. F. (2009). "Genome evolution and adaptation in a long-

term experiment with Escherichia coli." In: *Nature* 461.7268, pp. 1243–7. DOI: 10.1038/nature08480. URL: http://www.ncbi.nlm.nih.gov/pubmed/19838166.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005). "GenBank." In: *Nucleic acids research* 33.Database issue, pp. D34–8. DOI: 10.1093/nar/gki063. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=540017%5C&tool=pmcentrez%5C&rendertype=abstract.

Berg, C. M. and Rossi, J. J. (1974). "Proline excretion and indirect suppression in Escherichia coli and Salmonella typhimurium". en. In: *J. Bacteriol.* 118.3, pp. 928–934.

Bergthorsson, U., Andersson, D. I., and Roth, J. R. (2007). "Ohno's dilemma: evolution of new genes under continuous selection." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.43, pp. 17004–9. DOI: 10.1073/pnas.0707158104. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2040452%5C&tool=pmcentrez%5C&rendertype=abstract.

Bernier-Fébreau, C., Merle, L. du, Turlin, E., Labas, V., Ordonez, J., Gilles, A.-M., and Le Bouguénec*, C. (2004). "Use of Deoxyribose by Intestinal and Extraintestinal Pathogenic Escherichia coli Strains: a Metabolic Adaptation Involved in Competitiveness". In: *Infect. Immun.* 72.12, pp. 7381–7381.

Bonde, M. T., Klausen, M. S., Anderson, M. V., Wallin, A. I. N., Wang, H. H., and Sommer, M. O. A. (2014). "MODEST: a web-based design tool for oligonucleotide-mediated genome engineering and recombineering". en. In: *Nucleic Acids Res.* 42.Web Server issue, W408–15.

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). "Constraint-based models predict metabolic and associated cellular functions". In: *Nat. Rev. Genet.* 15.2, pp. 107–120. DOI: 10.1038/nrg3643.

Boronat, A., Britton, P., Jones-Mortimer, M. C., Kornberg, H. L., Lee, L. G., Murfitt, D., and Parra, F. (1984). "Location on the Escherichia coli genome of a gene specifying O-acetylserine (thiol)-lyase". en. In: *J. Gen. Microbiol.* 130.3, pp. 673–685.

Boughner, L. A. and Doerrler, W. T. (2012). "Multiple deletions reveal the essentiality of the DedA membrane protein family in Escherichia coli". en. In: *Microbiology* 158.Pt 5, pp. 1162–1171.

Bustos, S. A. and Schleif, R. F. (1993). "Functional domains of the AraC protein". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 90.12, pp. 5638–5642.

Chauvin, F., Brand, L., and Roseman, S. (1996). "Enzyme I: the first protein and potential regulator of the bacterial phosphoenolpyruvate: glycose phosphotransferase system". en. In: *Res. Microbiol.* 147.6-7, pp. 471–479.

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., Fero, M. J., McAdams, H. H., and Shapiro, L. (2011). "The essential genome of a bacterium". en. In: *Mol. Syst. Biol.* 7, p. 528.

Chung, B. K.-S., Dick, T., and Lee, D.-Y. (2013). "In silico analyses for the discovery of tuberculosis drug targets". en. In: *J. Antimicrob. Chemother.* 68.12, pp. 2701–2709.

Copley, S. D. (2000). "Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach". en. In: *Trends Biochem. Sci.* 25.6, pp. 261–265.

Dang, L., White, D. W., Gross, S., Bennett, B. D., Bittinger, M. A., Driggers, E. M., Fantin, V. R., Jang, H. G., Jin, S., Keenan, M. C., Marks, K. M., Prins, R. M., Ward, P. S., Yen, K. E., Liau, L. M., Rabinowitz, J. D., Cantley, L. C., Thompson, C. B., Vander Heiden, M. G., and Su, S. M. (2009). "Cancer-associated IDH1 mutations produce 2-hydroxyglutarate". en. In: *Nature* 462.7274, pp. 739–744.

D'Ari, R. and Casadesús, J. (1998). "Underground metabolism". en. In: *Bioessays* 20.2, pp. 181–186.

Datsenko, K. A. and Wanner, B. L. (2000). "One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products." In: *Proceedings of the National Academy of Sciences of the United States of America* 97.12, pp. 6640–5. DOI: 10.1073/pnas.120163297. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=18686%5C&tool=pmcentrez%5C&rendertype=abstract.

