UNIVERSITY OF CALIFORNIA, SAN DIEGO

Dynamics of genetic adaptation in *Escherichia coli* K12 MG1655

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

in

Chemistry & Biochemistry

by

Margaret Kenyon Applebee

Committee in charge:

Professor Bernhard Palsson, Chair
Professor Alexander Hoffmann, Co-chair
Professor Elizabeth Komives
Professor Andrew McCammon
Professor Milton Saier
Professor Roger Tsien

2010

The Dissertation of Margaret Kenyon Applebee is approved, and is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Co-Chair

_____

Chair

UNIVERSITY OF CALIFORNIA, SAN DIEGO

2010

DEDICATION

I dedicate this degree to my mother, Clydene Morgan Applebee. She may not have been able to see where I've gone since we lost her, but her encouragement and wise words still get me through when I'm down. We still miss you.

EPIGRAPH

It is good to have an end to journey towards, but it is the journey that matters, in the end.
Ursula K Le Guin

Nobody climbs mountains for scientific reasons. Science is used to raise money for the expeditions, but you really climb for the hell of it.
Edmund Hillary

Nothing has such power to broaden the mind as the ability to investigate systematically and truly all that comes under thy observation in life.
Marcus Aurelius

The most exciting phrase to hear in science, the one that heralds the most discoveries, is not "Eureka!" but "That's funny..."
Isaac Asimov

Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself.
Francis Crick

What is true for E. coli is true for an elephant.
Jacques Monod

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL DATA

ACKNOWLEDGEMENTS

I would first and foremost like to thank Professor Bernhard Ø. Palsson for his support as my advisor. His guidance has been invaluable, and I am very grateful for being given the opportunity to earn my PhD in his laboratory.

I would also like to acknowledge everyone I worked with in the Palsson lab - it has been a pleasure to meet and work with you all. I especially need to thank Anu Ragunathan, Markus Herrgard, Eric Knight, Dae Hee Lee, Young Seoub Park, Pep Charusanti, Karsten Zingler and Yu Qiu; I owe my degree to your willingness to answer my millions of questions and concerns in the lab, especially when things did not turn out as expected. Thank you for teaching and mentoring me. I would also like to acknowledge my fellow graduate students Jennie Reed, Andrew Joyce, Adam Feist, Vasiliy Portnoy, Nate Lewis, Tom Conrad, and Jeff Orth for their camaraderie, feedback, and assistance. In addition, I need to thank undergraduate assistant Jessica Na for her enthusiasm and hard work in the lab. Several projects would have taken much longer and been a lot less fun without her help.

I also need to thank Karsten Zengler, Jan Schellenberger, and Adam Feist for editing and providing feedback on the material in Chapter 2, which largely a reprint of the review, *Genome-scale Models and the Genetic Basis for E. coli Adaptation*, published as Chapter 12 of Systems Biology and Biotechnology of *Escherichia coli,(*edited by Sang Yup Lee, published by Springer Press 2009), which I wrote with Dr. Palsson.

Chapter 3 is also largely a reprint of the material as it appears in the Journal of Bacteriology article *Impact of individual mutations on increased fitness in adaptively-*

evolved strains of Escherichia coli, published in July 2008, 190(14). I would like to thank my co-authors, Dr. Markus Herrgård and Dr. Palsson. I am especially grateful to Markus for performing the high-throughput data analysis of the competitions results, and for his guidance during the writing of my first professional scientific manuscript. Parts of the discussion and introduction of this chapter were updated and include references to other chapters in the thesis. I also need to thank acknowledge Trina Patel for her help during the preliminary stages of this study, Dr. Christiane Honisch (Sequenom Inc., San Diego, CA) for sharing her invaluable technical insights and expertise, Tom Conrad for his assistance growing cultures, and Dr. Peter Andolfatto (UCSD Dept. of Biology) and Dr. Chris Herring (Mascoma Inc., Lebanon, NH) for insightful discussions.

Chapter 4 is a re-formatted version of a manuscript submitted for publication in the Journal of Bacteriology in 2010, titled *Catalytic and metabolic effects of adaptive GlpK mutations in Escherichia coli,* which I wrote with collaborator Dr. Donald W. Pettigrew (Texas A&M) and Dr. Palsson. I'm especially grateful to my co-author and collaborator, Donald W. Pettigrew, both for sharing his expertise and performing the sensitive kinetic experiments, and for his generous assistance and patience with a young scientist while writing this manuscript. I'd also like thank Dr. Milton Saier, who I consulted with several times during the development of this study. Im also indebted to Dr. Young Seoub Park and Dr. Dae Hee Lee for their invaluable technical advice, Megan Anderson (Quake Lab - Stanford) for sharing her GlpK cloning constructs, and Jessica Na, Pamela S. Miller, and Dr. Tzu-Wen Huang for expert technical assistance.

The cellulose project when through many ups and downs, and I asked and received help many times. I would like to thank Eric Knight for helping me design the strategy for inserting the heterologous genes, and Jessica Na who helped me

maintain the cellobiose evolutions. Also, a shout-out to the other members of the "The Metabolic Engineering Force" group (Adam Feist, Vasiliy Portnoy, Dae Hee Lee, and Jeff Orth) – it was great to troubleshoot together every other week, and maybe someday a grant committee will realize how awesome our strategies are. Also, thanks again to Karsten Zingler for helping me figure out how to tackle the *Paenibacillus* problem in the cellobiose-evolved cultures.

VITA

2003	B.S. in Biochemistry, University of Illinois, Urbana-Champaign
2006	M.S. in Chemistry & Biochemistry, University of California, San Diego
2010	Ph.D. in Chemistry & Biochemistry, University of California, San Diego

PUBLICATIONS

C. D. Herring, A. Raghunathan, C. Honisch, T Patel, M.K. Applebee, A.R. Joyce, T.J. Albert, F.R. Blattner, D. van den Boom, C.R. Cantor, B.Ø. Palsson. 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nature Genetics,* 38(12): 1406-12

J.L. Reed, T.R. Patel, K.H. Chen, A.R. Joyce, M.K. Applebee, C.D. Herring, O.T. Bui, E.M. Knight, S.S. Fong, B.Ø. Palsson. 2006. Systems approach to refining genome annotation. *Proceedings of the National Academy of Science,* 103(46):17480-4.

M.K. Applebee, M.J. Herrgård, B.Ø. Palsson. 2008. Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. *Journal of Bacteriology*. 190(14): 5087-94

M.K. Applebee, B.Ø. Palsson. 2009. Chapter 12: Genome-Scale Models and the Genetic Basis for *E. coli* Adaptation, in *Systems Biology and Biotechnology of Escherichia coli* Eds: Lee, S.Y., Springer Press. p237-56

M.K Applebee, D.W. Pettigrew, B.Ø. Palsson. Functional and metabolic effects of adaptive GlpK mutants in Escherichia coli. *Journal of Bacteriology*. *Manuscript submitted (2010)*.

ABSTRACT OF THE DISSERTATION

Dynamics of genetic adaptation in *Escherichia coli* K12 MG1655

by

Margaret Kenyon Applebee

Doctor of Philosophy in Chemistry & Biochemistry

University of California, San Diego 2010

Professor Bernhard Palsson, Chair

Professor Alexander Hoffmann, Co-Chair

Bacteria replicate quickly enough that adaptive changes can be observed on a timescale of weeks to months. Laboratory adaptive evolution experiments with bacteria, especially of *Escherichia coli*, have been used to study population dynamics and evolutionary processes. Recent advances in DNA sequencing now make it possible to identify all of the acquired mutations and interrogate the adaptive process at the molecular level.

The first two research studies in this dissertation analyze mutations acquired by *E. coli* during prolonged growth on glycerol minimal media. In the first study I measured fitness differences between independently-evolved lineages and between mutations acquired by the same gene in different lineages by head-to-head competitions. This allowed small fitness differences to be detected between independently-evolved strains and between mutations to the same genes. This also detected epistatic interactions between several mutations. Overall these results provide the basis for understanding how the mutations individually and cooperatively affect fitness.

The second study examines adaptive mutations to glycerol kinase (GlpK), which were acquired by almost all glycerol-adapted lineages. The relative ability of the GlpK mutations to increase fitness was used to identify the enzyme function responsible for improving phenotype, which was found to be sensitivity to the inhibitor fructose-1,6-bisphosphate. However, the GlpK mutants also induce autocatabolite repression which reduces *glpK* transcription, attenuating the ability of the mutations to increase glycerol metabolism. This likely occurs to prevent methylglyoxal toxicity. Additionally, the GlpK mutations altered growth on several sugars, but it is unclear what mechanism could be involved suggesting it is a novel function. This study demonstrates how adaptive mutations can be studied to identify constraints on expressed phenotype and to identify novel protein interactions.

The goal of the third study was to integrate genes for cellulose-degrading enzymes into the *E. coli* genome, and to use adaptive evolution to select for strains that improve use of the genes that allow growth on cellulose. This was both a metabolic engineering and basic research project, to produce a potentially useful

strain and to study the molecular dynamics involved in integrating a novel metabolic function into the existing metabolic and regulatory network.

Chapter 1

Introduction

This dissertation covers three studies involving laboratory adaptive evolution of *Escherichia coli*. The overarching goal of all of these studies was to understand the molecular and biochemical basis underlying the observed adaptation by determining how mutations increase fitness. The dynamics of adaptation are examined in the first two studies, which focus on mutations acquired in strains of *E. coli* that were adapted to improve growth on glycerol (26, 37). The first study (Chapter 3) measures and compares the fitness gains imparted by the glycerol-acquired adaptive mutations, while the second study (Chapter 4) examines the molecular mechanism that underlies the adaptive benefit of a set of those mutations. The goal of the third study (Chapter 5) was to improve usage of heterologous genes in *E. coli* using adaptation, and to characterize the mechanisms involved in the adaptation.

1.1 History of adaptive evolution

Laboratory adaptive evolution experiments with microorganisms have been used since the 1960's to study the phenotypic plasticity of microorganisms, discover new enzyme functions and metabolic capacities, and to test theories of population genetics (32). These studies involve growing a microorganism in a new environment for multiple generations to select for mutants that improve growth. Early studies primarily used adaptive evolution to generate new metabolic capabilities, which were analyzed by biochemically characterizing altered enzyme activity in cell lysate.

Bacteria and other microorganisms have several properties that make them very useful for studying adaptation and population dynamics (21). Most importantly, these organisms have short generation times so significant adaptive changes can be observed over the course of weeks or months. Large populations can also be easily and cheaply maintained, and evolution experiments can easily be replicated to study the distribution of possible outcomes. Additionally, at any point samples of the population can be preserved indefinitely as frozen stocks and later revived, allowing initial and intermediates states during the adaptation process to be accessed indefinitely.

Many adaptive evolution experiments have been conducted to study the dynamics of microbial evolution and the general properties that determine the rate of adaptation. For example, adaptive evolution experiments have been used to estimate mutation rates (43, 53), the distribution of beneficial mutations and selection coefficients (18, 30, 34, 43, 65), and the extent of genetic variation (43, 52, 74). Adaptive evolution has also been used to infer the frequency of epistatic interactions between mutations, how this effects the reproducibility of evolution and the shape of the fitness landscape, and how this determines robustness of intracellular interaction networks (5, 8, 33, 37, 52, 56, 73). Environmental conditions that produce coexisting phenotypes (42, 47, 55, 66) and speciation (16, 19), phenotypic variation within clonal populations (4), and phenotypic robustness (6, 40) have also been tested using adaptive evolution of bacteria.

1.2 Significance of adaptive mutations

While many important parameters controlling microbial evolution can be learned at the population level, it would also be very interesting to study the

underlying regulatory and metabolic changes involved. Adaptive evolution involves the accumulation of beneficial mutations that allow the organism to express a phenotype that was previously outside the range of behavior permitted by the genetically-encoded regulatory or metabolic network. This suggests the mutations can be used to discover regulatory or metabolic constraints that are especially important for determining phenotype under that growth condition.

E. coli is the most thoroughly studied free-living organism, but it is still far from being comprehensively understood. To date, sufficiently comprehensive knowledge about *E. coli* metabolism has been generated to construct genome-scale computational models that capture the range of potential metabolic states, and the optimum use of simulated resources (described further in chapter 2). However, regulatory and kinetic dynamics have not yet been sufficiently characterized at the systems level to create models that capture and predict true intracellular dynamics rather than the range of possible states. This is largely because methods for sufficiently characterizing *E. coli* at this level are just emerging (7, 45). Since adaptive mutations are selected because they alter key constraints on metabolism that result in significant phenotype changes, these perturbations may be ideal for studying dynamics of metabolic regulation at the systems level.

1.3 Adaptive evolution and genome resequencing

Many mutations acquired during adaptive evolution have been identified by selectively sequencing genes (2, 19, 49, 50, 53, 70) or by identifying IS element movements (17, 19, 70). However, selective screens likely miss many high-impact mutations to unexpected genetic targets such a global regulators (13, 37). Whole-genome resequencing is necessary if the goal is to understand the specific adaptive

process, rather than to make inferences about population dynamics or study adaptation of a particular protein or gene.

Rather than perform genome resequencing, several studies have attempted to understand the adaptive process by examining changes to the transcription profile (14, 15, 26) and protein co-regulation network(2, 50), in the absence of knowledge of the acquired mutations. Unfortunately, adaptation causes numerous changes at both of these levels, many of which are likely secondary responses rather than causal changes that reveal how phenotype is improved. In the absence of information to distinguish the causal changes, it has been difficult to reach strong conclusions about the mechanisms of adaptation from profiling data alone (26).

1.4 Analysis of adaptive mutations

As whole-genome resequencing has become more accessible, an increasing number of adaptive evolution studies have begun to take advantage of it(3, 13, 31, 37, 51, 72). Genome resequencing can identify the vast majority of causal mutations (13), but this is not sufficient to understand the mechanism of adaptation since the effect of a mutation on a protein or genetic element cannot be inferred from the sequence change alone. Identifying the mutations is only the first step for understanding the adaptive process.

1.4.1 Phenotypic effects of individual mutations

After identifying the mutations, the individual impact of each on phenotype in the adaptive environment should be measured. Not all mutations have a strong impact on fitness, and in fact some may be neutral or even slightly deleterious (57) (though far fewer neutral mutations are being discovered than expected from population theory (3)). The ability of mutations to independently improve the growth

phenotype can be determined by introducing them into parental strain and measuring the effect on growth in the selected environment (13, 37). Alternatively, many of the mutations with the strongest impact on fitness are likely those that fixed first in the population, since mutations that infer significant fitness gains fix quickly, and subsequent mutations are hypothesized to cause progressively smaller fitness gains (21, 57).

However, the ability of a mutation to impact fitness may be affected by co-acquired mutations. Epistatic interactions between mutations in *E. coli* are proving to be surprisingly common (10, 14, 64), and have been found between mutations in genes controlling very distant processes where it is unclear how their effects compound each other (ie, between mutations effecting peptidoglycan biosynthesis and RNA polymerase proteins in Chapter 3). Thus the adaptive function of many mutations may only be relevant in the context of a previously acquired mutation (10), and this may be important to recognize when trying to determine the molecular mechanism that lead to the selection of a particular mutation.

1.4.2 Determining the casual effect of adaptive mutations

As previously mentioned, the effect of a mutation on the altered gene product cannot be inferred from the sequence alone. The effect of the mutation can only be determined by measuring the effect on the function of the encoded element, usually a protein. However, mutations can alter multiple properties of the altered enzyme or genetic region, but were likely selected for their impact on only one critical parameter that was specifically limiting growth. Without another context, it can be impossible to identify this parameter.

As previously stated, one of the advantages of performing laboratory evolution experiments with bacteria is that the adaptation process can be replicated indefinitely by growing independent cultures under identical conditions from the same parent clone (52). As described further in Chapter 2 and elsewhere, many adapted populations acquire similar fitness gains in the selection environment, but these gains are often achieved by different sets of mutations (26, 69). However, given that all of the adapted strains are acquiring mutations to approach the same optimal metabolic state, it is safe to assume that the different sets of mutations are being acquired to change the same set of parameters that are constraining growth. It is also likely that mutations that affect the same gene or same function were acquired to alleviate the same constraint. As demonstrated in Chapter 4, this suggests it is possible to identify the molecular constraints that similar mutations are selected to alleviate by correlating their abilities to improve fitness to properties that they alter.

1.5 Final comments

This area of research draws upon the knowledge and expertise of many biological disciplines. The effect of the mutations must be analyzed to determine their effect on multiple scales, from their effect at the gene or protein level, to the consequences of that change on pathway regulation, metabolic flux, and feedback mechanisms, to the final phenotype. This type of research can be very taxing and projects should be chosen with care, but this strategy has the potential to identify meaningful connections between the different scales of biological characterization and significantly improve comprehensive understanding of *E. coli* regulation.

1.6 Dissertation Outline

This dissertation covers efforts to understand adaptive processes in *E. coli* at several different levels, including measuring the reproducibility of adaptive evolution, discovering biochemical and molecular mechanisms that underlie fitness improvements, and how growth conditions influence selection pressure.

Chapter 2 is largely a reprint of a literature review of previous research in the Palsson lab relevant to the adaptive evolution studies. Some sections not relevant to the dissertation material have been removed, and other sections have been edited for clarity. However, I am not the primary author of any of the work described in Chapter 2 – it is provided for background only.

Chapter 3 describes competition experiments performed between *E. coli* strains that had either been adapted to M9-glycerol, or that carried mutations discovered in those strains. This study demonstrates that different mutations acquired to the same gene in different lineages have similar but not identical abilities to improve fitness, and that some of the mutations effect fitness differently depending whether they are expressed with or without previously-acquired mutations.

Chapter 4 describes an in-dept study of four adaptive glycerol kinase (GlpK) mutations acquired in glycerol-adapted strains that had been competed in the study described in Chapter 3. This study represents an attempt to characterize the biochemical constraints on glycerol metabolism that the GlpK mutations were selected to alleviate. The competition experiments determined the relative abilities of the GlpK mutations to increase fitness, but fitness did not correlate to any of the mutant GlpK kinetic parameters measured at the time. In this study the relative effects of the mutations on fitness are used to identify sensitivity to fructose-1,6-bisphosphate as the predominant constraint that the mutations alleviate to improve

fitness. Increased activity of the mutant GlpK enzymes are also shown to increase glycerol-induced catabolite repression.

Chapter 5 describes work to develop a strain of *E. coli* that can grow on cellulose by expressing heterologous cellulase genes and using adaptive evolution to optimize growth. This work was not completed or published due to several experimental and funding roadblocks. However, the work that was completed is described, including adaptive evolution of *E. coli* to acquire growth on cellobiose, and construction of the plasmids containing the heterologous genes in line with their selection markers and the flanking regions necessary for integration into the *E. coli* genome. The research plans for completing the work are also described.

1.7 References

1. **Bantinaki, E., R. Kassen, C. Knight, Z. Robinson, A. Spiers, and P. Rainey.** 2007. Adaptive divergence in experimental populations of Pseudomonas fluorescens. III. Mutational origins of wrinkly spreader diversity. Genetics.

2. **Barrick, J. E., D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim.** 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature **461:**1243-7.

3. **Beaumont, H. J., J. Gallie, C. Kost, G. C. Ferguson, and P. B. Rainey.** 2009. Experimental evolution of bet hedging. Nature **462:**90-3.

4. **Beerenwinkel, N., L. Pachter, B. Sturmfels, S. F. Elena, and R. E. Lenski.** 2007. Analysis of epistatic interactions and fitness landscapes using a new geometric approach. BMC Evol Biol **7:**60.

5. **Bennett, A. F., and R. E. Lenski.** 2007. An experimental test of evolutionary trade-offs during temperature adaptation. Proc Natl Acad Sci U S A **104 Suppl 1:**8649-54.

6. **Bennett, B. D., E. H. Kimball, M. Gao, R. Osterhout, S. J. Van Dien, and J. D. Rabinowitz.** 2009. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol **5:**593-9.

7. **Blount, Z. D., C. Z. Borland, and R. E. Lenski.** 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc Natl Acad Sci U S A **105:**7899-906.

8. **Camps, M., A. Herman, E. Loh, and L. A. Loeb.** 2007. Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol **42:**313-26.

9. **Conrad, T. M., A. R. Joyce, M. K. Applebee, C. L. Barrett, B. Xie, Y. Gao, and B. O. Palsson.** 2009. Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. Genome Biol **10:**R118.

10. **Cooper, T. F., S. K. Remold, R. E. Lenski, and D. Schneider.** 2008. Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli. PLoS Genet **4:**e35.

11. **Cooper, T. F., D. E. Rozen, and R. E. Lenski.** 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proc Natl Acad Sci U S A **100:**1072-7.

12. **Cooper, V. S., and R. E. Lenski.** 2000. The population genetics of ecological specialization in evolving Escherichia coli populations. Nature **407:**736-9.

13. **Cooper, V. S., D. Schneider, M. Blot, and R. E. Lenski.** 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of Escherichia coli B. J Bacteriol **183:**2834-41.

14. **de Visser, J. A., and R. E. Lenski.** 2002. Long-term experimental evolution in Escherichia coli. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. BMC Evol Biol **2:**19.

15. **Dykhuizen, D. E., and A. M. Dean.** 2004. Evolution of specialists in an experimental microcosm. Genetics **167:**2015-26.

16. **Elena, S. F., and R. E. Lenski.** 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet **4:**457-69.

17. **Fong, S. S., A. R. Joyce, and B. O. Palsson.** 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res **15:**1365-72.

18. **Gerrish, P. J., and R. E. Lenski.** 1998. The fate of competing beneficial mutations in an asexual population. Genetica **102-103:**127-44.

19. **Gresham, D., M. M. Desai, C. M. Tucker, H. T. Jenq, D. A. Pai, A. Ward, C. G. DeSevo, D. Botstein, and M. J. Dunham.** 2008. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet **4:**e1000303.

20. **Hall, B. G.** 1989. Selection, adaptation, and bacterial operons. Genome **31:**265-71.

21. **Hayashi, Y., T. Aita, H. Toyota, Y. Husimi, I. Urabe, and T. Yomo.** 2006. Experimental rugged fitness landscape in protein sequence space. PLoS ONE **1:**e96.

22. **Hegreness, M., N. Shoresh, D. Hartl, and R. Kishony.** 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science **311:**1615-7.

23. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

24. **Hughes, B. S., A. J. Cullum, and A. F. Bennett.** 2007. Evolutionary adaptation to environmental pH in experimental lineages of Escherichia coli. Evolution **61:**1725-34.

25. **Imhof, M., and C. Schlotterer.** 2006. E. coli microcosms indicate a tight link between predictability of ecosystem dynamics and diversity. PLoS Genet **2:**e103.

26. **Imhof, M., and C. Schlotterer.** 2001. Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U S A **98:**1113-7.

27. **Jamshidi, N., and B. O. Palsson.** Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. Biophys J **98:**175-85.

28. **Kassen, R., and P. B. Rainey.** 2004. The ecology and genetics of microbial diversity. Annu Rev Microbiol **58:**207-31.

29. **King, T., S. Seeto, and T. Ferenci.** 2006. Genotype-by-environment interactions influencing the emergence of rpoS mutations in Escherichia coli populations. Genetics **172:**2071-9.

30. **Knight, C. G., N. Zitzmann, S. Prabhakar, R. Antrobus, R. Dwek, H. Hebestreit, and P. B. Rainey.** 2006. Unraveling adaptive evolution: how a single point mutation affects the protein coregulation network. Nat Genet **38:**1015-1022.

31. **Lee, D. H., and B. O. Palsson.** 2010. Adaptive Evolution Of Escherichia coli K-12 MG1655 On A Non-Native Carbon Source, L-1,2-Propanediol. Appl Environ Microbiol.

32. **Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler.** 1991. Long-Term Experimental Evolution in Escherichia Coli. I. Adaptation and Divergence During 2,000 Generations. American Naturalist **138:**1315-1341.

33. **Lenski, R. E., C. L. Winkworth, and M. A. Riley.** 2003. Rates of DNA sequence evolution in experimental populations of Escherichia coli during 20,000 generations. J Mol Evol **56:**498-508.

34. **Maharjan, R., S. Seeto, L. Notley-McRobb, and T. Ferenci.** 2006. Clonal adaptive radiation in a constant environment. Science **313:**514-7.

35. **Martin, G., S. F. Elena, and T. Lenormand.** 2007. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat Genet **39:**555-60.

36. **Orr, H. A.** 2005. Theories of adaptation: what they do and don't say. Genetica **123:**3-13.

37. **Remold, S. K., and R. E. Lenski.** 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. Nat Genet **36:**423-6.

38. **Rokyta, D. R., P. Joyce, S. B. Caudle, and H. A. Wichman.** 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat Genet **37:**441-4.

39. **Rozen, D. E., D. Schneider, and R. E. Lenski.** 2005. Long-term experimental evolution in Escherichia coli. XIII. Phylogenetic history of a balanced polymorphism. J Mol Evol **61:**171-80.

40. **Travisano, M., and R. E. Lenski.** 1996. Long-term experimental evolution in Escherichia coli. IV. Targets of selection and the specificity of adaptation. Genetics **143:**15-26.

41. **Treves, D. S., S. Manning, and J. Adams.** 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. Mol Biol Evol **15:**789-97.

42. **Velicer, G. J., G. Raddatz, H. Keller, S. Deiss, C. Lanz, I. Dinkelacker, and S. C. Schuster.** 2006. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. Proc Natl Acad Sci U S A **103:**8107-12.

43. **Wichman, H. A., M. R. Badgett, L. A. Scott, C. M. Boulianne, and J. J. Bull.** 1999. Different trajectories of parallel evolution during viral adaptation. Science **285:**422-4.

44. **Woods, R., D. Schneider, C. L. Winkworth, M. A. Riley, and R. E. Lenski.** 2006. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. PNAS **103:**9107-9112.

Chapter 2

Genome-Scale Models and the Genetic Basis for *E. coli* Adaptation – a Review

Majority of this material was previously published as Chapter 12 in Systems Biology and Biotechnology of *Escherichia coli* Ed. Lee SY, Springer, 2009 p237-56

The dissertation author was the primary author of this review, but not the research described by the review, and was co-authored by Dr. Bernhard Palsson.

2.1 Introduction to metabolic models of *E. coli*

E. coli is one of the most thoroughly annotated and understood organisms, even among model species (22, 46, 48). However, this organism is still far from being comprehensively understood given that the functions of 40% of the genes have still not been experimentally verified. The vast amount of accumulated knowledge on *E. coli* does however make it a leading candidate to study how networks of biological molecules function and interact to produce an organism's physiological behavior.

Computational models are often constructed to capture the dynamics of systems with numerous interacting elements, and can be used to identify emergent properties at levels between molecular biology of subsystems and phenotype

measurements (12). The models can be iteratively improved and lead to biological discovery by comparing their predictions of the cell's response to a simulated perturbation and the system's or organisms actual response; failed predictions can be analyzed to determine what may be missing or incorrect in the model and thus direct research towards critical elements or interactions.

2.1.1 Stoichiometric Models

Genome-scale *in silico* metabolic models simulate how nutrients are processed by an organism's metabolic network to harness energy and biomass, based on the set of enzymes the organism is known to harbor and the reactions they catalyze. The dynamics of enzyme-catalyzed reactions have traditionally been described using the enzyme's kinetic constants and the concentrations of reactants and products. However, there are often significant known differences between *in vivo* conditions and the assay conditions used to estimate enzyme activity, such as pH and ionic strength, making many values reported in the literature generally unreliable for inferring *in vivo* enzyme activity (61). Additionally, some parameters that effect enzyme activity, such as the concentration of metabolites and interacting enzymes within the cell, are not accurately known and can be highly condition-dependent. These factors have so far hampered efforts to construct robust metabolic models based on kinetic parameters (61, 71). However, it should be mentioned that the ability to accurately and comprehensively measure the concentrations of cellular metabolites has recently been improving (7, 44), spurring new efforts to build metabolic models that include kinetic constraints.

To avoid using kinetic parameters, some have built metabolic models simply constrained by the stoichiometry of the biochemical reactions in the metabolic network. In this approach, the metabolic network is represented by the set of

stoichiometrically-balanced metabolic equations representing the known set of enzymes harbored by the organism. This can be used to calculate all of the flux distributions that can result from passing a defined supply of simulated nutrients through the metabolic network that achieve a pre-defined set of requirements that simulate growth.

2.1.2 Using stoichiometric models to predict growth

The stoichiometric models can be used to simulate physiological behavior by applying biologically-relevant algorithms and setting resource constraints. For example, competition for resources during exponential growth of bacteria has been hypothesized to lead organisms to optimize the biomass they produce (52). The metabolic models have been used to make predictions about the exponential growth rate of the *E. coli* under various growth conditions by searching the solution space for flux distributions that use the simulated nutrients to produce the most simulated biomass (61), using a method known as flux balance analysis (FBA). FBA uses linear programming to find flux states that maximize a determined objective (63, 71). More information about FBA can be found elsewhere (23, 63), as its full description is beyond the scope of the research discussed in this dissertation. Biomass production is simulated by creating an objective function that defines the set of biological molecules (amino acids, nucleotides, lipid membrane, etc) that must be synthesized to generate another cell, which are based on experimental measurements of *E. coli* (61).

By applying FBA and an appropriate biomass objective function, one can predict the maximum growth rates possible given the availability oxygen and other limiting nutrients (71). The results of this type of calculation can be visualized as a three-dimensional phenotypic phase plane (PhPP) (**Figure 2.1**) that show the

maximum growth rate across the range of allowable uptake rates, such as for oxygen (OUR) and carbon substrate (SUR). These phase planes often have several faces, each of which represents a mode of growth such as aerobic growth or fermentation of a byproduct (41).

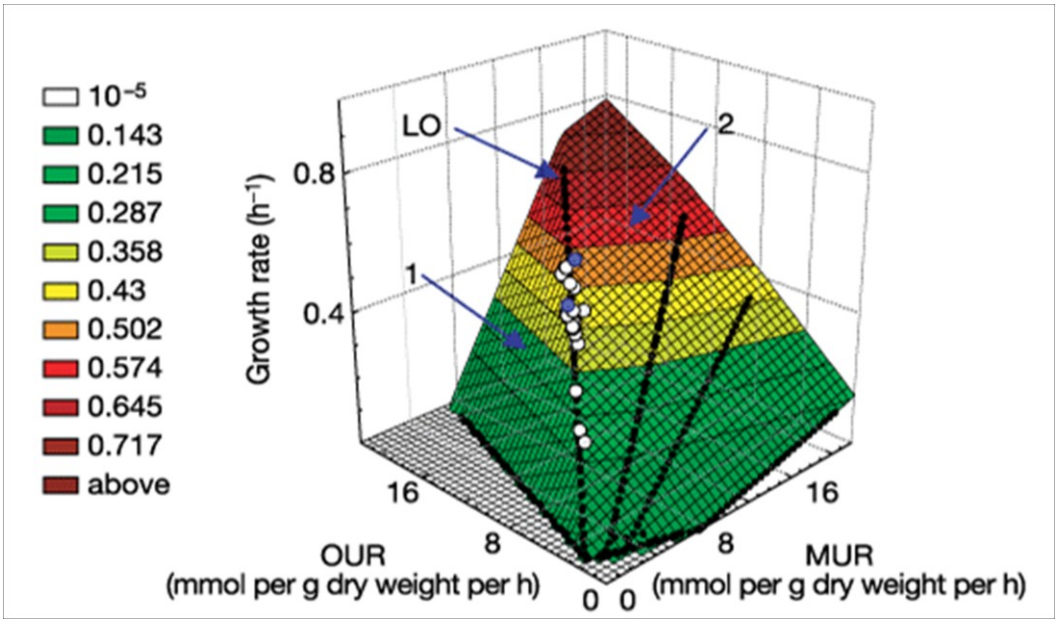


Figure 2.1 The Phenotype Phase Plane. This figure depicts the maximum growth rate that is predicted to be produced across the range of feasible oxygen and substrate uptake rates. This example shows predicted optimized growth rates of *E. coli* grown aerobically on malate. The line of optimality (LO) is highlighted. White and blue circles represent the experimentally-observed growth phenotypes of *E. coli* cultures, where blue circles represent (pre-evolution) wild type cultures, and white circles represent measurements of evolved cultures. Note that all experimentally-derived measurements cluster on the line of optimality. OUR, oxygen uptake rate; MUR, malate uptake rate. Figure originally published in (18).

The intersection separating oxygen-limited and carbon source-limited growth is known as the line of optimality (LO), and represents the growth mode in which the uptake rate of oxygen and the carbon source are stoichiometrically balanced, producing perfectly balanced aerobic growth without any energy loss to futile cycles (for example, due to concurrent activity of glycolytic and gluconeogenic pathways).

Interestingly, more than one set of uptake rates or even intracellular fluxes can produce the same predicted maximum growth rate (62).

The growth rate predictions made by genome-scale metabolic models of *E. coli* K12 MG1655 have generally been found to be accurate when compared to growth rates of cultures during exponential growth (41). The measured growth rate, SUR and OUR of growing cultures are also plotted onto the generated phenotypic phase plane shown in **Figure 2.1,** and generally align with the line of optimality. The plots include data from cultures grown under a variety of conditions, including variable temperatures, substrates (glucose, malate, succinate, and acetate), and substrate concentrations in minimal media. These results suggest that the metabolic model of *E. coli* K12 MG1655 is fairly accurate at predicting the metabolic capacity of the actual organism under many growth conditions (41).

2.2 Adaptive Evolution

2.2.1 Adaptation to a Substrate Challenge

However, experimentally-measured growth phenotypes do not always fall on the line of optimality predicted by biomass-optimized flux balance analysis. For example, *E. coli* K12 MG1655 was found to grow significantly more slowly than the model-predicted optimum on glycerol and lactate. It was not thought likely that these prediction failures were due to model errors since the metabolic pathways for utilizing these substrates have been well-characterized. Rather, the discrepancy was hypothesized to suggest that *E. coli* simply was not utilizing the substrates as effectively as possible.

Given that the metabolic pathways for these substrates exist, the growth capacity could be constrained by factors that regulate or can alter mechanisms that

are not included in the model, such as enzyme expression, kinetics, or feedback regulation. Such constraints, especially transcriptional regulation, are genetically malleable and are refined by evolution. The short generation time of *E. coli* makes it possible to observe the evolution of adaptive traits in long-term cultures (52). If the prediction failures are due to the laboratory *E. coli* strains not being well adapted, subjecting the strains to natural selection by long term exponential growth on the challenging substrate should select mutations that correct the regulatory constraint so that they approach or achieve the FBA-predicted optimal growth rates.

Laboratory adaptive evolution experiments to test this hypothesis were conducted by culturing *E. coli* K12 MG1655 on each of the challenging substrates for a prolonged (≥ 500 generations) period of time (27, 41). The resulting strains all grew significantly faster on the targeted substrate (**Table 2.1**) (41). Additionally, the growth profiles of each strain migrated towards the predicted line of optimality, and once the growth phenotype aligned with the predicted line of optimality generally it only migrated along it (**Table 2.1**, **Figure 2.2**). Further growth rate increases were accomplished by increasing both uptake rates proportionally so that the phenotype moved up the line of optimality (20, 41). However, it should be noted that several strains eventually increased their sugar uptake rates beyond what could be fully aerobically metabolized given physical limits on the oxygen uptake rate – these strains were driven by growth-rate dependent selection to ferment the excess sugar, and consequently migrated off of the predicted line of optimality.

Table 2.1 Summary of Substrate-Challenged Evolution Experiments

	Temp. (°C)	Generations Evolved	Growth Phenotype near LO? Wild type	Evolved Strains	% GR Increase	Ref.
Glucose	37	500	Yes	Yes	17%	(2)
Malate	37	500	Yes	Yes	19%	(2)
Acetate	37	700	Yes	Yes	20%	(2)
Lactate	30	950	No(acetate sec.)	Yes	147%	(1)
Lactate	30	950	No(acetate sec.)	Yes	132%	(1)
Glycerol	30	1000	No(futile cycles)	Yes	140%	(2)
Lactate	37	870	No(futile cycles)	No(O_2 maxed)	80%	(1)
Pyruvate	37	1200	No(acetate sec.)	No(O_2 maxed)	69%	(1)
Pyruvate	30	1000	Yes	No(O_2 maxed)	115%	(1)
αKG	37	625	Yes	No	41%	(1)
αKG	30	440	Yes	No	48%	(1)

This table compares outcomes of adaptive evolution experiments that challenged MG1655 wild type *E. coli* to increase growth on a variety of substrates with in-silico flux-balance analysis (FBA) growth rate predictions. The growth phenotype of wild type on some substrates did not meet the FBA prediction because of carbon or oxygen uptake imbalance, that resulted in acetate secretion (too much carbon uptake) or futile cycles (too much oxygen uptake). The growth phenotype of some evolved strains did not fall on the LO because they increased their carbon uptake beyond the limits of physiologically-available O_2 ("O_2 maxed"). αKG – alpha-ketoglutarate. LO – Line of Optimality. GR – Growth rate. sec. - secretion

Overall, the outcomes indicate that actual growth limits are captured by biomass-optimized flux-balance analysis of the stoichiometric reconstruction of *E. coli* metabolic network. Additionally, they suggest that the cellular architecture is efficient at relieving constraints aside from those imposed by chemical or catalytic limitations.

However, FBA alone cannot predict how the evolved strains metabolize nutrients, since each growth rate can be produced by multiple potential flux distributions (62). Quantitative metabolomic profiling is necessary to both identify and validate the flux distribution predictions, which is still a very challenging experimental endeavor – though as previously mentioned, progress is being made(7, 44). Additionally, these results do not begin to identify the mechanics of the adaptation, which will be addressed later in this chapter.

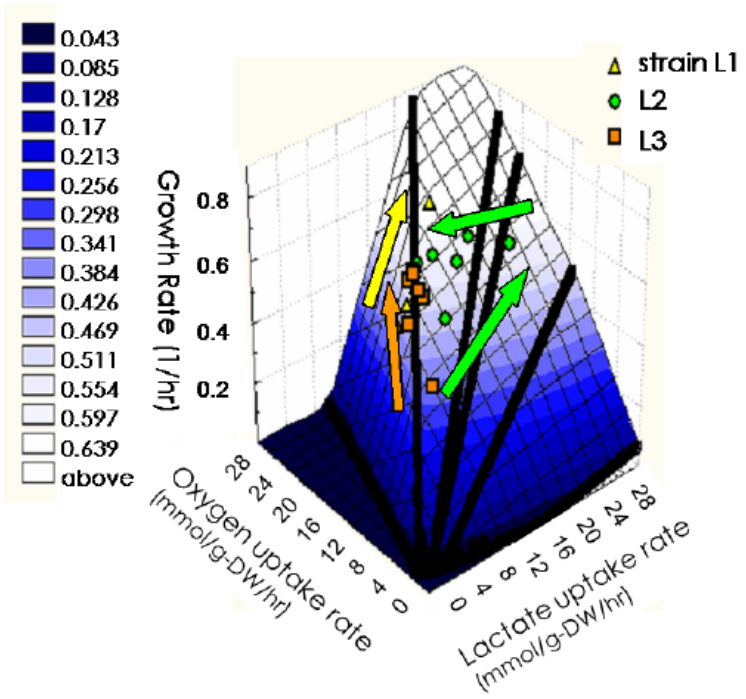


Figure 2.2 The growth phenotype of evolving *E. coli* strains migrate towards the line of optimality. This example shows the growth phenotype three replicate experimental evolutions of wild type on lactate (L1-L3), which all eventually migrate towards the line of optimality. Figure originally published in (12).

2.2.2 Adaptation following deletion of a metabolic gene

Deleting a metabolic gene also changes the optimum growth rate of a strain. This new growth rate can be predicted by removing the catalyzed reaction from the in silico metabolic model (63). The new line of optimality is composed of solutions that most effectively redistribute the metabolic flux around the lost reaction to optimize most biomass. However, *E. coli* strains with the gene deletion usually can not initially grow at the predicted growth rate if the missing gene would normally be utilized, since the regulatory response is adapted to impose optimum growth with the missing function. However, the strains should be able to re-set their regulatory response if provided the opportunity to adapt.

