UNIVERSITY OF CALIFORNIA, SAN DIEGO

Automation, Optimization, and Characterization of Adaptive Laboratory Evolution

A dissertation submitted in satisfaction of the

requirements for the degree of Doctor of Philosophy

in

Bioengineering

by

Ryan Alan LaCroix

Committee in charge:

Professor Bernhard Palsson, Chair
Professor Xiaohua Huang
Professor Scott Rifkin
Professor Milton Saier
Professor Kun Zhang

2016

The dissertation of Ryan Alan LaCroix is approved,

and it is acceptable in quality and form for publication on

microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2016

Dedication

I dedicate this work to the my family and friends that have supported me in during graduate school.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Professor Bernhard Palsson for his support as chair of my committee and advisor. He has created an environment that breeds success and welcomed me to be a part of it.

I would also like to acknowledge Dr. Adam Feist for his help, support, and mentoring. His input was integral to the success of my dissertation.

Chapter III, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in *Applied and Environmental Microbiology*. (LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM. 2015. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol 81:17-30.)

Chapter IV, in full, is a reprint that the dissertation author was the principal researcher and author of. The material has been submitted to *Applied and Environmental Microbiology*. (LaCroix, RA, Palsson, BO, Feist AM. 2016. Designing Adaptive Laboratory Evolution Experiments).

Vita

2010 Bachelors of Science, Bioengineering, University of California, Riverside

2016 Doctorate of Philosophy, Bioengineering, University of California, San Diego

PUBLICATIONS

LaCroix, Ryan A., et al. "The Dynamics and Design of Adaptive Laboratory Evolution Experiments" – submitted *2016*

LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM: Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. *Appl Environ Microbiol* 2015, 81(1):17-30

Sandberg TE, Pedersen M, **LaCroix RA**, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM: Evolution of Escherichia coli to 42 degrees C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. *Mol Biol Evol* 2014, 31(10):2647-2662.

FIELDS OF STUDY

Major Field: Bioengineering

Studies in Adaptive Laboratory Evolution and Laboratory Automation

Professor Bernhard Palsson

Abstract of the Dissertation

Automation, Optimization, and Characterization of Adaptive Laboratory Evolution

by

Ryan Alan LaCroix

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2016

Professor Bernhard Palsson, Chair

Adaptive laboratory evolution (ALE) has emerged as an effective tool for scientific discovery and addressing biotechnological needs. A typical ALE experiment requires significant attention over the course of the experiment and can last up to months. When designing such experiments much consideration is given to the logistics

of maintaining the experiment. Due to the difficult logistics, the consistency and throughput are often reduced. Overcoming these shortcoming is now possible with an automated platform. The automated platform was designed and built giving consideration to alleviating the design constraint that the time-sensitive processes impose on the experiment.

Much of ALE's utility is derived from reproducibly obtained fitness increases. Identifying causal genetic changes and their combinatorial effects is challenging and time-consuming. Understanding how these genetic changes enable increased fitness can be difficult. A series of approaches that address these challenges was developed and demonstrated using Escherichia coli K-12 MG1655 on glucose minimal media at 37ºC. By keeping E. coli in constant substrate-excess and exponential growth, fitness increases up to 1.6-fold were obtained over wild-type. These increases are comparable to previously-reported maximum growth rates in similar conditions but obtained over a shorter time frame. Across the 8 replicate ALE experiments performed, causal mutations were identified using three approaches: identifying mutations in the same gene/region across replicate experiments, sequencing strains before and after computationally-determined fitness jumps, and allelic replacement coupled with targeted ALE of reconstructed strains. Three genetic regions were most often mutated: the global transcription gene *rpoB*, an 82bp deletion between the metabolic *pyrE* gene and *rph*, and an IS element between the DNA structural gene *hns* and *tdk*. Model-derived classification of gene expression revealed a number of processes important for increased growth that were missed using a gene classification system alone. The

methods put forth here represent a powerful combination of technologies to increase the speed and efficiency of ALE studies. The identified mutations can be examined as genetic parts for increasing growth rate in a desired strain and for understanding rapid growth phenotypes.

The evolution process is increasingly being leveraged in laboratory settings for industrial and basic science applications. Despite an increasing deployment, there are no standardized procedures available for designing and performing adaptive laboratory evolution (ALE) experiments. Thus, there is a need to optimize the experimental design, specifically for determining termination criteria and for balancing outcomes with available resources (i.e., lab supplies, personnel, and time). To design and better understand ALE experiments, a simulator, ALEsim, was developed, validated, and applied to optimize ALE experimentation. The effects of various passage sizes were experimentally determined and subsequently evaluated with ALEsim to explain differences in experimental outcomes. Further, a beneficial mutation rate of $10^{-6.9}$-$10^{-8.4}$ mutations per cell division was derived. A retrospective analysis of ALE experiments revealed that passage sizes typically employed in batch culture ALE experiments led to inefficient production and fixation of beneficial mutations. ALEsim and the results herein will aid in the design of ALE experiments to fit the exact needs of the project while taking into account the tradeoff in resources required, and lower the barrier of entry to this experimental technique.