Deatherage, D. E. and Barrick, J. E. (2014). "Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq". en. In: *Methods Mol. Biol.* 1151, pp. 165–188.

D′Elia, M. A., Pereira, M. P., and Brown, E. D. (2009). "Are essential genes really essential?" In: *Trends Microbiol.* 17.10, pp. 433–438.

Di Nocera, P. P., Blasi, F., Di Lauro, R., Frunzio, R., and Bruni, C. B. (1978). "Nucleotide sequence of the attenuator region of the histidine operon of Escherichia coli K-12". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 75.9, pp. 4276–4280.

Doerrler, W. T., Sikdar, R., Kumar, S., and Boughner, L. A. (2013). "New functions for the ancient DedA membrane protein family". en. In: *J. Bacteriol.* 195.1, pp. 3–11.

Donath 2nd, M. J., Dominguez, M. A., and Withers 3rd, S. T. (2011). "Development of an automated platform for high-throughput P1-phage transduction of Escherichia coli". en. In: *J. Lab. Autom.* 16.2, pp. 141–147.

Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". en. In: *BMC Syst. Biol.* 7, p. 74.

El Qaidi, S., Allemand, F., Oberto, J., and Plumbridge, J. (2009). "Repression of galP, the galactose transporter in Escherichia coli, requires the specific regulator of N-acetylglucosamine metabolism". en. In: *Mol. Microbiol.* 71.1, pp. 146–157.

Falcoz-Kelly, F., Rapenbusch, R. van, and Cohen, G. N. (1969). "The methionine-repressible homoserine dehydrogenase and aspartokinase activities of Escherichia coli K 12. Preparation of the homogeneous protein catalyzing the two activities. Molecular weight of the native enzyme and of its subunits". en. In: *Eur. J. Biochem.* 8.1, pp. 146–152.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. (2007). "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information". en. In: *Mol. Syst. Biol.* 3.1, p. 121.

Feist, A. M. and Palsson, B. O. (2010). "The biomass objective function". en. In: *Curr. Opin. Microbiol.* 13.3, pp. 344–349.

Feist, A. M. and Palsson, B. Ø. (2008). "The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli". en. In: *Nat. Biotechnol.* 26.6, pp. 659–667.

Fic, E., Bonarek, P., Gorecki, A., Kedracka-Krok, S., Mikolajczak, J., Polit, A., Tworzydlo, M., Dziedzicka-Wasylewska, M., and Wasylewski, Z. (2009). "cAMP receptor protein from escherichia coli as a model of signal transduction in proteins–a review". en. In: *J. Mol. Microbiol. Biotechnol.* 17.1, pp. 1–11.

Finkel, S. E. (2006). "Long-term survival during stationary phase: evolution and the GASP phenotype." In: *Nature reviews. Microbiology* 4.2, pp. 113–20. DOI: 10.1038/nrmicro1340. URL: http://www.ncbi.nlm.nih.gov/pubmed/16415927.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). "The Pfam protein families database: towards a more sustainable future". en. In: *Nucleic Acids Res.* 44.D1, pp. D279–85.

Fischbach, M. A. and Walsh, C. T. (2009). "Antibiotics for emerging pathogens". en. In: *Science* 325.5944, pp. 1089–1093.

Fotheringham, I. G., Dacey, S. A., Taylor, P. P., Smith, T. J., Hunter, M. G., Finlay, M. E., Primrose, S. B., Parkert, D. M., and Edwardsi, R. M. (1986). "The cloning and sequence analysis of the aspC and tyrB genes from Escherichia coli K12". In: *Biochemistry* 234, pp. 593–604.

Franchini, A. G., Ihssen, J., and Egli, T. (2015). "Effect of Global Regulators RpoS and Cyclic-AMP/CRP on the Catabolome and Transcriptome of Escherichia coli K12 during Carbon- and Energy-Limited Growth". en. In: *PLoS One* 10.7, e0133793.

Frunzio, R., Bruni, C. B., and Blasi, F. (1981). "In vivo and in vitro detection of the leader RNA of the histidine operon of Escherichia coli K-12". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 78.5, pp. 2767–2771.