In a previous study by Fong et al, strains of *E. coli* with a single metabolic gene deleted were adaptively evolved to test whether the predicted growth phenotype reflects the actual ability of *E. coli* to adapt to gene loss (29). These experiments test the plasticity of cell regulation to allow the redistribution of metabolic fluxes. One of six genes was deleted from strains of *E. coli*, (acetate kinase A (ΔackA), fumarate reductase (Δfrd), phosphoenolpyruvate carboxykinase (Δppc), phosphoenol pyruvate carboxylase (Δppc), triosephosphate isomerase (Δtpi), or glucose 6-phospate-1-dehydrogenase (Δzwf)). These genes encode enzymes required for gluconeogenesis, fermentation or the pentose phosphate pathway (Figure 2.3). Adaptive evolution experiments were performed with these strains across a set of substrates that enter central metabolism at various points (29).

Approximately 80% of the adaptively-evolved gene deletion strains achieved growth rates within 10% of their respective FBA prediction. This high success rate indicated that the stoichiometric model captures many of the physiologically-relevant constraints on growth. The results of both adaptive evolution studies suggests that *E.*

coli is fairly adept at finding adaptations that reroute its metabolic flux to achieve the optimal growth phenotype available within the limits of it's metabolic catalytic capacity.

2.2.3 Adaptive evolution to optimize metabolically-engineered strains

The metabolic models have also been used to identify sets of genes that when

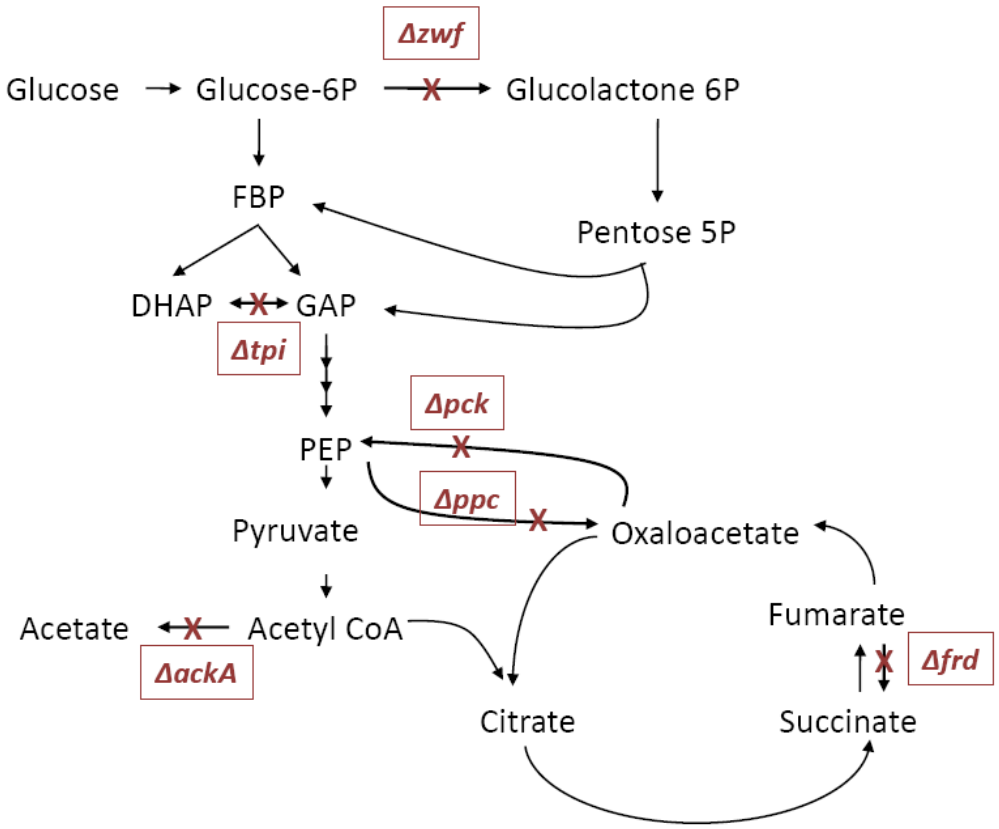


Figure 2.3 *E. coli* metabolic genes deleted to observe adaptation to their loss. Described in (14). Deleting *zwf* (glucose 6-phosphate dehydrogenase) prevents forward flux through the pentose phosphate pathway. Deleting *tpi* (triosphosphate isomerase) interferes with glycolysis and gluconeogenesis. Deleting *pck* (phosphoenolpyruvate carboxykinase) can interfere with gluconeogenesis from citric acid cycle intermediates. Deletion of *ppc* (phosphoenolpyruvate carboxylase) impedes the ability to replenish oxaloacetate to the citric acid cycle. Deletion of *frd* (fumarate reductase) obstructs utilization of the reductive pathway of the citric acid cycle. Deletion of *ackA* (acetate kinase A) blocks a pathway needed to secrete acetate. These strains were experimentally evolved on a variety of carbon sources to select for increased growth rate.

deleted make growth dependent on a secondary objective, such as product secretion. These algorithms include OptKnock (9, 59), OptStrain (60), and OptGene (58), which calculate the predicted growth phenotype across all possible gene deletion strains and identify permutations of the metabolic network that maximize both biomass formation and production of the secreted compound. *E. coli* growth becomes coupled to product secretion during under anaerobic conditions when metabolic reactions for secretion of alternative, more energetically favorable fermentation products are removed.

Since these designs involve deleting metabolic genes, adaptive evolution can be employed to drive recovery of the deletion strains and to optimize their growth rate and secretion rate. Manual examination of the metabolic network can also be used to predict gene deletions that will couple growth to product secretion; however, these algorithms have the advantage of also identifying designs that are not intuitively obvious.

A 2005 study by Fong et al optimized several growth-coupled designs using adaptive evolution. Three strains were generated based on OptKnock designs that are supposed to couple growth to lactate production (25). All three designs reached similar growth rates after adaptive evolution (0.24-0.26 hr^{-1}), and lactic acid production increased with growth rate in each lineage, indicating the designs were all successful. However, even though the rate of lactic acid secretion increased in all three designs during adaptation, the final lactate titers in the media recovered after culturing did not. This possibly suggests the existence of metabolic feedback mechanisms not included in the current model that arrests secretion beyond some threshold. The adaptations that allowed the strains to increase their growth and lactate secretion rates have not been studied.

2.3 Intracellular Mechanisms of Adaptation

Beyond illustrating that selection during exponential growth produces optimal phenotypes that converge with FBA predictions, the intracellular changes that facilitate the phenotype shifts are also of inherent interest. Adaptations can act through many intracellular activities, including but not limited to metabolism, transcriptional regulation, protein turnover, and intracellular feedback mechanisms. Thus identifying the mechanisms through which adaptive mutations act can require multiple high-throughput experimental methods, including mRNA transcription profiling, global metabolomic and flux profiling, and chromatin immunoprecipitation-on-chip – and requires an integrated analysis of such data sets. Additionally, the timescale required to change the growth phenotype implies the adaptations involve genetic mutations, rather than an adjustment to an existing mode of regulation. Therefore it is necessary to identify the acquired mutations in order to pursue the fundamental goal of understanding the mechanism that underlies the phenotypic change.

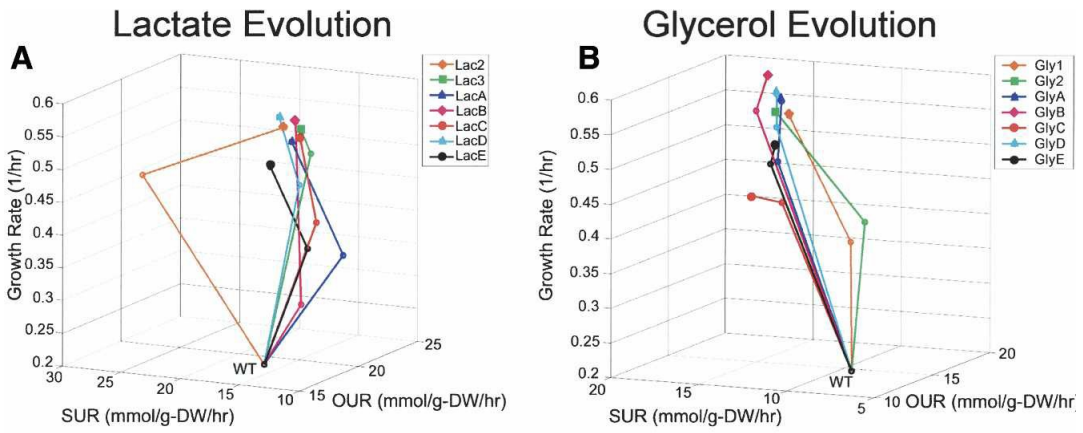
2.3.1 Phenotypic Characterization of Replicate Endpoints

The previously described studies demonstrated that regulatory mechanisms can readily be altered to allow optimal use of their metabolic network in a given condition. However, these studies provide no insight into how those changes are achieved, except to suggest that they require genetic mutations. As previously touched upon, flux balance analysis of an *in-silico* model cannot predict the intracellular state of the cell because multiple flux states can produce the same maximum biomass (however, solutions often differ by only a small set of reactions that

are allowed variable flux (62)). Thus any insight into the mechanism of adaptation must be experimentally derived.

Replicate lineages evolved under the same conditions often achieve nearly identical growth rates and nutrient uptake rates under the growth conditions of the adaptation. Glycerol- and lactate-adapted lineages derived from the same clonal parent strain all achieved growth rates, SURs, and OURs within 12% of each of the identically-treated replicates. However, these strains often differ in other phenotypic traits such as byproduct secretion and growth rates in other conditions (26, 27, 29). This suggests that replicate lineages acquire different adaptive changes that produce the same or similar adjustment to growth in the environment of the evolution, but respond differently in response to other growth conditions – ie, have different pleiotropic effects. It is possible that some of this variation results from different lineages adopting different equivalent flux states, though it could also result from different adaptations that result in changing regulation to allow the same flux state.

Interestingly, variation between replicate strains was significantly more pronounced at the mid-point of the adaptation period (day 20) than at the end of the adaptation period, as measured by differences in growth rate and substrate uptake rates (**Figure 2.4**). This variation has been suggested to indicate that adapting populations initially adopted multiple adaptive strategies that produced different patterns of adaptive radiation between replicate cultures; as better adaptations appeared, selection during the later period became more discriminate and caused the phenotypes to converge to a single optimal phenotype. Interestingly, the endpoint strains generally grow faster than the wild type strain on many other carbon sources in minimal media, suggesting that some of the acquired changes were generally beneficial.



Strain	SUR (mmol/g-DW/hr)	OUR (mmol/g-DW/hr)	GR (1/hr)
Wild-type	13.8 ± 0.4	15.9 ± 2.1	0.23
Lac2 D60	14.9 ± 0.2	19.9 ± 1.0	0.56
Lac3 D60	14.5 ± 0.04	20.2 ± 1.1	0.53
LacA D60	14.6 ± 0.03	18.9 ± 0.5	0.54
LacB D60	16.4 ± 1.4	21.2 ± 0.6	0.55
LacC D60	13.0 ± 0.2	17.7 ± 0.2	0.56
LacD D60	14.4 ± 0.4	17.8 ± 0.5	0.59
LacE D60	16.4 ± 0.1	18.9 ± 0.7	0.50

Strain	SUR (mmol/g-DW/hr)	OUR (mmol/g-DW/hr)	GR (1/hr)
Wild-type	7.6 ± 0.4	12.0 ± 0.7	0.22
Gly1 D44	14.7 ± 0.2	17.3 ± 0.7	0.54
Gly2 D44	14.9 ± 0.2	16.6 ± 0.4	0.55
GlyA D44	15.1 ± 0.2	17.3 ± 0.3	0.56
GlyB D44	17.2 ± 0.1	19.3 ± 0.2	0.58
GlyC D44	13.4 ± 0.1	12.2 ± 0.1	0.46
GlyD D44	14.5 ± 0.2	16.0 ± 0.2	0.58
GlyE D44	15.0 ± 0.2	16.6 ± 0.1	0.50

Figure 2.4 Phenotypic variability between replicate lactate- and glycerol-evolved strains. A. Lactate Evolution. The top figure shows the trajectory of each replicate as it evolved from wild type, in terms of changing growth, substrate uptake, and oxygen uptake rate. Note that most strains begin to converge to the same growth phenotype. The bottom table lists the values of the measured parameters for the wild type and the final (day 60) endpoint strains. B. Similar to A, except applies to strains evolved to glycerol. Figure originally published in Fong et al 2005 (11).

2.3.2 Metabolic flux changes during adaptation

^{13}C-labeling has been used to measure how metabolic fluxes changed before and after adaptation to lactate and following deletion of a metabolic gene. The ^{13}C-labeling experiments on the lactate evolved strains (24, 39, 67) showed that the first major flux change during the first 20 days of evolution was a dramatic increase in the uptake of lactate (up to 80%) and increased flux through most metabolic reactions. Also during this early period, all of the strains shift more metabolites into the TCA cycle from acetate fermentation, allowing the cells to generate more energy and anabolic precursors and thus achieve a faster growth

rate. After day 10 the flux changes among the replicate evolutions diverged. However, all strains eventually began partitioning approximately two-thirds of the lactate to the TCA cycle, indicating that regulation of gluconeogenic and catabolic flux partitioning is critical to the adaptation. This partitioning likely involves feedback mechanisms that recognize cellular concentrations of phosphoenolpyruvate (11).

The genes deleted in the adapted gene-deletion strains analyzed by flux balance analysis encode metabolic enzymes that catalyze key metabolic branch points in central metabolism, and were chosen because their loss was expected to dramatically change pathway utilization (28) (phosphoglucose isomerase (Δ*pgi*), phosphoenolpyruvate carboxylase (Δ*ppc*), triose-phosphate isomerase (Δ*tpi*), or phosphate transacetylase (Δ*pta*)) (**Figure 2.5**). The generated endpoint strains recovered near-wild type growth rates, and replicate endpoints differed in terms of phenotypic traits like byproduct secretion - suggesting they had acquired different specific adaptations. The ^{13}C-labeling experiments revealed that most of the gene deletion strains adapted by activating alternative pathways to locally reroute metabolites around the lost gene function. This rerouting was accomplished by activating pathways typically used for growth under other conditions, such as the pentose-phosphate pathway (Δpgi strains), glyoxylate shunt (Δppc strains), and the methyl-glyoxyl bypass (Δtpi strains) (**Figure 2.5**).

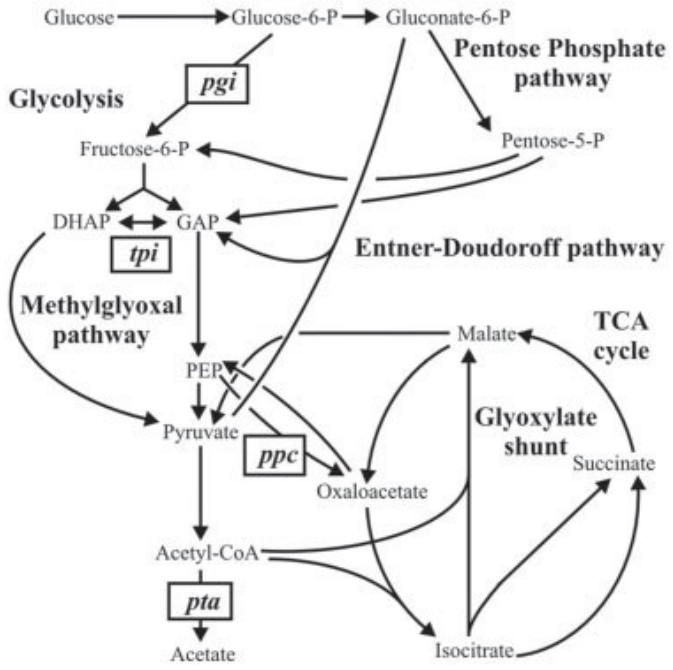


Figure 2.5 Metabolic genes deleted in evolved strains characterized by ^{13}C-labeling experiments. Deleting pgi (phosphoglucose isomerase) impedes glycolysis, forcing flux to be rerouted through the pentose phosphate pathway or Entner-Douderoff pathway. Deleting pta (phosphate acetyltransferase) impedes acetate fermentation. Deleting tpi (triose-phoshpate isomerase) forces DHAP to be metabolized through methylglyoxal pathway. The ppc deletion requires the glyoxylate shunt to be used to replenish oxaloacetate. This figure was originally published by Fong et al. 2006 (13).

Interestingly, the same pathways were used to circumvent the lost gene function both before and after adaptation, suggesting that adaptation functioned by increasing the flux through those pathways. This outcome suggests that *E. coli* immediately responds to a breakdown in metabolic capacity by activating repressed pathways to search for a way to redistribute flux through the metabolic network - a process that may favor short routes that cause the least disruption of the wild type flux configuration (68). This suggests that these adaptations involve optimizing the first

accessed and established solution rather than representing the culmination an exhaustive sampling of the genetic and phenotypic space for a complex but more effective mechanism. A more sophisticated adaptation could possibly appear if the populations were given more time to search their genotype/phenotype space, but since the recovered growth phenotype is near that of wild type, there may not be much potential for significant gains.

Additionally, even though replicate strains generally circumvented their metabolic lesions by the same latent pathway, this does not suggest that all evolved gene-deletion strains will eventually converge to utilize the same flux configuration. There were several significant differences in how some replicate strains utilized other pathways. For example, one evolved Δ*pgi* strains primarily utilizes the TCA cycle and secretes acetate while another has more flux through the glyoxylate shunt and secretes no acetate. These outcomes suggest that much of the variability between replicate evolutions may stem from variable means of making downstream metabolic adjustments that are necessary to refine usage of the major adaptive flux shift.

2.3.3 Gene expression changes during adaptation

The previous section described the flux changes associated with adaptation, and demonstrated that growth rate increase result from improved flux through the metabolic network. However, ^{13}C flux analysis provides no information about the mechanisms responsible for causing the uptake and flux increases, which result from refining the activity of metabolic elements. Metabolic activity is regulated at multiple levels, including allosteric control of enzymes, turnover rates, and transcriptional regulation.

A subsequent study examined changes to genome-wide mRNA transcription levels among the replicate glycerol- and lactate-evolved strains over the course of

Table 2.2 Number of Expression Changes across Evolutions on Glycerol and Lactate.

			Glycerol Evolved		Lactate Evolved
relative to wild type (glucose)	Day 1	39%	1687 genes	18%	756 genes
	Day 20	18%	770 genes	4%	194 genes
	Day 44	11%	498 genes	7%	323 genes

Table shows the average number of significant gene expression changes in adaptively evolved strains relative to wild type grown on glucose at different time points in the evolution experiments. Data from (11).

their evolutions (26). Global expression was also measured of cultures at day 1 and day 20 of each evolution, and of wild type grown on glucose. Interestingly, the transcription state of each evolved endpoint was distinct despite similarity in endpoint growth phenotypes, further indicating that endpoint cultures acquired different sets of adaptations.

Analysis of the transcription profiles revealed that the largest number of changes in gene expression were between wild-type during growth on glucose and wild-type grown on the substrate challenge (day 1 cultures), and that the course of adaptive evolution returned the expression of many of these genes to pre-evolution levels (**Table 2.2**). This could results from a large scale carbon-scavenging response intended to allow metabolism of any 'low-quality' carbon sources available in the absence of preferred substrates (54), in which case the adaptation process may involve refining the regulatory response to only activate genes specific to glycerol or lactate metabolism.

Expression changes that developed over the course of evolution that were common across all replicate lineages were also identified. Approximately 70 gene expression changes were identified in common among the glycerol-evolved strains, but only two among the lactate-evolved strains. The small number of genes identified across multiple lactate strains may be due to their adaptive pathways being more divergent compared to glycerol strains, which is discussed later in this chapter (**Section 2.6**).

Overall, it appears to be very difficult to identify adaptive mechanisms through examination of expression changes alone. Adaptation generally changes the expression of a large number of genes, and it is difficult without knowledge of changes to other intracellular systems, such as metabolism or of specific mutations, to identify critical expression shifts that mediate the adaptive mechanism rather than result as a secondary consequence.

Additionally, transcription changes in the metabolic gene deletion strains were also measured, and the results were analyzed in conjunction with the flux changes discussed in the previous section (**Section 2.3.2**) to identify gene expression changes responsible for the metabolic and phenotypic adaptations (28). However, as previously stated, not all of the flux changes are caused by expression changes as there are a significant number of other mechanisms that contribute to the regulation of metabolism. Additionally, not all gene expression changes directly alter the phenotype or metabolic flux, or are necessarily adaptive.

Nevertheless, many of the observed flux changes were linked to expression changes in the adapted gene deletion strains (28). Flux through the glyoxylate shunt, methylglyoxylate shunt, and TCA cycle correlated well to changes in expression of the relevant enzymes, suggesting that flux through these pathways are at least partially

controlled at the transcriptional level. On the other hand, no expression changes were identified that correlate to observed flux changes through glycolysis or the pentose phosphate pathway, suggesting that other mechanisms may predominantly regulate those pathways. Interestingly, no correlation was found between flux and gene expression changes in the pre-evolved deletion strains. This may indicate that the major, shared flux shifts were initially mostly dependent on other mechanisms of metabolic regulation, and that adaptation was necessary to acquire changes involving transcriptional regulation.

Additionally, the flux and expression data collected from the evolved gene-deletion strains has been used to change the metabolic model to account for several failed predictions between estimated and achieved growth rates (35). The developed method, called optimal metabolic network identification (OMNI), searches for metabolic reactions that, when eliminated, can produce flux distributions that are the closest fit to those actually measured in the relevant strains. The enzymes of these reactions may act as bottlenecks to optimal growth, that for some reason are disadvantageously regulated in a manner that had not been overcome by the point at which the experimental evolution was terminated. The validity of this approach is supported by the fact that expression of many of the genes identified as possibly causing flux bottlenecks can be seen to have reduced expression in the evolved strain compared to wild type.

2.4 Genome Resequencing

The altered phenotype capabilities of the evolved strains are ultimately caused by genetic changes. These must be identified to develop a comprehensive understanding of the adaptation process. Recent advances in high-throughput DNA

Table 2.3 Mutations Identified in Glycerol-evolved strains

Clone	Gene	Product or Function	Mutation	Gene position nt	Region	Genome Position
GB-1	glpK	Glycerol kinase	a->t	218	Coding	4115028
	rpoC	RNA polymerase β'	27 bp deletion	3132-3158	Coding	4186504-4186530
GC-1	glpK	Glycerol kinase	g->t	184	Coding	4115062
	n/a	All genes between insC-5 & insD6	1313kb duplication[1]	n/a	n/a	~3189209-4497523
GD-1	glpK	Glycerol kinase	g->a	816	Coding	4114430
	rpoB	RNA polymerase β	a->t	1685	Coding	4180952
	murE	peptidoglycan biosynthase	a->c	8	Coding	93173
GE-1	glpK	Glycerol kinase	a->c	113	Coding	4115133
	rpoC	RNA polymerase β'	c->t	2249	Coding	4185621
	dapF	Lysine / peptidoglycan biosynthase	c->a	512	Coding	3993293
G2-1	glpK	Glycerol kinase	9 bp duplication	705	Coding	4114541
	rph-pyrE	RNAse PH/ pyrimidine synthesis	82 bp deletion	rph: 610-end	Coding + Intergenic	3813882-3813963
	pdxK-crr	Pyridoxal kinase / enzyme IIA glucose	28 bp deletion	pdxK: 833-end	Coding + Intergenic	2534400-2534427

Mutations identified by whole-genome resequencing of *E. coli* strains adaptively-evolved to increase growth on glycerol minimal media. Previously published (16).
1. Evident in CGS mapping data; not independently validated.

sequencing have made resequencing the genomes of adapted *E. coli* strains economically feasible.

Conclusively identifying all of the mutations acquired during adaptive evolution requires a search of the entire genome that can find changes as small as of a single nucleotide (or a single nucleotide polymorphism - SNP). Six of the glycerol-evolved strains have undergone whole-genome resequencing (37). With one exception, the strains acquired two to three SNPs within the coding region of annotated genes, as described in **Table 2.3**. All five resequenced strains acquired a

different mutation within *glpK*, which encodes glycerol kinase which catalyzes the rate-limiting step in glycerol metabolism. Three of the strains also acquired mutations in RNA polymerase subunits β and β', encoded by *rpoB* and *rpoC*, which was surprising given the extensive influence these genes may have on global transcription regulation. Additionally, mutations were acquired by two strains that effect peptidoglycan biosynthesis (within genes *dapF* and *murE*).

The impact of individual mutations on the fitness phenotype was determined by assessing strains that had been altered to carry one or more of the discovered mutations using site-directed mutagenesis (37). Importantly, the set of mutations identified in each strain have been proven to be responsible for the phenotype change, since the endpoint phenotype has been shown to be reproduced by inducing the mutation sets into wild type by site-directed mutagenesis. The phenotype of mutant strains carrying only single mutations indicate that the *rpoB/C* mutations have the greatest impact on growth rate, followed by the *glpK* mutations. Additionally, results of competitions between strains with different induced mutations indicated that some of the *glpK* and *rpoB/C* mutations may have cooperative (epistatic) effects (1); possible mechanisms are still being investigated. Mutations to the peptidoglycan synthesis genes (*murE* & *dapF*) were also only shown to have a significant effect on growth rate in strains that also carried the co-acquired *glpK* and *rpoB/C* mutations.

The *rpoB/C* mutations clearly perform a critical function in optimizing the growth of *E. coli* on glycerol, and this function most likely involves adjusting transcriptional regulation. These mutations have the greatest impact on fitness among those acquired by glycerol-evolved stains, and they may additionally be responsible for the increased growth capacity in minimal media on a wide range of non-glycerol

substrates (unpublished results). One hypothesis suggests that these mutations improve growth by reducing the sensitivity of RNA polymerase to stress response signals, particularly ppGpp, that may be induced by the transition from rich to minimal media. This may prevent the expression of unnecessary or detrimental proteins that are associated with the stress response (54). The effect of these mutations on RNA polymerase activity and global gene expression is currently being investigated.

Strains evolved on lactate or in response to deletion of the *pgi* gene have also undergone whole-genome resequencing (unpublished data). Interestingly, there was more variation in the number of mutations these strains acquired (0-7 mutations per strain, versus 2-3 in glycerol), and they appeared across a more diverse set of genes than glycerol-evolved strains. It is not yet clear why these evolutions produced more genetically divergent replicates compared to the glycerol-evolved strains, though it suggests there may simply be more adaptive routes that these replicate evolutions can sample.

A developing pattern seen across the different experimental evolutions is that strains frequently acquire mutations to both a metabolic gene and a global transcription factor. The function of mutations to metabolic genes (PEP synthase in lactate-evolved strains, NADPH/NADH transhydrogenase genes in evolved Δ*pgi* strains) is assumed to increase the activity of rate-limiting enzymes under the growth conditions, as has been shown for the *glpK* mutations in the glycerol evolved strains (37). The function of mutated global regulatory or transcription factors is likely more complicated. Among the mutations discovered in lactate and Δ*pgi* strains are mutations to *crp, cyaA*, and *rpoS* (unpublished data). This trend suggests that the *E. coli* regulatory network is both robust to mutations that alter the function of major regulatory elements, and that these are more accessible or advantageous than

mutations that effect the transcriptional regulation of a smaller, more specific set of genes.

2.5 Summary Remarks

This chapter has described how stoichiometrically-constrained models of *E. coli* metabolism have been used to predict optimal growth rates on a variety of carbon sources. These predictions closely approximate the actual growth phenotype of *E. coli* K12 MG1655 observed on many substrates during exponential growth. However, *E. coli* was found to grow slower than predicted on some substrates, suggesting either a model defect or a regulatory deficiency that prevented optimal utilization of those substrates. Prolonged growth of *E. coli* on these substrates generates adaptations that cause the growth phenotype to converge towards the predicted optimum, validating the model. Additionally, the models have been successful at predicting the optimum growth phenotype that *E. coli* can acquire following loss of a metabolic gene function. These outcomes suggest that the metabolic model captures the critical constraints on growth capacity before considering metabolic regulation, without requiring kinetic constraints. Additionally, these results indicates that metabolic regulation is readily malleable to selection pressure to allow the optimum growth phenotype to be found in response to a wide range of environmental or metabolic challenges.

Each evolution experiment was performed multiple times, and although each of the replicate strains generally acquired the same growth phenotype (converging towards the predicted optimum), they generally differed in terms of other phenotypic traits such as byproduct secretion or growth capacity on alternative substrates. Significant differences between replicate strains were also validated by differences

between their flux and transcription profiles. The bases for those differences have now been genetically identified – no two replicate strains have acquired identical mutations. Thus different genetic changes can produce similar phenotypic changes. Flux balance analysis suggests that the different lineages could have also adopted different metabolic solutions, since the optimal phenotype is often calculated to result from multiple flux distributions (62). The degree of variation among replicate lineages may indicate how many adaptive routes exist to produce this phenotype.

Attempts to understand the intracellular changes that underlie the phenotype adaptation include measurements of changes in metabolic flux and mRNA expression over the course of evolution, and identifying acquired mutations. At this point the challenge involves deducing the mechanism by which the discovered mutations alter metabolic flux through regulatory mechanisms to produce increased growth rate. Among the most intriguing discoveries are the mutations to so-called "global regulators" like RNA polymerase subunits and elements of catabolite repression, that appear to play a significant role in inducing the adapted phenotype. It remains to be proven whether this truly is a general mechanism of adaptation, and if so how it functions.

These studies begin to highlight the potential of using adaptive evolution to discover important cellular dynamics that exist on the systems-biology level. The eventual goal is to be able to predict what regulatory and metabolic changes are necessary to produce a desired phenotype. We will have accomplished a true understanding of the relationship between phenotype and genotype when we understand the adaptive function of acquired mutations – and are able to predict them.

2.6 Acknowledgments

I would like to express my gratitude to Karsten Zengler, Jan Schellenberger, and Adam Feist for their insightful feedback, and to thank Jessica Na for her encouragement during the writing of this chapter.

2.7 References

1. **Applebee, M. K., M. J. Herrgard, and B. O. Palsson.** 2008. Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. J Bacteriol **190:**5087-94.

2. **Bennett, B. D., E. H. Kimball, M. Gao, R. Osterhout, S. J. Van Dien, and J. D. Rabinowitz.** 2009. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol **5:**593-9.

3. **Burgard, A. P., P. Pharkya, and C. D. Maranas.** 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng **84:**647-57.

4. **Chulavatnatol, M., and D. E. Atkinson.** 1973. Kinetic competition in vitro between phosphoenolpyruvate synthetase and the pyruvate dehydrogenase complex from Escherichia coli. J Biol Chem **248:**2716-21.

5. **Cohen, I. R., and D. Harel.** 2007. Explaining a complex living system: dynamics, multi-scaling and emergence. J R Soc Interface **4:**175-82.

6. **Edwards, J. S., R. U. Ibarra, and B. O. Palsson.** 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol **19:**125-30.

7. **Feist, A. M., C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. O. Palsson.** 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol **3:**121.

8. **Feist, A. M., and B. O. Palsson.** 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol **26:**659-67.

9. **Fischer, E., N. Zamboni, and U. Sauer.** 2004. High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. Anal Biochem **325:**308-16.

10. **Fong, S. S., A. P. Burgard, C. D. Herring, E. M. Knight, F. R. Blattner, C. D. Maranas, and B. O. Palsson.** 2005. In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol Bioeng **91:**643-8.

11. **Fong, S. S., A. R. Joyce, and B. O. Palsson.** 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res **15:**1365-72.

12. **Fong, S. S., J. Y. Marciniak, and B. O. Palsson.** 2003. Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. J Bacteriol **185:**6400-8.

13. **Fong, S. S., A. Nanchen, B. O. Palsson, and U. Sauer.** 2006. Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. J Biol Chem **281:**8024-33.

14. **Fong, S. S., and B. O. Palsson.** 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet **36:**1056-8.

15. **Herrgard, M. J., S. S. Fong, and B. O. Palsson.** 2006. Identification of genome-scale metabolic network models using experimentally measured flux profiles. PLoS Comput Biol **2:**e72.

16. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

17. **Hua, Q., A. R. Joyce, B. O. Palsson, and S. S. Fong.** 2007. Metabolic characterization of Escherichia coli strains adapted to growth on lactate. Appl Environ Microbiol **73:**4639-47.

18. **Ibarra, R. U., J. S. Edwards, and B. O. Palsson.** 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420:**186-9.

19. **Ishii, N., K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita.** 2007. Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science **316:**593-7.

20. **Karp, P. D., I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, C. Bonavides-Martinez, and J. Ingraham.** 2007. Multidimensional annotation of the Escherichia coli K-12 genome. Nucleic Acids Res **35:**7577-90.

21. **Keseler, I. M., C. Bonavides-Martinez, J. Collado-Vides, S. Gama-Castro, R. P. Gunsalus, D. A. Johnson, M. Krummenacker, L. M. Nolan, S. Paley, I. T. Paulsen, M. Peralta-Gil, A. Santos-Zavaleta, A. G. Shearer, and P. D. Karp.** 2009. EcoCyc: a comprehensive view of Escherichia coli biology. Nucleic Acids Res **37:**D464-70.

22. **Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler.** 1991. Long-Term Experimental Evolution in Escherichia Coli. I. Adaptation and Divergence During 2,000 Generations. American Naturalist **138:**1315-1341.

23. **Liu, M., T. Durfee, J. E. Cabrera, K. Zhao, D. J. Jin, and F. R. Blattner.** 2005. Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. J Biol Chem **280:**15921-7.

24. **Patil, K. R., I. Rocha, J. Forster, and J. Nielsen.** 2005. Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics **6:**308.

25. **Pharkya, P., A. P. Burgard, and C. D. Maranas.** 2003. Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. Biotechnol Bioeng **84:**887-99.

26. **Pharkya, P., A. P. Burgard, and C. D. Maranas.** 2004. OptStrain: a computational framework for redesign of microbial production systems. Genome Res **14:**2367-76.

27. **Pramanik, J., and J. D. Keasling.** 1997. Stoichiometric Model of *Escherichia coli* Metabolism: Incorporation of Growth-Rate Dependent Biomass Composition and Mechanistic Energy Requirements. Biotechnol Bioeng **56:**398-421.

28. **Reed, J. L., and B. O. Palsson.** 2004. Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res **14:**1797-805.

29. **Reed, J. L., and B. O. Palsson.** 2003. Thirteen years of building constraint-based in silico models of Escherichia coli. J Bacteriol **185:**2692-9.

30. **Sauer, U.** 2004. High-throughput phenomics: experimental methods for mapping fluxomes. Curr Opin Biotechnol **15:**58-63.

31. **Segre, D., D. Vitkup, and G. M. Church.** 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A **99:**15112-7.

32. **Varma, A., and B. O. Palsson.** 1994. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol **60:**3724-31.

Chapter 3

Impact of Individual Mutations on Increased Fitness in Adaptively-Evolved Strains of *Escherichia coli*

Majority of material was previously published in the Journal of Bacteriology 190(14):5087-94, July 2008.

The dissertation author was the primary author, co-authored by Dr. Markus Herrgård and Dr. Bernhard Palsson

3.1 Abstract

We measured the relative fitness among a set of experimentally-evolved *E. coli* strains differing by a small number of adaptive mutations by directly measuring allelic frequencies in head-to-head competitions using a mass spectrometry-based method. We compared the relative effects of mutations to the same or similar genes acquired in multiple strains, by comparing the effects of the mutations when expressed in allele replacement strains. We found that the strongest determinant of fitness among the evolved strains was the impact of beneficial mutations to the RNA polymerase β and β' subunit genes. We additionally identified several examples of epistatic interactions, between *rpoB/C* and *glpK* mutations, and identified two other mutations that are beneficial only in the presence of previously-acquired mutations but have little or no

adaptive benefit to the wild-type strain. Allele-frequency estimation is shown to be a highly-sensitive method for measuring selection rates during competitions between strains differing by as little as one single-nucleotide polymorphism, and may be of great use for investigating epistatic interactions.

3.2 Introduction

It is still difficult to understand the mechanisms employed to adapt to a defined metabolic challenge, even in well-characterized organisms like *Escherichia coli*. Genome-scale metabolic models have accurately predicted metabolic phenotypes such as growth and secretion rates after adaptation (5), but these models do not include the capacity to predict what genetic changes could induce the flux shifts that produce the end point phenotype. The mutations acquired during adaptation to a defined challenge have been identified following several adaptive evolution experiments (8, 23, 24), but in bacterial systems the mutations have been found to occur in multiple, distinct biological processes - including to genes without an easily-understood role in adaptation to the imposed pressure. Adding to the challenge of discovering the biochemical mechanism of a particular mutation's effect on phenotype, selection may involve pleiotropic consequences or epistatic interactions with other mutations (12, 20), which cannot generally be predicted. Epistatic interactions, or the dependence of the fitness value of some mutations on the presence or absence of other mutations, can be especially important to consider since they are thought to create the potential for divergent endpoint genotypes among replicate adaptive evolution experiments (11, 25).

In this study, we compared the fitness phenotype between similar sets of mutations acquired by four endpoint strains from parallel laboratory evolutions of

Escherichia coli K12 on glycerol minimal-medium for 44 days (4). The mutations fixed in these endpoint strains were previously identified by whole-genome sequencing, and were found to define the full genetic basis for adaptation (8). All four strains have acquired a mutation to glycerol kinase (*glpK*), three strains have mutations to one of two RNA polymerase genes (*rpoB/C*), and two strains have mutations to peptidoglycan synthesis genes (*dapF/murE*) (**Figure 3.1a**). The growth rates of the strains were relatively similar (0.46-0.58 hr^{-1}), indicating that they had achieved similar fitness gains over the wild type strain. However, none of these mutations are to the same loci, suggesting the endpoint strains had reached different peaks on the fitness landscape. We performed competition experiments, as described below, which measure the relative fitness among the endpoint strains to establish whether they have meaningfully-distinct phenotypes.

We also compared the fitness phenotype induced by mutations at similar loci. This allows us to compare how relative fitness among each type of mutation correlates to relative fitness among the endpoints. The outcomes also revealed several instances of synergistic epistasis between pairs of co-acquired adaptive mutations. This study represents the first time that relative fitness between parallel-evolved strains has been correlated to the fitness effects among the individual mutations that constitute the full genetic basis of adaptation. Though these results alone are not sufficient to explain the physiological mechanism mediating the genotype-phenotype relationship, they contribute significantly to the design of future studies that may – the relative fitness among similar mutations can be used to analyze which enzyme activity changes caused by the mutations correlate to increased fitness, and knowledge of epistatic interactions justifies study of the interaction networks of the effected proteins.

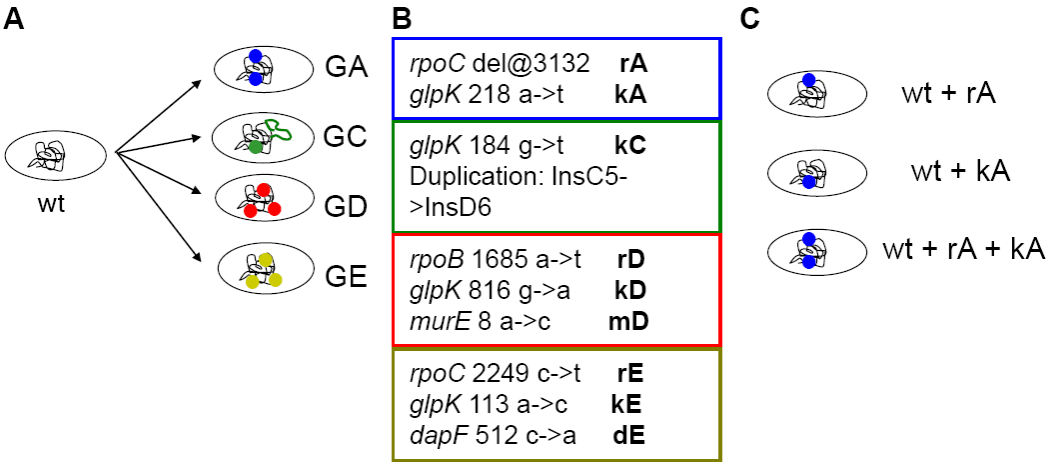


Figure 3.1 Origin of the strains utilized in this study (**a**) Parallel adaptive evolutions produced the four endpoint strains analyzed. (**b**) The mutations acquired by the endpoint strains were identified by whole-genome resequencing and are listed in the colored boxes. The short-hand notation for each mutation indicates the effected gene (r=*rpoB/C*, k=*glpK*, m=*murE*, d=*dapF*) and the endpoint strain it was found in (i.e., kA = *glpK* mutation identified in strain GA) (8). (**c**) Allele replacement strains were then constructed using gene gorging (7, 8) to introduce the mutations to the wildtype genetic background. An allele replacement strain was created that carries each mutation identified in isolation, as well as strains that carry multiple or all mutations identified in an endpoint strain.