With successful completion of an automated ALE platform and multiple applied cases, there became a need to expand the ALE protocol for variations of ALE.

Specifically to accommodate adaptation to environments that cannot initially sustain growth. Two algorithms were developed to implement two variations of ALE, pathway activation of latent enzymes (PALE ALE) and tolerization (TALE). The purpose of the PALE ALE protocol was to adapt and organism to growth using a substrate is it natively is unable to utilize. An algorithm was developed to put significant selection pressure on the population to adapt all while maximizing the amount of genetic diversity being created. The purpose of the TALE module is to adapt an organism to an increasing amount of stress (e.g. temperature, physical, chemical, etc…). The algorithm specifically targeted putting enough stress on the culture as reasonable but also ensuring that the culture is still able to grow. This is critical since if growth is arrested the genetic diversity in the culture drops off significantly. These two algorithms were successfully implemented into the automated ALE platform.

Chapter I - Introduction

1.1 Introduction

Natural selection is a powerful phenomenon that has shaped the entirely of life as we know it. It has shown the ability to entice a plethora of attributes out of an organisms building blocks—more specifically, the genome. Due its powerful nature, scientists began harnessing it in laboratories for scientific and industrial applications. This is often referred to as adaptive laboratory evolution (ALE). ALE has been used throughout the 20th century but a recent surge in popularity has occurred over the past decade. This surge in popularity is often attributed to advancements in technology associated with analyzing ALE experiments—specifically next-generation sequencing (NGS) technologies. Examining the genetic changes responsible for the adaptation is of great interest to the scientific community. As the analysis of these experiments has grown, the methodologies and design considerations in performing these experiments has not.

The basics of an ALE experiment are straightforward and well understood. Simply grow an organism in a given environment for multiple generations (e.g. >100 generations). Because of the number of generations needed, microorganisms make up the majority of applications. As the organism grows, mutations occur in the genome due to DNA replication errors. Though most mutations are thought to have negative effects on the organism, there is an off-chance that one of these mutations will confer a benefit over the rest of the population. Over a long enough time the lineage with this mutation will fix itself and become the dominant strain.

There are many ways to perform an experiment but the most popular methods are in continuous culture and batch culture. Continuous culture simply involves setting

up a chemostat-like device to constantly feed the culture with fresh growth media. Batch culture involved growing a culture in a batch and then at a certain time, taking an aliquot and passing it to a new batch with fresh media. Continuous culture requires a significant amount of upfront resources and planning whereas most microbiology laboratories have all the necessary equipment to perform a batch culture ALE experiment. As such, batch culture ALE experiments are more ubiquitous.

Getting a batch culture ALE experiment started fairly simple but there are experimental parameters that can drastically affect the outcome. The most apparent difference between experiments is the phase of growth the culture is passed from (e.g. exponential growth phase or stationary phase). When passed during stationary phase, the culture is subjected to an environment of alternating feast and famine, where exponential phase growth is followed by stationary phase. The issue arises when analyzing the results in terms of the selection pressure. In this case it becomes quite complex where the cells can increase fitness by affecting survivability in any stages of the growth curve as well as the transitions (1). Thus the resulting analysis becomes convoluted. Alternatively the cultures can be passed during exponential phase where nutrients are still in excess. This minimizes the any selection to alternating environment as the environment remains more consistent. Ultimately this allows a cleaner and more reproducible end result (2-4).

Passing cultures in exponential phase requires more time and resources than passing during stationary phase. When passing in stationary phase, it is common to pass the culture every 24 hours. Logistically, this is reasonable as the cultures can be passed during normal working hours and furthermore, a large number of cells can be

passed. This is important as a small passage size can reduce the genetic variation (chapter 4). Consequently, with exponential phase passaging, the logistic are much more complex. Passing in exponential phase requires strict timing and knowledge of the growth rate. Knowing the growth rate and cell density, a calculated number of cells can be passed to the next batch culture such that it will be in mid-exponential phase in 24 hours. The first issue is that unless biomass measurements are taken over the course of the culture, which they rarely are, calculating the growth rate becomes imprecise. This leads to incorrect timing and in 24 hours the culture can be not ready or has reached stationary phase. Additionally, the culture will begin to grow faster as evolution takes place. When this occurs, the number of cells propagated to the subsequent batch culture is decreased. This decrease often gets so low that the volume passage becomes sub micro-liter. At these volumes, the genetic diversity is severely bottlenecked and evolution is halted prematurely (5-7) (chapter 4).