Furnham, N., Beer, T. A. P. de, and Thornton, J. M. (2012). "Current challenges in genome annotation through structural biology and bioinformatics." In: *Current opinion in structural biology* 22.5, pp. 594–601. DOI: 10.1016/j.sbi.2012.07.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/22884875.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Del Moral-Chávez, V., Rinaldi, F., and Collado-Vides, J. (2016). "RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond". en. In: *Nucleic Acids Res.* 44.D1, pp. D133–43.

Gelfand, D. H. and Steinberg, R. A. (1977). "Escherichia coli mutants deficient in the aspartate and aromatic amino acid aminotransferases." In: *Journal of bacteriology* 130.1, pp. 429–40. URL: http://www.pubmedcentral.nih.gov/

articlerender.fcgi?artid=235221%5C&tool=pmcentrez%5C&rendertype=abstract.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A.-L., Oltvai, Z. N., and Osterman, A. L. (2003). "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655". en. In: *J. Bacteriol.* 185.19, pp. 5673–5684.

Gill, H. S. and Eisenberg, D. (2001). "The crystal structure of phosphinothricin in the active site of glutamine synthetase illuminates the mechanism of enzymatic inhibition". en. In: *Biochemistry* 40.7, pp. 1903–1912.

Ginsburg, A. and Peterkofsky, A. (2002). "Enzyme I: the gateway to the bacterial phosphoenolpyruvate:sugar phosphotransferase system". en. In: *Arch. Biochem. Biophys.* 397.2, pp. 273–278.

Greving, M., Cheng, X., Reindl, W., Bowen, B., Deng, K., Louie, K., Nyman, M., Cohen, J., Singh, A., Simmons, B., Adams, P., Siuzdak, G., and Northen, T. (2012). "Acoustic deposition with NIMS as a high-throughput enzyme activity assay". en. In: *Anal. Bioanal. Chem.* 403.3, pp. 707–711.

Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O., and Feist, A. M. (2015). "Model-driven discovery of underground metabolic functions in Escherichia coli". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.3, pp. 929–934.

Henderson, P. J., Giddens, R. A., and Jones-Mortimer, M. C. (1977). "Transport of galactose, glucose and their molecular analogues by Escherichia coli K12". en. In: *Biochem. J* 162.2, pp. 309–320.

Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., Boom, D. van den, Cantor, C. R., and Palsson, B. Ø. (2006). "Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale". en. In: *Nat. Genet.* 38.12, pp. 1406–1412.

Hildebrand, A., Remmert, M., Biegert, A., and Söding, J. (2009). "Fast and accurate automatic structure prediction with HHpred". In: *Proteins: Structure, Function, and Bioinformatics* 77.S9, pp. 128–132.

Holm, P. S. and Krupp, G. (1992). "The acceptor stem in pre-tRNAs determines the cleavage specificity of RNase P". en. In: *Nucleic Acids Res.* 20.3, pp. 421–423.

Huang, R., Hippauf, F., Rohrbeck, D., Haustein, M., Wenke, K., Feike, J., Sorrelle, N., Piechulla, B., and Barkman, T. J. (2012). "Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109.8, pp. 2966–2971.

Huerta, A. M. and Collado-Vides, J. (2003). "Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals". en. In: *J. Mol. Biol.* 333.2, pp. 261–278.

Hughes, D. and Andersson, D. I. (2016). "Evolutionary Trajectories to Antibiotic Resistance". In: *Annu. Rev. Microbiol.*

Hutchison 3rd, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., Glass, J. I., Merryman, C., Gibson, D. G., and Venter, J. C. (2016). "Design and synthesis of a minimal bacterial genome". en. In: *Science* 351.6280, aad6253.

Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2012). "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments". en. In: *Nat. Protoc.* 7.8, pp. 1534–1550.

Ish-Am, O., Kristensen, D. M., and Ruppin, E. (2015). "Evolutionary Conservation of Bacterial Essential Metabolic Genes across All Bacterial Culture Media". en. In: *PLoS One* 10.4, e0123785.

Itikawa, H., Baumberg, S., and Vogel, H. J. (1968). "Enzymic basis for a genetic suppression: accumulation and deacylation of N-acetylglutamic gamma-semialdehyde in enterobacterial mutants". In: *Biochimica et Biophysica Acta (BBA)-Enzymology* 159.3, pp. 547–550.