We measured relative fitness using head-to-head competition experiments between pairs of strains, and detected the change in population frequency by measuring the allele-frequency of each strain's known mutations. Competition experiments are the standard method for assessing relative fitness between closely-related strains of bacteria in a defined environment (9, 13, 16, 20), and entail culturing two strains together in the same vessel to compete directly for resources, while monitoring the rate of population frequency change as one strain selectively out-competes the other (**Figure 3.2a**).

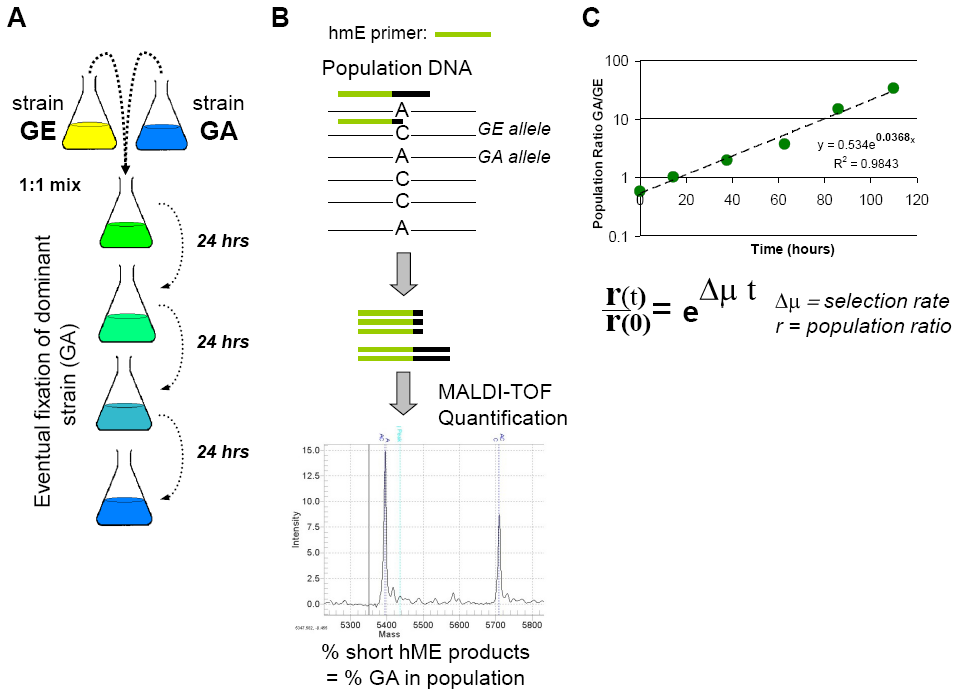


Figure 3.2 Design of the competition experiments. (**a**) The two strains to be competed are inoculated into a single vessel to compete directly for media resources. The culture is propagated for three days by serial passing, with samples collected daily to determine the current population frequency. (**b**) The population frequency at each time point is based on the allele frequency of the known mutations in each strain. DNA isolated daily from the competing culture is subjected to heterogeneous mass-extension assays (Sequenom Inc., San Diego, CA) to determine the frequency of the mutations that distinguish the competing strains. The 200 bp region around each mutation is PCR amplified (not shown), followed by PCR extension of an hME primer that binds directly upstream of the mutation. Extension of the primer is performed with NTP mix missing a species of nucleotide that complements one of the alleles, causing amplification of mutation-harboring alleles to terminate at different loci than the wildtype, producing products with distinct masses. All primers are listed in **Supplementary Table 3.1**. The relative amounts of the extension products are determined by MALDI-TOF mass spectrometry from the ratio of the product peak areas. (**c**) We calculate the estimated selection rate as the Malthusian parameter of the change in the population ratio over time (13). Assay bias was negligible in most cases **(Table 3.3**).

The population composition is often monitored using a detectible or selectable genetic marker in one of the competed strains to allow discriminatory colony counting. Because of the labor involved in generating strains with selectable markers,

often all measured selection rates are generated relative to a marked wildtype strain rather than by direct competition (13, 16). Other recently-developed methods that can be used to monitor competitions include fluorescent protein expression (6) and a plasmid sequence-based code (9) to trace lineages. In this study, competitions were monitored by directly measuring the frequencies of the mutant alleles, which allowed us to measure direct head-to-head competitions between strains that differed by as little as a single mutation and to use multiple allelic markers in instances where strains differed by multiple mutations. This approach provides significantly increased sensitivity to detect fitness differences as compared to optical-density based growth-rate measurements of individual strains that were previously used to study glycerol evolved strains (13).

3.3 Methods & Materials

3.3.1 Strains

Strains GA, GC, GD, and GE were studied because of the similarities between their sets of mutations. Further descriptions of these strains have previously been published (4, 8).

3.3.2 Competition Experiments

The two strains to be competed were cultured separately overnight from frozen stocks in M9-glycerol media. After sufficient culture density was reached by both inoculums (>0.2 OD), cultures were combined based on their relative optical densities to produce the desired population ratio. Competitions were initiated by inoculating flasks of 250 ml pre-warmed M9-glycerol media with the combined culture. The initial population ratio was typically prepared to be approximately 1:1 of each strain, though the initial ratio was occasionally varied to favor the less-fit strain to

allow more time-points to be sampled before fixation. The initial ratio was not observed to affect the observed selection rate in several independent competitions begun with varying initial ratios (1:1000, 1:100, 1:10, 1:1).

Competition cultures were typically maintained for 3 days by serial passage once per day. Cultures were maintained at 30°C with aeration, identical to evolution conditions (8). The volume of spent culture passed each day was estimated based on the previous day's growth-rate and the number of hours until the subsequent passage, to produce an OD of ~ 0.3. The population frequency of each strain was measured from their SNP allele frequency in samples of spent culture collected each day (5-20 ml). We recorded the exact time of collection (hour:minute). Competitions between a pair of strains were performed in duplicate or triplicate.

Competitions were limited to three days because the selection rates of competitions between allele replacement strains were often observed to change after longer periods of time. This was interpreted to be due to the emergence of new adaptive mutations within the competing cultures, since allele replacement strains with single mutations still have high adaptation potential. As described in the statistical methods section, data points collected after a clear selection rate shift were discarded. The selection rate shifts generally differed between biological replicates of the experiments, in terms of the time of emergence, the strain affected, and the magnitude of the shift.

3.3.3 Allelotyping

The DNA from competing cultures was isolated using Qiagen DNEasy kit, using ultra-pure water to elute DNA from the column. All samples were diluted to 5-10 ng/ul DNA concentration.

Allelotyping reactions were performed as described elsewhere (8) (**Figure 3.2b**) using Sequenom hME protocols. Assay primers are listed in **Supplementary Table 3.2**. All samples were assayed in triplicate or quadruplicate, and the measurements averaged. The accurate detection limit is reached at approximately 5% population frequency, putting a limit on the population frequency range that can be accurately sampled. Furthermore, some genome loci have proven difficult to develop working hME assays for, possibly due to local DNA conformation. Techniques for measuring frequencies of these loci, including *glpK* 113 a->c (kE), have been developed by modifying the related hMC resequencing protocols, as described elsewhere (8).

The small bias in some assays due to primer bias and/or other unidentified sources was corrected using normalization curves generated by assaying mixtures of purified endpoint and wild-type DNA manually combined to the desired ratios (1:9, 4:6, 6:4, 9:1). The best-fit curves between actual and detected allelic frequencies were exponential ("power-trend line" in MS Excel) with an average exponent value of 1.25±0.28.

3.3.4 Calculation of the Selection Rate

The Malthusian parameter is assumed to equal the difference in growth rate between the two competing strains, assuming maintained exponential growth in a constant environment with no quorum sensing or other cell-cell interactions (13). Correspondingly, strain fitness is based solely on growth rate, and the selection-rate, r, is equal to the difference in the growth rate between the two strains.

$$r_{AB} = m_A - m_B = [\ln(A_f / A_i) - \ln(B_f / B_i)] / \Delta t = [\ln(A_f/B_f) - \ln(A_i/B_i)] / \Delta t$$

The variables A and B refer to the population frequencies of the competed strains, at times initial (i) and final (f), and Δt is time passed. The selection rate of each competition is calculated as the slope of the best-fit line on a plot of the natural log of the population-frequency ratios versus time (**Figure 3.2c**). The data is presented in terms of the selection rate to allow values within a set of competitions to be directly compared, to facilitate analysis of epistatic and non-transitive interactions.

3.3.5 Data analysis and statistical methods

There were three different types of replication in the competition experiments: (1) biological replicates corresponding to separate competition experiments with the same two strains, (2) multiple assays used to measure the population frequencies within the same competition experiment, and (3) technical replicates corresponding to separate measurements of the individual allelic frequencies using the hME method. The raw population frequency data for each individual replicate competition experiment and allelotyping assay was first log transformed.

Biological replicate competitions that either didn't have sufficient number of reliable data points or strongly deviated from linearity were eliminated by manual inspection of individual population frequency curves. Similarly, outlier data points for individual competitions that were indicative of assay saturation or sudden slope-change due to selection to fix a mutation acquired during the experiment were eliminated. After filtering out outliers, the total number of individual data points used to estimate the selection rates ranged between 18 and 144 depending on the competition experiment. For each pair of strains competed there were two to five biological replicates, one to four separate assays depending on the number of mutations in the strains, and three technical replicates per biological replicate and assay (**Table 3.1**).

Table 3.1 Numbers of biological replicates and separate assays used in each competition experiment. The total number of data points refers to the number of separate time-population frequency points used to fit the mixed effect model for each pair-wise competition.

Competition	Strains	Biological	Assays	Data points
GA-GC	Endpoints	4	2	67
GA-GE	Endpoints	3	4	144
GD-GA	Endpoints	4	3	143
GD-GC	Endpoints	4	3	113
GD-GE	Endpoints	5	4	124
GE-GC	Endpoints	3	3	65
kA-kD	glpK	4	2	86
kA-kE	glpK	3	2	18
kC-kA	glpK	2	2	47
kC-kD	glpK	3	1	21
kC-kE	glpK	3	2	34
kD-kE	glpK	4	2	51
rkA-rkE	glpK+rpoB/C	4	4	130
rkD-rkA	glpK+rpoB/C	3	4	106
rkD-rkE	glpK+rpoB/C	4	4	120
dE-wt	murE/dapF	3	1	18
GD-rkD	murE/dapF	3	1	35
GE-rkE	murE/dapF	3	1	35
mD-wt	murE/dapF	3	1	33
rA-rE	rpoB/C	3	2	68
rD-rA	rpoB/C	4	2	62
rD-rE	rpoB/C	4	2	63

For each pair of strains competed we fitted a linear mixed effects model to all the data obtained for that pair simultaneously (i.e. all biological and technical replicates and assays). This allowed us to use all the data effectively to estimate the selection rates as well as standard errors for these rates. The model we used for estimating selection rates for individual competitions assumed that all replicates/assays have the same slope, but may have different intercepts corresponding to different initial ratios between the starting strains. We used a linear

Table 3.2 Summary of selection rates from competitions.

Competition	Strains	S x100	SE	df	t	P-value
GA-GC	Endpoints	9.13	0.32	61	28.93	2.5E-37
GA-GE	Endpoints	4.99	0.11	137	46.30	1.5E-85
GD-GA	Endpoints	2.63	0.11	136	23.48	1.1E-49
GD-GC	Endpoints	8.24	0.23	106	35.11	3.4E-60
GD-GE	Endpoints	7.13	0.28	115	25.54	3.3E-49
GE-GC	Endpoints	0.73	0.04	59	19.22	4.4E-27
kA-kD	glpK	0.74	0.10	80	7.82	1.8E-11
kA-kE	glpK	4.47	0.38	13	11.73	2.7E-08
kC-kA	glpK	0.49	0.06	43	7.62	1.7E-09
kC-kD	glpK	4.58	0.64	17	7.18	1.6E-06
kC-kE	glpK	4.83	0.63	29	7.70	1.7E-08
kD-kE	glpK	0.66	0.04	45	14.94	4.6E-19
rkA-rkE	glpK+rpoB/C	9.39	0.27	122	34.83	3.1E-65
rkD-rkA	glpK+rpoB/C	2.07	0.05	99	37.89	9.9E-61
rkD-rkE	glpK+rpoB/C	10.41	0.33	112	31.66	1.1E-57
dE-wt	murE/dapF	0.22	0.09	14	2.58	2.2E-02
GD-rkD	murE/dapF	4.98	0.21	31	23.33	3.2E-21
GE-rkE	murE/dapF	2.60	0.06	31	41.59	9.2E-29
mD-wt	murE/dapF	-0.06	0.05	29	-1.21	2.4E-01
rA-rE	rpoB/C	4.56	0.18	63	25.84	3.1E-35
rD-rA	rpoB/C	1.37	0.04	56	34.25	3.0E-39
rD-rE	rpoB/C	5.14	0.29	57	17.89	4.7E-25

S, selection rate (× 100); SE, standard error (95% confidence interval); df, degrees of freedom for the mixed effect model fit; t, Student's t value; P-value, t-test P-value (null hypothesis S=0).

mixed effect model with the assay effect treated as a fixed effect and the biological replicates treated as random effects(3). The mixed effect model was fit using the restricted maximum likelihood approach implemented in the nmle package in R (21).

A summary of the results obtained using the mixed effect model, including the estimated selection rates, is shown in **Table 3.2**. To detect possible assay-specific effects in individual competitions (i.e. cases where different assays give significantly different selection rates) we modified the mixed effect model so that the slope (or selection rate) was allowed to depend on the specific assay utilized. Statistically significant assay effects ($P < 0.01$) were detected in 9 of the 29 competitions. However, most of the assay effects were relatively small in magnitude (**Table 3.3**).

Table 3.3 Assay-specific effects. The selection rates estimated from each assay separately is shown. The P-values are obtained by a t-test using the null hypothesis that the selection coefficient estimated by a particular assay (e.g. R458 for GA-GC) is different from the selection coefficient estimated by the first assay listed for each competition (G326 for GA-GC). The cases where a significant assay specific effect was detected are indicated by red color. Data is only shown for competitions where more than one assay was used to estimate the selection rate (see **Table 3.2**).

GA-GC		Competition
G326	**R458**	**Assays**
0.076	0.096	Selection rates
	0.007	P-values

GA-GE				kA-kD	
DapF	**G360**	**R286**	**R458**	**G360**	**G480**
0.049	0.047	0.050	0.053	0.007	0.008
	0.333	0.763	0.128		0.535

GD-GA			kA-kE	
G360	**MurE**	**RpoB**	**G255**	**G360**
0.029	0.024	0.026	0.036	0.047
	0.042	0.283		0.281

GD-GC			kC-kA	
G326	**G480**	**MurE**	**G326**	**G360**
0.070	0.108	0.067	0.006	0.004
	3.40E-18	0.330		0.275

GD-GE				kC-kE	
DapF	**G480b**	**R286**	**RpoB**	**G255**	**G326**
0.081	0.055	0.069	0.082	0.033	0.053
	2.14E-05	6.25E-03	0.847		0.196

GE-GC			kD-kE	
G255	**G326**	**R286**	**G255**	**G480**
0.008	0.006	0.009	0.007	0.006
	0.061	0.475		0.485

rA-rE		rkA-rkE			
R286	**R458**	**G255**	**G360**	**R286**	**R458**
0.046	0.046	0.101	0.100	0.070	0.106
	0.989		0.911	0.022	0.723

rD-rA		rkD-rkA			
R458	**RpoB**	**G360**	**G480**	**R458**	**RpoB**
0.015	0.013	0.021	0.022	0.018	0.023
	0.003		0.291	0.052	0.133

rD-rE		rkD-rkE			
R286	**RpoB**	**G255**	**G480**	**R286**	**RpoB**
0.054	0.049	0.169	0.116	0.079	0.114
	0.414		2.90E-05	8.86E-12	1.45E-05

3.4 Results

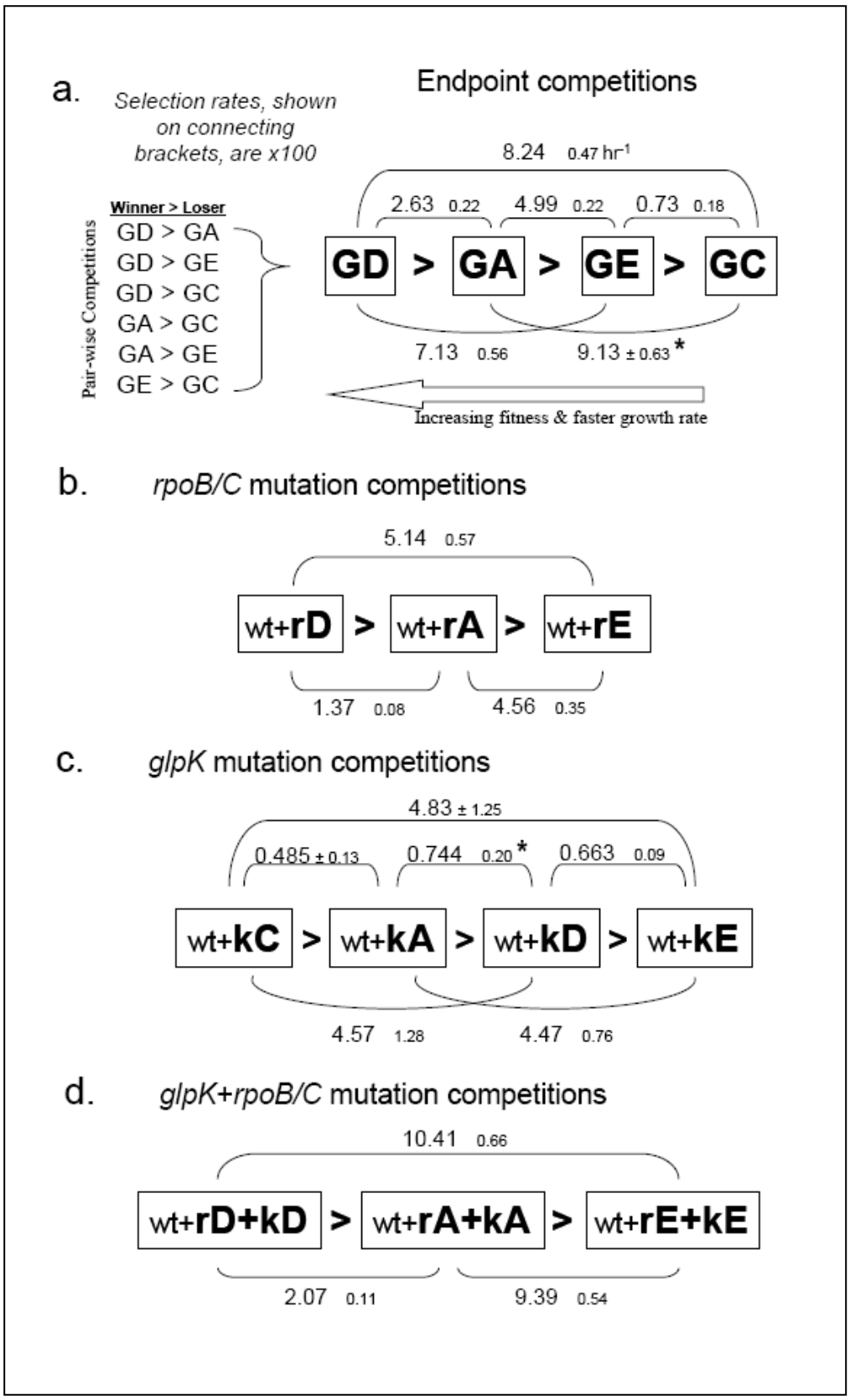
Competition experiments were performed between all pairs-wise combinations of glycerol-evolved endpoint strains, and between allele-replacement strains with similar mutations (7, 8) (**Figure 3.1**). The outcomes are reported as the selection rate (in units of 1/h) which corresponds to the estimated absolute growth-rate difference between the two strains **(Figure 3.3)**.

3.4.1 Fitness hierarchies determined by competitions

Competitions between the four endpoint strains yielded an internally consistent fitness hierarchy, with reproducible selection rates that were significantly different from zero in all cases (mixed effect linear model t-test $P < 10^{-26}$; see Methods for details). The relative fitness of each endpoint strain was distinctly different from the other strains. For example, the GE strain was found to be significantly less fit than the GA strain although the growth rates of these strains were essentially indistinguishable when measured by optical density (13).

We next compared the fitness effects of similar mutations acquired by different strains by competing allele-replacement strains that harbor each mutation individually. Like endpoint strains, the set of three *rpoB/C* mutation allele-replacement strains (**Figure 3.3b**) and set of four *glpK* mutation allele-replacement strains (**Figure 3.3c**) proved to have an internally-consistent fitness hierarchy, with selection rates that were reproducible and significantly different from zero ($P < 10^{-6}$). This demonstrates that the individual mutations each have a distinct impact on the fitness of the organism.

Figure 3.3 Head-to-Head Competition Results. The selection rate-constant between each pair-wise competition, ±2SE, is listed on the bracket connecting two strains. Displayed values have been increased x100. Each box represents a strain within the set. The outcomes within the four sets of competitions are internally-consistent. (**a**) Endpoint strain competition outcomes. Results of pair-wise competitions are on the left, and combined in the diagram on the right with the measured selection-rates displayed on the connecting brackets. The increased strain fitness correlates roughly to higher measured growth rate. The mutations acquired by each endpoint strain are listed in **Figure 3.1b**. (**b**) Outcome of competitions between allele replacement strains harboring one of the identified *rpoB/C* mutations. Mutations are described by the abbreviations shown in **Figure 3.1b**, that reference the source endpoint strain and the gene affected. (**c.**) Outcome of *glpK* allele-replacement strain competitions. (**d.**) Competitions between double-allele replacement strains that carry both the *glpK* and *rpoB/C* mutations identified in the same endpoint strain.
* Indicates the value does not agree well with estimates extrapolated from the rest of the set and may represent systematic measurement error (See Discussion).


a.
Endpoint competitions
Selection rates, shown on connecting brackets, are x100
8.24 0.47 hr-1
2.63 0.22
4.99 0.22
0.73 0.18
Winner > Loser
Pair-wise Competitions
GD > GA
GD > GE
GD > GC
GA > GC
GA > GE
GE > GC
GD > GA > GE > GC
7.13 0.56
9.13 ± 0.63 *
Increasing fitness & faster growth rate
b.
rpoB/C mutation competitions
5.14 0.57
wt+rD > wt+rA > wt+rE
1.37 0.08
4.56 0.35
c.
glpK mutation competitions
4.83 ± 1.25
0.485 ± 0.13
0.744 0.20 *
0.663 0.09
wt+kC > wt+kA > wt+kD > wt+kE
4.57 1.28
4.47 0.76
d.
glpK+rpoB/C mutation competitions
10.41 0.66
wt+rD+kD > wt+rA+kA > wt+rE+kE
2.07 0.11
9.39 0.54

The *rpoB/C+glpK* strains have a fitness hierarchy that matches both the *rpoB/C* and the endpoint strain set hierarchy in terms of the endpoint strains the mutations were identified from (i.e. D > A > E > C). However, the ordering of the *glpK* hierarchy differs significantly from the consensus order identified in the other experiments (**Figure 3.3**). This result suggests that the adaptive value of the acquired *rpoB/C* mutations

Table 3.4 Epistatic Interactions suggested by Additive Analysis. Epistatic interactions can be inferred when mutations cause different selection rate increases in different genetic backgrounds. This table shows the selection rate differences between mutations to *glpK* and *rpoB/C* Significant discrepancies between the selection coefficients of the same *glpK* and *rpoB/C* mutations when competed individually and in combination indicate epistatic interactions between the mutations of one or both of the involved endpoint strains. *Indicates the value does not agree well with estimates extrapolated from the rest of the set and may be questionable (See Discussion).

	competed allele-replacement strains	selection rate difference x100 (hr^{-1})
GA mutations vs. GE mutations	*glpK* mutation strains:	4.47 +/- 0.08
	rpoB/C mutation strains:	4.56 +/- 0.35
	sum	**9.03** +/- 0.43
		⇳ **no epistasis**
	rpoB/C+glpK strains:	**9.39** +/- 0.54
GD mutations vs. GE mutations	*glpK* mutation strains:	0.66 +/- 0.09
	rpoB/C mutation strains:	5.14 +/- 0.57
	sum	**5.8** +/- 0.66
		⇳ **indicates epistasis**
	rpoB/C+glpK strains:	**10.41** +/- 0.66
GD mutations vs. GA mutations	*glpK* mutation strains:	-0.74* +/- 0.2
	rpoB/C mutation strain:	1.37 +/- 0.08
	sum	**0.63** +/- 0.28
		⇳ **indicates epistasis**
	rpoB/C+glpK strains:	**2.07** +/- 0.11

largely determine the fitness of the endpoint strains, overshadowing the effects of the *glpK* mutations.

3.4.2 Evidence of epistatic interactions

We additionally analyzed the selection rates from allele-replacement strain competitions for evidence of epistatic interactions between co-acquired *glpK* and *rpoB/C* mutations in the three endpoints that carried mutations to both genes. If there are no epistatic interactions, we expect the selection rate between two *rpoB/C+glpK* double allele-replacement strains to approximately equal the sum of the selection rates of the same *glpK* and *rpoB/C* mutations when competed separately. This analysis is summarized in **Table 3.4**. The selection rates between allele-replacement strains carrying GA and GE's mutations are approximately additive, suggesting the pairs of mutations in neither of these strains have synergistic or antagonistic interactions. In contrast, when the mutations from GA or GE are similarly competed against GD's mutations, the fitness difference between the *rpoB/C+glpK* double mutant strains is significantly greater than the sum of the fitness differences found when the same *rpoB/C* mutations and *glpK* mutations are competed separately. This indicates epistatic interactions between the mutations in at least one strain – either synergism between *glpK* and *rpoB/C* mutations in the double-mutant of greater fitness (rkD in both cases), or antagonism between the mutations in the less-fit double-mutant (rkA or rkE) in each competition. If the later case is true, the additive outcome observed from competitions between strains rkA and rkE could be produced if the antagonistic interactions between the *glpK* and *rpoC* mutations in strain GA and GE are approximately equal.

The GD and GE strains both acquired an additional mutation apart from the *glpK* and *rpoB/C* mutations. These mutations were in the *murE* and *dapF* genes involved in peptidoglycan synthesis (**Figure 3.1**) and they were the last mutations these strains acquired (8). The growth rates of allele replacement strains harboring these mutations suggested that they have little to no adaptive value as single mutations to wild type, though they increased the growth rate when added to strains carrying their co-acquired *glpK* and *rpoB/C* mutations. To test for potential epistatic relationships, we performed competition experiments between strains that differed only by the *dapF* or *murE* mutations in both the wild type and endpoint genetic backgrounds (**Figure 3.4**), which confirmed that both the *dapF* and *murE* mutations

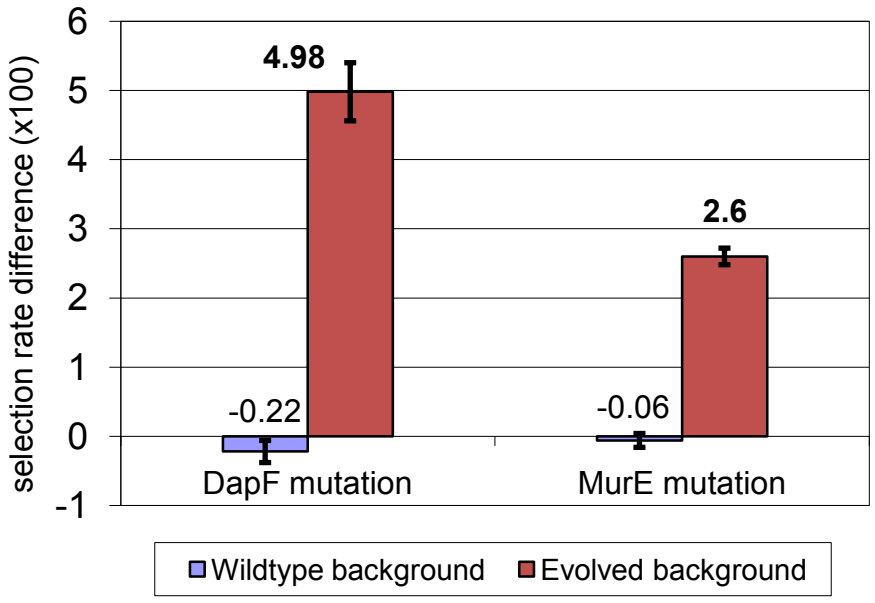


Figure 3.4 Effect of *dapF* and *murE* mutations on fitness changed by co-acquired mutations. Competitions were performed in both the wildtype and evolved genetic background, between strains that differed only by ± *murE* or *dapF* mutations. Both mutations had effectively no adaptive value when added to the wildtype genetic background (violet bars), but were clearly adaptive when added to strains harboring their respective co-acquired *rpoB/C* and *glpK* mutations (red bars). This strongly suggests the *dapF* and *murE* mutations epistatically-interact with at least one of their co-acquired mutations.

considerably increase relative fitness of strains with the co-acquired *glp+/rpoB/C* mutations, but have little impact on wildtype.

3.5 Discussion

Previous experimental adaptive evolution studies have demonstrated strong parallelism between independent endpoints. Viral adaptation experiments have produced independent endpoints in which the majority of adaptive mutations are at identical loci (23). Although bacterial laboratory evolution studies have found far fewer exact allele changes in common between replicate evolutions, specific genes have frequently been identified that acquired adaptive mutations in all or most of the replicate evolved strains (8, 24, 27). This parallelism is presumed to be due to the strains following slightly divergent adaptive paths towards the same improved phenotype. By competing the endpoint strains and allele-replacement strains, we have measured the relative effectiveness of the four apparently-parallel adaptive paths taken by these strains.

3.5.1 Fitness hierarchies and adaptive mechanisms

The relative fitness phenotype hierarchy among the *glpK* and *rpoB/C* mutations will be used to help interpret results from assays of mutant enzyme activity. The fact that the majority of the replicate evolutions repeatedly acquire mutations to the same genes suggests that the mutations are selected to alter the same functional attribute. If so, fitness gains may be largely proportional to the mutation's affect on that function (though this may also be complicated by pleiotropic and epistatic interactions (20)). However, mutations can have multiple effects on enzyme activity, especially in proteins with complex functions like the β and β'-RNA polymerase subunits. The relative fitness hierarchy can provide a criterion for identifying the

activity changes that correspond to increasing the fitness phenotype. The ability of the *glpK* mutations to increase fitness has been found to be primarily determined by their ability to reduce sensitivity to the inhibitor fructose-1,6-bisphosphate, and this is the subject of Chapter 4. Efforts to profile the activity of the mutated β- and β'-RNA polymerase subunits (encoded by *rpoB* and *rpoC*, respectively) are underway and indicate significant the mutations generally increase fitness in minimal medium and that there are many intriguing kinetic differences between wild type and mutated polymerases (Conrad T et al 2010, forthcoming publication).

3.5.2 *rpoB/C* mutations have largest impact on fitness

The results of this study further indicates that the impact of the *rpoB/C* mutations largely determine endpoint fitness. The order of the *rpoB/C* mutations in terms of increased fitness correlates to the rank order among the *rpoB/C*+*glpK* double mutant strains and the endpoint strains. This, and the fact that the allele replacement strains harboring *rpoB/C* mutations have higher growth rates than strains harboring their co-acquired *glpK* mutations (8) indicates that they contribute more to the fitness of the endpoint strains. This conclusion is further supported by the evolutionary trajectory of each strain, which show that the *rpoB/C* mutations were the first acquired in strains GA, GD, and GE (1, 8, 18, 19). Interestingly, the *rpoB/C* mutations did not become fixed until the lineage also acquired the *glpK* mutations.

The role of the *rpoB/C* mutations is believed to be linked to stabilizing the initially large-scale changes in the transcription profile observed in each evolution's population, that eventually stabilized and largely returned to pre-perturbation expression state by the end of the adaptation period (4). The *rpoB/C* mutations may accomplish this by influencing catabolite repression (8) or the stringent response to minimal media. The biological effects of the *rpoB/C* mutations on global

transcriptional activity are currently being investigated to clarify how the adaptive benefits of these mutations are realized.

3.5.3 Analysis of competition results

All four sets of competitions had internally consistent outcomes, and their hierarchies correlate to the order of increasing growth rates among the strains (8) as predicted for selection under continuous exponential growth (13). The competition experiments performed here also resolved the relative fitness relationships between strains with indistinguishable growth rates (13).

The competition experiments also produce strong evidence that, except in the case of strain GA, every causal mutation acquired by the studied endpoint strains was

Table 3.5 Test for fitness differences between evolved strains and allele-replacement strains that carry all identified mutations. These experiments were done to verify whether all causative mutations had been found. Allele-replacement 'reconstruction' strains are labeled 'rGX', ie, reconstruction of evolved strain GX.

Competition	Selection Rate (S) (x 100)	Difference in S	SE	P-value	Outcome
GD-GA	3.050	0.0042	0.0021	0.0446	**GD = rGD**
rGD-GA	2.625				
GE-GD	-7.334	0.0048	0.0077	0.5400	**GE = rGE**
rGE-GD	-7.128				
GC-rGA	-6.702	0.0753	0.0041	**7.06E-35**	**GC > rGC**
rGC-rGA	-14.351				
GA-GC	9.135	-0.0245	0.0040	**1.18E-08**	**GA > rGA**
rGA-GC	6.702				

identified by resequencing during the previous study. We performed competitions against a common competitor against each endpoint strain and an allele replacement strain harboring all of the endpoint's known mutations (except for the large duplication in strain GC). No significant fitness differences were identified when the competitions were performed simultaneously with GD and GE strains, but this was not the case GA and GC (**Table 3.5**). Resequencing of GA by Solexa technology has since discovered an additional mutation in an uncharacterized gene (T. Conrad, personal communication). The fitness inequality found between GC and its reconstruction is attributed to the duplicated region in the evolved strain.

In analyzing the competition outcomes, we discovered that selection rates are not universally transitive within each competition set (ie, if GX>GY>GZ, the selection rates are transitive when the selection rate of GX vs. GZ equals the sum of the selection rates from GX vs. GY and GY vs. GZ). We determined that all major discrepancies could be attributed to the selection rates from the GA vs. GC and wt+kA vs. wt+kD competitions. The magnitudes of these selection rates appear to be over and underestimated respectively in the competition assays, based on analyzing transitivity of the end-point strain (**Figure 3.2a**) and *glpK* allelic replacement strain (**Figure 3.2c**) competition sets. The reason why these selection rate measurements appear to have systematic errors has not been determined. Possible error sources could be due to interactions between the strains in the form of cross feeding, or minor day-to-day differences in culture conditions such as media preparation that specifically affect one of the strains. Despite these considerations, we are confident that the indicated direction of the fitness relationships are still valid in both cases as they are still statistically significant (i.e. the selection rates are significantly different from zero) and are supported by the outcomes of the rest of the competitions.

3.5.4 Implications of discovered epistatic interactions

We identified several epistatic interactions between co-acquired adaptive mutations. At least one endpoint's *glpK* and *rpoB/C* mutations have epistatic interactions, as indicated in **Table 3.4**. The origin of this interaction is not yet known. Chapter 4 discusses a study that determined that the *glpK* mutations primarily determine fitness by reducing sensitivity to inhibition by fructose-1,6-bisphosphate, but that increased *glpK* activity activates negative feedback on *glpK* expression mediated by catabolite repression. This negative feedback is necessary to prevent methylglyoxal toxicity(14). However, GlpK expression is not significantly depressed in the endpoint strains compared to wild type (4, 10), suggesting that other mutations alleviate repression of *glpK* and prevent methylglyoxal toxicity. It seems more likely that the *rpoB/C* mutations accomplish this rather than the peptidoglycan synthesis gene mutations, given that they are acquired more frequently and have a function more likely able to cause this effect. However, this relationship between the *glpK* and *rpoB/C* mutations has yet to be fully explored.

Additionally, the *murE* and *dapF* mutations both synergistically interact with at least one of the mutations previously acquired by their respective endpoint strains (**Figure 3.5**); because the DapF and MurE enzymes have no known relationship to glycerol metabolism or RNA polymerase/transcription regulation, it seems likely that the mutations are adaptive due to indirect consequences of the previous mutations on general cellular properties. We suspect the *rpoB/C* mutations largely facilitate most of the synergistic interactions, since it has been proposed that synergism is an emergent property of complex regulatory networks (22). Additionally, while antagonistic interactions are known to be relatively common in organisms with small

genomes (2, 20), there has previously been little evidence that synergistic interactions occur frequently enough to effect the evolution of organisms like *E. coli* (22).

3.5.5 Need for methods to quantitatively measure relative fitness

Measuring the fitness phenotype differences between single mutations may be applicable to the development of adaptation models. Fitness landscapes of individual enzymes have recently been mapped and successfully used to alter enzyme function by predicting the most effective sets of mutations (15, 26). However, these models require prior knowledge of the fitness value of each mutation since their fitness effects are not easily predicted - and additionally assume that the effects of multiple mutations combine in an additive, non-epistatic manner (17, 26). Developing predictive models of adaptation on a genome scale promises to be much more complicated, since there are many more components and levels of interaction and, consequently, a non-negligible frequency of epistatic interactions (2). Studying the biochemical basis of epistatic interactions could become a valuable strategy for discovering subtle but important functional interactions between distinct cellular components and systems. Regardless, developing models of specific adaptation processes will likely require accurate fitness effect measurements of many single mutations.

3.5.6 Conclusions

This study delineated the relative fitness among four fully-resequenced, parallel-evolved *E. coli* endpoint strains by head-to-head competitions. Competition sets among endpoint strains and constituent-mutation allele-replacement strains were all internally consistent, confirm the distinct fitness phenotype of all of the strains, and indicate the range of fitness gains that can be expected from adaptation to this particular challenge. The relative fitness of the first acquired mutation, to the *rpoB/C*

gene in three strains, largely determines relative fitness among the endpoints. The adaptive benefits of two mutations, to genes *murE* and *dapF*, have little to no benefit in the wildtype background and are epistatically-dependent on co-acquired mutations, though the nature of their dependence is not yet known. The selection rates during each competition were measured with high sensitivity, suggesting this method has great potential for examining the fitness contributions of specific mutations and identifying epistatic interactions.

3.6 Acknowledgements

We gratefully acknowledge Trina Patel for her help during the development of this study, Dr. Christiane Honisch (Sequenom Inc., San Diego, CA) for sharing her invaluable technical insights and expertise, Tom Conrad for his assistance growing cultures, and Dr. Peter Andolfatto (UCSD Dept. of Biology) and Dr. Chris Herring (Mascoma Inc., Lebanon, NH) for insightful discussions.