Solving the problems with exponential phase passaging ALEs is not a scientific limitation but a technological one. If the process of ALE could be automated the pressures of when and how much to passage could be alleviated. Not only will this allow these experiments to be performed without lowering the passage size but it will also expand the parameter space in designing ALE experiments. The expansion of the parameter space can then open up new capabilities where selection pressure can be applied at unprecedented ranges.

1.2 References

1. **Vasi F, Travisano M, Lenski RE.** 1994. Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. American Naturalist**:**432-456.

2. **LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM.** 2015. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol **81:**17-30.

3. **Sandberg TE, Long CP, Gonzalez JE, Feist AM, Antoniewicz MR, Palsson BO.** 2016. Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology. PLoS One **11:**e0151130.

4. **Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM.** 2014. Evolution of Escherichia coli to 42 degrees C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Mol Biol Evol **31:**2647-2662.

5. **Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BO.** 2010. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet **6:**e1001186.

6. **Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

7. **Ibarra RU, Edwards JS, Palsson BO.** 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420:**186-189.

Chapter II - Adaptive Laboratory Evolution Automation

2.1 Introduction

Adaptive laboratory evolution (ALE) experiments has a wide range of scientific and industrial applications(1-11) . The popularity of these experiments is growing as is the demand for them. Though fruitful in their outcomes, performing these experiments requires a considerable amount of time, resources, and dedication. ALE experiments often go in excess of months and require constant monitoring. Because constant monitoring is required, the experimental parameters are often chosen to allow for about 24 hours of walk-away time. (e.g. passage OD, passage volume, and culture volume). This 24 hour period can be attained by passing a small number of cells or letting the culture reach stationary phase. Though experiments performed in such a fashion have proven to be fruitful, there is a large experimental space that has yet to be explored as the manual process by which ALE experiments are currently performed prohibit their feasibility.

The most limiting experimental parameter in a manual ALE experiment is the passage OD. Passing at a given OD it determines the selection pressure as well as the amount of genetic diversity. Ideally, the higher the passage OD the greater cell count and thus greater genetic diversity exists. With more genetic diversity beneficial mutation are more likely to occur at a faster rate. Passing at a low OD can attenuate the rate of adaptation (12, 13). However there is an upper limit on the passage OD as cultures will eventually reach stationary phase. If the culture is allowed to reach stationary phase then the cells have a secondary method of adaptation where they can outcompete during stationary phase, or even in lag phase after the culture is passed

(11, 14, 15). Depending on the desired outcome, this can be non-ideal as analyzing the end result of the experiment becomes convoluted. Ultimately what is desired is a passage OD that is as high as possible without a transition to stationary phase beginning. This window of time with which to pass the culture can often be small. Being able to predict this window of time is contingent on two properties, the current OD and the growth rate. Without knowing the values of these properties accurately it is easy to miss such a window. In a typical ALE experiment the researcher is typically only present at inoculation and at the passage time. An OD measurement can theoretically be taken at inoculation but these values are often below the detection limit of spectrophotometers or on par with noise in the measurement. Exponentially extrapolating from this value also exponentially increases the noise and the resulting prediction is often wrong. This furthermore requires an accurate growth rate. Determining the growth rate requires additional OD measurements throughout the lifetime of the culture. This would require the researcher to be present through the lifetime of the culture which is often not possible. This is further confounded by replicate experiments where variations between cultures require that the passage OD for each culture is reached at different times. Thus ensuring the culture is passed at the desired OD require near constant attention, including throughout the night.

The constant attention needed to pass the culture at the desired OD can be alleviated by controlling the passage size but changes in the passage size can have further effects on the results. Since consistent surveillance of the culture is basically infeasible in a standard laboratory environment, a common technique to make sure the culture does not hit stationary phase throughout the night is to pass a small amount

such that even with the errors in starting OD and estimated growth rate, it is safe to conclude that the passage OD will not be reached until working hours the next day. This can be effective for catching the passage OD but passing a small amount limits the genetic variation in the culture thereby limiting the number of mutations being selected for. As the culture evolves and the growth rate increases, the passage size must be further reduced to compensate. Ultimately the passage size is reduced to a point where the chances of any beneficial mutation getting captured or next to nil (16, 17). When ALE experiments stop increasing in growth rate the experiment is often terminated as a reasonable endpoint is presumed to have been reached. The problem is that with a reduced passage size there may be myriad beneficial mutations available they are simply thrown away due to the small passage size and a leveling off of the growth rate is simply an artifact of the experimental procedure.