Itoh, T., Mikami, B., Hashimoto, W., and Murata, K. (2008). "Crystal structure of YihS in complex with D-mannose: structural annotation of Escherichia coli and Salmonella enterica yihS-encoded proteins to an aldose-ketose isomerase". en. In: *J. Mol. Biol.* 377.5, pp. 1443–1459.

Jahn, M., Rogers, M. J., and Söll, D. (1991). "Anticodon and acceptor stem nucleotides in tRNA(Gln) are major recognition elements for E. coli glutaminyl-tRNA synthetase". en. In: *Nature* 352.6332, pp. 258–260.

Jensen, R. A. (1976). "Enzyme recruitment in evolution of new function". en. In: *Annu. Rev. Microbiol.* 30, pp. 409–425.

Johnston, H. M., Barnes, W. M., Chumley, F. G., Bossi, L., and Roth, J. R. (1980). "Model for regulation of the histidine operon of Salmonella". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 77.1, pp. 508–512.

Joyce, A. R., Reed, J. L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S. A., Palsson, B. Ø., and Agarwalla, S. (2006). "Experimental and computational assessment of conditionally essential genes in Escherichia coli". en. In: *J. Bacteriol.* 188.23, pp. 8259–8271.

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). "Advances in flux balance analysis". In: *Current Opinion in Biotechnology* 14.5, pp. 491–496. DOI: 10.1016/j.copbio.2003.08.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S0958166903001174.

Kelley, L. A. and Sternberg, M. J. (2009). "Protein structure prediction on the Web: a case study using the Phyre server". In: *Nature protocols* 4.3, pp. 363–371.

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A. M., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Schröder, I., Shearer, A. G., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R. P., Paulsen, I., and Karp, P. D. (2013). "EcoCyc: fusing model organism databases with systems biology". en. In: *Nucleic Acids Res.* 41.Database issue, pp. D605–12.

Khersonsky, O. and Tawfik, D. S. (2010). "Enzyme promiscuity: a mechanistic and evolutionary perspective". en. In: *Annu. Rev. Biochem.* 79, pp. 471–505.

Kim, D., Seo, S. W., Gao, Y., Nam, H., Guzman, G. I., Cho, B.-K., and Palsson, B. O. (2018). "Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP". en. In: *Nucleic Acids Res.*

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015). "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". en. In: *PLoS Comput. Biol.* 11.8, e1004321.

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". en. In: *Nucleic Acids Res.* 44.D1, pp. D515–22.

Kirsebom, L. A. and Svärd, S. G. (1992). "The kinetics and specificity of cleavage by RNase P is mainly dependent on the structure of the amino acid acceptor stem". en. In: *Nucleic Acids Res.* 20.3, pp. 425–432.

Koser, S. A. (1923). "Correlation of citrate utilization by members of the colon-aerogenes group with other differential characteristics and with habitat." In: *Journal of bacteriology* 9, pp. 59–77.

Kredich, N. M. and Tomkins, G. M. (1966). "The enzymic synthesis of L-cysteine in Escherichia coli and Salmonella typhimurium". en. In: *J. Biol. Chem.* 241.21, pp. 4955–4965.

Kumar, V. S. and Maranas, C. D. (2009). "GrowMatch: an automated method for reconciling in silico/in vivo growth predictions." In: *PLoS computational biology* 5.3, e1000308. DOI: 10.1371/journal.pcbi.1000308. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2645679%5C&tool=pmcentrez%5C&rendertype=abstract.

Kuo, T. T. and Stocker, B. A. (1969). "Suppression of proline requirement of proA and proAB deletion mutants in Salmonella typhimurium by mutation to arginine requirement". en. In: *J. Bacteriol.* 98.2, pp. 593–598.

LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O., and Feist, A. M. (2015). "Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium". en. In: *Appl. Environ. Microbiol.* 81.1, pp. 17–30.

Langmead, B. and Salzberg, S. L. (2012). "Fast gapped-read alignment with Bowtie 2". en. In: *Nat. Methods* 9.4, pp. 357–359.

Lässig, M., Mustonen, V., and Walczak, A. M. (2017). "Predicting evolution". en. In: *Nature Ecology & Evolution* 1.3, s41559–017–0077.