3.7 Appendix

Supplementary Table 3.1 Primers used in allele frequency assays. PCR primers were used to amplify the ~200 bp region around each mutation. hME primers were used with the listed termination mix to generate allele-specific extension products

	Mutation		PCR Primers		hME primer	Termination Mix
GA	*rpoC*	del27@3132	F:	cactatagggaaggaactGTATGTGCGCACTGCTACGG	TGATCGACGGCCAGACC	ACT
			R:	cactatagggagaaggtcATACGCGCCAGGGTGTCA		
	glpK	218 a->t	F:	cactatagggaaggaactGGCGCGATTGTAGGTGCATTT	CCGATAGCTGCAATTTGA	CGT
			R:	cactatagggagaaggtcTCGCGAGAGCCTTCCACA		
GC	*glpK*	184 g->t	F:	cactatagggaaggaactGGCGCGATTGTAGGTGCATTT	TTTCGCCAGCACTTCTA	CGT
			R:	cactatagggagaaggtcTCGCGAGAGCCTTCCACA		
GD	*rpoB*	1685 a->t	F:	cactatagggaaggaactGCCGATTTCCGCAGCAG	TCGGACCTTCAGGGGTT	CGT
			R:	cactatagggagaaggtcAGCGCGCAGAGTCGGAA		
	glpK	816 g->a	F:	cactatagggaaggaactTGCGAGCATTTAAAACGTGAC	TCGCCAGTGTTCATCAG	ACG
			R:	cactatagggagaaggtcGCGCATACGGGTCCCAG		
	murE	8 a->c	F:	cactatagggaaggaactTTCGCGCTGGTTGTTGTTAT	GAGGGGACAGGTGGCAG	CGT
			R:	cactatagggagaaggtcGCCCACGAGACGTAAATTGTC		
GE	*rpoC*	2249 c->t	F:	cactatagggaaggaactTCCGAGGCAGAAGCAGAAGT	TGGTCTGATGGCGAAGC	ACG
			R:	cactatagggagaaggtcCGCGCCAGGTCACGAC		
	glpK	113 a->c	F:	cactatagggaaggaactGGCGCGATTGTAGGTGCATTT	N/A (see Materials & Methods)	
			R:	cactatagggagaaggtcTCGCGAGAGCCTTCCACA		
	dapF	512 c->a	F:	cactatagggaaggaactCTTCCGCCGTGCCGTT	CAAGCGTTTCTACCGCC	ACT
			R:	cactatagggagaaggtcATGTACCGCCGGGCCA		

Supplementary Table 3.2 Names given to hME-Assays for each mutation.

	genes mutated	mutated loci	hME-Assay Name
GA	*rpoC*	del27@3132	R458
	glpK	218 a->t	G360
GC	*glpK*	184 g->t	G326
		InsC5->InsD6 dup.	*no assay*
GD	*rpoB*	1685 a->t	RpoB
	glpK	816 g->a	G480
	murE	8 a->c	MurE
GE	*rpoC*	2249 c->t	R286
	glpK	113 a->c	G255
	dapF	512 c->a	DapF

3.8 References

1. **Betancourt, A. J., and J. P. Bollback.** 2006. Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev **16:**618-23.

2. **Elena, S. F., and R. E. Lenski.** 1997. Test of synergistic interactions among deleterious mutations in bacteria. Nature **390:**395-8.

3. **Fitzmaurice, G. M., N. M. Laird, and J. H. Ware.** 2004. Applied Lognitudinal Analysis, vol. John Wiley & Sons, New York.

4. **Fong, S. S., A. R. Joyce, and B. O. Palsson.** 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res **15:**1365-72.

5. **Fong, S. S., J. Y. Marciniak, and B. O. Palsson.** 2003. Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. J Bacteriol **185:**6400-8.

6. **Hegreness, M., N. Shoresh, D. Hartl, and R. Kishony.** 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science **311:**1615-7.

7. **Herring, C. D., J. D. Glasner, and F. R. Blattner.** 2003. Gene replacement without selection: regulated suppression of amber mutations in Escherichia coli. Gene **311:**153-63.

8. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-12.

9. **Imhof, M., and C. Schlotterer.** 2001. Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U S A **98:**1113-7.

10. **Joyce, A. R.** 2007. Modeling and Analysis of the E. coli Transcriptional Regulatory Network: An Assessment of its Properties, Plasticity, and Role in Adaptive Evolution. University of California, San Diego, La Jolla.

11. **Kauffman, S. A., and E. D. Weinberger.** 1989. The NK model of rugged fitness landscapes and its application to maturation of the immune response. J Theor Biol **141:**211-45.

12. **Knight, C. G., N. Zitzmann, S. Prabhakar, R. Antrobus, R. Dwek, H. Hebestreit, and P. B. Rainey.** 2006. Unraveling adaptive evolution: how a single point mutation affects the protein coregulation network. Nat Genet **38:**1015-22.

13. **Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler.** 1991. Long-Term Experimental Evolution in Escherichia Coli. I. Adaptation and Divergence During 2,000 Generations. American Naturalist **138:**1315-1341.

14. **Lin, E. C.** 1976. Glycerol dissimilation and its regulation in bacteria. Annu Rev Microbiol **30:**535-78.

15. **Lunzer, M., S. P. Miller, R. Felsheim, and A. M. Dean.** 2005. The biochemical architecture of an ancient adaptive landscape. Science **310:**499-501.

16. **Maharjan, R., S. Seeto, L. Notley-McRobb, and T. Ferenci.** 2006. Clonal adaptive radiation in a constant environment. Science **313:**514-7.

17. **Mildvan, A. S.** 2004. Inverse thinking about double mutants of enzymes. Biochemistry **43:**14517-20.

18. **Orr, H. A.** 2005. The genetic theory of adaptation: a brief history. Nat Rev Genet **6:**119-27.

19. **Orr, H. A.** 2002. The Population Genetics of Adaptation: The Adaptation of DNA Sequences. Evolution **56:**1317-1330.

20. **Pepin, K. M., M. A. Samuel, and H. A. Wichman.** 2006. Variable pleiotropic effects from mutations at the same locus hamper prediction of fitness from a fitness component. Genetics **172:**2047-56.

21. **Pinheiro, J. C., and D. M. Bates.** 2002. Mixed-Effects Models in S and S-Plus, vol. Springer, New York.

22. **Sanjuan, R., and S. F. Elena.** 2006. Epistasis correlates to genomic complexity. Proc Natl Acad Sci U S A **103:**14402-5.

23. **Wichman, H. A., M. R. Badgett, L. A. Scott, C. M. Boulianne, and J. J. Bull.** 1999. Different trajectories of parallel evolution during viral adaptation. Science **285:**422-4.

24. **Woods, R., D. Schneider, C. L. Winkworth, M. A. Riley, and R. E. Lenski.** 2006. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proc Natl Acad Sci U S A **103:**9107-12.

25. **Wright, S.** 1931. Evolution in Mendelian Populations. Genetics **16:**97-159.

26. **Yoshikuni, Y., and J. D. Keasling.** 2007. Pathway engineering by designed divergent evolution. Curr Opin Chem Biol **11:**233-9.

27. **Zhong, S., A. Khodursky, D. E. Dykhuizen, and A. M. Dean.** 2004. Evolutionary genomics of ecological specialization. Proc Natl Acad Sci U S A **101:**11719-24.

Chapter 4

Functional and metabolic effects of adaptive GlpK mutants in *Escherichia coli*

Majority of this material has been submitted to the Journal of Bacteriology for publication as a research article (April 2010).

The dissertation author was the primary author, co-authored by Dr. Donald Pettigrew and Dr. Bernhard Palsson.

4.1 Abstract

We have investigated the ability of four adaptive non-synonymous glycerol kinase mutations to improve fitness during growth on glycerol media. Though each of the residue changes caused by mutations alter multiple enzyme properties, their most significant impact is reduced affinity for the inhibitor fructose-1,6-bisphosphate (FBP). Furthermore, FBP sensitivity reductions correlate to increased fitness and growth in glycerol media across the GlpK mutants, indicating this parameter drives growth improvement. We also found evidence that the ability of the *glpK* mutations to increase GlpK activity in isogenic strains is attenuated by a negative feedback; the *glpK* mutations significantly increase catabolite repression in response to glycerol when expressed in isogenic strains, which

leads to decreased *glpK* expression since *glpK* is CRP-regulated. In addition, the *glpK* mutations reduce growth on several sugars but not organic acids, suggesting that the GlpK mutants interfere with either glycolysis or sugar uptake.

4.2 Introduction

Adaptive mutations can be seen as highly-informative perturbations. They are selected specifically because they alter a genetic element with an activity that is critical for determining the fitness of the organism. A growing number of studies have used adaptive mutations to identify critical biochemical and regulatory constraints, such as the structural mechanism underlying enzyme cofactor preference , the function of uncharacterized protein domains (4), thermodynamic constraints on protein structure (11, 12, 17), and epistatic interactions that limit the development of antibiotic resistance (23, 40, 43). These studies demonstrate that determining how adaptive mutations improve growth can lead to the discovery of subtle interactions that have a large impact on the resulting phenotype, and thus significantly contribute to our understanding of the genotype-phenotype relationship.

Wild type *Escherichia coli* K-12 MG1655 does not grow as well in glycerol-supplemented minimal medium as expected given its complement of metabolic enzymes (30). However, this difference can be overcome by prolonged laboratory adaptation (40+ days) in glycerol M9 minimal medium (20, 30). The adapting cultures were passed each day before they entered stationary phase to select for strains with increased growth. Every lineages produced grows at least twice as fast as wild type in glycerol minimal media (30). The rapid adaptations to the predicted growth rate suggests

that glycerol metabolism in *E. coli* is constrained by suboptimal regulation of glycerol metabolism rather than a change in the chemistry underlying the metabolic pathway.

The mutations that impart the adaption have been identified in seven of these strains by whole genome resequencing (25), which revealed that *E. coli* MG1655 repeatedly acquires non-synonymous mutations to the *glpK* gene, encoding glycerol kinase, and the *rpoB* and *rpoC* genes, encoding the β and β' subunits of RNA polymerase. However, different lineages acquired mutations to different loci within these genes.

The four glycerol kinase mutations profiled in this study were originally identified in glycerol-adapted lineages GA, GC, GD and GE (1, 20, 25) (Table 1). In this study we examine their impact on the catalytic and allosteric properties of glycerol kinase enzyme, expression of *glpK*, and phenotype during growth on other substrates. To determine the impact of the *glpK* mutations on enzyme function, we have produced highly purified variants of each mutant. To study the phenotype induced by the *glpK* mutations independently of other mutations co-acquired during adaptation to glycerol media, we have used strains in where the *glpK* mutations were introduced into the pre-evolved wild type genetic background (25). These strains previously showed that each *glpK* mutation improves growth rate and fitness during growth on glycerol to a similar degree compared to wild type (+~0.07 hr^{-1}). The growth rates of these strains are too similar to clearly distinguish relative fitness differences, but head-to-head competition experiments have been performed and measured small but consistent fitness differences between the effects of the different mutations (1), as shown in **Table 4.1**. The mutations in **Table 4.1** are listed in descending order of relative fitness, and the measured selection rate is between the listed mutation and the next-ranked one.

Table 4.1 Description of examined adaptive GlpK mutants

GlpK residue change[1]	*glpK* genetic locus	Lineage of origin[2]	Mutation isolation strain[2,3]	Effected GlpK domain	Relative fitness rank[3]	Growth Rate of mutation strain(hr^{-1})[4]	Selection rate over next-fittest mutant strain[3] ($x10^{-3}$ hr^{-1})
V-61-I	g184t	GC	BOP 234	Tetramer formation surface	1	0.326 ±0.028	4.85 ± 1.3
D-72-V	a218t	GA & GB	BOP 21	Tetramer formation surface	2	0.334 ±0.026	7.44 ± 2.0
M-271-I	g816a	GD	BOP 154	Conserved ATPase core domain II	3	0.319 ±0.020	6.63 ± 0.9
Q-37-P	a113c	GE	BOP 252	Active site loop & Putative GlpR site on *glp*[5]	4	0.314 ±0.022	

Footnotes:

1. Loci of changed residues differ from their description in Herring et al (2006)(25) to conform to the convention of not counting the N-terminal methionine residue cleaved following protein translation.
2. Strains further described in Herring et al (2006) (25).
3. Determined by head-to-head competitions in Applebee et al (2008) (1), of which Chapter 3 is a reprint.
4. Average and standard deviation from five independent measurements. Notice that the error associated with the measurements is as large as the differences in growth rate between the strains, which is why competitions experiments were necessary to determine fitness differences.
5. Reported by Larson et al (1987) (32).

Glycerol kinase catalyzes the phosphorylation of glycerol to glycerol-3-phosphate, which is considered the rate-limiting step in glycerol metabolism (47). GlpK is inhibited by both fructose-1,6-bisphosphate (FBP) (48) and IIA^{Glc} (of the PTS system) (37, 41), both of which inhibit GlpK activity during growth on glucose and other preferred substrates (though inhibition by FBP may be more predominant (28)). The enzyme exists as a multimeric complex as an active dimer or an inactive tetramer (14, 15), and the subunits display negative cooperativity with respect to ATP binding that cause its kinetics to significantly diverge from the Michaelis-Menten model (39). FBP inhibits GlpK by preventing dissociation of the inactive tetrameric form of the enzyme when bound (15, 35, 37).

Growth can be improved by alleviating a limiting constraint. However, the constraint alleviated by an adaptive mutation can be difficult to identify since mutations often have multiple pleiotropic effects – even in relatively straight-forward adaptations, such as single-residue changes to glycerol kinase. It is known that glycerol kinase catalyzes the rate-limiting step of glycerol metabolism (47), suggesting that the *glpK* mutations are selected to improve expressed GlpK activity. However, it is not clear what enzyme properties the mutations alter that most effectively achieve this goal. Though previous kinetic measurements suggest that the *glpK* mutations increase V_{max}, these increases do not correlate to the relative ability of each mutation to increase fitness (1, 25), suggesting that changes to other properties may be more effective. In this study, we examine the relative effect of adaptive GlpK mutants on both enzyme function and metabolic phenotype to identify parameters that correlate to improved fitness.

4.3 Materials and Methods

4.3.1 Purification of mutant Glycerol Kinase enzymes (GlpK)

4.3.1.1 Cloning into pGEX-6P-1

Mutant and wild type glycerol kinase sequences were cloned into vector pGEX-6P-1 (GE Healthcare). This plasmid allows IPTG-inducible expression of the inserted coding region with an added N-terminal glutathione-S-transferase domain with a built-in protease cleavage site (PreScission protease – GE Healthcare). The genes were cloned by amplifying the mutant GlpK sequences from glycerol-evolved endpoint colonies (26) by PCR using primers with EcoRI and XhoI restriction sites on the 5' ends of the F- and R-primers, respectively, which were then digested with the respective enzymes and ligated into similarly pre-digested pGEX-6P-1. The *glpK* sequence in the generated vector for each GlpK variant was validated by Sanger sequencing to insure that the expected mutation was present in each case and that no other mutations had been acquired. In one case (vector carrying *glpK* g184t), a novel non-synonymous mutation had been acquired, likely during gene amplification. This vector was re-generated by using the QuikChange II site-directed mutagenesis kit (Stratagene) to introduce the *glpK* g184t mutation into the pGEX vector carrying the wild type *glpK* sequence.

The EcoRI restriction site introduces the *glpK* start codon eight codons downstream from the PreScission protease cleavage site, introducing the 8 residues Gly-Pro-Leu-Gly-Ser-Pro-Glu-Phe- to the N-terminus of the final version of the purified protein. The pGEX-6P-1 vector also carries a BamHI cloning site that is only five codons away from the protease cleavage site, however, this site was not used because of a BamHI restriction site within the *glpK* gene.

4.3.1.2 GlpK Expression

Each mutant GlpK protein was expressed from pGEX-6P-1 transformed into One Shot BL21 star (DE3) *E. coli* (Invitrogen). Expression from ampicillin-selected clones was induced in 2x 200 ml cultures of LB ampicillin (100µg/ml) by addition of 3 mM IPTG once culture density exceeded 0.6 OD. The IPTG-induced cultures were grown at 37C for an additional 4-6 hours. High expression of GlpK was verified by SDS-PAGE of samples of culture collected periodically during the expression period.

4.3.1.3 GlpK Purification

Pellets were re-suspended in 2 ml of 1xPBS with 10 mM glycerol and 2 mM β-mercaptoethanol. Lysis was achieved by lysozyme digestion and sonication. The GST-tagged GlpK enzymes were purified using 1 ml GSTrap FF columns (GE Healthcare). The sonicated lysate was loaded onto the column after centrifugation to clear insoluble cellular debris, and filtration of the soluble fraction with a 0.45 µm syringe filter to remove debris that could clog the column. A separate GSTrap FF column was used for each glycerol kinase variant. Each column was used as described by the manufacturer, except that 10 mM glycerol was added to both the binding buffer and the PreScission cleavage buffer to stabilize GlpK conformation, and 2 mM β-mercaptoethanol was added to the binding buffer to minimize oxidative damage. Glycerol kinase protein was eluted from the column by cleavage from the glutathione-S-transferase domain using PreScission protease (GE Healthcare). Protein homogeneity was verified by SDS-PAGE. Ethylene glycol was added to eluted protein fractions to a final concentration of 30% total volume before being stored at -80C.

4.3.2 Kinetic Assays

For assays of catalytic and allosteric properties, the enzymes were equilibrated with standard buffer (0.1 M triethanolamine-HCl, 2 mM glycerol, 1 mM EDTA, 1 mM β-

mercaptoethanol) by using gel permeation chromatography. Enzyme concentrations and glycerol kinase activities at pH 7.0 and 25°C were determined and catalytic and allosteric parameters were obtained from fits of initial-velocity enzyme kinetics data as described (36). Enzyme concentrations in the assays were 0.5 µg/mL, unless otherwise noted. The concentration of glycerol was 2 mM in all assays. The concentration of ATP was 2.5 mM for studies of FBP and IIA^{Glc} inhibition. The results are presented as the best fit parameter ± standard error as given from fits obtained with Kaleidagraph v. 3.51 (Synergy Software), unless otherwise noted.

Inhibition by FBP was analyzed by fitting the dependence of the SA on FBP concentration to the equation

$$\mathbf{SA = SA_0 - ((SA_0(1-W)[FBP]^{nH})/((K_{0.5})^{nH} + [FBP]^{nH}))}$$

where, SA is the SA at the indicated concentration of FBP, SA_0 is the SA at 0 FBP, W is SA_∞/SA_0 with SA_∞ the SA in the saturating presence of FBP, $K_{0.5}$ is the FBP concentration that gives one-half maximal inhibition, and nH is the Hill coefficient. The inhibition parameters could not be independently estimated from fitting the data because the apparent FBP affinities were reduced so dramatically (see Results). To enable data for the variant enzymes to be fitted, the values for the Hill coefficient and W were fixed to the values obtained from the fits to the N-terminal extension native enzyme.

Inhibition by IIA^{Glc} was analyzed by fitting the dependence of the specific activity (SA) on the concentration of concentration of IIA^{Glc} to the equation

$$\mathbf{SA = SA_0 - ((SA_0(1-W)[IIA^{Glc}])/(K_{0.5} + [IIA^{Glc}]))}$$

where, SA is the SA at the indicated concentration of IIA^{Glc}, SA_0 is the SA at 0 IIA^{Glc}, W is SA_∞/SA_0 with SA_∞ the SA in the saturating presence of IIA^{Glc}, and $K_{0.5}$ is the IIA^{Glc} concentration that gives one-half maximal inhibition.

4.3.3 *E. coli* strains

All strains utilized in this study are derived from *E. coli* K-12 MG1655 (ATCC #47076), with the exception of those used for vector maintenance and protein expression described above. "Wild-type" *E. coli* refers to genetically-unmodified MG1655 (ATCC #47076) stock strain. The GlpK mutant strains were derived in a previous study from the wild type by introducing individual *glpK* mutations into the genome by λ^{red} recombination (24, 25).

4.3.4 Growth Rate Measurements

Each culture was grown in 200-250 ml of M9 medium + 0.2% of the relevant carbon source, with an initial optical density (OD) at 600 nm of 0.02 to 0.05. Cultures were grown in a 500 ml Erlenmeyer flask and stirred at 1200-1400 rpm using a 6.3 cm Teflon stir-bar in a 30°C water bath. The OD600 of the cultures were measured every 40-80 minutes. Optical density was measured in a Thermo Spectronic BioMate3 spectrophotometer. Data were plotted in semi-log to evaluate the linearity of the readings and to identify the logarithmic phase of growth. Growth rates based on natural logarithm were calculated with the Microsoft Excel SLOPE function, using at least three data points within the exponential growth phase (linear fit R^2 >0.99). The growth rate of the GlpK mutant strains were measured at least three independent times in each substrate.

4.3.5 Intracellular cAMP Assays

Samples of culture were collected during logarithmic growth, between 0.10-0.30 OD(600). These samples were rapidly vacuum filtered onto a Millipore® 25mm diameter, 0.45 µm pore triton-free nitrocellulose filter. The volume of culture passed through the filter was based on the optical density, so that the volume that would contain approximately the same number of cells as 1 ml of culture with an OD(600) of 1 (ie, 5 ml collected from a culture at 0.2OD). Cells on the filter were immediately washed with 10 ml of 30°C fresh media to rinse away extracellular cAMP, and then the filter was quickly submerged in 5 ml ice cold 65% ethanol to quench cellular activity. The samples were then vortexed using the maximum setting for 5-10 seconds and stored at -20°C (16). To prepare the cellular debris for the cAMP ELISA assay, the 65% ethanol solution was evaporated using a speedivac and the dried residue was re-dissolved in cAMP assay buffer. The cAMP levels were assayed using an enzyme-linked immunoassay kit (dual range, GE Healthcare), following the manufacturer's instructions using the non-acetylation protocol. One fifth of each sample was used per assay. Each strain was cultured in triplicate, and each sample was assayed twice.

4.3.6 Quantitative PCR

Samples for RNA transcript analysis were taken from exponentially growing cultures collected between 0.2-0.3 OD(600). Methods used prepare RNA and cDNA have previously been described (13). The cDNA was prepared from three independent cultures of each strain. Each quantitative PCR (qPCR) reaction contained 5-10 ng of cDNA, 0.3 µM of each primer, and 12.5 µl 2X SYBR Master Mix (Qiagen). Each qPCR reaction was performed in triplicate. All qPCR reactions were performed on a Bio-Rad Lightcycler. PCR efficiency of each primer pair was calculated by generating a standard curve generated by qPCR with *E. coli* genomic DNA diluted over a range of 10^2-10^6 fold.

The relative quantity of cDNA of each gene was calculated by normalizing it to the quantity of *dmsA* (assay for *glpK* in strain kA and kC versus wild type) or *rrbS* (all other assays).

4.4 Results

4.4.1 Mutant Glycerol Kinase Catalytic and Allosteric Properties

We measured the kinetic activity of each of the four GlpK allozymes, including initial velocity, ATP affinity, and sensitivity to allosteric inhibitors FBP and IIA^{Glc}, to assess the relationship between changes to enzyme function and fitness caused by the mutations. Results of these studies are shown in **Figure 4.1** and **Table 4.2**.

Expression and purification of the variant GlpK enzymes added eight amino acid residues to the N-terminus of each enzyme. We assessed how this N-terminal extension affects catalytic and allosteric properties by comparing the activities native GlpK with and without the N-terminal extension, as shown in **Table 4.2**. These results indicate that the added amino acids do not significantly alter GlpK.

Figure 4.1. Kinetics of GlpK variants. (A) and (B) show the dependencies of the initial velocities (V_o) of each GlpK variant on the concentration of ATP, with the values normalized to the V_{max} obtained from fits of the data to the Michaelis-Menten equation. The lines show the fit to the Michaelis-Menten equation. The values for V_{max} and $K_{m\text{-}ATP}$ are shown in **Table 4.2**. (C) FBP inhibition. The points show the specific activity of each variant at the indicated concentration of FBP is normalized to the specific activity. The lines show the fit to the equation that models IIA^{Glc} inhibition described in **section 4.3.2**. Values for SA_0 and parameters from the fits are shown in **Table 4.2**.

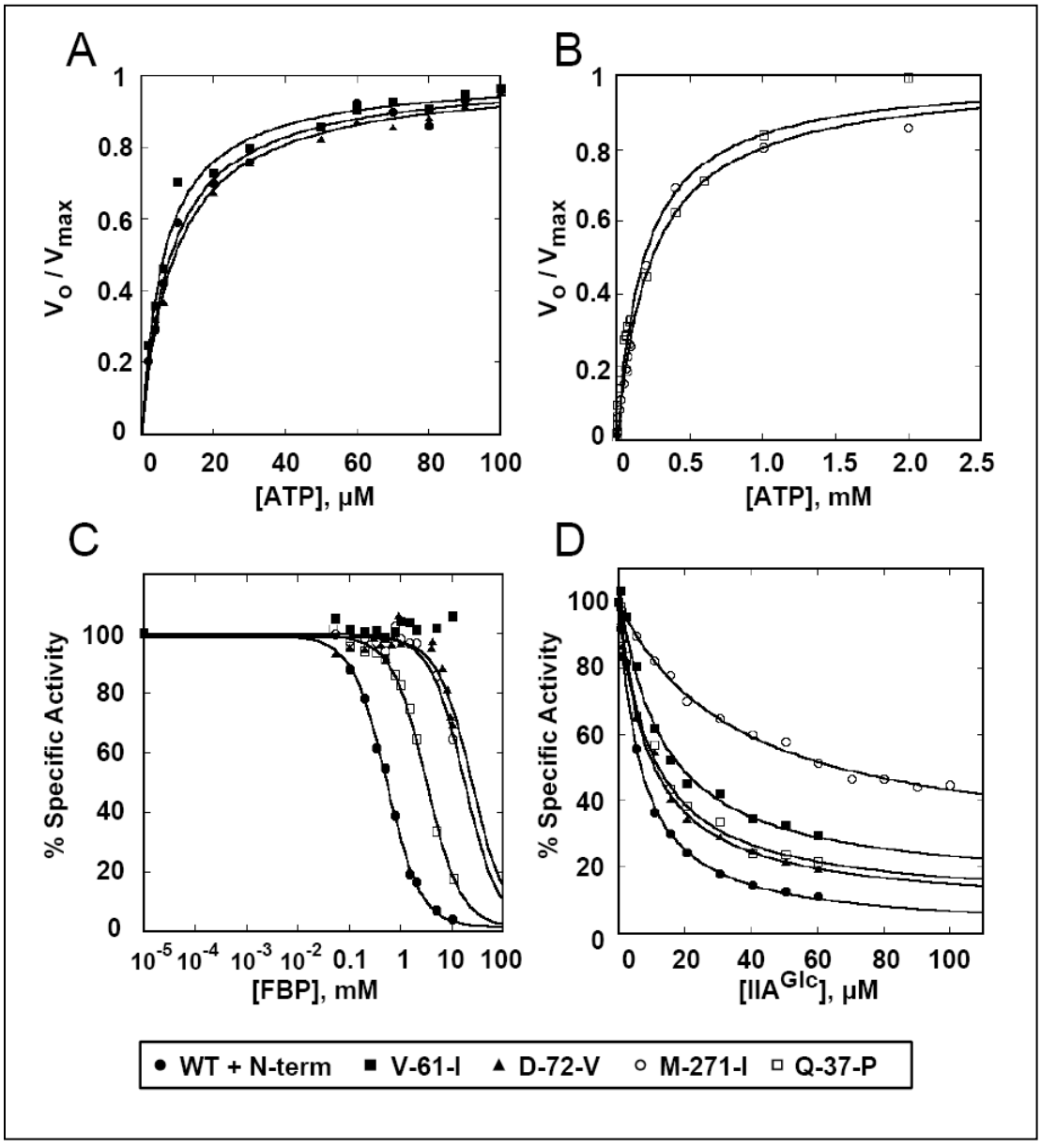


Table 4.2 Catalytic and Allosteric Properties of GlpK mutants

GlpK residue change	Catalytic parameters[1]			FBP inhibition[2]				IIA^{Glc} inhibition[2]		
	V_{max}, U/mg	K_m ATP, µM	V_{max}/K_m, L* $min^{-1}mg^{-1}$	SA_o, U/mg	W	n_H	$K_{0.5}$, mM	SA_o, U/mg	W	$K_{0.5}$, µM
V-61-I	18 ±0.4	6 ±0.6	3 ±0.1	55 ±1	na	na	>100	56 ±1	0.12 ±0.04	13 ±2
D-72-V	24 ±0.4	9 ±1	2.7 ±0.1	53 ±1	0.02	1.3	21 ±2	53 ±1	0.07 ±0.01	10 ±1
M-271-I	77 ±3	250 ±20	0.31 ±0.1	54 ±3	0.02	1.3	16 ±1	69 ±1	0.23 ±0.03	37 ±4
Q-37-P	86 ±3	190 ±20	0.45 ±0.1	80 ±1	0.02	1.3	3.2 ±0.1	80 ±1	0.08 ±0.03	10 ±1
Native + N-term[3]	16.2 ±0.4	8.5 ±0.8	1.9 ±0.1	62 ±1	0.02 ±0.02	1.3 ±0.1	0.52 ±0.01	62 ±1	0.01 ±0.01	6 ±0.6
Native GlpK	18.1 ±0.4	7.8 ±0.7	2.3 ±0.1	62 ±2	0.03 ±0.03	1.7 ±0.2	0.52 ±0.04	64 ±1	0.09 ±0.02	6 ±0.6

Footnotes:

1. Catalytic parameters are obtained from fits of the dependence of the initial velocity on [ATP] in the saturating presence of glycerol (10 mM) to the Michaelis-Menten equation.

2. FBP and IIA^{Glc} inhibition are determined at 2.5 mM ATP and 10 mM glycerol. SA_o is the specific activity at 0 allosteric inhibitor and W is given by the ratio SA_∞/SA_o, where SA_∞ is the specific activity in the saturating presence of the inhibitor. FBP inhibition shows cooperative homotropic effects, given by the Hill coefficient, n_H. $K_{0.5}$ is the concentration of inhibitor that gives one-half of the maximum inhibition. For fits of FBP inhibition for the variant enzymes, the values for W and n_H were fixed to the values obtained for the enzyme with the N-terminal extension.

3. Native GlpK enzyme carrying the same N-terminal extension (Gly-Pro-Leu-Gly-Ser-Pro-Glu-Phe) as the GlpK mutant enzymes, which remains after cleavage of the GST tag used to purify the proteins (**See Section 4.3.1**).

Native GlpK has two apparent Michaelis constants that describe how the initial velocity depends on ATP concentration – a high affinity constant of about 10 μM and a lower affinity constant of about 2 mM (39). Estimating the apparent lower affinity constant is problematic because GlpK activity also becomes affected by excess substrate inhibition when ATP concentration is that high. For this reason, we only examined the effect of the amino acid substitutions on the catalytic parameters (K_m and V_{max}) at ATP concentrations targeted to reveal changes to the high affinity constant. This revealed that while two of the GlpK allozymes (V-61-I and D-72-V) retain both apparent Michaelis constants (their fits to the Michaelis-Menten equation over the high-affinity ATP-binding constant is shown in **Figure 4.1A**), while the other two allozymes (M-272-I and Q-32-P) have only a single K_m-ATP of intermediate affinity (200 μM, Figure 1B – note difference in range of [ATP] shown).

Three of the adaptive GlpK residues changes increase the V_{max} compared to native enzyme. The D-72-V substitution increases the V_{max} associated with the lower $K_{m\text{-}ATP}$ by about one-third, while M-271-I and Q-37-P increase the apparent V_{max} of GlpK by approximately 4-fold.

4.4.1.1 Allosteric inhibition

Strikingly, all four of the fitness-improving GlpK variants also greatly reduce affinity for the allosteric inhibitor FBP, as shown in **Figure 4.1C** and **Table 4.2**. FBP sensitivity was most dramatically altered by the V-61-I variant, which does not show any appreciable inhibition even with 10 mM FBP (higher concentrations of FBP were not evaluated because of its weak binding to the dimer, which reverses the inhibition (46)). The other GlpK allozymes all retain detectable sensitivity to FBP, with $K_{0.5}$ decreasing in the order: D-72-V > M-271-I > Q-37-P > native GlpK. Additionally, the substitutions only cause small

changes in SA_o, specific activity at 2.5 mM ATP without FBP, suggesting the residue changes do not significantly alter catalytic rates at near-saturating concentrations of ATP.

Allosteric control by IIA^{Glc} is also modestly reduced in all of the GlpK allozymes (**Figure 4.1D**). The M-271-I residue change causes the biggest change in IIA^{Glc} inhibition as it reduces IIA^{Glc} affinity three-fold and allows the enzyme to retains approximately 40% of its specific activity when it nears saturation with IIA^{Glc}, compared to native GlpK that retains less than 10% activity. The other residue changes have significantly less impact on sensitivity to IIA^{Glc}.

The parameter most consistently and strongly altered across the GlpK variants appears to be affinity for FBP ($K_{0.5}$). Additionally, changes to this parameter correlate to the relative fitness imparted by the variants, as can be seen in **Table 4.2** where the variants are listed in descending order of relative fitness. This correlation suggests that the ability of the mutations to improve fitness may be derived from to their ability to reduce FBP inhibition.

4.4.2 Effect of mutations on phenotype

In addition to examining how the GlpK variants alter enzyme properties, we have analyzed how they affect phenotype by characterizing four GlpK mutant strains, which each carry one of the *glpK* mutations introduced into the pre-evolved *E. coli* MG1655 wild type strain (1, 25).

4.4.2.1 GlpK Expression

Loss of FBP inhibition by GlpK has been associated with reduced expression of the *glpK* gene (28, 47). We measured *glpK* expression in wild type parent and four GlpK mutant strains during logarithmic growth by quantitative PCR, as shown in **Table 4.3**. These data reveal that glycerol kinase expression is greatly reduced in each of the

Table 4.3 Expression changes of catabolite regulated genes in GlpK mutant strains

GlpK residue change	*glpK* (repressed)[1]	*aceE* (activated)	*aspA* (repressed)	*crr* (repressed)
V61I	12.7% ± 2.8%	149% ± 14%	15.6% ± 1.5%	89.0% ± 7.8%
D72V	14.8% ± 4.7%	118% ± 12%	15.5% ± 1.5%	60.0% ± 5.9%
M271I	30.8% ± 4.2%	270% ± 68%	25.1% ± 9.3%	61.5% ± 15.1%
Q37P	65.7% ± 7.6%	291% ± 73%	42.5% ± 15.5%	89.3% ± 11.6%

1. Indicates effect of catabolite repression on transcription of gene

isogenic GlpK variant strains, ranging from a one-third reduction in the Q-37-P variant to 8-fold reduction in the V-61-I variant compared to wild type. Interestingly, reductions in *glpK* expression correlate to greater ability of the mutants to increase fitness.

4.4.2.2 *glpK* mutations induce autocatabolite repression

Expression of the *glpFKX* operon is repressed by GlpR in absence of intracellular glycerol (specific repression) and by CRP when levels of cAMP are low (catabolite repression). Since GlpR repression should be alleviated by the glycerol present in the medium, the observed reduction in GlpK expression is suspected to be caused by catabolite repression. To test this, we measured intracellular cAMP levels and transcription levels of several other genes predominately regulated by catabolite repression. As shown in **Figure 4.2**, intracellular cAMP levels in the four isogenic GlpK mutant strains are significantly reduced, showing only 30-50% of the amount measured in the parent wild type strain during growth on glycerol M9 minimal medium. Additionally, expression of four cAMP-controlled genes (*aspA* - aspartate ammonia-lyase, *aceE* - pyruvate dehydrogenase EI component, *crr* – enzyme IIA^{Glc} (22)) change in a manner largely consistent with induction of catabolite repression (**Table 4.3**). These results strongly

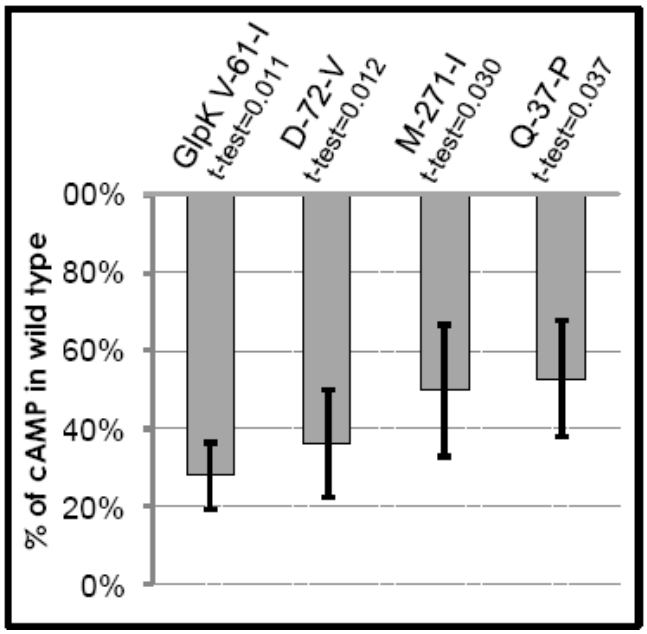


Figure 4.2 Glycerol kinase mutant strains have reduced intracellular cAMP. Intracellular cAMP concentrations in four glycerol kinase mutant strains are reduced compared to wild type (100% on graph). Reported t-tests indicate significance of difference between cAMP concentration in wild type and the GlpK mutant strain. Bars represent average of three independent measurements. Error bars: +/- SD.

indicate the GlpK mutant strains have increased autocatabolite repression during growth on glycerol.

4.4.2.3 glpK mutations impact growth on non-glycerol substrates

The GlpK mutants were also found to reduce growth rates on glycolytic substrates. As shown in Figure 3, the GlpK mutants are adaptive to growth on glycerol, but reduce the logarithmic growth rate on glucose, N-acetyl glucosamine (NAG), and maltose, and have no detectable impact to growth on gluconeogenic substrates (acetate, pyruvate, and succinate). Additionally, the *glpK* mutations did not consistently alter cAMP concentrations during growth on these substrates (Figure 4) (except possibly during growth on maltose), indicating that catabolite repression mechanisms are not responsible for reduced growth rate.

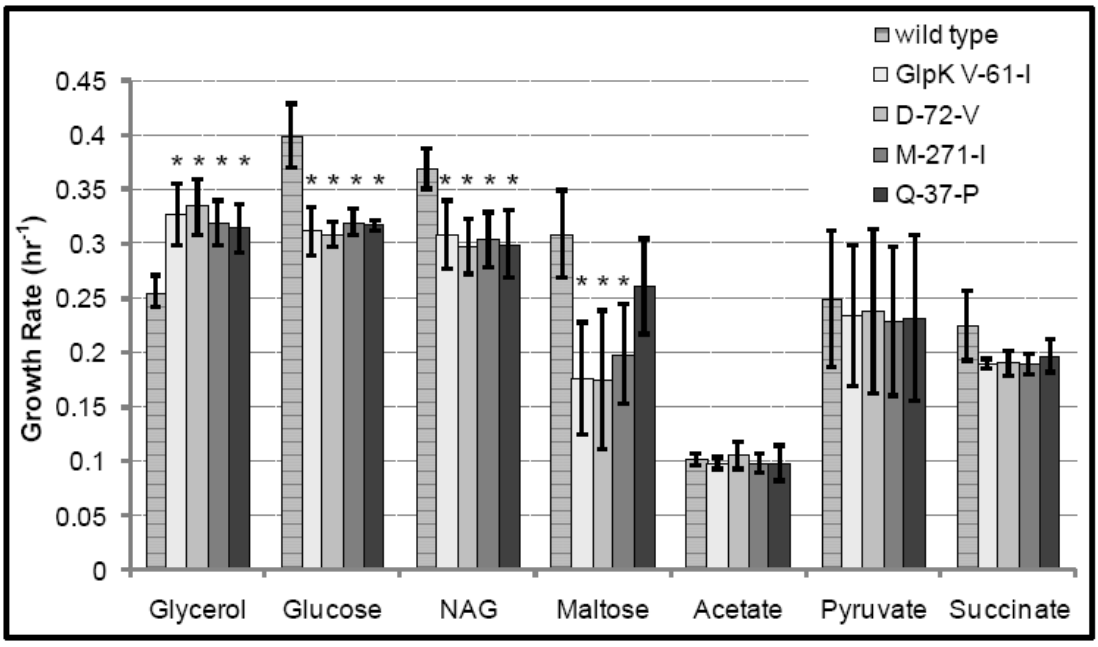


Figure 4.3 Glycerol kinase mutations alter growth on some non-glycerol substrates. GlpK mutants increase the growth rate on M9-glycerol, but unexpectedly, also decrease growth on the sugars glucose, N-acetyl glucosamine (NAG) and maltose. Growth on the organic acids acetate, pyruvate, and succinate are unaffected. Error bars represent +/- SD. *Indicate t-test of $p<0.05$ between mutant strain and wild type growth rate measurements under the same condition. Independent measurements: Glycerol – 5, Glucose – 4, NAG – 3, Maltose – 7, Acetate – 4, Pyruvate – 3, Succinate – 3.