A further confounding factor when varying the passage size or passage OD to properly time an experiment is that it differentiates replicate experiments. As is typical in experimentation, ALE experiments are often run with biological replicates. This provides points of comparison and contrast as well as statistical power in downstream analysis when making conclusions (7). It becomes increasingly important to have biological replicates when trying to identify causality among mutated genes. When the parameters of ALE experiments are varied between themselves for the sake of performing the experiment, the replicate nature of the experiments can be lost wherein one culture has been able to sample and select for more mutations than another.

Ultimately the issue with performing manual ALE experiments is that the space of feasible design parameters (e.g. passage OD and passage size) is significantly

limited. Accordingly, this presents a limitation of the field as a whole where the experiments that are designed to find optimal states are themselves inhibiting the finding of these optimal states. With all the technological advancements in automation it has become quite feasible to automate the ALE process. Automation would not only alleviate the self-imposed limitation but allow researchers to explore a design space that was previously unreached for new scientific findings.

2.2 The Automation Platform

As with any automation platform, there are tradeoffs to consider when automating ALE. There are many processes required for completing a long term evolution experiment that must all be coordinated. These processes range from media preparation, to culture passaging, and even dish washing. Of the processes involved in ALE not all of them would add benefit if automated. Primarily those processes that are time sensitive are candidates for automation. For example, media preparation, though integral to completing an experiment, does not have a stringent time constraint. Excess media can easily be prepped in advance and stored until needed. Automating this process would be mostly a convenience to the researcher but ultimately would not lead to a more productive experiment. On the other hand, culture passaging has a strict time constraint and as such would add significant benefit if automated. Whether late or premature, it can affect the overall result of the experiment. Thus culture passaging is a good candidate for automation. Considering all processes the primary focus of automation was on processes that are time sensitive. The two main time sensitive process are culture measurement and passaging.

A liquid handling platform was used to create an automated platform (ALE machine) that would automate OD measurement and passaging. With measurement and passaging automated it is possible to accurately measure the growth rate for each batch culture. Accurately measuring growth rates and tracking the OD allows for passing the cultures during a well-defined yet narrow window of time where the OD is high for increased genetic diversity yet low enough that the transition to stationary phase has not begun.

A software platform was developed to control the ALE machine. Key design features of the ALE machine include: robust control algorithms, expandable algorithms, sample tracking, and cloud enabled. A robust control algorithm was designed to accomplish batch culture ALE experiments. As with any automated platform obtaining primary functionality is straightforward. Under most circumstances accurate measurements are taken and processes are performed at their intended times. However, measurements taken are subject to noise and artifacts. The software makes use of measurement data to make predictions about when to measure and pass the cultures. Any inaccuracies in measurements could lead to early or delayed measurements or passaging. Either are undesired as it can lead to inaccurate growth rate calculations or missing the window of opportunity to pass to culture. As such the algorithm was developed to identify such errors in measurement. The breadth of errors sufficiently managed by the algorithm are: inaccurate pipetting into the detector, random noise in pipetting and detector measurements, and poor timing due to backlog.

Beyond a robust control algorithm, the software was developed to allow expansion of the algorithms for variations of ALE experiments. As the ALE field

grows, new ideas and experimental designs are envisioned and tested. The software platform allows a developer to use a documented application programming interface (API) to create new experimental designs using the basic building blocks already in place. This lets the developer focus on the experimental design instead of how best to implement it in the context of the software platform.

The software platform further tracks all samples, tubes, and cultures through the entire process. It is common practice to take samples throughout an experiment for additional tests (e.g. HPLC, sequencing, frozen stocks, etc…). The software is able to create short identifiable tags to be used on the samples that would allow all data and meta-data to be looked up at a later date. This creates a straightforward working environment for the user where mistakes in labeling and identification are minimized.

Since the runtime of ALE experiments can often be on the order of months, it is important that all interested parties are able to watch and monitor the experiment. As such, the ALE machine is completely cloud enabled. All data generated as well as analysis of such data is available on a responsive website accessible from any modern computer, tablet, or phone. Based on the data generated parameters can be changed and designs updated as the cultures evolve in often unexpected ways.

Overall the ALE machine platform was designed to eliminate many of the constraints that manual ALE experiments are subjected. In doing so it allows experiments that were previously impractical to be performed allowing for further understanding and application of adaptive evolution. The ALE machine was developed to scale alongside of the growing field, not only in number but with new

algorithms and experimental designs. Ultimately, this is a tool to push to frontier of evolutionary science.

2.3 References

1. **Bachmann H, Fischlechner M, Rabbers I, Barfa N, Branco dos Santos F, Molenaar D, Teusink B.** 2013. Availability of public goods shapes the evolution of competing metabolic strategies. Proc Natl Acad Sci U S A **110:**14302-14307.