Latif, H., Federowicz, S., Ebrahim, A., Tarasova, J., Szubin, R., Utrilla, J., Zengler, K., and Palsson, B. (2016). "ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions". en.

Lazcano, A. and Miller, S. L. (1996). "The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time." In: *Cell* 85, pp. 793–798. URL: http://www.ncbi.nlm.nih.gov/pubmed/8681375.

Ledwidge, R. and Blanchard, J. S. (1999). "The Dual Biosynthetic Capability of N -Acetylornithine Aminotransferase in Arginine and Lysine Biosynthesis". In: *Biochemistry* 38, pp. 3019–3024.

Lee, D.-H., Feist, A. M., Barrett, C. L., and Palsson, B. Ø. (2011). "Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of Escherichia coli". en. In: *PLoS One* 6.10, e26172.

Lee, D.-H. and Palsson, B. Ø. (2010). "Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol." In: *Applied and environmental microbiology* 76.13, pp. 4158–4168. DOI: 10.1128/AEM.00373-10. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2897412%5C&tool=pmcentrez%5C&rendertype=abstract.

Lee, L. V., Gerratana, B., and Cleland, W. W. (2001). "Substrate specificity and kinetic mechanism of Escherichia coli ribulokinase". en. In: *Arch. Biochem. Biophys.* 396.2, pp. 219–224.

Lee, P. T., Hsu, A. Y., Ha, H. T., and Clarke, C. F. (1997). "A C-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the Escherichia coli ubiE gene". en. In: *J. Bacteriol.* 179.5, pp. 1748–1754.

Lee-Peng, F.-C., Hermodson, M. A., and Kohlhaw, G. B. (1979). "Transaminase B from Escherichia coli : Quaternary Structure , Amino-Terminal Sequence, Substrate Specificity, and Absence of a Separate Valine-alpha-Ketoglutarate Activity". In: *Journal of bacteriology* 139.2, pp. 339–345.

Liu, R., Blackwell, T. W., and States, D. J. (2001). "Conformational model for binding site recognition by the E.coli MetJ transcription factor". en. In: *Bioinformatics* 17.7, pp. 622–633.

Loo, B. van, Jonas, S., Babtie, A. C., Benjdia, A., Berteau, O., Hyvönen, M., and Hollfelder, F. (2010). "An efficient, multiply promiscuous hydrolase in the alkaline phosphatase superfamily." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.7, pp. 2740–5. DOI: 10.1073/pnas.0903951107. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2840280%5C&tool=pmcentrez%5C&rendertype=abstract.

Manzer, L. E., Waal, J. C. van der, and Imhof, P. (2013). "The Industrial Playing Field for the Conversion of Biomass to Renewable Fuels and Chemicals". In: *Catalytic Process Development for Renewable Materials*. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 1–24.

McCloskey, D. and Palsson Bernhard Øand Feist, A. M. (2013). "Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli." In: *Molecular systems biology* 9.661, p. 661. DOI: 10.1038/msb.2013.18. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3658273%5C&tool=pmcentrez%5C&rendertype=abstract.

McCloskey, D., Palsson, B. Ø., and Feist, A. M. (2013). "Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*". In: *Mol. Syst. Biol.* 9.1, p. 661. DOI: 10.1038/msb.2013.18.

Megchelenbrink, W., Huynen, M., and Marchiori, E. (2014). "optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks". en. In: *PLoS One* 9.2, e86587.

Mobegi, F. M., Zomer, A., Jonge, M. I. de, and Hijum, S. A. F. T. van (2017). "Advances and perspectives in computational prediction of microbial gene essentiality". en. In: *Brief. Funct. Genomics* 16.2, pp. 70–79.

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M., and Palsson, B. Ø. (2013). "Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.50, pp. 20338–20343.

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M., and Palsson, B. O. (2017). "iML1515, a knowledgebase that computes Escherichia coli traits". en. In: *Nat. Biotechnol.* 35.10, pp. 904–908.

Mortlock, R. (2013). *Microorganisms as Model Systems for Studying Evolution*. en. Springer Science & Business Media.

Mundhada, H., Seoane, J. M., Schneider, K., Koza, A., Christensen, H. B., Klein, T., Phaneuf, P. V., Herrgard, M., Feist, A. M., and Nielsen, A. T. (2017). "Increased production of L-serine in Escherichia coli through Adaptive Laboratory Evolution". en. In: *Metab. Eng.* 39, pp. 141–150.