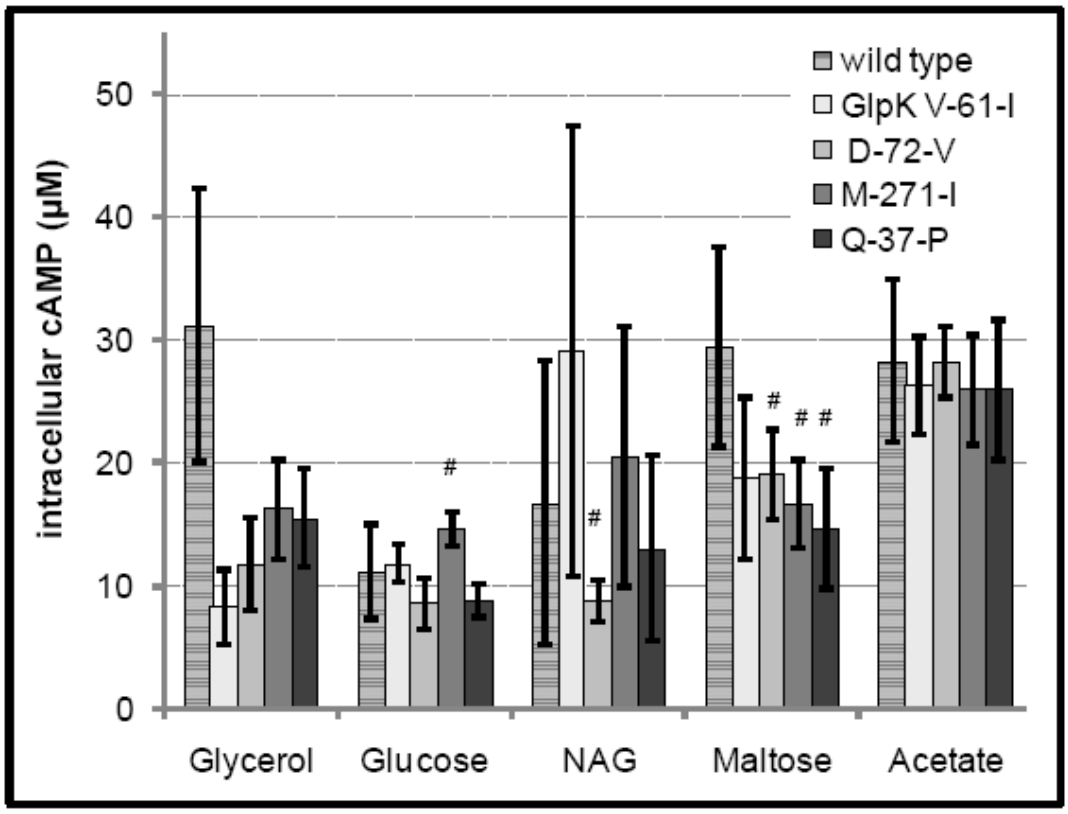


Figure 4.4 cAMP levels in GlpK mutant strains during logarithmic growth on a variety of substrates. The concentration of cAMP in GlpK mutant strains was only measured to be significantly different from wild type during growth on glycerol (t-test $p<0.05$). # - Indicates average of only two independent measurements instead of three. Error bars represent +/- SD. NAG – N-acetyl glucosamine

4.5 Discussion

The goal of this study is to understand how four selection-acquired GlpK mutants improve fitness during growth on glycerol minimal medium. We had previously found that the four GlpK variants have small but significant differences in their ability to improve fitness and growth rate on glycerol M9-minimal media (1), as summarized in **Table 4.1**. Briefly, the variants can be rank-ordered V-61-I > D-72-A > M-272-I > Q-32-P in terms of ability to improve fitness (1). The impact of the mutations on enzyme properties were measured, including kinetic activity, sensitivity to inhibitors, and expression, which reveal that the allozymes most significantly have decreased sensitivity to FBP and increase autocatabolite repression during growth on glycerol. Moreover, fitness of strains harboring only the GlpK mutants correlates to the loss of GlpK sensitivity to FBP inhibition, suggesting that fitness is most strongly determined by sensitivity of GlpK to FBP. In addition, we have discovered that the mutations have an unexpected effect on growth on some non-glycerol substrates.

4.5.1 Catalytic and allosteric properties of the variant GlpK enzymes

4.5.1.1 Comparison to previous kinetic measurements

Estimates of the V_{max} and approximate $K_{0.5\text{-}FBP}$ of the GlpK mutants reported in this study differ considerably from those previously reported in Herring et al (25). Specifically, the previous study estimated that the mutations generally cause much larger increases to V_{max} than the results in this study indicate. The previous results are based on assays of the GlpK allozymes that were partially-purified from lysates of cells induced to over express the enzyme. The assay results reported in this study are considered more reliable since the GlpK enzymes are of higher purity and were assayed over a wider range of substrate concentrations. In addition, this study was

able to quantitatively measure how the mutations affected the affinity for ATP and the apparent inhibition constants associated with FBP and IIA^{Glc} inhibiton.

4.5.1.2 GlpK V-61-I

The V-61-I residue change imparts the largest fitness increase during growth on glycerol minimal medium among the four variants studied (**Table 4.1**). This residue change modestly increases enzyme catalytic efficiency ($V_{max}/K_{m\text{-}ATP}$) and reduces inhibition by the allosteric inhibitor IIA^{Glc}. It also causes the largest reduction in FBP inhibition, essentially abolishing sensitivity to FBP since the FBP titration data could not be to the model (**Figure 4.1C**). Residue 61 sits within an alpha helix that is critical to tetramer stability, between two residues that when exchanged (A-65-T and S-58-W) are known to disrupt tetramer formation and abolish FBP inhibition (6, 35). The V-61-I alteration likely causes a similar disturbance. Mutations that disrupt tetramer formation preclude FBP inhibition because FBP only binds and inhibits glycerol kinase in the tetramer form (37, 46).

4.5.1.3 GlpK D-72-V

The GlpK D-72-V mutation imparts the second largest fitness increase among the four GlpK variants studied. The altered residue is located near the C-terminal end of the alpha helix critical for tetramer formation also impacted by V-61-I. The carboxylate group of Asp72 forms a salt bridge with Lys232, which is located in the loop that binds to FBP. An allozyme with a different residue change at the same locus (asparagine instead of valine) has also previously been characterized, and the change from a carboxyl to an amide side chain was hypothesized to disrupt the electrostatic interactions of the region important to tetramer formation (35). Valine's hydrophobic side chain may have a similar effect. However, this GlpK variant has

significantly reduced rather than abolished sensitivity to FBP, suggesting that it is less effective at destabilizing tetramer formation than V-61-I.

Interestingly, Asp72 has also been observed to be the most frequently mutated GlpK locus. Forty-five additional glycerol-adapted lineages have been generated and screened for GlpK mutations, and the Asp72 residue was altered in 18 of those lineages. In total, Asp72 has changed to valine in 7 lineages and alanine in 13 lineages adapted in glycerol minimal medium (31).

4.5.1.4 GlpK M-271-I

The M-271-I variant not only has significantly reduced FBP affinity, but its kinetic properties and sensitivity to inhibitor IIA^{Glc} are also significantly affected. Met271 is located in the conserved ATPase core in a β-sheet of domain II, where its carbonyl oxygen hydrogen is bonded to the backbone amide of Gly304. This location suggests that consequences of the change may be similar to those caused by the G-304-S substitution, which also effects many enzyme properties and is associated with reduced allosteric control by both FBP and IIA^{Glc} (38).

The V_{max} of the M-271-I enzyme is almost five-fold higher than wild type GlpK. However, this is also one of the residue changes that alter the ATP-binding profile to suggest binding behaves one intermediate Michaelis constant instead of the two observed in native enzyme. This single Michaelis constant suggests the enzyme's affinity for ATP is reduced almost 20 fold compared to the high-affinity constant of native GlpK. This suggests the enzymatic efficiency of the M-271-I variant is reduced overall compared to native GlpK, which could result in less turnover at low concentrations of ATP.

4.5.1.5 GlpK Q-37-P

The GlpK Q-37-P variant causes the smallest fitness increase among the four GlpK variants examined. However, it still significantly improves growth on glycerol minimal medium relative to native GlpK. The altered residue is part of the ATPase conserved catalytic core on a loop in domain I (29). This loop has been shown to undergoe a conformational change when glycerol binds in the glycerol kinase of *Enterococcus casseliflavus*, which shows 78% sequence identity and a very similar structure to the *E. coli* enzyme (45). The location of the nucleotide change in the *glpK* gene has also been identified as a putative GlpR-binding site (44); however, it is unlikely this function effects phenotype since GlpR repression is not active during growth on glycerol.

Like the M-271-I variant, the Q-37-P variant strongly impacts multiple catalytic properties in addition to FBP inhibition. Though it has the weakest impact on FBP sensitivity of the four variants, Q-37-P still increases $K_{0.5\text{-}FBP}$ by over six-fold compared to native enzyme. It also increases V_{max} by over five-fold, which is the largest V_{max} increase among the variants. However, like the M-271-I mutant, increased velocity is coupled with decreased ATP affinity; the $K_{m\text{-}ATP}$ increases to almost 200 μM compared to the 8 μM of the high-affinity site of the native enzyme. This indicates the enzyme efficiency of the Q-37-P enzyme is only about a quarter of native enzyme.

4.5.2 Loss of sensitivity to FBP most significantly altered parameter

Recent advances in metabolomic profiling has produced data from *E. coli* growing on various substrates in minimal medium (2). This data suggests that during growth on glycerol that the intracellular FBP concentration is around 5 mM. At this FBP concentration, native GlpK is expected to be significantly inhibited as this is tenfold higher than the affinity constant of native enzyme (~0.5 mM, Table 2) and **Figure 4.1C** suggests GlpK would retain less than 10% of its specific activity. This indicates there is

significant potential for mutations that reduce FBP sensitivity to increase GlpK activity and glycerol metabolism. The GlpK variants retain at least 30% of their activity at 5 mM FBP, increasing GlpK activity at least 3-fold compared to wild type, even after accounting for small differences in specific activity between the allozymes (SA_o in **Table 4.2**).

Furthermore, the impact of the mutations on FBP sensitivity qualitatively correlates to their ability to improve fitness during growth on glycerol medium, which suggests the two traits are linked. *E. coli* growth on glycerol is limited by GlpK activity (47), indicating that the mutations were selected for their ability to increase intracellular GlpK activity. The correlation to FBP sensitivity suggests that GlpK activity is most effectively increased by reducing FBP inhibition. This is not to say that other GlpK properties that the residue changes alter, such as reduced IIA^{Glc} inhibition of M-271-I and increased V_{max} of Q-37-P, do not contribute to increasing specific activity, but that the correlation to the fitness trend suggests that the effect on $K_{0.5\text{-}FBP}$ is the dominant factor that determines *in vivo* specific activity and relative fitness.

4.5.3 Catabolite repression by GlpK mutations

The GlpK mutations also have consequences on metabolism and phenotype beyond their direct effects on enzyme activity. When the GlpK variants were independently expressed in the parental genetic background, the strains were found to have significantly reduced levels of intracellular cAMP (**Figure 4.2**). Glycerol metabolism with the native GlpK enzyme also induces partial catabolite repression (8, 18, 42), but GlpK mutants cause cAMP levels to drop significantly more – to levels that are only marginally higher than those observed during growth on glucose (**Figure 4. 4**).

This effect was first observed almost 40 years ago in response to another FBP-desensitized GlpK mutant (47), but the mechanism by which the mutants caused

cAMP levels to plummet could not be explained at the time. Since then it has been hypothesized that cAMP production during growth on glycerol is reduced because increased expression of GlpK in response to glycerol leads to most of the IIA^{Glc} becoming sequestered; this prevents the accumulation of phosphorylated IIA^{Glc} that serves as the signal to adenylate cyclase to produce cAMP (8, 42). However, it has since been shown that reduction of cAMP during growth on glycerol is dependent on glycerol-3-phosphate (G3P) not IIA^{Glc} (18). This lead to further work that indicated that G3P, the product of glycerol kinase, induces catabolite repression by inhibiting adenylate cyclase stimulation by phosphorylated IIA^{Glc} (18).

The effects of the glycerol kinase mutations are more consistent with the second hypothesis. We have inferred that the mutations increase fitness by increasing *in vivo* GlpK activity and glycerol metabolism, including corresponding increases in G3P production that could inhibit adenylate cyclase, leading to the observed decreases in cAMP.

Our results also cast further doubt on the IIA^{Glc}-sequestering hypothesis since the GlpK mutants do not generally show increased IIA^{Glc} affinity and because the GlpK mutant strains actually have lower GlpK expression (**Table 4.3**). The GlpK mutant strains also have minorly reduced expression of *crr* that encodes IIA^{Glc} (**Table 4.3**), which could reduce adenylate cyclase activity, but this effect is likely to be in response to catabolite repression rather than a cause of it.

Levels of cAMP levels have also been shown to decline with increasing growth rates across a wide range of growth conditions (3), suggesting the GlpK mutant strains may have reduced cAMP levels as a function of their increased growth on glycerol. The study by Bettenbrock et al suggests that growth rate is linked to cAMP by the PEP:pyruvate ratio, which determines the ratio of phosphorylated IIA^{Glc} and thus

adenylate cyclase stimulation. This mechanism appears sound, however, it is not clear that it sufficiently accounts for these results since the cAMP levels in the GlpK mutant strains are reduced more than expected for the corresponding growth rate increase; ie, the cAMP concentration in the GlpK mutant strains during growth on glycerol are similar to the wild type parent strain during growth in glucose medium (**Figure 4.4**), but wild type grows significantly faster on glucose than the GlpK mutant strains in glycerol (**Figure 4.3**). This suggests that cAMP generation is repressed by an additional mechanism.

4.5.4 Loss of *glpK* expression

Counter-intuitively, the GlpK mutants cause reduced GlpK expression in the GlpK mutant strains (Table 3). This is likely because the *glpFKX* operon is itself a target of catabolite repression (27). This indicates that GlpK activity is also regulated by a negative feedback response where accumulation of G3P reduces *glpK* expression until a new equilibrium between GlpK activity and cAMP concentration is reached. The ability of the *glpK* mutations to increase growth on glycerol suggests the feedback loop attenuates but does not completely counteract the ability of the mutations to increase intracellular GlpK activity.

The existence of a negative feedback mechanism that limits use of a substrate seems counterproductive to growth. However, a similar negative feedback mechanism is known to regulate growth on lactose (5), and such mechanisms may exist to prevent accumulation of toxic metabolites that result from unrestricted substrate uptake (21). In the case of glycerol, unrestricted GlpK activity has long been known to lead to synthesis of methylglyoxal (10, 19), which can cause cell death and forms from over-accumulation of G3P and DHAP (34).

Since G3P is required for phospholipid biosynthesis under all growth conditions, many mechanisms exist to maintain its intracellular concentration at around 200 μM under all growth conditions (32), while somewhat higher amounts accumulate during growth on glycerol that relieve GlpR repression (32). However, methylglyoxal synthesis is known to lead to cell toxicity when both GlpK inhibition by FBP and catabolite-induced glpK repression are undermined during growth on glycerol (34), strongly suggesting that a limit on tolerable GlpK activity exists, which may represent an additional constraint on glycerol adaptation, and also explain why mutations that directly target CRP-repression of *glpFKX* were never selected.

4.5.5 Implications on evolved strains and epistatic interactions

As described in Chapter 4, the analysis of competition experiments suggested that the *glpK* and *rpoB/C* mutations may synergistically interact. The endpoint strains have significantly greater growth rates on glycerol than the *glpK* mutant strains, suggesting the other mutations allow increased glycerol kinase uptake. The *rpoB/C* mutations have been shown to increase glycerol uptake by 20% relative to wild type, but this does not involve increased expression from the *glpFKX* operon compared to wild type (9). This may suggest that the *rpoB/C* mutations increase native GlpK activity and thus uptake by reducing the concentration of FBP. However, an important discovered effect of the *rpoB/C* mutations is that they improve biomass yields from glycerol, suggesting that they increase flux through gluconeogenesis. It is possible that they increase biomass while simultaneously reducing FBP concentrations by improving flux through a rate-limiting step downstream of FBP during growth, or by increasing utilization of other anabolic pathways. It would be interesting to generate ^{13}C-tracing or metabolomic profiles on strains with different combinations of these mutations to examine this possible relationship.

At this point it is also not completely clear whether the endpoint strains also strongly induce the catabolite response during growth on glycerol. Analysis of proteomic data from the endpoint strains found significant enrichment of changes to CRP-regulated genes compared to wild type, while expression profiling data did not (likely due to differences in sensitivity)(33). This enrichment largely suggests increased catabolite repression in the evolved strains compared to wild type, including reduced levels of GlpK; however, the fold change may not be as large as those caused by the GlpK mutations alone. It would be interesting to examine cAMP or GlpK expression levels in *rpoB/C+glpK* double-mutation strains and the evolved strains to determine if any of the additional adaptations alleviate glycerol-induced catabolite repression or otherwise allow increased mutant *glpK* expression. If the GlpK mutants strongly induce catabolite repression in the endpoint strains, it may also be interesting to determine whether the *rpoB/C* mutations affect the catabolite response and *vis versa*, and whether this contributes to the two mutations being repeatedly acquired in rapid succession (25).

4.5.6 Altered growth on glycolytic substrates

The glycerol kinase mutations also significantly reduce growth on several non-glycerol substrates. The underlying molecular mechanism responsible is not clear since GlpK has no known function during growth on non-glycerol substrates. The mutations alter growth on glucose, N-acetyl glucosamine (NAG), and maltose, but have no apparent effect during growth on organic acids (**Figure 4.4**). This suggests that the GlpK mutants interact and interfere with some aspect of growth on glycolytic substrates.

One clear difference between the substrates on which the GlpK mutations do and do not effect growth on is the direction of the glycolytic pathway, since sugar

metabolism employs glycolysis while organic acid metabolism requires gluconeogenesis. This suggests that the GlpK mutants could cause the observed effect by impairing glycolysis. The types of uptake mechanisms also differ between the two sets of substrates; glucose and NAG are both imported by PTS transporters, and maltose is predominately imported by an ATP-binding cassette transporter, while acetate, succinate, and pyruvate are transported by different symporters. These differences suggest that the GlpK mutant strains could also cause the observed effect by impairing certain types of transport. It may also be important to note that GlpK expression is minimal during growth on glucose and NAG since under these growth conditions GlpK expression is repressed by both catabolite repression and GlpR (34), suggesting that mutations are able to cause the growth rate defect even when there is very little mutant GlpK enzyme.

Though the specific mechanism by which the GlpK mutants interfere with growth on glycolytic substrates is beyond the scope of this manuscript, its discovery is interesting since it points to an unrecognized trade-off between optimized glycolytic metabolism and glycerol metabolism. If such a trade-off exists, it could explain why laboratory strains of *E. coli* K-12 MG1655 maintain highly-FBP sensitive variants of GlpK since these strains are grown in sugars more frequently than glycerol.

4.5.7 Reproducibility of GlpK adaptations

Growth on glycerol consistently selects GlpK mutants with reduced sensitivity to FBP inhibition. Non-synonymous *glpK* mutations have been found in 93% (46 of 49) of lineages independently adapted in glycerol minimal media (31), and 29 of those mutations effect the same protein domains as the mutants analyzed in this study, suggesting they reduce FBP sensitivity by similar mechanisms. Additionally, GlpK mutants with reduced FBP sensitivity have also been found following mutagenesis and

screening for reduced glucose control and loss of diauxic growth (35, 48). This suggests that FBP sensitivity is especially accessible to modulation by mutation, both because it is strongly selected during growth on glycerol and because there are many mutations that reduce it that only mildly affect other GlpK properties.

As previously discussed, the persistence of the highly FBP-sensitive GlpK in the wild type, despite the wide availability of variants with reduced sensitivity, may suggest that it is positively selected. However, the number of single-nucleotide *glpK* mutations available that reduce FBP sensitivity may also suggest that the native enzyme specifically retains the ability to quickly adapt in the presence of glycerol (7).

4.5.8 Conclusions

This study demonstrates that examining adaptive mutations to metabolic genes can lead to the discovery of critical metabolic interactions and regulatory dynamics that are otherwise difficult to identify. We have shown that the GlpK variants increase fitness by decreasing affinity for the inhibitor fructose-1,6-bisphosphate (FBP). Additionally, increased catabolite repression, likely caused by increased intracellular GlpK activity, imposes negative feedback on *glpK* expression that most likely exists to prevent methylglyoxal toxicity. In addition, the growth defect of GlpK mutants on glycolytic substrates suggests there is an additional trade-off between specialization on glucose and glycerol. The mechanism behind this growth defect is unknown, but likely interferes with either glycolysis or sugar uptake. This study has contributed to the understanding of the constraints on glycerol metabolism by *E. coli*, and demonstrates the genotype-phenotype relationship of an adaptive mechanism.

4.6 Acknowledgements

I'd like to thank my co-author and collaborator, Donald W. Pettigrew, both for sharing his expertise and performing the kinetic experiments I could not, but also for his generous assistance and patience with a young scientist while writing this manuscript. I'd also like thank Milton Saier for his generous time and feedback, Young Seoub Park and Dae Hee Lee for their invaluable technical advice, Megan Anderson (Quake Lab - Stanford) for sharing her GlpK cloning constructs, and Jessica Na, Pamela S. Miller, and Tzu-Wen Huang for expert technical assistance.

4.7 References

1. **Applebee, M. K., M. J. Herrgard, and B. O. Palsson.** 2008. Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. J Bacteriol **190:**5087-94.

2. **Bennett, B. D., E. H. Kimball, M. Gao, R. Osterhout, S. J. Van Dien, and J. D. Rabinowitz.** 2009. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol **5:**593-9.

3. **Bettenbrock, K., T. Sauter, K. Jahreis, A. Kremling, J. W. Lengeler, and E. D. Gilles.** 2007. Correlation between growth rates, EIIACrr phosphorylation, and intracellular cyclic AMP levels in Escherichia coli K-12. J Bacteriol **189:**6891-900.

4. **Bokma, E., E. Koronakis, S. Lobedanz, C. Hughes, and V. Koronakis.** 2006. Directed evolution of a bacterial efflux pump: adaptation of the E. coli TolC exit duct to the Pseudomonas MexAB translocase. FEBS Lett **580:**5339-43.

5. **Bruckner, R., and F. Titgemeyer.** 2002. Carbon catabolite repression in bacteria: choice of the carbon source and autoregulatory limitation of sugar utilization. FEMS Microbiol Lett **209:**141-8.

6. **Bystrom, C. E., D. W. Pettigrew, B. P. Branchaud, P. O'Brien, and S. J. Remington.** 1999. Crystal structures of Escherichia coli glycerol kinase variant S58-->W in complex with nonhydrolyzable ATP analogues reveal a putative active conformation of the enzyme as a result of domain motion. Biochemistry **38:**3508-18.

7. **Camps, M., A. Herman, E. Loh, and L. A. Loeb.** 2007. Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol **42:**313-26.

8. **Chagneau, C., M. Heyde, S. Alonso, R. Portalier, and P. Laloi.** 2001. External-pH-dependent expression of the maltose regulon and ompF gene in Escherichia coli is affected by the level of glycerol kinase, encoded by glpK. J Bacteriol **183:**5675-83.

9. **Conrad, T. M., M. Frazier, A. R. Joyce, N. E. Lewis, R. Landick, and B. O. Palsson.** (In Preparation). RNA Polymerase Mutants Found Through Adaptive Evolution Re-program Escherichia coli K-12 MG1655 for Optimal Growth in Minimal Media. PNAS.

10. **Cooper, R. A.** 1984. Metabolism of methylglyoxal in microorganisms. Annu Rev Microbiol **38:**49-68.

11. **Counago, R., S. Chen, and Y. Shamoo.** 2006. In vivo molecular evolution reveals biophysical origins of organismal fitness. Mol Cell **22:**441-9.

12. **Counago, R., C. J. Wilson, M. I. Pena, P. Wittung-Stafshede, and Y. Shamoo.** 2008. An adaptive mutation in adenylate kinase that increases organismal fitness is linked to stability-activity trade-offs. Protein Eng Des Sel **21:**19-27.

13. **Covert, M. W., E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson.** 2004. Integrating high-throughput and computational data elucidates bacterial networks. Nature **429:**92-6.

14. **de Riel, J. K., and H. Paulus.** 1978. Subunit dissociation in the allosteric regulation of glycerol kinase from Escherichia coli. 1. Kinetic evidence. Biochemistry **17:**5134-40.

15. **de Riel, J. K., and H. Paulus.** 1978. Subunit dissociation in the allosteric regulation of glycerol kinase from Escherichia coli. 2. Physical evidence. Biochemistry **17:**5141-6.

16. **Death, A., and T. Ferenci.** 1994. Between feast and famine: endogenous inducer synthesis in the adaptation of Escherichia coli to growth with limiting carbohydrates. J Bacteriol **176:**5101-7.

17. **DePristo, M. A., D. M. Weinreich, and D. L. Hartl.** 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet **6:**678-87.

18. **Eppler, T., P. Postma, A. Schutz, U. Volker, and W. Boos.** 2002. Glycerol-3-phosphate-induced catabolite repression in Escherichia coli. J Bacteriol **184:**3044-52.

19. **Ferguson, G. P., S. Totemeyer, M. J. MacLean, and I. R. Booth.** 1998. Methylglyoxal production in bacteria: suicide or survival? Arch Microbiol **170:**209-18.

20. **Fong, S. S., A. R. Joyce, and B. O. Palsson.** 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res **15:**1365-72.

21. **Freedberg, W. B., W. S. Kistler, and E. C. Lin.** 1971. Lethal synthesis of methylglyoxal by Escherichia coli during unregulated glycerol metabolism. J Bacteriol **108:**137-44.

22. **Gosset, G., Z. Zhang, S. Nayyar, W. A. Cuevas, and M. H. Saier, Jr.** 2004. Transcriptome analysis of Crp-dependent catabolite control of gene expression in Escherichia coli. J Bacteriol **186:**3516-24.

23. **Hall, B. G.** 2002. Predicting evolution by in vitro evolution requires determining evolutionary pathways. Antimicrob Agents Chemother **46:**3035-8.

24. **Herring, C. D., J. D. Glasner, and F. R. Blattner.** 2003. Gene replacement without selection: regulated suppression of amber mutations in Escherichia coli. Gene **311:**153-63.

25. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-12.

26. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

27. **Hogema, B. M., J. C. Arents, R. Bader, and P. W. Postma.** 1999. Autoregulation of lactose uptake through the LacY permease by enzyme IIAGlc of the PTS in Escherichia coli K-12. Mol Microbiol **31:**1825-33.

28. **Holtman, C. K., A. C. Pawlyk, N. D. Meadow, and D. W. Pettigrew.** 2001. Reverse genetics of Escherichia coli glycerol kinase allosteric regulation and glucose control of glycerol utilization in vivo. J Bacteriol **183:**3336-44.

29. **Hurley, J. H., H. R. Faber, D. Worthylake, N. D. Meadow, S. Roseman, D. W. Pettigrew, and S. J. Remington.** 1993. Structure of the regulatory complex of Escherichia coli IIIGlc with glycerol kinase. Science **259:**673-7.

30. **Ibarra, R. U., J. S. Edwards, and B. O. Palsson.** 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420:**186-9.

31. **Joyce, A. R.** 2007. Modeling and Analysis of the E. coli Transcriptional Regulatory Network: An Assessment of its Properties, Plasticity, and Role in Adaptive Evolution. University of California, San Diego, La Jolla.

32. **Larson, T. J., S. Z. Ye, D. L. Weissenborn, H. J. Hoffmann, and H. Schweizer.** 1987. Purification and characterization of the repressor for the sn-glycerol 3-phosphate regulon of Escherichia coli K12. J Biol Chem **262:**15869-74.

33. **Lewis, N. E., K. Kixon, T. M. Conrad, J. Lerman, P. Charusanti, A. Polpitiya, J. Adkins, G. Schramm, S. Purvine, D. Lopez-Ferrer, K. Weitz, R. Eils, R. Konig, R. Smith, and B. O. Palsson.** 2010. Omic Data from Evolved E. coli are Consistent with Computed Optimal Growth from Genome-scale Models. Mol Sys Bio**:**(Accepted 6/1/10).

34. **Lin, E. C.** 1976. Glycerol dissimilation and its regulation in bacteria. Annu Rev Microbiol **30:**535-78.

35. **Liu, W. Z., R. Faber, M. Feese, S. J. Remington, and D. W. Pettigrew.** 1994. Escherichia coli glycerol kinase: role of a tetramer interface in regulation by fructose 1,6-bisphosphate and phosphotransferase system regulatory protein IIIglc. Biochemistry **33:**10120-6.

36. **Miller, S. P., M. Lunzer, and A. M. Dean.** 2006. Direct demonstration of an adaptive constraint. Science **314:**458-61.

37. **Novotny, M. J., W. L. Frederickson, E. B. Waygood, and M. H. Saier, Jr.** 1985. Allosteric regulation of glycerol kinase by enzyme IIIglc of the phosphotransferase system in Escherichia coli and Salmonella typhimurium. J Bacteriol **162:**810-6.

38. **Pettigrew, D. W., W. Z. Liu, C. Holmes, N. D. Meadow, and S. Roseman.** 1996. A single amino acid change in Escherichia coli glycerol kinase abolishes glucose control of glycerol utilization in vivo. J Bacteriol **178:**2846-52.

39. **Pettigrew, D. W., G. J. Yu, and Y. Liu.** 1990. Nucleotide regulation of Escherichia coli glycerol kinase: initial-velocity and substrate binding studies. Biochemistry **29:**8620-7.

40. **Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans.** 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature **445:**383-6.

41. **Postma, P. W., W. Epstein, A. R. Schuitema, and S. O. Nelson.** 1984. Interaction between IIIGlc of the phosphoenolpyruvate:sugar phosphotransferase system and glycerol kinase of Salmonella typhimurium. J Bacteriol **158:**351-3.

42. **Rohwer, J. M., R. Bader, H. V. Westerhoff, and P. W. Postma.** 1998. Limits to inducer exclusion: inhibition of the bacterial phosphotransferase system by glycerol kinase. Mol Microbiol **29:**641-52.

43. **Weinreich, D. M., N. F. Delaney, M. A. Depristo, and D. L. Hartl.** 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science **312:**111-4.

44. **Weissenborn, D. L., N. Wittekindt, and T. J. Larson.** 1992. Structure and regulation of the glpFK operon encoding glycerol diffusion facilitator and glycerol kinase of Escherichia coli K-12. J Biol Chem **267:**6122-31.

45. **Yeh, J. I., V. Charrier, J. Paulo, L. Hou, E. Darbon, A. Claiborne, W. G. Hol, and J. Deutscher.** 2004. Structures of enterococcal glycerol kinase in the absence and presence of glycerol: correlation of conformation to substrate binding and a mechanism of activation by phosphorylation. Biochemistry **43:**362-73.

46. **Yu, P., and D. W. Pettigrew.** 2003. Linkage between fructose 1,6-bisphosphate binding and the dimer-tetramer equilibrium of Escherichia coli glycerol kinase: critical behavior arising from change of ligand stoichiometry. Biochemistry **42:**4243-52.

47. **Zwaig, N., W. S. Kistler, and E. C. Lin.** 1970. Glycerol kinase, the pacemaker for the dissimilation of glycerol in Escherichia coli. J Bacteriol **102:**753-9.

48. **Zwaig, N., and E. C. Lin.** 1966. Feedback inhibition of glycerol kinase, a catabolic enzyme in Escherichia coli. Science **153:**755-7.

Chapter 5

Strategy to develop a cellulose-consuming strain of *E. coli*

5.1 Introduction

As described in Chapter 2, laboratory evolution has been used to improve growth of *E. coli* on poorly-utilized substrates (16, 17, 31) and following the deletion of a variety of metabolic genes (15, 18, 19). This chapter discusses a project designed to apply laboratory adaptive evolution to optimize usage of gnomically-integrated heterologous genes to allow a novel carbon source to be utilized – cellulose.

The ability to use cellulose and convert it into energy or valuable chemical products has become very important because of the incredible need to find replacements for fossil fuels, since cellulose is one of the most abundant forms of carbon on Earth (40). The challenge lays in harnessing methods to breakdown the cellulose, pectin, and lignin polymers to use as raw materials. The most obvious strategy is to utilize the biological capacity of microbes responsible for degrading these products in the biosphere.

Full-scale industrial applications are unlikely to use *E. coli* or the genes from *Erwinia chrysanthemi* described here, since there have already been significant efforts to isolate strains and enzymes that are much better suited to the extreme conditions involved in large-scale cellulose processing. However, this study would have gauged

the potential of adaptive evolution to improve cellulose-dependent growth and expression of heterologous genes, which may be of significant value in many metabolic engineering applications.

Additionally, the results of this study were expected to have significant basic research value as well. The heterologous genes would have functioned like those recently acquired by horizontal gene transfer. Because these newly-acquired genes would be essential for growth, there would be strong selection pressure for adaptations that optimize their expression and integrate them into the existing metabolic and regulatory network. Thus this study would have provided an opportunity to directly observe how regulatory networks develop and adapt.

Further, as described in the next section, cellulose polymers are too large to be engulfed by bacteria, so many cellulases are secreted into the medium. This makes growth dependent on protein secretion, which is a novel dynamic not previously explored by adaptive evolution studies. Additionally, adaptations that change the regulation of biomass partitioning in the cell to increase protein secretion or biosynthesis could be of significant value to the many research and industrial applications that use *E. coli* for protein expression.

5.1.1 Project design background

Cellulose is a mostly linear polymer of β-linked glucose molecules, and these polymers usually self-assemble into tight crystalline structures (40). These crystalline structures make cellulose difficult to break down, both chemically and with enzymes. Natural sources of cellulose often also contain various quantities of hemicellulose, lignin, and pectins, which require separate enzymes to be broken-down but are also potential carbon sources.

Three types of enzymes are typically required for cellulose degradation; (a) endoglucanase (EC 3.2.1.4), which cleaves at random inside the cellulose polysaccharide chain, (b) exoglucanase (EC 3.2.1.91), which progressively cleave mono- or disaccharide (cellobiose) fragments off of the ends of the polysaccharide chains, and (c) β-glucosidase (E.C. 3.2.1.21), which hydrolyze the β-bond between the disaccharide units to produce glucose and glucose-6-phosphate (40).

Since adaptive traits must be functionally expressed to improve growth and be selected, the project needed cellulase genes that can be functionally expressed in *E. coli* when first introduced. For this reason, cellulase enzymes from other gram-negative bacteria were preferred. Most gram-negative cellulose-degrading species secrete cellulase enzymes rather than constructing cellulose-degrading complexes on their cell surface(40). The secreted cellulases degrade the cellulose polymers to cellobiose (β-linked glucose dimer) which can be transported into the cell and cleaved to glucose by a β-glucosidase. Thus this system also requires an enzyme secretion system and a cellobiose-uptake transporter.

Both the cellulases and type II secretion system for transporting the cellulases from gram-negative *Erwinia chrystanthemi* have been functionally expressed in *E. coli* (25, 39, 67, 68), and it was this set of genes that were chosen for this study. *Er. chrysanthemi* is a plant pathogen that causes soft-rot disease, and the CelY and CelZ cellulases are two of many enzymes secreted to break down plant tissues to this end (4, 23, 25). CelZ and CelY work synergistically to digest long cellulose molecules to into cellobiose (67). CelZ is the cellulose exonuclease (E.C. 3.2.1.4) that cleaves progressively from the ends of cellulose chains, and is responsible for the majority of the cellulase activity in *Er. chrysanthemi* (66-68). The CelY enzyme is the endoglucanase (EC 3.2.1.4) that cleaves long cellulose molecules at random into

shorter chains, and increases CelZ activity by providing additional ends to act as substrate(23, 66, 67).

CelZ and CelY are transported across the outer membrane by the type II secretion system (T2SS) from *Er. chrysanthemi* encoded by the *outCDEFGHIJKLMO* (*out*) operon (39, 68). These enzymes are transported across the inner membrane into the periplasm by the Sec pathway (6). *E. coli* also encodes a type II secretion system, encoded by the *gsp* operon (21), however, it is unlikely that it could perform the same function as the *out* operon because type II secretion systems are very selective about their targets, and the secretion system from one species will not secrete the proteins recognized by the homologous system another - even within closely-related γ-proteobacteria (22). The mechanism by which type II secretion systems recognize their targets is still being studied, as are the specific functions of many of the 12 to 14 proteins necessary to form the secretion complex (53).

As previously stated, the CelY and CelZ cellulases from *Er. chrysanthemi* degrade cellulose polymers into cellobiose. However, most strains of *E. coli* do not grow on cellobiose. Nevertheless, cellobiose-competent *E. coli* strains have been isolated following prolonged growth on agar plates in which cellobiose is the only provided carbon source (33, 48, 49). Following this discovery, genes that allow *E. coli* to utilize cellobiose were found that encode a cellobiose transporter (CelBC) and β-glucosidase (CelF) that can hydrolyze cellobiose into glucose and glucose-6-phosphate. This '*cel* operon' was classified as 'cryptic' since no growth conditions were identified that could induce expression in wild type that did not involve mutations.

The existence of these genes in *E. coli* was not entirely surprising since many non-cellulytic organisms that inhabit the rumen harbor them to utilize the cellobiose

produced by endo- and exoglucanases secreted by cellulolytic organisms (40). However, since then other work has shown that this operon is induced by and functions to metabolize chitobiose, a dimer of β-linked N-acetyl glucosamine sugars (34), and the operon was renamed the *chb* operon after that function.

5.1.2 Project Design

This project can be divided into the following stages.

Stage 1: Develop Cellulose-competent *E. coli* strain

A. Adaptive evolution to generate *E. coli* with optimized growth on cellobiose
B. Insert 12.5 kb *out* operon from *Er. chystanthemi* into genome in place of *gsp* operon
C. Insert cellulase genes from *Er. chrysanthemi* into genome in place of *chiA* gene

Stage 3: Adaptive evolution to develop cellulose-dependent growth

Stage 4: Sequence genome of developed strains

Stage 5 & 6 (post-doctoral work): characterize mutations and evolved strains, determine mechanisms of adaptation

Significant progress was made towards the first stage, the development of the cellulose-competent strain, as described in the following sections. However, unexpected difficulties were also encountered that slowed progress, and this project was terminated before the platform strain could be generated for adaptive evolution on cellulose (stage 2) due to a lack of funding.

5.2 Results and Discussion

5.2.1 Adaptive Evolution on Cellobiose

5.2.1.1 Activation of the *chb* operon

The first stage of the cellulose project was to develop a strain of *E. coli* that can grow on cellobiose. As previously described, wild type *E. coli* does not grow on cellobiose, but it does carry genes that can utilize cellobiose encoded on the *chb* operon (34, 48, 49) (**Figure 5.1**). Cellobiose-competent *E. coli* strains have been isolated after prolonged incubation on cellobiose-supplemented agar plates.

Mutations that allow cellobiose utilization in wild type *E. coli* all involve the *chbBCARFG* operon. In the first study of these studies, strains were found with one of two type of mutations – either the insertion of an IS element into the promoter region or a non-synonymous point mutation to the substrate-specific repressor protein *chbR* that allow it to recognize cellobiose and several other bi-sacharides as inducers (49). However, Plumbridge et al (2004) later found that the *chb* operon is also strongly

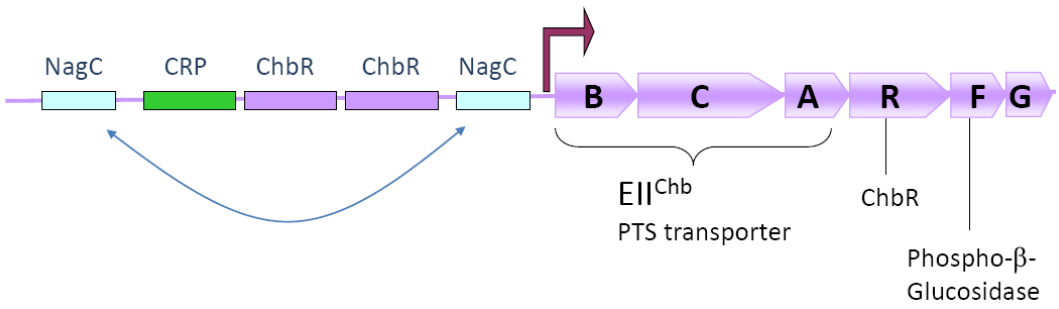


Figure 5.1 Structure of the *chb* operon. Figure identifies the function of the encoded genes and the known transcription factors controlling expression,. In the presence of cellobiose, the *chb* operon is normally repressed by ChbR, the specific repressor of the operon that is alleviated by chitobiose binding, and NagC, which is alleviated by the product of chitobiose hydrolysis, N-acetyl glucosamine (NAG). Repression by NagC is especially strong when both sites are bound because the proteins then dimerize (blue arrow), forming the promoter region into a DNA loop that prevents RNA polymerase binding(50). Expression in response to cellobiose has previously been found to require either an IS element insertion to the promoter to disrupt repressor binding, or mutations to both NagC and ChbR that destroy function and change inducer recognition to cellobiose from chitobiose, respectively(33)

repressed by NagC, which regulates metabolism of amino-sugars like N-acetyl glucosamine produced by chitobiose hydrolysis, and by CRP which regulates preferred substrate utilization (50) (**Figure 5.1**). Following this, later study of cellobiose adaptation found that ChbR mutations alone cannot impart cellobiose competence, but that mutations that deactivate NagC repression are also required(33).