2. **Bacun-Druzina V, Cagalj Z, Gjuracic K.** 2007. The growth advantage in stationary-phase (GASP) phenomenon in mixed cultures of enterobacteria. FEMS Microbiol Lett **266:**119-127.

3. **Conrad TM, Lewis NE, Palsson BO.** 2011. Microbial laboratory evolution in the era of genome-scale science. Mol Syst Biol **7:**509.

4. **Dragosits M, Mattanovich D.** 2013. Adaptive laboratory evolution -- principles and applications for biotechnology. Microb Cell Fact **12:**64.

5. **Gresham D, Hong J.** 2015. The functional basis of adaptive evolution in chemostats. FEMS Microbiol Rev **39:**2-16.

6. **Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ.** 2013. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. PLoS Comput Biol **9:**e1003091.

7. **LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM.** 2015. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol **81:**17-30.

8. **Palsson B.** 2010. Adaptive laboratory evolution. Microbe.

9. **Sandberg TE, Long CP, Gonzalez JE, Feist AM, Antoniewicz MR, Palsson BO.** 2016. Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology. PLoS One **11:**e0151130.

10. **Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM.** 2014. Evolution of Escherichia coli to 42 degrees C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Mol Biol Evol **31:**2647-2662.

11. **Vasi F, Travisano M, Lenski RE.** 1994. Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. American Naturalist**:**432-456.

12. **Wahl LM, Gerrish PJ.** 2001. The probability that beneficial mutations are lost in populations with periodic bottlenecks. Evolution **55:**2606-2610.

13. **Wahl LM, Zhu AD.** 2015. Survival probability of beneficial mutations in bacterial batch culture. Genetics **200:**309-320.

14. **Blom EJ, Ridder AN, Lulko AT, Roerdink JB, Kuipers OP.** 2011. Time-resolved transcriptomics and bioinformatic analyses reveal intrinsic stress responses during batch culture of Bacillus subtilis. PLoS One **6:**e27160.

15. **Clark ME, He Q, He Z, Huang KH, Alm EJ, Wan XF, Hazen TC, Arkin AP, Wall JD, Zhou JZ, Fields MW.** 2006. Temporal transcriptomic analysis as Desulfovibrio vulgaris Hildenborough transitions into stationary phase during electron donor depletion. Appl Environ Microbiol **72:**5578-5588.

16. **Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BO.** 2010. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet **6:**e1001186.

17. **Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

Chapter III - Automation Test Case

3.1 Introduction

Adaptive laboratory evolution (ALE) is a growing field facilitated by whole genome sequencing. The process of ALE involves the continuous culturing of an organism over multiple generations. During an ALE experiment, mutations arise and those beneficial to the selection pressure are fixed over time in the population. Most ALE experiments analyze a perturbation from a reference state to another (e.g., environmental (1, 2) or genetic (3)). After adaptation, understanding what genetic changes enabled an increase in fitness is often desirable (4). Generally there are two methods of evolving microorganisms – batch cultures and chemostats. Each method has its own advantages and disadvantages, in terms of maintenance, growth environment, and selection pressures (5). Applications of ALE are numerous and include those for biotechnological goals, such as improving tolerance to a given compound of interest (6-8), or more progressive uses such as improving electrical current consumption in an organism (9). Additionally, there has been a significant focus on using ALE to understand antibiotic resistance to given compounds (i.e., drugs) in order to combat clinical resistance (10). A number of in depth reviews on ALE have appeared as the field continues to grow (5, 11, 12).

The methodology utilized for conducting an ALE experiment needs to be carefully considered. A critical characteristic of ALE experiments is that they have long timescales, on the order of months, and often require daily attention (1, 5). The timescale is typically determined by culture size, amount of cells propagated to the next culture (i.e., passage size), and the growth phase under which it is passed. When passing strictly in exponential phase (3, 13-15), the timescale becomes restrictive as

there is only a small window of time in which to aliquot from the culture and propagate it. The amount passed significantly influences when the next window will occur. Thus, it is often the case that the passage size is adjusted according to the experimenter's schedule (3, 16). An unfortunate consequence of this is that as the growth rate increases, the passage size is generally decreased. This allows for fewer potentially beneficial mutations to advance to the next flask, possibly slowing evolution. An alternate approach is to pass a fixed amount at a regular time interval, generally once per day. This time frame allows the cells to reach stationary phase, where they remain for the majority of the time. This approach has been used in a notable study where *E. coli* B strains were evolved in glucose minimal media batch cultures for over 25 years (17). Passing cells after they have reached stationary phase creates a more complex selection pressure than strictly passing cells during exponential growth (18), favoring both growth rate increases and decreases in lag-phase duration (19). Thus, experimental setup should be tailored to the desired selection pressure of the experiment.