Nam, H., Lewis, N. E., Lerman, J. A., Lee, D.-H., Chang, R. L., Kim, D., and Palsson, B. O. (2012). "Network context and selection in the evolution to enzyme specificity". en. In: *Science* 337.6098, pp. 1101–1104.

Näsvall, J., Sun, L., Roth, J. R., and Andersson, D. I. (2012). "Real-time evolution of new genes by innovation, amplification, and divergence". en. In: *Science* 338.6105, pp. 384–387.

Nemoto, N., Kurihara, S., Kitahara, Y., Asada, K., Kato, K., and Suzuki, H. (2012). "Mechanism for regulation of the putrescine utilization pathway by the transcription factor PuuR in Escherichia coli K-12." In: *Journal of bacteriology* 194.13, pp. 3437–47. DOI: 10.1128/JB.00097-12. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3434745%5C&tool=pmcentrez%5C&rendertype=abstract.

Newman, J., Seabrook, S., Surjadi, R., Williams, C. C., Lucent, D., Wilding, M., Scott, C., and Peat, T. S. (2013). "Determination of the structure of the catabolic N-succinylornithine transaminase (AstC) from Escherichia coli." In: *PloS one* 8.3, pp. 1–11. DOI: 10.1371/journal.pone.0058298. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3590144%5C&tool=pmcentrez%5C&rendertype=abstract.

Nichols, R. J., Sen, S., Choo, Y. J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K. M., Lee, K. J., Wong, A., Shales, M., Lovett, S., Winkler, M. E., Krogan, N. J., Typas, A., and Gross, C. A. (2011). "Phenotypic landscape of a bacterial cell". en. In: *Cell* 144.1, pp. 143–156.

Notebaart, R. A., Szappanos, B., Kintses, B., Pal, F., Gyorkei, A., Bogos, B., Lazar, V., Spohn, R., Csorg, B., Wagner, A., Ruppin, E., Pal, C., and Papp, B. (2014). "Network-level architecture and the evolutionary potential of underground metabolism". In: *Proceedings of the National Academy of Sciences* 111.32, pp. 11762–11767.

Notebaart, R. A., Kintses, B., Feist, A. M., and Papp, B. (2017). "Underground metabolism: network-level perspective and biotechnological potential". en. In: *Curr. Opin. Biotechnol.* 49, pp. 108–114.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment". en. In: *J. Mol. Biol.* 302.1, pp. 205–217.

Nyerges, Á., Csörgő, B., Nagy, I., Bálint, B., Bihari, P., Lázár, V., Apjok, G., Umenhoffer, K., Bogos, B., Pósfai, G., and Pál, C. (2016). "A highly precise and portable

genome engineering method allows comparison of mutational effects across bacterial species". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.9, pp. 2502–2507.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. Ø. (2011). "A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011". en. In: *Mol. Syst. Biol.* 7, p. 535.

Orth, J. D. and Palsson, B. (2012). "Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions". en. In: *BMC Syst. Biol.* 6, p. 30.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). "What is flux balance analysis?" In: *Nat. Biotechnol.* 28.3, pp. 245–248. DOI: 10.1038/nbt.1614.

Papp, B., Notebaart, R. A., and Pál, C. (2011). "Systems-biology approaches for predicting genomic evolution". en. In: *Nat. Rev. Genet.* 12.9, pp. 591–602.

Patrick, W. M., Quandt, E. M., Swartzlander, D. B., and Matsumura, I. (2007a). "Multicopy suppression underpins metabolic evolvability". en. In: *Mol. Biol. Evol.* 24.12, pp. 2716–2722.

— (2007b). "Multicopy suppression underpins metabolic evolvability". In: *Molecular biology and evolution* 24.12, pp. 2716–22. DOI: 10.1093/molbev/msm204. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2678898%5C&tool=pmcentrez%5C&rendertype=abstract.

Patton, A. J., Hough, D. W., Towner, P., and Danson, M. J. (1993). "Does Escherichia coli possess a second citrate synthase gene?" In: *European journal of biochemistry / FEBS* 214.1, pp. 75–81. URL: http://www.ncbi.nlm.nih.gov/pubmed/8508809.