5.2.1.2 Solid vs. liquid media

Both of the previous studies generated cellobiose-utilizing *E. coli* strains by prolonged incubation on plates containing cellobiose (24, 33, 49). This method was reported to produce cellobiose-competent papillae in less than 30 days. However, in this study I chose to develop cellobiose-competent strains by growth in liquid media instead of on plates.

All previous adaptive evolution studies in the Palsson lab have been performed in liquid media, and achieved considerable success improving the growth phenotype (**Chapter 2**). By using this culturing method, I sought to avoid introducing novel variables that might further complicate comparisons to our other adaptive evolution studies. Additionally, while adaptive evolution with plates has been shown to be effective at generating cellobiose-utilizing mutants, it is not likely to have further selected for the combination of mutations that best optimize growth on cellobiose. Adaptive evolution in liquid media is superior to plates for selecting for optimal growth rates for several reasons. First, liquid media can support much larger populations, which increases the rate at which adaptive mutations are found. Further, in liquid culture the resources available to each cell are homogenous, which facilitates competition for resources and take-over by advantageous genotypes.

5.2.1.3 Selection for the cellobiose-utilization phenotype

Six sets of lineages, CB1-6, were seeded from wild type and grown under conditions designed to select for cellobiose growth. The strains were grown in M9-minimal medium containing 2 g/L or more of cellobiose, and decreasing amounts of a secondary substrate, started at 1 g/L. The secondary substrate was added to allow the cell divisions necessary to sample mutations and acquire those that allow growth on cellobiose.

Cellobiose was included in the media throughout the adaptation period to provide cellobiose-utilizing mutants an advantage and thus create selection. However, strains that acquire the ability to use cellobiose may still preferentially grow on the secondary substrate when it is available. Also, if growth on cellobiose is slower than on a secondary substrate there will not be selection for cellobiose-utilization until all of the secondary substrate has been utilized and the non-cellobiose utilizing strains are locked in stationary phase. This suggests that both exponential and stationary phase growth are important for acquiring cellobiose-utilizing mutants.

The first set of lineages, CB1-3, were initially grown using glycerol, lactate, or acetate as the secondary substrate. These substrates do not induce catabolite repression in wild type *E. coli* and so are less likely to be preferentially utilized over available cellobiose. However, after day 10 only N-acetyl glucosamine (NAG) was used as the secondary substrate. This decision was made after considering the results of Kachroo et al (2007) who found that dual mutations to ChbR and NagC are necessary to achieve cellobiose competence. It is considered exponentially more difficult to acquire a trait that requires two mutations to produce the improved phenotype (46). To overcome this possible obstacle, growth on N-acetyl glucosamine (NAG) was used to try to alleviate NagC repression of the chb operon, which would

allow mutations to *chbR* to be selected first. NAG was the only secondary substrate used during the evolution of lineages CB4-6.

5.2.1.4 Culture and population management during adaptation

During previous adaptation studies in this research group, the volume transferred each day to fresh medium was calculated to allow exponential growth until the next passage, which created consistent selection pressure for increased growth rate. However, the volumes passed were often small, on average around 0.01% of the population, which created a considerable population bottleneck each day. Several computational modeling studies have suggested that passing 10-20% of the population maximizes the rate of adaptation by optimally balancing the need to preserve acquired population diversity and to allow sufficient population growth to acquire new beneficial mutations (30, 62). For this reason, larger passage volumes (5-50 ml, or 2-20% of culture volume) were often utilized during this study.

However, the size of population bottlenecks was not the only variable considered when determining the passage volume. The passage volume and substrate concentration also determines the time the cultures spend in exponential and stationary phase. During the development of the cellobiose-utilization phenotype it was considered necessary for the NAG-utilizing cultures to spend some time in stationary phase to provide a period during which cellobiose-utilizing strains can propagate when all others must stop, increasing their probability of being passed and further adapting.

However, this tactic is only useful once the desired phenotype has appeared. If it has not, nutrient limitation in stationary phase changes the growth objective to select for optimal biomass instead of optimal growth rate(55), and the stress can generate strains with increased mutation rates (57) that may over accumulate neutral

mutations (1, 44). Since the primary goal of this project was to develop the phenotype, not to specifically study the dynamics if this variable, this balance was not systematically studied but rather tested and varied somewhat haphazardly during the course of the adaptations until the desired phenotype was found.

During the development of CB1-3, the cultures were consistently passed once a day, with the volume transferred ranging from 200 ul to 20 ml (0.8-8% of culture volume), with the median volume passed being 5 ml (2% of total culture volume) (This adaptation experiment was performed before publications on optimal passage ratios were found (30, 62)). After 24 days of adaptation, with NAG reduced to 0.025 g/L, no cellobiose-independent growth could be detected after 24 hours, and the adaptation of these lineages was at that time stopped.

As a result, the CB4-6 cultures were developed by imposing conditions

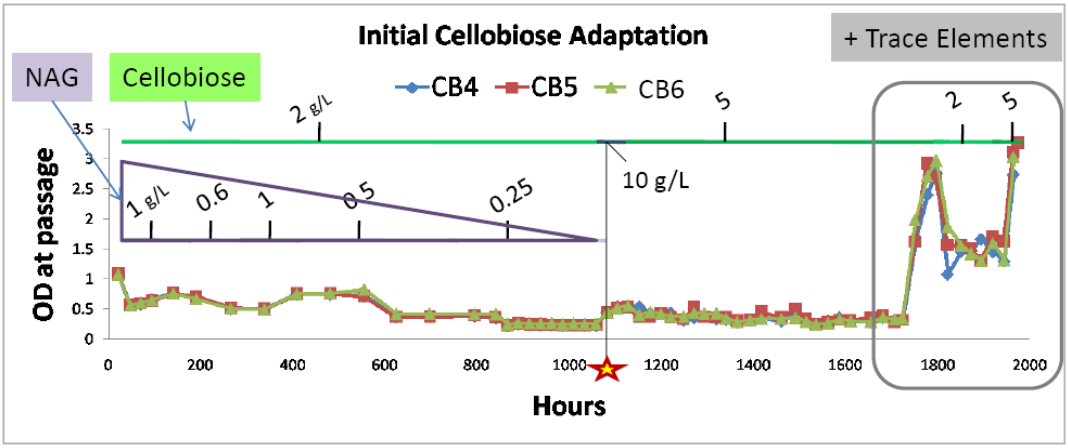


Figure 5.2 Adaptive Evolution of strains CB4-6. The above figure is a plot of the optical density (600 nm) in the spent media measured at the time of transfer to fresh media over the course of the adaptive evolution experiment, measured in hours. The concentration of NAG and cellobiose in the M9-minimal media is shown above the plot in purple and green, respectively. During the first ~1100 hours / 46 days, optical density in spent media appeared largely dependent on the concentration of NAG, which was slowly decreased. Cellobiose-dependent growth was finally detected when the concentration of cellobiose was increased to 10 g/L (timepoint marked by red star), at which point the cultures were grown with cellobiose as the only carbon source. Culture density was further increased when Sauer's trace elements were included in the media (grey box).

designed to create much stronger selection for cellobiose growth (**Figure 5.2**). The cultures were passed daily for the first three days to allow population diversity to accumulate, but were then only passed once every 72 hours for the next eight passages, in which 5 ml of culture was transferred each day. This left the cultures in stationary phase for at least 2 days, during which time cellobiose-utilizing strains would have a significant growth advantage. When that failed to produce cellobiose-independent growth after the eighth passage (detected by development of culture density in M9-minimal media supplemented with only cellobiose inoculated with culture from the previous day), the incubation period was shorted to 24 hours to allow more frequent periods of cell division. However, the concentration of NAG was reduced to 0.5 g/L for four days, and then to 0.25 g/L, and the transferred volume increased to 10 ml. After an additional nine days there was still no detected cellobiose-independent growth in any of the cultures, and no significant increase in the optical density at the time of culture transfer.

To further increase the selection pressure for cellobiose growth, the concentration of cellobiose was increased from 2 g/L to 10 g/L (**Figure 5.2**). The next day the optical density of the spent culture was increased to 0.4 compared to 0.23 on previous days despite no change in the incubation period or provided NAG, suggesting that cellobiose had been converted into biomass. This was confirmed when passage into M9-minimal media containing only cellobiose (5 g/L) produced optical density overnight, with an estimated growth rate of 0.08 hr^{-1}. Further laboratory evolution to optimize growth on cellobiose is described in the next section.

Additionally, cellobiose utilization was also eventually discovered in the CB1-3 lineages. Since these strains had never been grown for longer than a day before transfer, there was a chance that a cellobiose-utilizing strain had developed but was

not selected strongly enough to be detected based on optical density. Frozen stocks of the endpoint cultures from the initial adaptation were grown up in 4ml M9-cellobiose without secondary substrate and incubated for 48 hours, at which point detectable cell density was detected. These cultures were subsequently passed to larger flasks for growth rate measurements. Strangely, at that point the CB-2 strain failed to grow. However, the CB-1 and -3 were measured to grow at 0.06 and 0.055 doublings per hour, respectively.

5.2.1.5 Optimizing cellobiose utilization

After acquiring strains with the ability to utilize cellobiose, the adaptive evolution experiments were continued to select for optimized growth. Given that cellobiose is composed of two linked glucose molecules, optimal growth rates and yields should potentially be similar to growth on glucose (0.6 hr^{-1}, max OD_{600} >3).

After discovering cellobiose-dependent growth in CB1 and CB3 endpoints, the faster of the two strains, CB1, was used to inoculate two additional lineages named CB1a and CB1b. These lineages were grown in 2 g/L cellobiose M9 medium for an additional 23 days. The goal of this adaptation was to increase growth rate, so the cultures were passed to maximize their time in the exponential growth phase. This was accomplished by calculating the amount of volume to transfer based on the desired final OD the next day, the previous day's growth rate, the optical density of cells in the source culture, and the estimated incubation time. The target final OD was set to 0.2-0.3, which is about half the maximum OD_{600} reached by CB1 after 48 hours, and was also the OD target used in the glycerol evolutions. At the observed growth rates, passage volumes were often large – 15-50 ml, or 5-20% of the culture volume, suggesting that little acquired population diversity was lost by passing. After 23 days of

sustained growth, both CB1a and –b doubled their growth rate from ~0.06 to ~0.13 doublings per hour.

The CB4-6 lineages were also further adapted in cellobiose-M9 after acquiring the ability to utilize cellobiose. These lineages were grown with 5 g/L cellobiose and were transferred using volumes calculated to maintain exponential growth (generally 5-20% of total volume passed). At the end of 24 days of growth, the growth rates of the lineages had also increased to approximately 0.2 hr^{-1}.

However, during this adaptation it also became clear that the strains were not able to utilize the majority of the cellobiose, since the cultures were entering stationary phase with an optical density of ~0.5 despite being grown with 5 g/L. This suggested that something else was limiting growth. Trace elements as described by Sauer (54) were added to the cultures, which immediately increased the growth rate to around 0.3 hr^{-1} and the maximum optical density to 2.7-3.2. This suggested that cellobiose utilization may be limited by an enzyme that requires an ion cofactor included in the supplement, which contains Fe^{3+}, Mn^{2+}, Zn^{2+}, Cu^{2+}, Co^{2+}, and MoO_4^{2-}. At this point the concentration of cellobiose was reduced back down to 2 g/L, which reduced the observed growth rate of the strains to ~0.2 hr^{-1} and the maximum optical density (600 nm) to ~1.6.

This growth rate still seemed somewhat slow given that cellobiose hydrolyzes into glucose, and growth rates of *E. coli* on 2 g/L glucose are reported to be closer to 0.6 hr^{-1}. There was also concern that this growth rate might be too slow to allow adaptation to occur in a reasonable amount of time when the heterologous cellulase genes were introduced.

5.2.1.6 Treatment with mutagen to speed rate of adaptation

There is research that suggests the mutation rate of *E. coli* may limit its speed of adaptation (1, 7). To try to further increase the growth rate of these strains, on day 70 the CB4-6 evolutions the lineages were split, and one set of cultures were treated with the mutagen N- methyl-N -nitro-N-nitrosoguanidine (NTG) (28, 32, 65). This type of treatment only increases the mutation rate of the treated generation, suggesting that it would increase the rate at which advantageous mutations were found without acquiring as many hitchhiker neutral mutations as 'mutator' strains with consistently increased mutation rates. The mutagenized cultures are designated with an "m" (ie, CB4m).

Evolution of the cellobiose-adapted strains on M9-cellobiose was resumed for

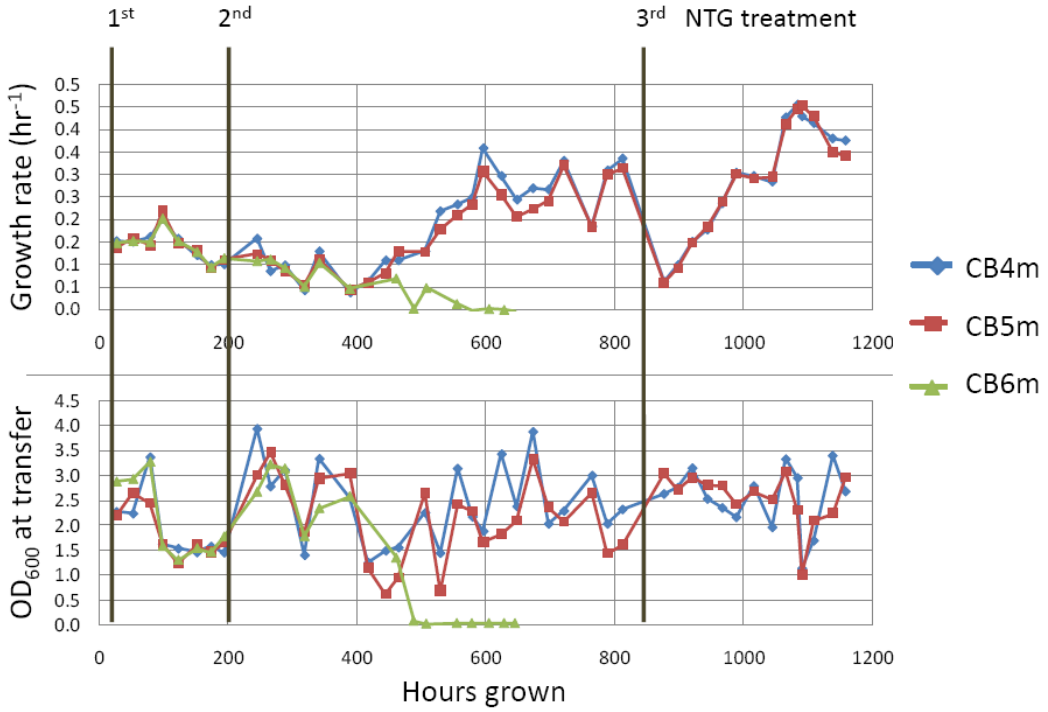


Figure 5.3 Optimization of Cellobiose-utilization in strains CB4-6m using NTG mutagen. Top panel shows estimated growth rate of culture during time between transfers, and bottom panel shows optical density of culture at time of transfer. The three timepoints at which the cultures were treated with N- methyl-N'-nitro-N-nitrosoguanidine (NTG) are highlighted by black lines across both plots. NTG treatments generally increased the observe growth rate after an initial recovery period, but had little effect on maximum culture density reached. Strain CB6m crashed following second mutagen treatment for unknown reasons.

another 50 days (1.56×10^{11} divisions), punctuated by two additional treatments with the mutagen on day 10 and 35 (**Figure 5.3**). Interestingly, lineage CB6m survived the first NTG treatment, but the population subsequently crashed after the second (**Figure 5.3**). The CB4m and -5m lineages also showed a decline in growth rate and accumulated cell-density following the second treatment due to the effect of strains with deleterious mutations, but their populations subsequently recovered and appeared to be taken over by a phenotype with improved growth. For reasons that are not entirely clear, the CB6m population continued to deteriorate until no growth could be detected. Several attempts were made to revive the lineage from frozen stocks of culture collected before the crash but after the second mutagen treatment also failed.

It is also interesting to note that during this same period of time, the colors of all of the cultures at the time of transfer were observed to shift to a darker brown or grey color (**Figure 5.4**). The culture color was not consistent day-to-day or the same between cultures. The first response was to verify that the culture had not been contaminated, which was initially tested by colony color on McConkey agar plates and later 16S RNA and *rpoB* sequencing, which indicated the culture was predominately *E. coli* (see **Section 5.2.1.9**). To test whether the color change was a property of the whole population or a subpopulation, the cultures were streaked onto M9-glucose plates and colonies were selected and propagated in the same liquid media. Not all of the colony-inoculations grew (only 3 of 10 did), but the ones that did also had various grey hues when they reached high culture density. The color of CB4m and -5m populations grown in M9 glucose was unchanged from wild type, but returned to a grey-purple color when subsequently grown in M9-cellobiose. The lack of a color shift in M9-glucose may suggest it may involve oxidation of trace elements

in a manner specific to growth in cellobiose medium. This effect was also transient, as it disappeared after about a week (by day 30).

Following a third NTG treatment on day 35, evolution of lineages CB4m and -5m was terminated on day 50. The growth rates at day 49 of CB4m and -5m were 0.47 and 0.45 hr^{-1}, respectively, which is indistinguishable from their growth rates on day 33 before the second mutagenesis (0.47 and 0.46 hr^{-1}, respectively). This suggests that the majority of adaptive mutations accessible to NTG mutagenesis had been acquired during the first two rounds, and that further adaptation and mutagenesis was unlikely to produce significant gains in the near future.

5.2.1.7 Characterization of endpoint strains

At the end of adaptation the cultures were plated out on cellobiose-M9 agar plates with trace elements, and 5 well-spaced colonies were selected. It was not known whether the each of the populations had fixed a single genotype and phenotype or if some diversity remained. It was also possible that several ecotypes had emerged; an adaptive evolution study of *E. coli* on glucose found only some members of the population specialized to growth on glucose while others become specialized to use fermentation products such as acetate (60).

Several colonies were picked to identify and isolate genotypes with the fastest growth rate on cellobiose for further genetic manipulation. **Table 5.1** shows the measured growth rates of five cultures grown in parallel, including the endpoint population and three clonal isolates. Growth rate differences between these strains are small (0.45-0.48 hr-1), which may suggest that the populations are largely fixed, though could also result if the isolation method (growth on M9-cellobiose supplemented agar plates) predominately favors the fast-cellobiose growing ecotype. Not enough measurements were made to determine whether the slightly

higher growth rates of the endpoint populations versus the clones are incidental or if they signify an advantage of the mixed population due to cross-feeding or the presence of a variant with even higher fitness.

Table 5.1 Growth rates of final mutagenized cellobiose adapted strains.

		Growth Rate	R^2 of fit
CB4m	endpoint pop.	0.475	0.996
	clone A	0.458	0.997
	clone B	0.459	0.998
CB5m	endpoint pop.	0.468	0.996
	clone A	0.450	0.996

Growth rates and R-square values calculated from six optical density readings of each culture, taken between 0.3-3.0 OD_{600}.
Strains grown on 5 g/L cellobiose M9-minimal medium with 1x Sauer's trace elements.

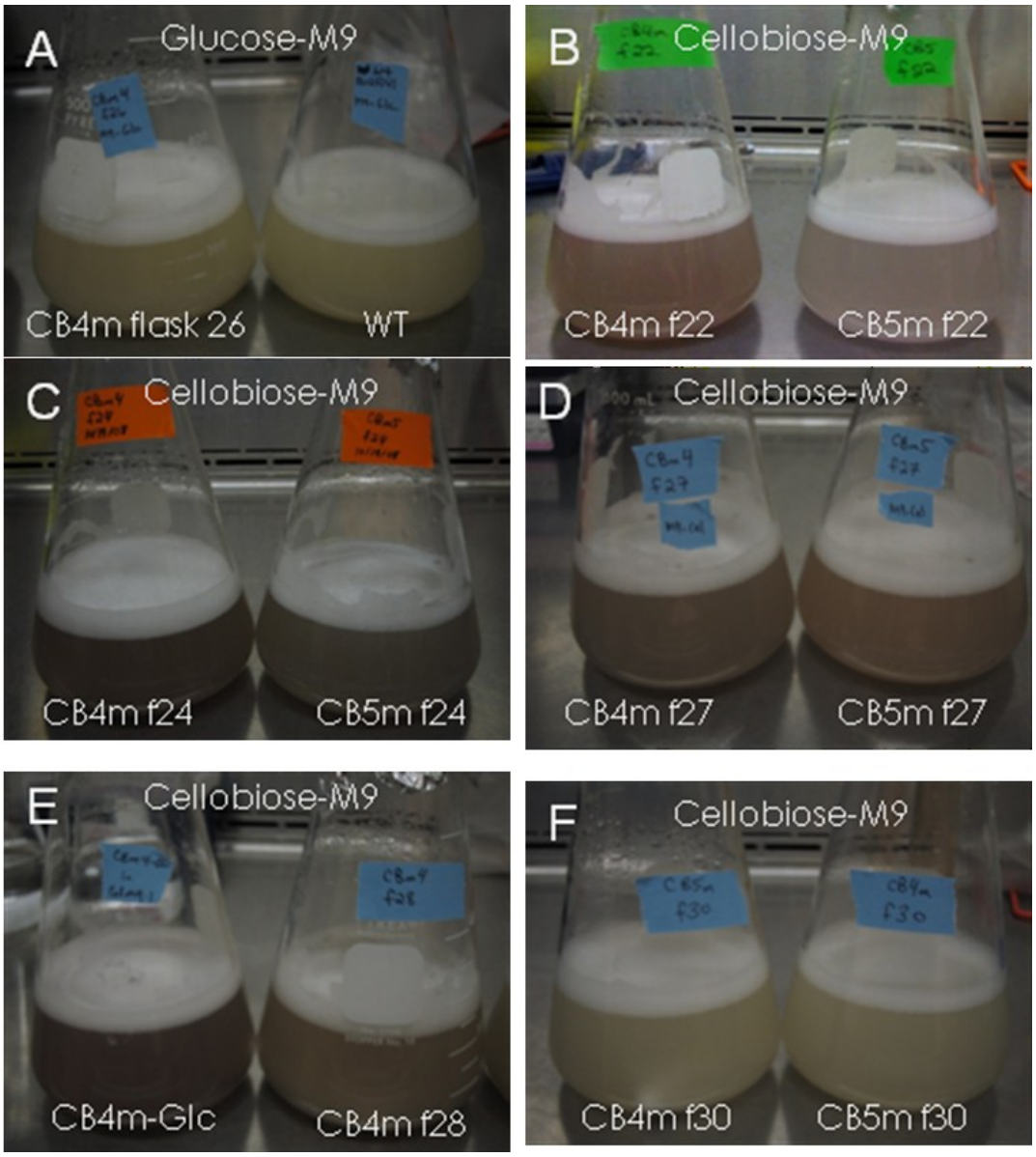


Figure 5.4 Color shift in CB4m and -5m cultures. (A) Culture CB4m from flask 26 ("f26") and wild type *E. coli* after overnight growth on M9-glucose (2 g/L + trace elements); shows expected color of culture. Culture in CB4m-f26 do not cause color shift in M9-glucose, but do in M9-cellobiose. (B) CB4m and CB5m culture in flask 22 of NTG-treated adaptation. Both cultures have acquired a rosy hue. (C) Same cultures after two subsequent transfers (flask 24); both cultures now dark grey/brown. (D) Cultures in flask 27 also brownish grey. (E) Flask on left was grown by inoculated M9-cellobiose with CB4m-f26 that had been grown in M9-glucose (panel A). Shows that color phenotype returns when culture is again grown on cellobiose. Flask on right shows CB4m-f28 that was never grown on glucose, which is still a grayish brown. (F) CB4m and -5m after transfer to flask 30, at which point the color-shifting phenotype has abated.

5.2.1.8 Stability of the cellobiose-utilizing phenotype

The ability of the adapted strains to utilize cellobiose has also proven to be rather unstable. Examples of this include the two adapted strains, CB2 and CB6m, shown in **Table 5.2**, which both failed to grow on cellobiose for undetermined reasons after initially showing growth. The frozen stock of culture CB2 initially grew in M9-cellobiose, but failed to survive following transfer. Additionally, the CB6m culture crashed after the second mutagen treatment and failed to be resurrected from the most recent frozen stock collected. Furthermore, there were many additional instances in which colonies on M9-cellobiose agar plates failed to propagate when inoculated into M9-cellobiose liquid media, or stopped being able to grow on cellobiose after being grown on another substrate such as LB or M9-glucose. One

Table 5.2 Summary of cellobiose-adapted cultures.

Cellobiose-adapted lineage	Growth phases between transfers	Days evolved	Population doublings	Approx. no. of cell divisions ($x10^9$)	Growth rate on M9-cellobiose (hr^{-1})
CB1	log	23	141	10.5	0.06
CB2	log	23	130	7.3	n/a
CB3	log	15	91	6.6	0.055
CB1a	log	23 (46)†	74 (215)	8.06 (18.6)	0.13
CB1b	log	23 (46)	79 (220)	9.07 (19.6)	0.13
CB4	log & stat	82	295	63	0.2*
CB5	log & stat	82	304	67	0.18*
CB6	log & stat	82	309	67	0.19*
CB4m	log & stat	49 (116)	322 (554)	160 (191)	0.47*
CB5m	log & stat	49 (116)	317 (560)	155 (179)	0.45*
CB6m	log & stat	49 (116)	62 (307)	53 (83)	n/a

† Values in parentheses are the total since adaptation of strain began, including previous legs of adaptation

* growth rate when provided 5 g/L cellobiose (all others measured with 2 g/L cellobiose)

dramatic example is shown in **Figure 5.5,** in which colonies were propagated in 4 ml of M9-cellobiose overnight and then used to inoculate a 96-well plate for overnight growth in an incubated plate reader. Although none of the clones had trouble growing on glucose, LB, or NAG (not shown), only two of the cultures sustained grown on cellobiose media. It is possible that growth on cellobiose is particularly sensitive to growth conditions in the plate reader, including reduced aeration compared to other growth methods or possibly uneven temperature incubation. However, this was just one instance of many encountered where ability to utilize cellobiose rapidly disappeared.

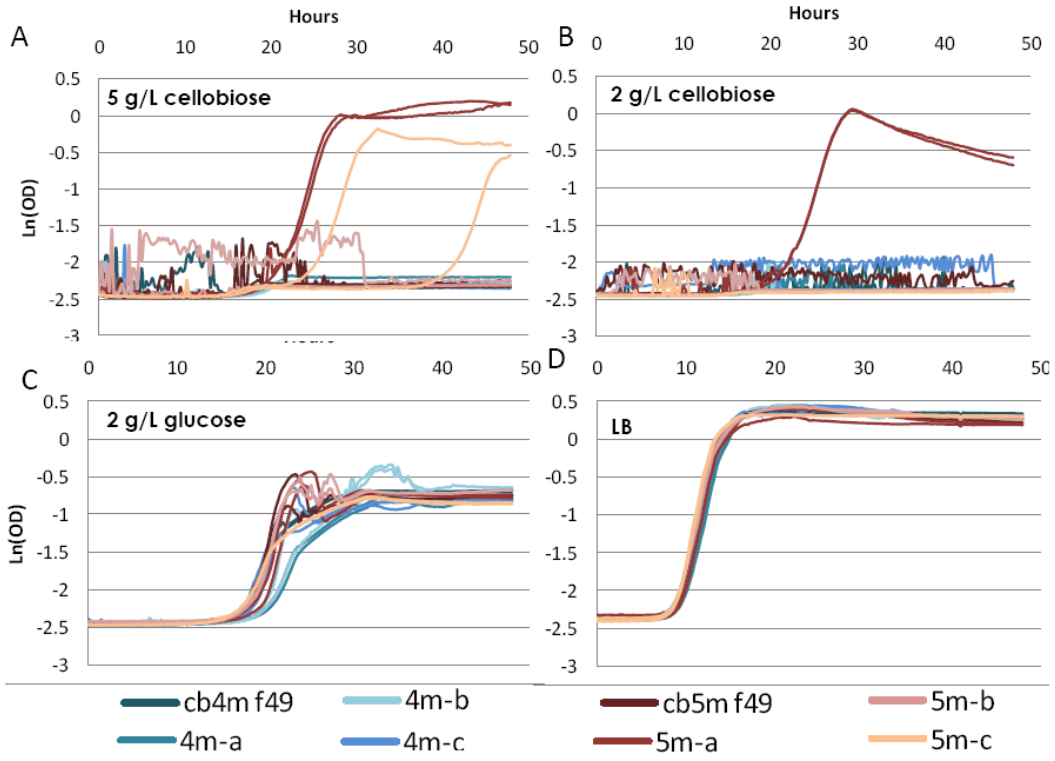


Figure 5.5 High-throughput growth measurements of cellobiose strains shows instability of phenotype. Eight pre-cultures were used to inoculate wells on the same 96-well plate containing various media. (A) Despite the fact that all cultures had been precultured on 5 g/L cellobiose M9 + trace elements, only two strains (CB5m clone A and C) continued to grow on that media in the 96-well plate. (B) On media with reduced cellobiose concentration (2 g/L), only one strain – CB5m clone A – developed optical density. However, none of these strains showed a growth defect on M9-glucose, LB, or N-acetyl glucosamine (not shown). Noise in (A)-(C) likely due to condensation on lid of plate.

5.2.1.9 Discovery of *Paenibacillus* contamination

The observed color shifts of the cultures raised concerns that the evolving cultures had been contaminated by another organism during the long term adaptation. It was assumed that a contaminating organism would take over the culture – especially one that changed the appearance of the culture.

At the time of the color shift, the cultures were spread onto MacConkey agar indicator plates supplemented with glucose, on which only gram-negative bacteria can grow and which contains a pH indicator that turns the colonies red if they can

ferment the substrate. These plates developed red colonies, indicating that the cultures contained gram-negative bacteria that can ferment glucose, as expected. However, this did not prove that the colonies were *E. coli*. To prove this, a 200 bp segment of the 16S rRNA gene (110-320bp) was amplified using colony PCR and sequenced by Sanger sequencing on five colonies grown on M9-cellobiose agar plates from each of the three evolving cultures. All but one of the resulting fifteen 16S rRNA sequences was homologous to the expected *E. coli* sequence. However, one outlier, a 16S rRNA sequence amplified from a CB5m colony, was identified by a BLAST search against known bacterial sequences to belong to a species of gram-positive *Paenibacillus* from taxon E29. It was unclear how this organism was able to grow on the McConkey agar plates, which suggested that this result most likely indicated contamination of the PCR reaction rather than the culture. As previously described, the adaptive evolutions of these lineages were then continued to day 50.

However, *Paenibacillus* with the same 16S rRNA sequence was again discovered in media that had not been inoculated. After the adaptive evolution experiments were concluded, it was discovered that bottles of M9-cellobiose consistently develop growth after sitting on the shelf for several weeks. A 500 ml flask of media containing 250 ml of fresh M9-cellboise media was incubated at 30◦C, which developed optical density after about a week. The organism was subsequently identified as *Paenibacillus* by 16S RNA sequencing. Species of *Panibacillus* are known to be able to grow on cellulose and cellobiose(35-37). However, the growth rate of the *Paenibacillus* in M9-cellobiose media was found to be very slow (<0.02 hr^{-1}), suggesting that the strain found in un-inoculated media should not have been a threat to the cellobiose-adapted *E. coli* strains.

5.2.1.9.1 Source of Paenibacillus contamination

The source of this contamination has not definitively been determined, but it most likely comes from a non-autoclaved component of the media. Four components of M9-cellobiose are not autoclaved but are instead sterile-filtered using 0.2 µm vacuum-driven bottle top filters - 100 g/L cellobiose, 1 M $CaCl_2$,1 M $MgSO_4$, and Sauer's 100X Trace Element solution. Many species within the *Paenibacillus* genus are known to form spores (35-37, 43), which may be small enough to penetrate the filters. The *Paenibacillus* spores most likely originate in the cellobiose stock since the contaminating organism has not been identified in any other media made using the other three reagents. They may only germinate after being introduced to an environment with sufficient resources such as ammonia, since growth was only seen in M9-cellobiose media but not 100 g/L cellobiose stock solution.

Though the strain of *Paenibacillus* found in the un-inoculated M9-cellobiose media is not competitive with the adapted strains, if this organism has been in the media throughout the adaptive evolution study, it is likely that strains had the opportunity to adapt alongside *E. coli*. While the initial PCR screen suggested that *Paenibacillus* had not overrun *E. coli* in the evolving culture, it is possible they developed a cross-feeding relationship – in which case *E. coli* could be growing on fermentation products produced by *Paenibacillus* from cellobiose, rather than on cellobiose directly.

5.2.1.9.2 Possibility of Paenibacillus-E. coli *co-culture*

This issue had not been resolved by the time this study was terminated. However, several observations and results suggest a co-culture may have developed. Before the full extent of *Paenibacillus* contamination were discovered and fully comprehended, several other experiments were performed on the isolated colonies and adapted cultures.

For instance, at various time points after acquiring the cellobiose-consuming phenotype, the adapting strains were screened for mutations to ChbR, NagC, and the promoter of the *chb* operon, but none were ever found. Eventually some endpoint colonies were even screened for mutations to *bglGFB,* another operon containing a phospho-β-glucosidase gene (*bglB*) in *E. coli* that could conceivably catalyze cellobiose hydrolysis for metabolism (24, 47). *E. coli* colonies had not acquired mutations to any of the genes or genetic regions previously shown to allow cellobiose utilization (*chb* promoter, *nagC*, *chbR*, *bglGFB*). This was initially interpreted to suggest that growth in liquid culture selected a different cellobiose-utilization mechanism, since several known and hypothetical phospho-β-glucosidases are known. However, this could result from cellobiose utilization in *E. coli* simply not being developed.

It was also noticed that cellobiose-evolved cultures grown on M9-cellobiose plates always developed two types of colonies – large well-spaced colonies, and tiny colonies that always form around the large colony. Efforts to select the small colonies and grow them separately of the large colonies always failed, even in glucose media. The large colonies could be picked and propagated, but when re-plated always developed with the tiny colonies.

Additionally, PCR sequencing of 16S rRNA genes from these colonies from M9-cellobiose plates always produced multiple overlapping peaks. Eventually the 16S rRNA gene in CB5m-clone A was amplified and then ligated into TOPO cloning vectors so that individual sequences in the library could be analyzed. Two out of three sequences came back as *E. coli*, while the third was *Paenibacillus*. These results suggest a cross-feeding relationship between the strains composing the large and

small colonies, and that the large colonies only form when both types of cells are deposited together.

When this project was terminated, I was testing whether I could isolate *E. coli* colonies on plates made with cellobiose media filtered with 10,000 kDa centrifugal filters (Millipore® Amicon Ultra-15), which should eliminate any spore contamination. Colonies streaked from frozen stock of CB6 culture (flask 60, pre-mutagen) again showed two morphologies, large and small, though in this case they were well-spaced. However, the identities of these colonies were never determined.

5.2.1.9.3 Possible strategies to isolate cellobiose-utilizing E. coli

If ever given the opportunity to continue this project, I would briefly continue to determine whether cellobiose-consuming *E. coli* could be isolated from the adapted endpoints. This would require confirmed *Paenibacillus*-free media. While using protein-grade centrifugal filter devices to remove spores is not a good long-term solution given the volume of media needed, HPLC-purified cellobiose has since the start of this study become available from Sigma Aldrich(Fluka 22150), and would be far less likely to harbor spores than the source previously used (Sigma D7252, for which purification details are not readily available). Media prepared with this source of cellobiose would then be incubated for several weeks to observe whether *Paenibacillus* growth again accumulates without inoculation.

Once *Paenibacillus*-free cellobiose media can be generated, I would again try to isolate cellobiose-utilizing *E. coli* from frozen stocks of the adapted lineages and clones. This would most likely be done by streaking the frozen stock strains onto MacConkey agar plates supplemented with cellobiose as the carbon source. This medium is supposed to prevent growth of gram positive bacteria like *Paenibacillus*, though it is interesting that it was first discovered from a colony on a MacConkey-

glucose plate. Colonies confirmed to be *E. coli* could then be cultured and re-plated on M9-cellobiose to confirm consistent colony morphology. The optical density of culture grown from colonies and the volume plated could also be used to estimate the number of resulting colonies compared to wild type *E. coli*; if colonies only form where both *E. coli* and any residual *Paenibacillus* are plated together, the number of resulting colonies will be far below expectation.

If this were discovered, these lineages would not be usable for any further application in the cellulose project. However, the development of a cross-feeding relationship between *E. coli* and *Paenibacillus* during the course of adaptation may still be very interesting to study in terms of the development of microbial communities – especially since the complete development of this relationship is preserved in the frozen stocks collected throughout the course of the adaptation study.

If the cellobiose-utilizing phenotype in the adapted lineages were found not to belong to *E. coli*, adaptive evolution for this trait would have to be repeated. I would probably consider developing these strains by the same method previously described to generate this phenotype, incubating cells on cellobiose-containing plates for a month (24, 33), before applying adaptive evolution in liquid culture to any colonies that developed to optimize growth rate.

5.2.1.10 Conclusions from cellobiose adaptations

Despite the serious issue of *Paenibacillus* contamination, this stage of the project still produced some valuable discoveries. Adaptive evolution successfully produced strains able to grow well on cellobiose (0.2-0.4 hr^{-1}), which represents significant growth improvement for either species given that non-adapted *Paenibacillus* was measured to grow at <0.02 hr^{-1}. Additionally, this study explored different methods of applying selection pressure for a new trait.

It is unclear exactly when variants able to utilize cellobiose were first acquired since the trait was discovered in both sets of lineages when the selection pressure was amplified. Cellobiose utilization was discovered in the CB1-3 lineages after 24 days of growth in medium containing cellobiose, but these cultures were maintained in exponential growth on secondary substrates where cellobiose utilization may have lent too small of an advantage for mutants with this trait to become dominant. However, the discovery of cellobiose utilization when the cultures were grown into stationary phase suggests that this variant did nevertheless emerge in the population.

It took much longer to detect cellobiose utilization in the CB4-6 lineages, which was found after 46 days of adaptation, and after ten-fold more cell divisions than the CB1-3 lineages (**Table 5.2**). It is unclear why the significant periods of growth in stationary phase throughout the adaptation period were not sufficient to develop detectable cellobiose utilization under the given growth conditions. Cellobiose utilization was finally detected when the amount of cellobiose was increased to 10 g/L from 2 g/L. This suggests that the ability to utilize cellobiose had, again, been acquired at an earlier time point within the population, but that the growth of the cellobiose-competent cells was insufficient on 2 g/L cellobiose to be detected relative to NAG utilization or in overnight tests in M9-cellobiose.

Another interesting discovery was that the addition of trace elements to the media significantly increased the growth rate of on cellobiose. This indicates that the cellobiose-utilizing organism was limited by a metal ion. This enzyme was likely ChbF/CelF, which has been characterized as a member of the family 4 hydrosylases that work with PTS-transport systems are known to require divalent metal ions, often preferably Mn2+, for catalysis (58, 61). This family of enzymes is somewhat rare, suggesting that the growth increase from adding trace elements is most likely due to

increased ChbF activity in *E. coli* than a phospho-β-glucosidase in *Paenibacillus* – though other explanations are also possible.