Next generation sequencing has eased the process of finding mutations in ALE studies, however tying specific components of the genotype to the phenotype remains difficult. Strains generated using ALE often have multiple mutations (20, 21) and if one wants to determine causality for a phenotype, it can require a significant effort (22-24). Despite the growing availability of genome engineering tools (22, 25, 26), determining causality is still a time consuming process. An alternative approach to speed in the discovery of causal mutations would be to perform multiple independent experiments and examine mutations that occur most frequently. Performing multiple

experiments under strict identical conditions can help filter casual mutation candidates encountered during ALE.

Along with understanding causal genetic changes in ALE experiments, there is also a need to understand changes at the cellular pathway level. Omics characterization coupled with systems modeling approaches enable the mechanistic interpretation of data based on reconstructed metabolic network content (27). Constraint-based modeling, which is a bottom up approach based on network interactions and overall physiochemical constraints, has been shown to be a valuable systematic approach for analyzing omics data (28, 29). This approach has largely been pioneered using *E. coli* K-12 MG1655 as the organism of choice for validation and comparison of *in silico* predictions to experimental data (30, 31). In short, integration of omics data types with genome-scale constraint-based models has provided a context in which such data can be integrated and interpreted.

In an effort to demonstrate the power of using strict selection pressure to understand the process of ALE, *E. coli* K-12 MG1655 was adaptively evolved in minimal media at 37ºC with excess glucose in eight parallel experiments. At the end of the ALE experiments, clones from the final populations were characterized in terms of their growth rate, metabolic uptake and secretion rates, genome sequence, and transcriptome. These multi-omics data types were then integrated and further categorized with genome-scale models to examine how the cells adapted to the conditions and how their physiology and genomes changed.

3.2 Materials and Methods

3.2.1 Adaptive Laboratory Evolution

Primary adaptive evolutions were started from wild type *E. coli* strain MG1655 (ATCC47076) frozen stock and grown up overnight in 500mL Erlenmeyer flask with 200mL of minimal media. 8 aliquots of 900µL were passed into eight flasks containing 25mL of media and magnetic stir discs for aeration. 800µL of culture was serially passed during mid-exponential phase (3.2% of the total culture volume was propagated to the next culture). Cultures were not allowed to reach stationary phase before passage. Four OD_{600nm} measurements were taken between ODs of 0.05 and 0.30 to determine growth rates. Periodically, aliquots of samples were frozen in 25% glycerol solution and stored at -80ºC for future analysis. Glucose M9 minimal media consisted of 4g/L Glucose, 0.1mM $CaCl_2$, 2.0mM $MgSO_4$, Trace element solution and M9 salts. 4000X Trace element solution consisted of 27g/L $FeCl_3*6H_2O$, 2g/L $ZnCl_2*4H_2O$, 2g/L $CoCl_2*6H_2O$, 2g/L $NaMoO_4*2H_2O$, 1g/L $CaCl_2*H_2O$, 1.3g/L $CuCl_2*6H_2O$, 0.5g/L H_3BO_3, and Concentrated HCl dissolved in ddH_2O and sterile filtered. 10x M9 Salts solution consisted of 68g/L Na_2HPO_4 anhydrous, 30g/L KH_2PO_4, 5g/L NaCl, and 10g/L NH_4Cl dissolved ddH_2O and autoclaved. Final concentrations in the media were 1x. The validation was performed under the same conditions as above except 0.7% of the culture was passed.

3.2.2 Physiological characterizations

Growth rates of clones isolated from the primary ALE experiments were screened by inoculating cells from an overnight culture to a low optical density (OD) and sampling the OD_{600nm} until stationary phase was reached. A linear regression of

the log-linear region was computed using ‘polyfit’ in MATLAB and the growth rate (slope) was determined. Growth rates of clones isolated from the follow-up validation ALE were similarly started but passed serially three times in late exponential phase. The growth rates of each culture were computed as above and the average of the three cultures was taken. The first culture was omitted due to physiological characterization (32).

Growth rates of populations were determined by the output of the interpolated cubic spline used, unless stated otherwise.

Extra-Cellular by-products were determined by HPLC. Cell cultures were first sampled and then sterile filtered. The filtrate was injected into an HPLC column (Aminex HPX-87H Column #125-0140). Concentrations of detected compounds were determined by comparison to a normalized curve of known concentrations.

Substrate uptake and secretion rates were calculated from the product of the growth rate and the slope from a linear regression of gDW vs substrate concentration.

Biomass Yield (Y_{X/S_ss}) was calculated as the quotient of the growth rate and glucose uptake rates during the exponential growth phase.