Peterkofsky, A., Svenson, I., and Amin, N. (1989). "Regulation of Escherichia coli adenylate cyclase activity by the phosphoenolpyruvate:sugar phosphotransferase system". en. In: *FEMS Microbiol. Rev.* 5.1-2, pp. 103–108.

Piatigorsky, J., O'Brien, W. E., Norman, B. L., Kalumuck, K., Wistow, G. J., Borras, T., Nickerson, J. M., and Wawrousek, E. F. (1988). "Gene sharing by delta-crystallin and argininosuccinate lyase". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 85.10, pp. 3479–3483.

Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., and Schomburg, D. (2017). "BRENDA in 2017: new perspectives and new tools in BRENDA". en. In: *Nucleic Acids Res.* 45.D1, pp. D380–D388.

Postma, P. W., Lengeler, J. W., and Jacobson, G. R. (1993). "Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria". en. In: *Microbiol. Rev.* 57.3, pp. 543–594.

Powell, J. T. and Morrison, J. F. (1978). "Role of the Escherichia coli aromatic amino acid aminotransferase in leucine biosynthesis." In: *Journal of bacteriology* 136.1, pp. 1–4. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=218624%5C&tool=pmcentrez%5C&rendertype=abstract.

Raad, M. d., Modavi, C., Sukovich, D. J., and Anderson, J. C. (2017). "Observing Biosynthetic Activity Utilizing Next Generation Sequencing and the DNA Linked Enzyme Coupled Assay". en. In: *ACS Chem. Biol.* 12.1, pp. 191–199.

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., and Palsson, B. O. (2006). "Systems approach to refining genome annotation." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.46, pp. 17480–4. DOI: 10.1073/pnas.0603364103. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1859954%5C&tool=pmcentrez%5C&rendertype=abstract.

Reidl, J. and Boos, W. (1991). "The malX malY operon of Escherichia coli encodes a novel enzyme II of the phosphotransferase system recognizing glucose and maltose and an enzyme abolishing the endogenous induction of the maltose system". en. In: *J. Bacteriol.* 173.15, pp. 4862–4876.

Reidl, J., Römisch, K., Ehrmann, M., and Boos, W. (1989). "MalI, a novel protein involved in regulation of the maltose system of Escherichia coli, is highly homologous to the repressor proteins GalR, CytR, and LacI". en. In: *J. Bacteriol.* 171.9, pp. 4888–4899.

Rison, S. C. G. and Thornton, J. M. (2002). "Pathway evolution, structurally speaking." In: *Current opinion in structural biology* 12, pp. 374–382. URL: http://www.ncbi.nlm.nih.gov/pubmed/12127458.

Rodríguez-Zavala, J. S., Allali-Hassani, A., and Weiner, H. (2006). "Characterization of E. coli tetrameric aldehyde dehydrogenases with atypical properties compared to other aldehyde dehydrogenases". en. In: *Protein Sci.* 15.6, pp. 1387–1396.

Roy, A., Kucukural, A., and Zhang, Y. (2010). "I-TASSER: a unified platform for automated protein structure and function prediction". In: *Nature protocols* 5.4, pp. 725–738.

Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., Palsson, B. O., Sommer, M., and Feist, A. M. (2014). "Evolution of Escherichia coli to 42 C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations". en. In: *Mol. Biol. Evol.* 31.10, pp. 2647–2662.

Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). "Metabolites: a helping hand for pathway evolution?" en. In: *Trends Biochem. Sci.* 28.6, pp. 336–341.

Sévin, D. C., Fuhrer, T., Zamboni, N., and Sauer, U. (2017). "Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli". en. In: *Nat. Methods* 14.2, pp. 187–194.

Sharma, P., Teixeira de Mattos, M. J., Hellingwerf, K. J., and Bekker, M. (2012). "On the function of the various quinone species in Escherichia coli : The role of DMK in the electron transfer chains of E. coli". In: *FEBS J.* 279.18, pp. 3364–3373.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. (2001). "Cell-free translation reconstituted with purified components". en. In: *Nat. Biotechnol.* 19.8, pp. 751–755.

Soisson, S. M., MacDougall-Shackleton, B., Schleif, R., and Wolberger, C. (1997). "Structural basis for ligand-regulated oligomerization of AraC". en. In: *Science* 276.5311, pp. 421–425.