Several other lessons learned from this study involve decisions that would not be repeated if given the opportunity. Clearly stronger efforts would be made to prevent and test for contaminating organisms. Additionally, I might not choose to use NTG mutagen again. The sequenced genome from another adaptively-evolved strain treated with NTG was recently received, and the preliminary data suggests hundreds of mutations were accumulated by each lineage (Dae Hee Lee, personal communication), which is far too to reasonably screen to identify those few that are beneficial rather than neutral or slightly deleterious. Since one of the goals of this study is not only to generate cellulose-utilizing *E. coli*, but to identify mechanisms of adaptation, this problem outweighs the benefits of achieving a higher growth rate.

5.2.2 Insertion of cellulase-secretion system encoded by the *out* operon

As described in the introduction (**Section 5.1**), the *out* operon encodes the type II secretion system (T2SS) that transports CelY and CelZ cellulases across the outer membrane (39, 68) in *Erwinia chrysanthemi*. These cellulases must be secreted to the extracellular space to gain access to their substrate, cellulose. *E. coli* also encodes a type II secretion system, encoded by the *gsp* operon (21). However, the *gsp* operon is unlikely to recognize the CelY and CelZ enzymes as secretion targets because type II secretion systems are very selective, and the secretion system from one species will not secrete the proteins recognized by the homologous system another even between closely-related gamma-proteobacteria (22). An important reason that the secretion and cellulase system from *Er. chrysanthemi* was chosen is that the entire system has already been shown to be functionally expressed in *E. coli* (39, 68).

This operon was going to be introduced into the *E. coli* genome in place of the silent *E. coli gspCDEFGHIJKLMO* operon. This location was chosen in order to minimize disruption of the chromatin structure and transcription regulation of other nearby genes, so that the adaptive value of acquired mutations could more confidently be attributed to their impact on cellulose metabolism rather than their possible role in compensating for genomic disruptions.

It was additionally decided that the promoter of the *gsp* operon would be preserved. *E. coli*'s *gsp* operon is normally repressed by heat-stable nucleoid structural protein (H-NS), a histone-like protein that binds to AT-rich, curved DNA and causes repression through chromosome condensation (13). H-NS is believed to repress many horizontally-transferred gene cassettes due to their higher AT content, producing the theory that H-NS acts to protect *E. coli* from the expression of unnecessary or harmful foreign genes (13). This study provides a unique opportunity to observe how this mechanism of repression can be repealed to allow advantageous foreign genes to be expressed and integrated, which is currently not well understood. Previous characterization of the *gsp* operon suggests that the existing repression motif is leaky, possibly allowing sufficient expression for minimal prerequisite viability. If this failed, modifications to the promoter would have been made to produce a sufficiently viable strain to initiate adaptive evolution on cellulose.

Integrating the *out* operon into the *E. coli* genome was expected to be the most challenging part of this part of the project. The plan was to integrate the 12.5 kb *out* operon into the *E. coli* genome using λ^{red} recombination (12), which has been shown to be capable of recombining fragments as large as 55 kb within *E. coli* (52). If attempts to integrate the *out* operon into the *E. coli* genome repeatedly failed, the backup plan was to express the *out* operon from the pCPP2006 plasmid (25, 68).

However, genomic integration is preferred since it is likely that selection on a plasmid-borne version will differ from selection in the genome since it has different replication and copy number dynamics.

Although the goal was to integrate the *out* operon into a strain of *E. coli* adapted to use cellobiose, extensive work was put into a pilot study to integrate the *out* operon into wild type *E. coli* to develop the method while the cellobiose-consuming strains were in development.

5.2.2.1 Design of *out* operon sequence for homologous recombination

The 12.5 kb *out* operon was obtained on plasmid pCPP2006, which was provided as a generous gift from Dr. Alan Collmer (Cornel University), who's research group first isolated and identified the operon (25).

In order to integrate the out operon, several regions of sequence needed to be added. The λ^{red} recombination system identifies double-stranded breaks and anneals them to regions with sequence homology at the ends (51). To insert a new sequence, the process works best when the sequence is presented in linear form with the flanking regions homologous to the desired insertion sites. In addition, a selection marker is needed to isolate strains that successfully integrated the new sequence.

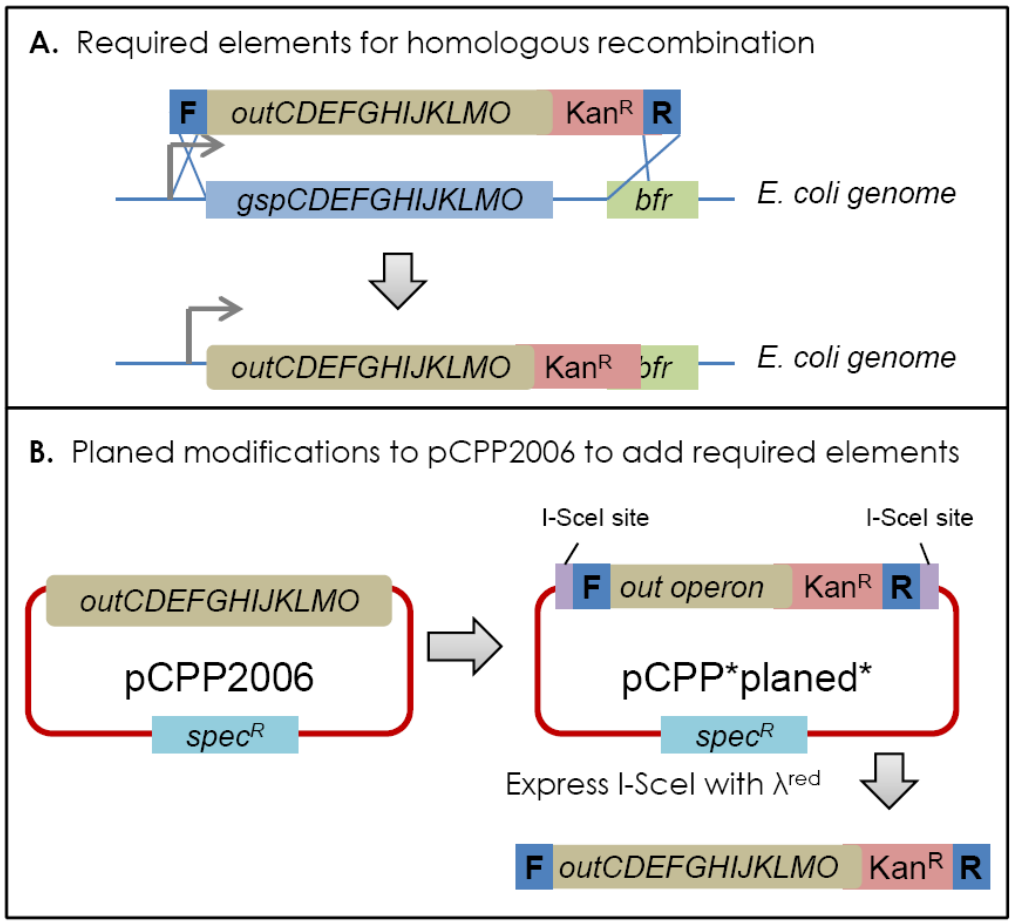


Figure 5.6 Plan to integrate the *out* operon by homologous recombination. (A) Required elements for homologous recombination of the outCDEFGHIJKLMO over the *gsp* operon. This includes addition of the kanamycin resistance gene downstream of the operon to allow selection of transformants, and ~50 bp flanking regions homologous to the insertion sites. The front gsp homology arm (dark blue box labeled "F") is homologous to the 50 bp region immediately upstream of the *gsp* start codon, and is designed to seamlessly integrate the *out* genes without disrupting the *gsp* promoter. The rear homology arm has homology to the 50bp region starting from the sequence 20 bp downstream of the end of the last transcribed *gspO* codon. This overlaps the *bfr* gene which encodes bacterioferritin, one of several nonessential iron storage proteins in *E. coli*. (B) A plan was devised to add the necessary elements to the out operon by modifying the pCPP2006 plasmid. In addition to adding the front and rear arms homologous to the gsp insertion sites and kanamycin resistance gene, I-SceI restriction sites were also added (purple boxes). This would allow the plasmid to be linearized in vivo using the pACBSR plasmid, which has both the λ^{red} and I-SceI gene on L-arabinose inducible promoters. *Note: None of the diagrams are drawn to scale.*

The design of the linear DNA fragment intended to integrate the *out* operon over the *gsp* operon is shown in **Figure 5.6A**. The front *gsp* homology arm sequence was designed to carry the 50 bp sequence immediately upstream of the start codon of the *gspC* gene, which leaves the native *gsp* promoter in-tact. The rear gsp homology arm sequence is comprised of the first 50 bp of the *E. coli bfr* gene, which is encoded 20 bp after the termination codon of *gspO*. A kanamycin gene was selected as the selection marker.

The additional sequencing regions were added to the *out* operon by modifying the pCPP2006 plasmid (**Figure 5.6B**). I-SceI restriction sites were also added to allow the plasmid to be linearized *in vivo* by co-expressing λ^{red} genes and I-SceI restriction enzyme from the same plasmid (pACBSR) (26) (the I-SceI restriction enzyme does not have any other digestion sites in the *E. coli* genome).

5.2.2.2 Construction of the modified pCPP2006 plasmid

Modifications made to the pCPP2006 plasmid are outlined in **Figure 5.7**. Homologous recombination was used to add the necessary flanking regions to the *out* operon sequence. Each *gsp* homology arms were added in separate phases so that the same kanamycin resistance gene could be used each time as the selection marker for the insertion of both. This was because the kanamycin resistance gene used contains FRT sites that allow one-step deletion when propogated in a strain that is also transformed with pCP20, which carries FLP recombinase (12).

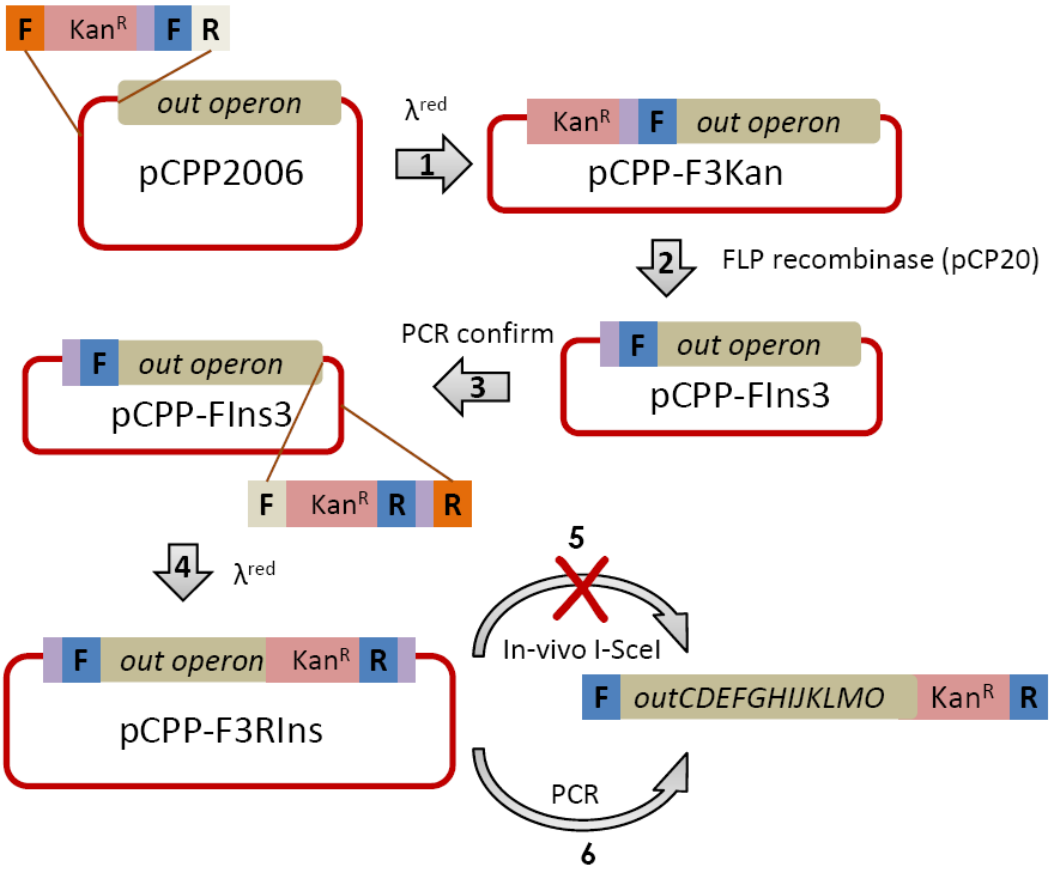


Figure 5.7 Generation of pCPP-F3R1 from pCPP2006. The *gsp* homology arms and kanamycin resistance gene were added to the flaking regions of the *out* operon on pCPP2006 by two rounds of λ^{red} recombination. Step Integrate the kanamycin gene, I-SceI restriction site, and the front *gsp* homology arm. This DNA fragment was generated by PCR of the kanamycin resistance gene from pKD13 using overextended primers that included the sequences for the other elements, including the arms homologous to the insertion sites on the pCPP2006 plasmid (orange and beige "F" and "R" boxes). Step 2. Used FRT sites within the kanamycin resistance gene to delete it by expressing FLP recombinase on pCP20 in *E. coli* carrying both plasmids. The new plasmid was renamed pCPP-F3. Step 3. Used PCR to confirm successful removal of kanamycin resistance gene. Step 4. Used λ^{red} recombination to re-insert the kanamycin resistance gene, the rear gsp homology arm, and I-SceI site. Generated plasmid named pCPP-F3R1. The proper insertion of all elements was confirmed by PCR and Sanger sequencing. Step 5. Homologous recombination of the sequence between the I-SceI sites into the *E. coli* genome was attempted using pABSCR, which expresses the I-SceI restriction enzyme and λ^{red} recombination proteins when induced by L-arabinose. However, all of these attempts failed as all tested KanR-selected colonies were positive for the pCPP-F3R1 plasmid instead of for genome integration. Step 6. Instead, the linear fragment was generated by PCR using high-fidelity polymerase specific for generating long PCR products (Takara LA-Taq polymerase).

The forward *gsp* homology arm was inserted first by amplifying the KanR gene from pkD13 to generate the necessary linear insertion fragment. The homologous regions to the insertion site on the pCPP2006 plasmid, the front homology arm to *gsp* operon, and the I-SceI restriction sites were synthesized as extensions on the primers used to amplify the kanamycin gene from pKD46. Following PCR with these extended primers, the product was isolated by cutting the appropriate band following agarose gel electrophoresis and agarose gel purification.

Following homologous recombination and selection for colonies by kanamycin resistance, proper insertion of the inserted sequence was confirmed by PCR using primers designed inside and outside of the inserted region. This produced a single band of the expected size (500 bp).

After confirming the presence of the forward insert, the KanR gene was removed by transforming the plasmid into *E. coli* also carrying pCP20 and growing 43^{o}C. The pCP20 plasmid expresses genes under these conditions that delete the KanR gene from pKD13 by recognition of FRT sites within the KanR gene (12, 26). The resulting version of pCPP2006 was renamed pCPP-FIns3.

The rear GSP homology arm was inserted similarly, but the KanR resistance gene was left intact to serve as a selection marker for the *out* operon when it is inserted into the *E. coli* genome. The KanR gene, rear-*gsp* homology arm, and I-SceI site were inserted into pCPP-HAinsF 225 bp downstream of the stop codon of the last out operon gene (*outO*). The final plasmid was named pCPP-F3RIns.

5.2.2.3 Homologous recombination of *out* operon into the *E. coli* genome

5.2.2.3.1 By in-vivo I-SceI cleavage

The first attempt to insert the *out* operon into the *E. coli* genome involved transforming the modified plasmid into *E. coli* MG1655 (CGSC 7740) carrying pACBSR.

The pACBSR plasmid carries the λ^{red} genes and I-SceI restriction enzyme on arabinose-inducible promoters (26). *E. coli* carrying both plasmids were treated with arabinose to induce linearization of the modified *out* operon fragment from the I-SceI sites, which could then be exchanged for the *gsp* operon by homologous recombination. However, unfortunately this method failed to produce colonies with the *gsp* insert, even after multiple attempts.

5.2.2.3.2 PCR amplification of the modified out operon

A second strategy to integrate the *out* operon was attempted by using PCR to generate linear copies of the modified out operon. The ~12kb product was purified by gel extraction following electrophoresis, and then electroporated into λ^{red}-expressing wild type *E. coli*. Cells that had undergone recombination were selected for kanamycin resistance on LB plates. Several kanamycin colonies were subsequently isolated. Integration of the kanamycin gene and downstream region of the out operon was then verified by generation of PCR products of the expected size from primers homologous to regions inside and outside of the insertion region. However, attempts to similarly verify insertion of the upstream region of the *out* operon consistently yielded negative results.

This suggested only partial integration of the *out* operon, and prompted an analysis of possible homology between the out and gsp operon that could have provided alternative splicing sites. A BLAST comparison found several such regions in the central regions of each of the operons (**Figure 5.8**).

5.2.2.3.3 Disruption of homologous regions in gsp *operon*

In response to this discovery, I decided to try to delete the region in the *gsp* operon that harbors most of the homology with the *out* operon. Since kanamycin resistance was already being used to select for *out* operon integration, and amplicilin

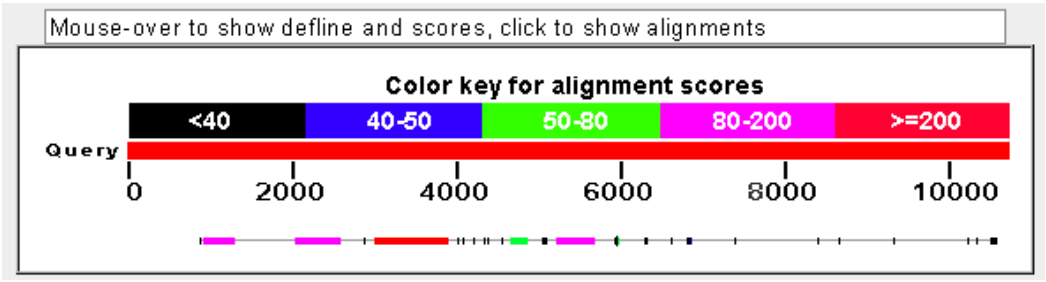


Figure 5.8 Blast-n alignment of the out and gsp operon sequences. A blastn comparison of the *gsp* and *out* operon found several regions with significant homology. Regions of high homology could act as homology arms that result in incomplete integration of the *out* operon over the *gsp* operon. Image is a screenshot of results from the NCBI blastn webpage (http://blast.ncbi.nlm.nih.gov/).

resistance is needed to maintain the pKD46 plasmid, a chloramphinicol resistance gene was selected. This gene was amplified from pACBSR(26) using primers synthesized to include 45bp extensions homologous to the ends of the region in the *gsp* operon being deleted (between 816-7346 bp from *gspC* start codon, to create a 6.5 Kb deletion). Homologous recombination of this purified PCR product was attempted in wild type MG1655 as previously described, with subsequent selection on chloramphenicol plates. This produced numerous colonies. However, none of the subsequent colonies were found to have the chloramphenicol gene inserted in the *gsp* operon.

This is as far as this leg of the project progressed before it was terminated. If it had been continued, the next step would have been to try deleting a smaller region of the *gsp* operon, with the hopes that the difficulty arose from the discrepancy in size between the chloramphenicol insert (1.3 kb) and the deleted region (6.5 kb). The kanamycin resistance gene with the FRT-sites also could have been used instead and then deleted before integrating the *out* operon fragment, which contains the same selection marker.

Alternatively, I could have attempted to insert the *out* operon in two steps by deleting the integrated Kan^r selection marker between subsequent steps. The only problem with this tactic is the possibility that two separate regions of homology between the *out* and *gsp* operon would be used in each integration step, leaving a gap between them of native *gsp* sequence which could damage the ability of the expressed type II secretion system to recognize the CelZ and CelY cellulases.

The final last resort was to express the *out* operon from the pCPP2006 plasmid itself, as previously mentioned. This would have possibly changed the selection dynamics on the genes. Additionally, it would have required spectinomycin to be included in all growth media until cellulose-dependent growth was achieved to maintain the presence of the plasmid. This situation also would have had the benefit of providing circumstances under spontaneous horizontal transfer might have been observed. Ultimately though, I think it is likely that this option would have been unnecessary and that the *out* operon would have been integrated given a few more weeks.

5.2.3 Integration of the CelZ and CelY cellulases

As described in the introduction to Chapter 5, the CelY and CelZ cellulases are two of many enzymes secreted by the plant pathogen *Er. chrysanthemi* to break down plant tissues to this end (4, 23, 25). Both the *celY* and *celZ* genes are needed to efficiently digest long cellulose molecules to into cellobiose. CelZ is a cellulose exonuclease (E.C. 3.2.1.4) that cleaves progressively from the ends of cellulose chains, and is responsible for the majority of the cellulase activity in *E. chrysanthemi* (66-68). The CelY enzyme is an endoglucanase (EC 3.2.1.4) that cleaves long cellulose molecules at random into shorter chains, and increases CelZ activity by providing additional ends to act as substrate(23, 66, 67).

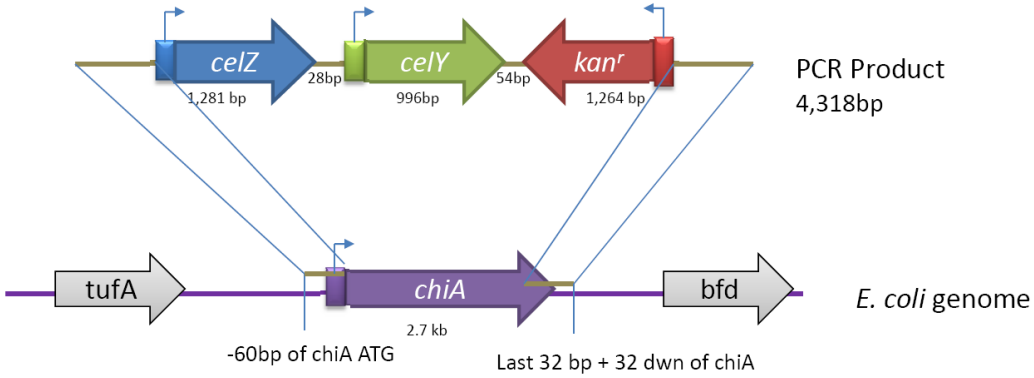


Figure 5.9 Design for integration of *celY* and *celZ* into genome.

5.2.3.1 Design of celZ and celY genome integration

The *celZ* and *celY* genes were going to be integrated into the position occupied by the *chiA* gene. The *chiA* gene encodes an endochitinase that has been shown to be secreted by the type II secretion system encoded by the *gsp* operon (20, 21). It is encoded approximately 1 kb downstream from the *gsp* operon, in the opposite orientation. It is also a cryptic gene as it is strongly repressed by H-NS (20).

As shown in **Figure 5.9**, the celY and celZ genes were going to be inserted over the *chiA* gene under the control of their promoters from *Er. chrysanthemi*. These promoters have been shown to effectively induce expression of *celZ* and *celY* in *E. coli* (25, 67, 68).

5.2.3.2 celZ on pLOI1620

The *celZ* gene was obtained on pLOI1620 (68) (**Figure 5.10**), which was provided by a materials transfer agreement with Dr. Ingram and the University of Florida – Gainesville. This gene was sequenced by PCR and Sanger sequencing of overlapping 600-900 bp regions. The amino acid sequence of the CelZ protein has been previously reported twice, by the laboratories of Lonnie Ingram (68) and by Federick Barras (6).

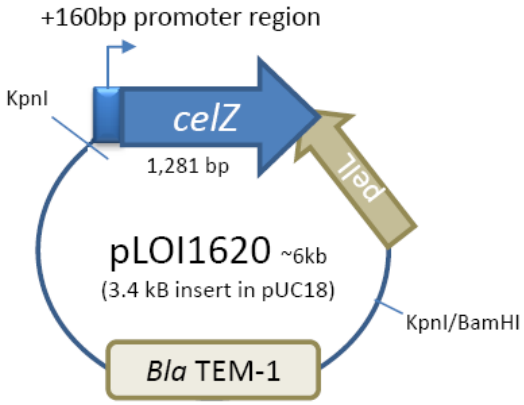
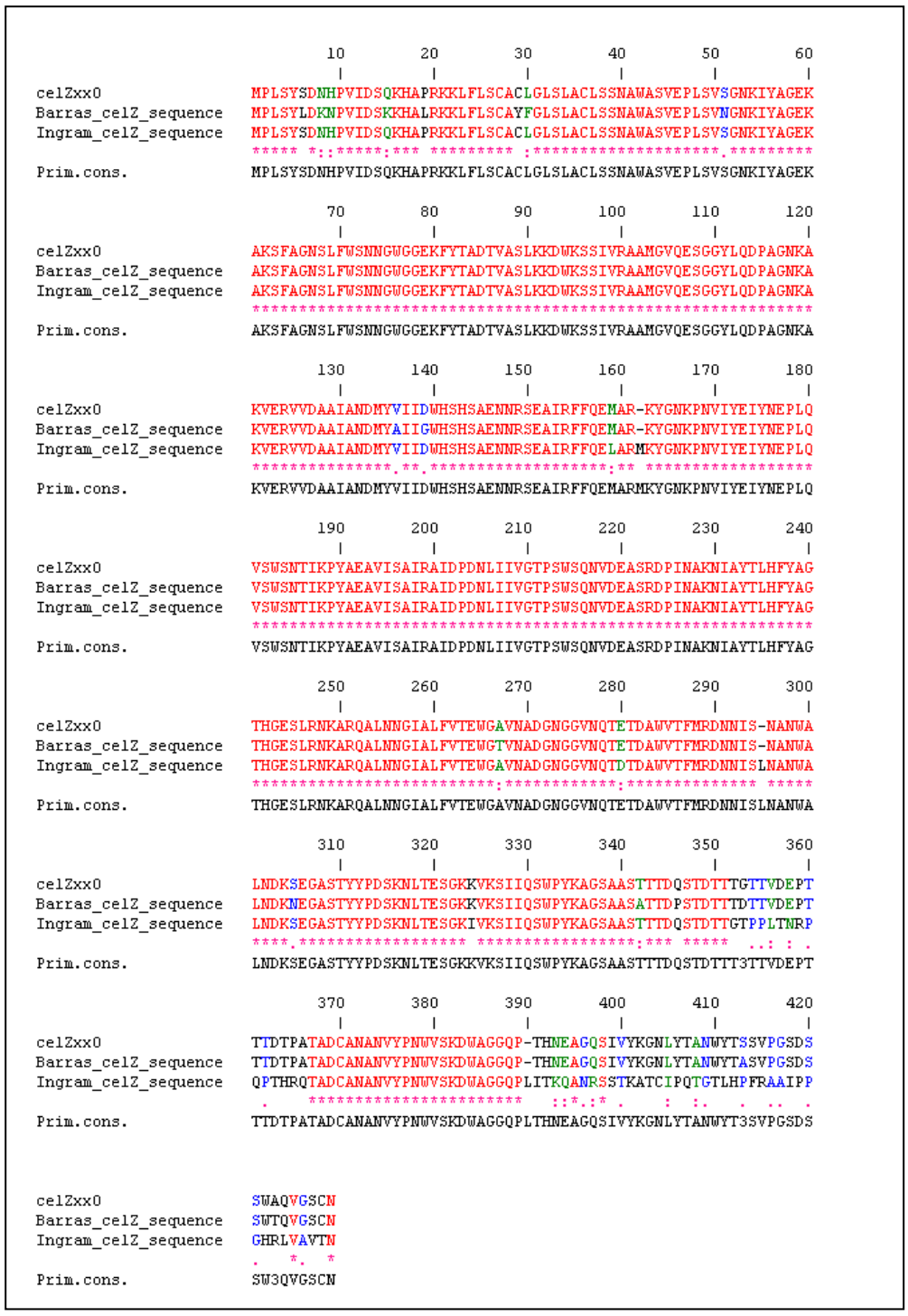


Figure 5.10 Diagram of pLOI1620. This vector was provided courtesy of Dr. Lonnie Ingram and the University of Florida-Gainesville. As shown, it consists of a 3.4 kb fragment of genomic DNA from *Er. chrysanthemi* that has been ligated into pUC18 at KpnI digestion sites (68). This fragment harbors the 1.2 kb *celZ* gene and the preceding 160 bp that includes the *celZ* promoter. The fragment also contains the pelL gene, which partially overlaps the *celZ* gene.

The translated protein sequences was aligned with both of the previously-reported sequences (**Figure 5.11**), which revealed many coding differences (celZ sequence provided in **Supplementary Data 5.1**). There were 14 residue differences with the sequence published by the Barras lab. It is unclear whether the sequenced gene is from the same strain of *Er. chrysanthemi*, so these differences could be due to population heterogeneity. However, it is more difficult to explain differences from the sequence reported by the Ingram group since they reported the sequence from the very same plasmid, but 53 residue differences were discovered – the vast majority in the C-terminal region. It seems possible that this resulted from a sequencing error. However, after receiving advice from committee members that this type of inconsistency is not rare, it was decided to continue without attempting to re-acquire the plasmid or to modify it to conform better to the previously reported sequences.

Figure 5.11 Alignment of CelZ amino acid sequences. The sequence of celZ from pLOI1620 was translated into to protein sequence (celZxx0) and aligned with the CelZ sequence reported by the Barras group(6) and Ingram group(68). Identical residues in all three sequences are shown in red. There are 14 differences between the translation of in-house sequenced *celZ* and the protein sequence reported by the Barras group, and 53 differences from the sequence published by the Ingram group.


10 20 30 40 50 60
celZxx0 MPLSYSDNHPVIDSQKHAPRKKLFLSCACLGLSLACLSSNAWASVEPLSVSGNKIYAGEK
Barras_celZ_sequence MPLSYLDKNPVIDSKKHALRKKLFLSCAYFGLSLACLSSNAWASVEPLSVNGNKIYAGEK
Ingram_celZ_sequence MPLSYSDNHPVIDSQKHAPRKKLFLSCACLGLSLACLSSNAWASVEPLSVSGNKIYAGEK
Prim.cons. MPLSYSDNHPVIDSQKHAPRKKLFLSCACLGLSLACLSSNAWASVEPLSVSGNKIYAGEK
70 80 90 100 110 120
celZxx0 AKSFAGNSLFWSNNGWGGEKFYTADTVASLKKDWKSSIVRAAMGVQESGGYLQDPAGNKA
Barras_celZ_sequence AKSFAGNSLFWSNNGWGGEKFYTADTVASLKKDWKSSIVRAAMGVQESGGYLQDPAGNKA
Ingram_celZ_sequence AKSFAGNSLFWSNNGWGGEKFYTADTVASLKKDWKSSIVRAAMGVQESGGYLQDPAGNKA
Prim.cons. AKSFAGNSLFWSNNGWGGEKFYTADTVASLKKDWKSSIVRAAMGVQESGGYLQDPAGNKA
130 140 150 160 170 180
celZxx0 KVERVVDAAIANDMYVIIDWHSHSAENNRSEAIRFFQEMAR-KYGNKPNVIYEIYNEPLQ
Barras_celZ_sequence KVERVVDAAIANDMYAIIGWHSHSAENNRSEAIRFFQEMAR-KYGNKPNVIYEIYNEPLQ
Ingram_celZ_sequence KVERVVDAAIANDMYVIIDWHSHSAENNRSEAIRFFQELARMKYGNKPNVIYEIYNEPLQ
Prim.cons. KVERVVDAAIANDMYVIIDWHSHSAENNRSEAIRFFQEMARMKYGNKPNVIYEIYNEPLQ
190 200 210 220 230 240
celZxx0 VSWSNTIKPYAEAVISAIRAIDPDNLIIVGTPSWSQNVDEASRDPINAKNIAYTLHFYAG
Barras_celZ_sequence VSWSNTIKPYAEAVISAIRAIDPDNLIIVGTPSWSQNVDEASRDPINAKNIAYTLHFYAG
Ingram_celZ_sequence VSWSNTIKPYAEAVISAIRAIDPDNLIIVGTPSWSQNVDEASRDPINAKNIAYTLHFYAG
Prim.cons. VSWSNTIKPYAEAVISAIRAIDPDNLIIVGTPSWSQNVDEASRDPINAKNIAYTLHFYAG
250 260 270 280 290 300
celZxx0 THGESLRNKARQALNNGIALFVTEWGAVNADGNGGVNQTETDAWVTFMRDNNIS-NANWA
Barras_celZ_sequence THGESLRNKARQALNNGIALFVTEWGTVNADGNGGVNQTETDAWVTFMRDNNIS-NANWA
Ingram_celZ_sequence THGESLRNKARQALNNGIALFVTEWGAVNADGNGGVNQTDTDAWVTFMRDNNISLNANWA
Prim.cons. THGESLRNKARQALNNGIALFVTEWGAVNADGNGGVNQTETDAWVTFMRDNNISLNANWA
310 320 330 340 350 360
celZxx0 LNDKSEGASTYYPDSKNLTESGKKVKSIIQSWPYKAGSAASTTTDQSTDTTTGTTVDEPT
Barras_celZ_sequence LNDKNEGASTYYPDSKNLTESGKKVKSIIQSWPYKAGSAASATTDPSTDTTTDTTVDEPT
Ingram_celZ_sequence LNDKSEGASTYYPDSKNLTESGKIVKSIIQSWPYKAGSAASTTTDQSTDTTGTPPLTNRP
Prim.cons. LNDKSEGASTYYPDSKNLTESGKKVKSIIQSWPYKAGSAASTTTDQSTDTTT3TTVDEPT
370 380 390 400 410 420
celZxx0 TTDTPATADCANANVYPNWVSKDWAGGQP-THNEAGQSIVYKGNLYTANWYTSSVPGSDS
Barras_celZ_sequence TTDTPATADCANANVYPNWVSKDWAGGQP-THNEAGQSIVYKGNLYTANWYTASVPGSDS
Ingram_celZ_sequence QPTHRQTADCANANVYPNWVSKDWAGGQPLITKQANRSSTKATCIPQTGTLHPFRAAIPP
Prim.cons. TTDTPATADCANANVYPNWVSKDWAGGQPLTHNEAGQSIVYKGNLYTANWYT3SVPGSDS
celZxx0 SWAQVGSCN
Barras_celZ_sequence SWTQVGSCN
Ingram_celZ_sequence GHRLVAVTN
Prim.cons. SW3QVGSCN

5.2.3.3 Synthesis of the *celY* gene and generation of pLOI-CelZY

It was not possible to obtain the *celY* gene from another research group, so it was synthesized (GeneArt). The synthesized sequence was based on the *celY* gene reported by Guiseppi et al (23). However, 37 synonymous nucleotide changes were made to increase codon optimization for expression in *E. coli* and to reduce the GC-content in a region (between 1000-1050 bp) that may have interfered with gene synthesis. Codon optimization was performed somewhat conservatively – instead of changing every codon to the one most-frequently used in *E. coli*, only codons used less than 20% of the time to encode a particular amino acid were changed, and they were changed to the next most-frequently used codon for that amino acid. Frequencies of codon usage in *E. coli* were obtained from the codon usage databank maintained by the Kazusa DNA Research Institute, Japan (45). The sequence of the synthesized DNA fragment is provided as **Supplementary Data 5.2**.

The fragment of DNA synthesized also included the 280 bp region upstream of

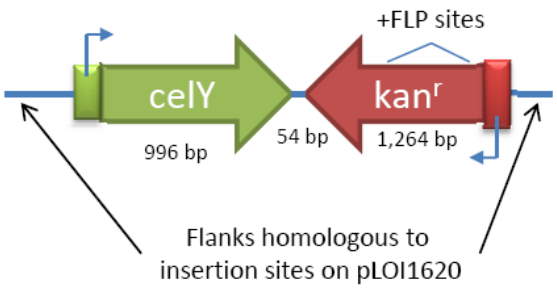


Figure 5.12 Design of synthesized *celY* fragment. The synthesized fragment of DNA was synthesized to harbor the celY gene and associated promoter region, the kanamycin resistance gene, and 50 bp arms homologous to the desired insertion sites on the pLOI1620 plasmid. Its total length was 2893 bp, and was synthesized by GeneArt®.

the *celY* gene, a kanamycin resistance gene with FRT sites, and 50 bp homology arms for insertion into pLOI1620 (**Figure 5.12**). After synthesis, it was amplified by PCR (LA-Taq, TaKara®) and inserted into pLOI1620 by λ^{red} recombination (**Figure 5.13**) to generate pLOI-CelZY. This plasmid was isolated from colonies selected for kanamycin-resistant growth, and the correct insertion of the *celY* and kanamycin resistance gene was verified by PCR using primers both in and outside the inserted gene sequence.

5.2.3.4 Next step - integration of CelZ & CelY into *E. coli* genome

Had the cellulose project continued, the next step would have been to use homologous recombination to introduce the *celZ-celY-kan*^{r} construct to the *E. coli* genome in place of the *chiA* gene, as previously described (**Figure 5.9**). These genes can be integrated into the *E. coli* genome by PCR amplification and λ^{red} recombination, using primers with 40-60 bp extensions homologous to the desired insertion sites on the *chiA* gene. This step would have been performed once a cellobiose-competent strain of *E. coli* had been isolated (**Section 5.2.1.6).**

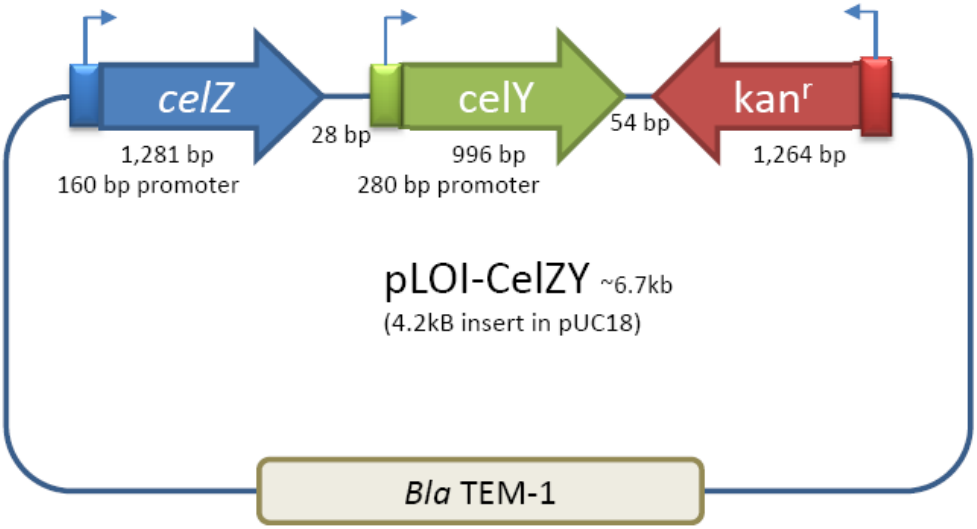


Figure 5.13 Diagram of pLOI-CelZY. This plasmid was derived by λ^{red} recombination of the synthesized fragment containing *celY* **(Figure 5.11)** into pLOI1620.

5.3 Uncompleted Work - Plans for remaining project stages

5.3.1 Validate secretion of cellulases in genetically-modified strains

Once the cellulase genes and type II secretion system had been integrated into the genome of a cellobiose-utilizing strain of *E. coli* (or, in the case of the out operon, possibly expressed from a vector), the first step would have been to validated that active cellulases were being expressed and secreted. This would have involved assaying media and cell lysate for release of reduced sugars using a combination of the dinitrosalycilic acid (DNS) method (64, 66) and congo red plate staining (5, 68). Once this had been validated, selection for cellulose-dependent growth could have begun.

5.3.2 Adaptive evolution on cellulose

Adaptive evolution of the cellulase-secreting strain would have involved continuous growth on M9 minimal media supplemented with amorphous cellulose (most likely phosphoric-acid swollen cellulose) (10 g/L) (63). A single clone would have been selected and used to initiate 8-10 independent lineages. Adaptations would have been maintained for at least 60 days and for no more than 120 days, depending on whether the growth phenotypes became stable or continued to change towards the end of the adaptation period. Frozen stocks of each population would be made every 5-10 days.

It was also considered likely that the adapting lineages may not initially secrete cellulases sufficiently to grow on cellulose. As in the cellobiose adaptations, the initial cultures may have been supplemented with cellobiose, which would have gradually been reduced to select for mutants with improved cellulose-consuming

abilities. Once growth is independent of supplemented cellobiose, adaptive evolution will continue to select for improvements in cellulose-based growth.