3.2.3 DNA Sequencing

Genomic DNA was isolated using Promega’s Wizard DNA Purification Kit. The quality of DNA was assessed with UV absorbance ratios using a Nano drop. DNA was quantified using Qubit dsDNA High Sensitivity assay. Paired-end resequencing libraries were generated using Illumina’s Nextera XT kit with 1 ng of input DNA total. Sequences were obtained using an Illumina Miseq with a PE500v2 kit. The breseq pipeline (33) version 0.23 with bowtie2 was used to map sequencing reads and

identify mutations relative to the E. Coli K12 MG1655 genome (NCBI accession NC_000913.2). These runs were performed on the National Energy Research Scientific Computing Center carver supercomputer. The identified mutations were then entered into an SQL database to track mutations along each evolution. All samples had an average mapped coverage of at least 25x.

3.2.4 RNA-Sequencing

RNA-sequencing data was generated under conditions of exponential and aerobic growth in M9 minimal media with a glucose carbon source. Cells were washed with Qiagen RNA-protect Bacteria Reagent and pelleted for storage at -80°C prior to RNA extraction. Cell pellets were thawed and incubated with Read-Lyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 minutes at 37°C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase-treatment was performed for 30 minutes at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA-nano chip on a bioanalyzer. Paired-end, strand-specific RNA-seq was performed following a modified dUTP method (34). A majority of rRNA was removed using Epicentre's Ribo-Zero rRNA removal kit for Gram Negative Bacteria.

Reads were mapped with bowtie2 (35). Expression levels in units fragments per kilobase per million fragments mapped (FPKM) were found with cufflinks 2.0.2 (36). Gene expression fold change (with respect to the wild-type strain) was found using cuffdiff; a q-value cutoff of 0.05 was used to call significant differential expression. Gene annotation from EcoCyc version 15.0 was used for all analysis (37).

3.2.5 Commonly differentially expressed genes

A statistical model was used to determine how many genes are expected to be commonly differentially expressed in the same direction (up or down) across multiple strains. In the null model, each gene in each strain can have one of three states: up-regulated, down-regulated, or not significantly differentially expressed compared to the wild-type. For each gene in a given strain, the probability of the three states follows a multinomial distribution parameterized empirically by the differential expression calls in the processed RNA-seq data (see RNA-Sequencing). The genes that are differentially expressed in each strain are assumed independent in the null model, so the probability that a gene is differentially expressed in multiple strains is determined by the product rule of probability. Commonly differentially expressed genes are then called when no genes are expected to be differentially expressed in the same direction across that number of strains (i.e., expected value is less than 1). For this dataset, no genes are expected to be commonly differentially expressed (in either direction) across 6 or more strains.

3.2.6 ME-Model simulation and gene classification

The ME-model as published in O'Brien et al. was used for all simulations (38). 20 distinct glucose uptake rates, evenly spaced between 0 and the optimal substrate uptake rate (when glucose is unbounded) were simulated as described in O'Brien et al. (38). Any gene predicted to be expressed in any of the 20 simulations are classified as 'Utilized ME'; genes within the scope of the ME-Model, but not expressed in any of the 20 simulations are classified as 'Non-utilized ME'; genes outside the scope of the ME-Model are classified as 'Outside scope ME'. These gene groups are then

compared to COGs and the identified commonly differentially expressed genes in the end-point strains (see Commonly differentially expressed genes) (39).

3.2.7 Jump Finding

Growth rates were calculated for each batch during the course of evolution using a least-squares linear regression. The following criteria were used to determine whether to accept or reject the computed growth rate

- Number of OD samples ≥ 3
- Range of OD measurements must be $\geq .02$
- Passage OD within 50% of targeted passage OD

The accepted growth rates were fit with a monotonically increasing piecewise cubic spline. Regions with a slope greater than $4.2x10^{-15}hr^{-1}CCD^{-1}$ were considered jumps with a few exceptions. The spline was created using 'slmtools' function in MATLAB available on the MATLAB file exchange. The number of spline segments (#knots-1) was varied to capture the upward trends in growth rates.

3.2.8 Knock in Procedure

The single point mutation introduction in *rpoB* was done by 'gene gorging' as described previously (22). Briefly, the mutation in *rpoB* was amplified by PCR from the genomic DNA of the ALE clone where it was originally found. Amplification was done with primers approximately 500 bp upstream and downstream of the mutation and flanked by the 18 bp I-*SceI* site, and PCR product was cloned in a pCR-Blunt II-Topo vector (Invitrogen, Carlsbad, CA) to create a donor plasmid. The donor plasmid was co-transformed along with the pACBSR plasmid harboring an arabinose induced

lamda-red system and the I-SceI endonuclease on a compatible replicon. A colony of the strain transformed with both plasmids was grown with arabinose as an inducer and after 7-12h several dilutions of culture were plated with and without antibiotics to verify the loss of the donor plasmid. The initial screening of positive clones was carried out by PCR using a 3’ specific primer to the introduced mutation (40). The positive colonies were confirmed by Sanger sequencing.