St Martin, E. J. and Mortlock, R. P. (1977). "A comparison of alternate metabolic strategies for the utilization of D-arabinose". en. In: *J. Mol. Evol.* 10.2, pp. 111–122.

Szappanos, B., Fritzemeier, J., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R. A., Lercher, M. J., Pál, C., and Papp, B. (2016). "Adaptive evolution of complex innovations through stepwise metabolic niche expansion". en. In: *Nat. Commun.* 7, p. 11607.

Tawfik, D. S. (2014). "Accuracy-rate tradeoffs: how do enzymes meet demands of selectivity and catalytic efficiency?" In: *Current opinion in chemical biology* 21, pp. 73–80. DOI: 10.1016/j.cbpa.2014.05.008. URL: http://www.ncbi.nlm.nih.gov/pubmed/24954689.

Tian, Q., Wang, C., Liu, Y., and Xie, W. (2015). "Structural basis for recognition of G-1-containing tRNA by histidyl-tRNA synthetase". en. In: *Nucleic Acids Res.* 43.5, pp. 2980–2990.

Toll-Riera, M., San Millan, A., Wagner, A., and MacLean, R. C. (2016). "The Genomic Basis of Evolutionary Innovation in Pseudomonas aeruginosa". en. In: *PLoS Genet.* 12.5, e1006005.

Tourneux, L., Bucurenci, N., Saveanu, C., Kaminski, P. A., Bouzon, M., Pistotnik, E., Namane, A., Marlière, P., Bârzu, O., Li De La Sierra, I., Neuhard, J., and Gilles, A. M. (2000). "Genetic and biochemical characterization of Salmonella enterica serovar typhi deoxyribokinase". en. In: *J. Bacteriol.* 182.4, pp. 869–873.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq". en. In: *Nat. Biotechnol.* 31.1, pp. 46–53.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J. van, Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". en. In: *Nat. Biotechnol.* 28.5, pp. 511–515.

Utrilla, J., O'Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M., and Palsson, B. O. (2016). "Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution". en. In: *Cell Syst* 2.4, pp. 260–271.

Ventola, C. L. (2015). "The antibiotic resistance crisis: part 1: causes and threats". en. In: *P T* 40.4, pp. 277–283.

Visser, J. A. G. M. de and Krug, J. (2014). "Empirical fitness landscapes and the predictability of evolution". en. In: *Nat. Rev. Genet.* 15.7, pp. 480–490.

Voordeckers, K., Brown, C. A., Vanneste, K., Zande, E. van der, Voet, A., Maere, S., and Verstrepen, K. J. (2012). "Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication". en. In: *PLoS Biol.* 10.12, e1001446.

Wagner, A. (2011). *The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems*. en. OUP Oxford.

Wissenbach, U., Ternes, D., and Unden, G. (1992). "An Escherichia coli mutant containing only demethylmenaquinone, but no menaquinone: effects on fumarate, dimethylsulfoxide, trimethylamine N-oxide and nitrate respiration". en. In: *Arch. Microbiol.* 158.1, pp. 68–73.

Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A., and Lenski, R. E. (2006). "Tests of parallel molecular evolution in a long-term experiment with Escherichia coli". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 103.24, pp. 9107–9112.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). "The I-TASSER Suite: protein structure and function prediction". en. In: *Nat. Methods* 12.1, pp. 7–8.

Yip, S. H.-C. and Matsumura, I. (2013). "Substrate ambiguous enzymes within the Escherichia coli proteome offer different evolutionary solutions to the same problem". en. In: *Mol. Biol. Evol.* 30.9, pp. 2001–2012.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. (2012). "Structure-based prediction of protein-protein interactions on a genome-wide scale." In: *Nature* 490.7421, pp. 556–60. DOI: 10.1038/nature11503. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3482288%5C&tool=pmcentrez%5C&rendertype=abstract.

Zhou, K., Zhou, L., Lim, Q. '., Zou, R., Stephanopoulos, G., and Too, H.-P. (2011). "Novel reference genes for quantifying transcriptional responses of Escherichia coli to protein overexpression by quantitative PCR." In: *BMC molecular biology* 12.1, p. 18. DOI: 10.1186/1471-2199-12-18. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3110127%5C&tool=pmcentrez%5C&rendertype=abstract.