5.3.2.1 Selection of growth medium

To our knowledge, this would have been the first laboratory evolution to select for growth dependent on protein secretion. Because of this, it is unclear whether we can grow the cultures in liquid media since it would make it difficult for cells to directly benefit from the enzymes they produce as they would be able to diffuse away. Modeling studies have also found that cooperative behavior is only advantageous when there is limited dispersal of the organisms and their products (11). This suggests that selection for increased cellulase secretion and degradation is more likely to be effective if the cells are grown on M9-cellulose agar pales.

To replenish the resources available to cells and simulate transfers, the plan was to scrape the cells off the surface of each plate into liquid media each day. The cell colonies would be dispersed by vortexing and by providing several hours of incubation for cell clumps to disperse. This would hopefully maintain sufficient selection pressure for increased growth rate as cells from larger colonies would presumably compose a larger fraction of the population transferred to the new plate. Meanwhile, once plated the cells could directly benefit from their own cellulases. If this method proves to be too impractical or fails to select for increased growth rate, methods that alternate growth in liquid media and on plates could be attempted.

5.3.2.2 Characterization of evolved cultures

During the course of adaptive evolution, the growth rate of each of the populations in liquid M9-cellulose and M9-cellobiose would have periodically been measured. Cellulase activity and accumulation of byproducts in the spent media would also be monitored. This information would be used to gauge the rate of

adaptation and determine whether a stable phenotype has emerged in each population.

Once a stable phenotype emerged or sufficient time had elapsed (>120 days), the adaptive period would be terminated and the adaptations in the resulting strains would be assessed. These changes would have occurred on the population, cellular, molecular, and genetic level. However, this study is mostly focused on connecting genetic changes to cellular fitness. This can only reasonably done for a single genotype, so a single clone from each population would need to be isolated for further study.

The selection of these clones may need to be done carefully, since it is unclear whether at the end of the adaptive evolution process the populations would have fixed a single genotype (16, 27), still harbor many competing genotypes with similar phenotypes (42), or have developed frequency-dependent selection for cross-feeding specialists (with some specialized to consume byproducts such as acetate) (14, 60). To gauge this, approximately five colonies would have been selected from the endpoint population of each independent lineage to assess the heterogeneity present in the population. Each colony would be screened for growth rate on cellulose, intracellular cellulase activity (DNS assay of cell lysate (64)), and secreted cellulase activity (congo red halos (68)) to gauge their relative phenotype improvements relative to growth on cellulose.

We would also determine whether the adaptations affected the growth rate of each clone on a range of alternative substrates. Previous adaptive evolution studies have shown that adapted strains frequently have altered growth capacities on other substrates compared to the initial strain and each other (16, 29, 59), caused by acquired mutations impacting metabolism growth on other substrates (pleiotropy).

The number of conditions where growth is effected and its severity can indicate how strongly metabolic regulation has been altered by the acquired mutations. Additionally, these growth profiles can distinguish colonies that have acquired different genetic changes that lead to similar abilities to grow on cellulose, since different sets of mutations are likely to have different impacts to growth on other substrates (16).

These results would have been used to select one colony from each population for in-depth study through genome resequencing and proteomic analysis. It is most likely that colonies with the highest cellulase secretion rate or cellulose growth rate would be primarily selected, though other traits would likely be used to pick between colonies from the same population that are similar in that regard.

5.2.4.3 Genome Resequencing

The genome of each colony would have then been sequenced using next-generation Illumina sequencing (2, 3, 56). This method is very proficient at discovering single-nucleotide polymorphisms and deletions, and is somewhat less reliable at identifying insertions (8). Restriction fragment length polymorphism (RFLP) analysis may also have been employed to screen for large-scale rearrangements or insertions/deletions.

The genome of the selected cellobiose-competent colony with the cellulase and type II secretion genes would also have been sequenced to serve as a reference strain for mutations acquired during adaptive evolution to cellulose. The genome of the original wild type strain, *E. coli* K12 MG1655 (BOP27) has already been sequenced several times, and its sequence could be compared to the pre-cellulose adaptation to discover mutations acquired during adaptation to cellobiose. The resulting list of genetic differences would have been validated by Sanger sequencing.

5.3.3 Analysis of Adaptive mechanisms and mutations

The phenotype characterization and genome resequencing would have completed the first major stage of this project, at which point I would have prepared the first manuscript and considered my PhD research complete. The work described in the following section outlines plans for post-doctoral research had the project acquired funding.

The list of acquired mutations alone has not been sufficient to understand mechanisms of adaptation, and this is the point at which the real interesting questions can begin to be answered. Adaptive mutations have been found in genes with no clear relationship to the phenotype, as well as in global regulators that have so many functions it is unclear which are targeted or how they cause their adaptive effect (8-10, 27, 38). As demonstrated in **Chapter 4,** even when the gene clearly encodes an enzyme that clearly has a function related to adapting to the selection pressure, such as glycerol kinase mutations in response to growth on glycerol, the mechanism underlying adaptation can be difficult to predict. Thus once the strains and the list of mutations have been acquired, the genotype-phenotype relationships that result from previously unrecognized but critical molecular relationships can be examined.

The explicit goal of these studies would be to understand the molecular mechanisms underlying the ability of the mutations to increase fitness. Mutations increase fitness by alleviating constraints on growth, but the constraints they alleviate are usually not known – even in the most comprehensively studied model organisms like *E. coli*. These constraints are targeted by selection because of their importance in determining the phenotype, and the mutations represent opportunities to identify parameters to which phenotype is sensitive.

In this study, several specific constraints are known. Mutations that are likely to increase expression and function of the heterologous genes would have been of special interest, especially those involved in reducing H-NS repression of the gsp promoter (13, 21). Mutations that may affect biomass partitioning to protein synthesis would also be considered high-priority targets

5.3.3.1 Systems-level Profiling

The effect of adaptation on metabolism and gene regulation would have been determined, and correlated to the effects of specific mutations later if possible. However, this type of data has been difficult to understand because many of the changes seen at the systems-level in adapted strains are secondary responses to the critical adjustments, but do not significantly effect phenotype (16).

Broad changes to metabolism and gene regulation over the course of adaptation would have been measured by generating profiles of byproduct secretion and gene expression. The byproduct secretion profiles would indicate whether strains had adopted different metabolic modes to optimize growth on cellulose (16, 41). Metabolomic profiling or ^{13}C-tracing experiments may have also been performed for this purpose, depending on accessibility. Gene expression profiles (and possibly proteomic profiles) would have been used to identify transcriptional changes common across all evolved strains relative to the pre-evolved cellulose construct, and in common between those that appear to have acquired the same mode of metabolic adaptation. This would have been used to observe whether particular regulons have been significantly affected. Previous adaptive evolution studies have found changes to the Crp, RpoS, Fis, and ppGpp regulons (9, 10, 27).

5.3.3.2 Biochemical characterization of mutations

Following resequencing and systems-level profiling of the endpoint strains, a group of up to 15 mutations were to be chosen to study at the biochemical level. Based on previous metabolic adaptation studies, at least four mutations were expected to have been acquired per strain on average (8, 27) (this estimate increases significantly if mutagen is used). This suggests that there would be at least 40 mutations across the 10 proposed clones. Unfortunately, this is too many to study in detail within a reasonable timeframe.

Mutations would primarily be selected based on the gene affected. Mutations to the heterologous cellulase genes or type II secretions system would be considered particularly interesting, as would non-synonymous mutations to enzymes known to be involved in cellobiose metabolism. Mutations to global regulatory genes would also be particularly interesting. Another factor likely to be considered is how often different lineages acquired mutations to the same genes or to genes involved in the same or similar processes, as this can signal the mutations alter processes important to cellulose or cellobiose adaptation.

The selected mutants would be introduced into either the wild type or pre-cellulose adapted strains by site-directed mutagenesis (26) to observe their ability to independently alter phenotype. Mutations that have a strong effect on phenotype and the ability to grow on cellulose would be considered good targets for in-depth molecular profiling to determine the molecular mechanism by which the mutation improves fitness. The types of experiments this would involve are difficult to speculate on, as they would be highly dependent on the mutation loci and the affected phenotype. It is also likely that the assistance of collaborators with in-depth experience with particular proteins or assays would be sought.

5.4 Summary and Concluding Remarks

The goal of this project was to develop a strain of *E. coli* that can utilize cellulose as a carbon source by introducing heterologous genes encoding cellulases and their affiliated secretion system and using adaptive evolution to optimize their expression and growth on cellulose. This project was not completed, but significant progress was made towards constructing the initial cellulose-competent strain of *E. coli*.

5.4.1 Summary of Completed Work

5.4.1.1 Adaptive evolution to generate growth on cellobiose

Growth on cellobiose in M9 medium was found after 20 or 46 days of propagation, depending on the lineage (all grew at ~0.06 hr^{-1}). The development of this trait required the cultures to be in stationary phase for significant periods of time to create a sufficient advantage for cellobiose utilization over the secondary substrate (N-acetyl glucosamine) used to propagate the population. Once growth on cellobiose was detected, the strains were cultured in M9-cellobiose without secondary substrate for another 24-25 days to allow further adaptations to acquire that optimize cellobiose utilization. This increased the growth rates of the lineages to ~0.12 hr^{-1}. Addition of trace elements further increased the growth rates to 0.2 hr^{-1}. The rate of adaptation was then increased by treating the cultures with the mutagen NTG, which caused one of the cultures to crash but increased the growth rate of the other two lineages to 0.45 hr^{-1} after three treatments.

The NTG treatments also caused the color of the cultures to change dramatically, which prompted a check for contamination that revealed both *E. coli* and a gram-positive strain of *Paenibacillus* in the cultures. *Paenibacillus,* which is known to grow on cellobiose, was subsequently found in un-inoculated cellobiose

media suggesting that the filtration method used to sterilize the cellobiose stock was insufficient to eliminate *Paenibacillus* spores. However, the *Paenibacillus* in the un-inoculated cellobiose media was very slow growing (<0.01 hr^{-1}), indicating that this strain was not responsible for the observed growth and not competitive with the developed cultures. But it was also possible that strains of *Paenibacillus* had co-evolved with *E. coli* throughout the adaptation period, and that adapted *Paenibacillus* was the primary cellobiose consumer while *E. coli* consumed the resulting byproducts.

At the time work on this project was terminated, measures were being taken to isolate *E. coli* from the cultures and to screen them for the ability to independently grow on cellulose.

5.4.1.2 Integration of the *out* operon

The 12.5 kb *out* operon, encoding the type II secretion system, was acquired on plasmid pCPP2006. This operon was significantly modified using λ^{red} recombination to incorporate a kanamycin resistance selection marker, I-SceI restriction sites, and flanking regions homologous to the desired integration sites on the *gsp* operon in *E. coli*. In-vivo linearization of the construct at the I-SceI restriction sites failed to produce colonies with the integrated *out* operon. Instead, λ^{red} recombination was attempted using linear copies of the region generated by PCR with a taq polymerase able to amplify long (up to 30 kb) fragments. Homologous recombination using λ^{red} was again attempted and partially successful – half of the operon was integrated. At this point several regions of significant homology between the *out* and *gsp* operons were discovered, which likely served as intermediate insertion sites.

At the time the project was terminated, efforts were underway to delete the regions of homology in the *gsp* operon by replacing them with a chloramphenicol

resistance gene. Once deleted, insertion of the *out* operon would have been attempted again. If subsequent attempts to integrate the *out* operon continued to fail, it would have been expressed from pCPP2006 plasmid.

5.4.1.3 CelY and CelZ

The *celY* and *celZ* gene were both integrated onto plasmid pLOI-CelZY in tandem with a kanamycin resistance gene. The *celY* gene could not be obtained, so it was synthesized in tandem with a kanamycin resistance gene and integrated into the pLOI1620 plasmid containing *celZ*. At the time the project was terminated, these genes were ready for integration into a cellobiose-utilizing *E. coli* strain.

5.4.2 Parting Comments

Significant progress was made towards completing this project, but not as much was accomplished as anticipated when it was begun. From the start it was considered an "ambitious" project, though at the time I was not experienced enough to know exactly how difficult it can be to translate a plan from paper to practice.

Many of the problems encountered were, admittedly, due to inexperience and experimenter error. The cellobiose adaptations took much longer than expected, and probably should have been circumvented by either using recombination techniques to add mutations previously found to induce cellobiose competence, or by performing the adaptations on plates as previously described – growth on cellobiose could have been optimized after the phenotype had been acquired. The most serious problem, the *Paenibacillus* contamination issue, could have possibly been avoided if higher-grade cellobiose had originally been purchased, the limits of the sterilization filters had been recognized, or if thorough checks for culture contamination had been regularly planned and performed.

Additionally, homology between the *gsp* and *out* operons should have been expected given their shared function.

However, all of these problems could have been addressed and likely overcome, as described more thoroughly in previous sections. Frustratingly, due to financial constraints, this project was terminated shortly before accomplishing the first major goal – construction of a cellulose-competent strain.

I still think that this project is viable and worthy of pursuit if it found funding. The surface goal of the project, to develop a cellulose-consuming strain of *E. coli*, has practical value as a platform for metabolic engineering of strains that convert cellulose to other high-value chemicals. However, the most interesting questions are those that would have been addressed in the later stages of the project, after all of the mutations had been identified and assessed. These questions, as discussed in the introduction, address how regulatory networks adapt to integrate new metabolic functions, if *E. coli* can change how it partitions its metabolic resources to optimize protein secretion, and how H-NS repression can be overcome.

This type of study exemplifies how adaptive evolution can be used in the post-genomic era to identify molecular mechanisms that have otherwise eluded study, and to identify the critical biochemical and regulatory constraints that determine phenotype. Ideally, this research should fill in the gaps between the different scales of biological function currently studied, by following how a single adaptive (and thus high-impact) genetic mutation changes the function or expression of the encoded target, and how that change propagates through the metabolic and regulation network to finally produce the observed phenotype. Such efforts can be very demanding experimentally, and requires knowledge of a broad range of biological fields including population genetics, systems biology, molecular biology and protein

biochemistry. However, only by addressing these types of questions will we be able to develop a comprehensive understanding of *E. coli,* and of biological systems in general.

5.5 Acknowledgements

I would like to thank Eric Knight for helping me design the DNA integration scheme, Dae Hee Lee for always being available to help me troubleshoot and for sharing his experience with NTG mutagenesis, and Karsten Zingler for sharing his microbiology expertise while investigating the contamination issues in the cellobiose adaptive evolutions. I'd also like to thank the other members of the Metabolic Engineering FORCE – Adam Feist, Vasiliy Portnoy, Dae Hee Lee, and Jeff Orth – it was great being able to go over our daily concerns and headaches together, as well as our successes.

5.6 Appendix

Supplemental Data 5.1 – Sequence of celZ insert found on pLOI1655

Vector insertion site (1-14bp):
TCGAGCTCGGTACC
Promoter region (15-175 bp):
CGGGGATCTACTGTAGAAAATCCACTATAACCTAGGGAGCTTCATCATGTGCGAGTATT
GCAATATTGGTGAACACATCGACTGAATAGAATAAAACCAATCACCTATTCTATTATTACA
GTTAACACTGACACATCGTCATATGAAATGGAGATTCATTT
celZ (176-1,436 bp) (length:1,281 bp):
ATGCCTCTCTCTTATTCGGATAACCATCCAGTCATCGATAGCCAAAAACACGCCCCAC
GTAAAAAACTGTTTCTATCTTGTGCCTGTTTAGGATTAAGCCTTGCCTGCCTTTCCAGTAATG
CCTGGGCGAGTGTTGAGCCGTTATCCGTTAGCGGCAATAAAATCTACGCAGGTGAAA
AAGCCAAAAGTTTTGCCGGCAACAGCTTATTCTGGAGTAATAATGGTTGGGGTGGGGA
AAAATTCTACACAGCCGATACCGTTGCGTCGCTGAAAAAAGACTGGAAATCCAGCATT
GTTCGCGCCGCTATGGGCGTTCAGGAAAGCGGTGGTTATCTGCAGGACCCGGCTGG
CAACAAGGCCAAAGTTGAAAGAGTGGTGGATGCCGCAATCGCCAACGATATGTATGT
GATTATTGACTGGCACTCACATTCTGCAGAAAACAATCGCAGTGAAGCCATTCGCTTCTT
CCAGGAAATGGCGCGCAAATATGGCAACAAGCCGAATGTCATTTATGAAATCTACAA

CGAGCCGCTTCAGGTTTCATGGAGCAATACCATTAAACCTTATGCCGAAGCCGTGATTT
CCGCCATTCGCGCCATTGACCCGGATAACCTGATTATTGTCGGTACGCCCAGTTGGTC
GCAAAACGTTGATGAAGCGTCGCGCGATCCAATCAACGCCAAGAATATCGCCTATA
CGCTGCATTTCTACGCGGGAACCCATGGTGAGTCATTACGCAATAAAGCCCGCCAG
GCGTTAAATAACGGTATTGCGCTTTTCGTCACCGAGTGGGGCGCCGTTAACGCGGAC
GGCAATGGCGGAGTGAACCAGACAGAAACCGACGCCTGGGTAACGTTCATGCGTG
ACAACAACATCAGCAACGCAAACTGGGCGTTAAATGATAAAAGCGAAGGGGCATC
AACCTATTATCCGGACTCTAAAAACCTGACCGAGTCGGGTAAAAAAGTAAAATCGATC
ATTCAAAGCTGGCCATATAAAGCGGGCAGCGCCGCCAGTACAACAACCGATCAGT
CAACCGATACCACCACTGGCACCACCGTTGACGAACCGACCACAACCGACACAC
CGGCAACCGCTGATTGCGCCAATGCCAACGTTTACCCCAACTGGGTTAGCAAAGACT
GGGCGGGCGGGCAGCCGACCCATAACGAAGCGGGCCAATCGATCGTCTACAAA
GGCAACCTGTATACCGCAAACTGGTACACTTCATCCGTTCCGGGCAGCGATTCCTCCT
GGGCACAGGTTGGTAGCTGTAACTAA

Post-translation region (1436-3424bp):

TTGATTAATCTTTTCACCCCCAAAAAAACAGGGCTGCGATTGCAGCCCTGATATGCAAC
ATTCCATTACTTAATTGCGTTCAAAAGCGCCCAAATCCGGTGCGCTGCCTGAATAACTG
ATATTGCTCTCTTTCGTACCCGCGTTAATCAGCTTTGAGTTAGCCGACAGACGGAACAG
CGAGGTTTCCGGCAACGTGCCGTCATTATCACGAGATACGGTAGCCAGCGAGGTGT
CCAGGCTGACGAAATCGGACGCGGAAGCCGCCGGTCCCGTATCCCATGAGTTGG
ACTTCGCATCCGCATTACTGACCGTTGCAGAAGCAGACAGAGACACGTTGTTGCGGA
AGTAATGTTTCTGTCCTGACTGGACGTTGCTCCCGAAACCATAATTAATGCCGTTTTTATAT
GACGTGTTATTTATTACCGTGACGCCGCCGGCGTTATTGTTCTGGTCAAAACCTTTGCTCA
CGTTGCCAAACGCGACGGAACGGGTAATGCGATGATTGCCGACCGCCTGGTTTCCT
CCCAGTTTGAACCCGTTGCCATTGCCGGCGAACGCGCTGTCATTCCAGTAATTAATAC
CGTTACGGAAAGCCCAGCTGTTTTCAATCACCACTTTTTGCGGGCTGTCGAATAAGTCA
AATCCGTCATCCGAGTTTTCCCAGGCACGGCAACCGACAAAACGGTTACCCGGCCC
CTGTTTTTGTTTTGGTCCAAAGCCATCGGCCATGCTGCCGTTCTTTTTAGGATCATAGTTACG
GTAGGCGTCTGAGTTAATTACCGTGTTGTATGAACCACCGTTATTGATTTCAAGGCCGGT
ATTTCGGTTATGATGGAACGCCGTATTCTCAAAGGTATTGTGGTTACCAATCACATAAGCC
CCCTGATAACCGGCTCGGGTTACCTCAACGCCTTTGAAATACCAGTAGTCGCCGGTC
ACATAAAATCCGTAAGACGCCTGTACCCATTGGCTGTCAGGGAATGAGAAGTCAAATA
CCGCCCGGCCGCAATTGGCGGTAGCCACATATATCGGCGCTCCCTCTTTCCCGGATT
TATTAAATGTGATCGTATTCCCTTTACCCTGAGTATAGGGAATGGTATAGGTTCCCGGTTTC
AACAGAATCAGCTCGCCGGGGGTTAACAGCCGCCATGGCGGCGGAAAAACTCAT
CGGCGCATTAAAACTGCTGCCGTTATTGCTGCTGGTGCCATTCGGCGCGACATAATAA
ATACGCTTGGTGCTAATTCCGCTAGTGAGATCCGAACTGCAATCTGCGGCCAGAACTG
ATGTGCTATTTACCAGAAAGAATGCGGCTAACCCGGTACTGATAAAGCAATTTAAATATTT
CATCCTTTGCCTCATGATGTTGTAAAAATCATTTATCCATTGTCCTCAAGATGCATGCTTCG
GCAAATGTCGTCAGATGGGGAAACGGCGGCGATCCGCTGCCTCTAAAATCGATAGC
CCAATGAATCAGCCTGTCACTGTGAAAAATTACGTTGGAAGGGTACCCGGGGATCCT
CTAGAGTCGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTCGTTTTACAACGTCGT
GACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTC
GCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGC
GCAGCCTGAATGGCGAATGGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGG
TATTTCACACCGCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAG
CCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCT
CCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGA
GGTTTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTA
TTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCA

Supplemental Data 5.2 – Sequence of synthesized *celY* DNA fragment

5' pLOI1620 Homology Arm
CAGGGCTGCGATTGCAGCCCTGATATGCAACATTCCATTACTTAATTGCGTTCAAAAGC
GCCC

celY promoter region
CGCGGTGACGGTGGTGGATCTGGAACGGGAAATTAATCGGGTGCTGTCGCTGTTCCA
TTGGGGATGCGTGGACCTGCGCCCGTATGAAAATCGGCTGGAGATTTATCATCTGCGTT
GCCCGCCTCGGTGAACTCGTCGGGCAGCGTCAGATGGCGCATGGCGATGGCGGC
GGTGTTGCAGGGGCTCTACAGCCGCTGGCTGCGTGAGCAGGGCGGCGTCGAGGT
CTGTACCGCTGTCCTGCGAAGAAACCGACAGCGAATCGACCCTGCTGTTCCGTTACC
AACACTGAGCCGGACAGTA

celY sequence
ATGGGTAAGCCGATGTGGCGTTGTTGGGCGTTAATGCTGATGGTGTGGTTCAGCGCGT
CGGCAACGGCGGCGAACGGCTGGGAAATCTATAAAAGCCGTTTCATGACCACGGA
CGGTCGCATTCAGGATACCGGCAATAAGAATGTTAGCCACACCGAAGGTCAGGGTTT
CGCCATGCTGATGGCGGTGCATTACGATGACCGCATCGCGTTCGATAACCTGTGGAA
CTGGACGCAAAGCCACCTGCGGAACACGACCAGCGGCTTATTCTACTGGCGTTACG
ATCCGTCGGCGGCCAATCCGGTGGTGGATAAGAACAACGCCTCGGATGGCGATGT
GCTGATTGCCTGGGCGCTGTTAAAAGCGGGTAATAAGTGGCAGGACAACCGTTACCT
GCAGGCGTCGGACAGCATCCAGAAAGCGATCATCGCCAGCAATATCATTCAGTTTGC
GGGCCGCACCGTGATGTTACCGGGCGCCTATGGTTTCAACAAGAACAGCTATGTGAT
CCTGAACCCGTCGTATTTCCTGTTCCCGGCCTGGCGCGACTTTGCAAACCGCAGCCA
TCTGCAGGTGTGGCGTCAACTGATTGACGACAGCCTGTCATTAGTCGGTGAAATGCGTT
TTGGTCAAGTCGGTCTGCCGACGGATTGGGCGGCGCTGAATGCGGATGGTTCAATGG
CGCCGGCGACGGCCTGGCCGTCACGTTTTAGCTACGACGCCATTCGTATCCCGCTGT
ATTTATACTGGTATGACGCCAAAACCACGGCGCTGGTGCCGTTCCAGCTGTACTGGC
GTAACTATCCGCGCCTGACGACGCCGGCCTGGGTTGATGTGCTGAGCAGTAACACC
GCGACTTACAATATGCAGGGCGGTTTACTGGCGGTGCGCGACCTGACGATGGGCAA
CCTGGACGGTCTGAGCGATCTGCCGGGCGCATCGGAAGATTACTACTCGTCGAGCC
TGCGCCTGCTGGTGATGTTAGCGCGCGGTAAA

Linking region (downstream of *celY* in *Er. chrysanthemi*)
TAACCTTATTCTTGCGGTACACATGGCGAGGACGGTGTAGGCTGGAGCTGCTTC

kanR gene (reverse complement)
GAAGTTCCTATACTTTCTAGAGAATAGGAACTTCGGAATAGGAACTTCAAGATCCCCTTA
TTAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTGCGAATCGGGAGCG
GCGATACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCAAGCTCTTCAG
CAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGCCACACCCAGCCG
GCCACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCA
GGCATCGCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAG
CCTGGCGAACAGTTCGGCTGGCGCGAGCCCCTGATGCTCTTCGTCCAGATCATCCTG
ATCGACAAGACCGGCTTCCATCCGAGTACGTGCTCGCTCGATGCGATGTTTCGCTTGG
TGGTCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGCATCAG
CCATGATGGATACTTTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCC
GGCACTTCGCCCAATAGCAGCCAGTCCCTTCCCGCTTCAGTGACAACGTCGAGCAC

AGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGATAGCCGCGCTGCCTCGTC
CTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGC
CCCTGCGCTGACAGCCGGAACACGGCGGCATCAGAGCAGCCGATTGTCTGTTGTG
CCCAGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAA
TCCATCTTGTTCAATCATGCGAAACGATCCTCATCCTGTCTCTTGATCAGATCTTGATCCC
CTGCGCCATCAGATCCTTGGCGGCAAGAAAGCCATCCAGTTTACTTTGCAGGGCTTCC
CAACCTTACCAGAGGGCGCCCCAGCTGGCAATTCCGGTTCGCTTGCTGTCCATAAAA
CCGCCCAGTCTAGCTATCGCCATGTAAGCCCACTGCAAGCTACCTGCTTTCTCTTTGCG
CTTGCGTTTTCCCTTGTCCAGATAGCCCAGTAGCTGACATTCATCCGGGGTCAGCACC
GTTTCTGCGGACTGGCTTTCTACGTGTTCCGCTTCCTTTAGCAGCCCTTGCGCCCTGAGT
GCTTGCGGCAGCGTGAGCTTCAAAAGCGCTCTGAAGTTCCTATACTTTCTAGAGAATAG
GAACTTCGAACTGCAGGTCGACGGATCCCCGGAAT

3' pLOI1620 homology arm
GGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCA
TAGGGATAACAGGGTAAT

5.7 References

1. **Arjan, J. A., M. Visser, C. W. Zeyl, P. J. Gerrish, J. L. Blanchard, and R. E. Lenski.** 1999. Diminishing returns from mutation supply rate in asexual populations. Science **283:**404-6.

2. **Aury, J. M., C. Cruaud, V. Barbe, O. Rogier, S. Mangenot, G. Samson, J. Poulain, V. Anthouard, C. Scarpelli, F. Artiguenave, and P. Wincker.** 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics **9:**603.

3. **Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, E. C. M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen,**

K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456:**53-9.

4. **Boyer, M. H., B. Cami, J. P. Chambost, M. Magnan, and J. Cattaneo.** 1987. Characterization of a new endoglucanase from Erwinia chrysanthemi. Eur J Biochem **162:**311-6.

5. **Carder, J. H.** 1986. Detection and quantitation of cellulase by Congo red staining of substrates in a cup-plate diffusion assay. Anal Biochem **153:**75-9.

6. **Chapon, V., M. Czjzek, M. El Hassouni, B. Py, M. Juy, and F. Barras.** 2001. Type II protein secretion in gram-negative pathogenic bacteria: the study of the structure/secretion relationships of the cellulase Cel5 (formerly EGZ) from Erwinia chrysanthemi. J Mol Biol **310:**1055-66.

7. **Clune, J., D. Misevic, C. Ofria, R. E. Lenski, S. F. Elena, and R. Sanjuan.** 2008. Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. PLoS Comput Biol **4:**e1000187.

8. **Conrad, T. M., A. R. Joyce, M. K. Applebee, C. L. Barrett, B. Xie, Y. Gao, and B. O. Palsson.** 2009. Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. Genome Biol **10:**R118.

9. **Cooper, T. F., D. E. Rozen, and R. E. Lenski.** 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proc Natl Acad Sci U S A **100:**1072-7.

10. **Crozat, E., C. Winkworth, J. Gaffe, P. F. Hallin, M. A. Riley, R. E. Lenski, and D. Schneider.** 2010. Parallel Genetic and Phenotypic Evolution of DNA Superhelicity in Experimental Populations of Escherichia coli. Mol Biol Evol.

11. **Czaran, T., and R. F. Hoekstra.** 2009. Microbial communication, cooperation and cheating: quorum sensing drives the evolution of cooperation in bacteria. PLoS One **4:**e6655.

12. **Datsenko, K. A., and B. L. Wanner.** 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A **97:**6640-5.

13. **Dorman, C. J.** 2007. H-NS, the genome sentinel. Nat Rev Microbiol **5:**157-61.

14. **Dykhuizen, D. E., and A. M. Dean.** 2004. Evolution of specialists in an experimental microcosm. Genetics **167:**2015-26.

15. **Fong, S. S., A. P. Burgard, C. D. Herring, E. M. Knight, F. R. Blattner, C. D. Maranas, and B. O. Palsson.** 2005. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. Biotechnology and Bioengineering **91:**643-8.

16. **Fong, S. S., A. R. Joyce, and B. O. Palsson.** 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res **15:**1365-72.

17. **Fong, S. S., J. Y. Marciniak, and B. O. Palsson.** 2003. Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. J Bacteriol **185:**6400-8.

18. **Fong, S. S., A. Nanchen, B. O. Palsson, and U. Sauer.** 2006. Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. J Biol Chem **281:**8024-33.

19. **Fong, S. S., and B. O. Palsson.** 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet **36:**1056-8.

20. **Francetic, O., C. Badaut, S. Rimsky, and A. P. Pugsley.** 2000. The ChiA (YheB) protein of Escherichia coli K-12 is an endochitinase whose gene is negatively controlled by the nucleoid-structuring protein H-NS. Mol Microbiol **35:**1506-17.

21. **Francetic, O., D. Belin, C. Badaut, and A. P. Pugsley.** 2000. Expression of the endogenous type II secretion pathway in Escherichia coli leads to chitinase secretion. Embo J **19:**6697-703.

22. **Francetic, O., and A. P. Pugsley.** 2005. Towards the identification of type II secretion signals in a nonacylated variant of pullulanase from Klebsiella oxytoca. J Bacteriol **187:**7045-55.

23. **Guiseppi, A., J. L. Aymeric, B. Cami, F. Barras, and N. Creuzet.** 1991. Sequence analysis of the cellulase-encoding celY gene of Erwinia chrysanthemi: a possible case of interspecies gene transfer. Gene **106:**109-14.

24. **Hall, B. G.** 1988. Adaptive evolution that requires multiple spontaneous mutations. I. Mutations involving an insertion sequence. Genetics **120:**887-97.

25. **He, S. Y., M. Lindeberg, A. K. Chatterjee, and A. Collmer.** 1991. Cloned Erwinia chrysanthemi out genes enable Escherichia coli to selectively secrete a diverse family of heterologous proteins to its milieu. Proc Natl Acad Sci U S A **88:**1079-83.

26. **Herring, C. D., J. D. Glasner, and F. R. Blattner.** 2003. Gene replacement without selection: regulated suppression of amber mutations in Escherichia coli. Gene **311:**153-63.

27. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

28. **Hince, T. A., and S. Neale.** 1977. Physiological modification of alkylating agent induced mutagenesis. II. Influence of the numbers of chromosome replicating forks and gene copies on the frequency of mutations induced in Escherichia coli. Mutat Res **43:**11-24.

29. **Hua, Q., A. R. Joyce, B. O. Palsson, and S. S. Fong.** 2007. Metabolic characterization of Escherichia coli strains adapted to growth on lactate. Appl Environ Microbiol **73:**4639-47.

30. **Hubbarde, J. E., and L. M. Wahl.** 2008. Estimating the optimal bottleneck ratio for experimental evolution: the burst-death model. Math Biosci **213:**113-8.

31. **Ibarra, R. U., J. S. Edwards, and B. O. Palsson.** 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420:**186-9.

32. **Jimenez-Sanchez, A., and E. Cerda-Olmedo.** 1975. Mutation and DNA replication in Escherichia coli treated with low concentrations of N-methyl-N'-nitro-N-nitrosoguanidine. Mutat Res **28:**337-45.

33. **Kachroo, A. H., A. K. Kancherla, N. S. Singh, U. Varshney, and S. Mahadevan.** 2007. Mutations that alter the regulation of the chb operon of Escherichia coli allow utilization of cellobiose. Mol Microbiol **66:**1382-95.

34. **Keyhani, N. O., and S. Roseman.** 1997. Wild-type Escherichia coli grows on the chitin disaccharide, N,N'-diacetylchitobiose, by expressing the cel operon. Proc Natl Acad Sci U S A **94:**14367-71.

35. **Khianngam, S., S. Tanasupawat, A. Akaracharanya, K. K. Kim, K. C. Lee, and J. S. Lee.** Paenibacillus xylanisolvens sp. nov., a xylan-degrading bacterium from Thai soil. Int J Syst Evol Microbiol.

36. **Kim, B. C., K. H. Lee, M. N. Kim, E. M. Kim, S. R. Min, H. S. Kim, and K. S. Shin.** 2009. Paenibacillus pini sp. nov., a cellulolytic bacterium isolated from the rhizosphere of pine tree. J Microbiol **47:**699-704.

37. **Kim, K. K., K. C. Lee, H. Yu, S. Ryoo, Y. Park, and J. S. Lee.** 2009. Paenibacillus sputi sp. nov., isolated from the sputum of a patient with pulmonary disease. Int J Syst Evol Microbiol.

38. **King, T., S. Seeto, and T. Ferenci.** 2006. Genotype-by-environment interactions influencing the emergence of rpoS mutations in Escherichia coli populations. Genetics **172:**2071-9.

39. **Lindeberg, M., and A. Collmer.** 1992. Analysis of eight out genes in a cluster required for pectic enzyme secretion by Erwinia chrysanthemi: sequence comparison with secretion genes from other gram-negative bacteria. J Bacteriol **174:**7385-97.

40. **Lynd, L. R., P. J. Weimer, W. H. van Zyl, and I. S. Pretorius.** 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol Mol Biol Rev **66:**506-77, table of contents.

41. **Mahadevan, R., and C. H. Schilling.** 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng **5:**264-76.

42. **Maharjan, R., S. Seeto, L. Notley-McRobb, and T. Ferenci.** 2006. Clonal adaptive radiation in a constant environment. Science **313:**514-7.

43. **Martinez, J., V. Simon, B. Gonzalez, and P. Conget.** A real-time PCR-based strategy for the detection of Paenibacillus larvae vegetative cells and spores to improve the diagnosis and the screening of American foulbrood. Lett Appl Microbiol.

44. **McKenzie, G. J., and S. M. Rosenberg.** 2001. Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens. Curr Opin Microbiol **4:**586-94.

45. **Nakamura, Y., T. Gojobori, and T. Ikemura.** 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res **28:**292.

46. **Orr, H. A.** 2005. Theories of adaptation: what they do and don't say. Genetica **123:**3-13.

47. **Park, Y. W., and H. D. Yun.** 1999. Cloning of the Escherichia coli endo-1,4-D-glucanase gene and identification of its product. Mol Gen Genet **261:**236-41.

48. **Parker, L. L., and B. G. Hall.** 1990. Characterization and nucleotide sequence of the cryptic cel operon of Escherichia coli K12. Genetics **124:**455-71.

49. **Parker, L. L., and B. G. Hall.** 1990. Mechanisms of activation of the cryptic cel operon of Escherichia coli K12. Genetics **124:**473-82.

50. **Plumbridge, J., and O. Pellegrini.** 2004. Expression of the chitobiose operon of Escherichia coli is regulated by three transcription factors: NagC, ChbR and CAP. Mol Microbiol **52:**437-49.

51. **Poteete, A. R.** 2001. What makes the bacteriophage lambda Red system useful for genetic engineering: molecular mechanism and biological function. FEMS Microbiol Lett **201:**9-14.

52. **Rivero-Muller, A., S. Lajic, and I. Huhtaniemi.** 2007. Assisted large fragment insertion by Red/ET-recombination (ALFIRE)--an alternative and enhanced method for large fragment recombineering. Nucleic Acids Res **35:**e78.

53. **Russel, M.** 1998. Macromolecular assembly and secretion across the bacterial cell envelope: type II protein secretion systems. J Mol Biol **279:**485-99.

54. **Sauer, U., D. R. Lasko, J. Fiaux, M. Hochuli, R. Glaser, T. Szyperski, K. Wuthrich, and J. E. Bailey.** 1999. Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. J Bacteriol **181:**6679-88.

55. **Schuetz, R., L. Kuepfer, and U. Sauer.** 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol **3:**119.

56. **Shen, Y., S. Sarin, Y. Liu, O. Hobert, and I. Pe'er.** 2008. Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. PLoS One **3:**e4012.

57. **Sniegowski, P. D., P. J. Gerrish, and R. E. Lenski.** 1997. Evolution of high mutation rates in experimental populations of E. coli. Nature **387:**703-5.

58. **Thompson, J., S. B. Ruvinov, D. I. Freedberg, and B. G. Hall.** 1999. Cellobiose-6-phosphate hydrolase (CelF) of Escherichia coli: characterization and assignment to the unusual family 4 of glycosylhydrolases. J Bacteriol **181:**7339-45.

59. **Travisano, M., and R. E. Lenski.** 1996. Long-term experimental evolution in Escherichia coli. IV. Targets of selection and the specificity of adaptation. Genetics **143:**15-26.

60. **Treves, D. S., S. Manning, and J. Adams.** 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. Mol Biol Evol **15:**789-97.

61. **Varrot, A., V. L. Y. Yip, Y. Li, S. S. Rajan, X. Yang, W. F. Anderson, J. Thompson, S. G. Withers, and G. J. Davies.** 2005. NAD+ and Metal-ion Dependent Hydrolysis by Family 4 Glycosidases: Structural Insight into Specificity for Phospho-[beta]-d-glucosides. Journal of Molecular Biology **346:**423-435.

62. **Wahl, L. M., P. J. Gerrish, and I. Saika-Voivod.** 2002. Evaluating the impact of population bottlenecks in experimental evolution. Genetics **162:**961-71.

63. **Wood, T. M.** 1988. Preparation of crystalline, amorphous, and dyed cellulase substrates. Methods Enzymol. **160:**19-25.

64. **Wood, T. M., and K. M. Bhat.** 1988. Methods of measuring cellulase activities. Methods Enzymol. **160:**87-112.

65. **Yoon, S. H., E. G. Lee, A. Das, S. H. Lee, C. Li, H. K. Ryu, M. S. Choi, W. T. Seo, and S. W. Kim.** 2007. Enhanced vanillin production from recombinant E. coli using NTG mutagenesis and adsorbent resin. Biotechnol Prog **23:**1143-8.

66. **Zhou, S., F. C. Davis, and L. O. Ingram.** 2001. Gene integration and expression and extracellular secretion of Erwinia chrysanthemi endoglucanase CelY (celY) and CelZ (celZ) in ethanologenic Klebsiella oxytoca P2. Appl Environ Microbiol **67:**6-14.

67. **Zhou, S., and L. O. Ingram.** 2000. Synergistic hydrolysis of carboxymethyl cellulose and acid-swollen cellulose by two endoglucanases (CelZ and CelY) from Erwinia chrysanthemi. J Bacteriol **182:**5676-82.

68. **Zhou, S., L. P. Yomano, A. Z. Saleh, F. C. Davis, H. C. Aldrich, and L. O. Ingram.** 1999. Enhancement of expression and apparent secretion of Erwinia chrysanthemi endoglucanase (encoded by celZ) in Escherichia coli B. Appl Environ Microbiol **65:**2439-45.