3.3 Results

3.3.1 Characterization of the Evolution Process and the Endpoint Strains

Adaptive laboratory evolution was used to examine *E. coli*'s physiological and genetic adaptation to simple media conditions under a strict selection pressure. Eight independent populations of wild-type *E. coli* K-12 MG1655 from the same seed culture were adaptively evolved in parallel under continuous exponential growth for a time period of 39-81 days. During this time, the cultures underwent approximately $8.3x10^{12}$-$18.3x10^{12}$ cumulative cell divisions (CCD) (Table 3.1) (41, 42). The use of CCD as a coordinate allows for incorporation of the number of cells passed in an ALE experiment along with generations of a culture (41). Variations in time courses and CCD are due to re-inoculations from frozen stocks (taken throughout the experiment) and occasional unexpected losses of cultures or suspected contamination as determined using 16S ribosomal sequencing. The fitness trajectories (i.e., population growth rates) as fit by a spline over the course of the evolution are given in Figure 3.1. Each of the evolved populations increased in fitness from the starting strain (Table 3.1). The growth rate increases were 1.47±0.05 (standard deviation, n=3) fold faster than the starting strain and ranged from 1.42-1.59. One of the populations (determined to be a hypermutator strain, see below) was statistically faster than the rest and increased 1.59 fold (p-value ≤ 0.01).

There was a significant increase in fitness from the first flask to the second in each of the independent experiments (Figure 3.1, insert). This phenomenon has been

previously observed and described through an examination of growth when cells are repeatedly passed during their exponential growth phase (32). An initial 'physiologically-adapted' growth rate was determined for the starting wild-type strain of 0.824±0.036 hr^{-1} and was determined using growth rates recorded for flasks 2-4 across all of the independent ALE experiments. This repeated exponential phase growth rate is 19% faster than the average growth rate of flask 1 from each experiment (0.69±0.02hr^{-1}). It should be noted that this increase in growth rate is not expected to be a result of a beneficial mutation.

Clones were isolated from the last flask of each of the evolved populations, phenotypically characterized (growth rates, glucose update rates (GUR) and acetate productions rates (APR)), and compared to the starting wild-type strain to understand how their behavior changed after evolution (Figure 3.2). Nine clones isolated from the experiments were analyzed (six isolated from the non-hypermutator populations, and three isolated from the hypermutating linage were analyzed as it possessed a significantly higher population fitness). The increase in fitness (i.e., growth rate) was 1.29-1.46 fold. To quality control the data, the phenotype of the wild-type strain was compared with other studies and found to be in good agreement with previous characterizations (43). The clone growth rates were compared to the population from which they were derived, and the Pearson correlation coefficient between them was 0.16. The isolated hypermutator clones diverged more significantly from the population growth rates (1.10-1.20) than did the non-hypermutator strains (1.02-1.11).

The physiological properties of each of the clones isolated from the independent ALE experiments were compared to examine if there were any conserved

trends across the different experimental outcomes. There was a similar increase in growth rate across the isolates from different experiments, but a larger variation in the glucose uptake rates and biomass yields (Figure 3.2A). The glucose uptake rates (GUR) and acetate production rates (APR) increased in the endpoint strains compared to wild-type (except for one strain where the APR decreased). There is a correlation (r^2= 0.70) between the increase observed in the GUR and APR (Figure 3.2B). Of the characterized strains, the hypermutators accounted for three of the four lowest APRs and highest steady-state biomass yields (Y_{X/S_ss}). No other common fermentation products of *E. coli* K-12 MG1655 (i.e., formate, ethanol, succinate, lactate) were detected as secretion products in any of the endpoints, thus indicating that these the three hypermutator strains strains generally metabolized glucose more efficiently. A similar correlation was also seen between biomass yield and APR (r^2=0.57, Figure 3.2C). Thus, clones in the independent ALE experiments converged to a similar optimal fitness by either becoming more efficient in their biomass yield or increasing GUR and overflow metabolism in the form of acetate secretion. A tradeoff between GUR and Y_{X/S_SS} was observed in that higher glucose uptake rates led to lower Y_{X/S_SS} (i.e., they are inversely correlated, r^2=0.93). However, it should be noted that the Y_{X/S_SS} calculation involves GUR as a factor.

3.3.2 Analysis of Mutations Identified in the Evolved Strains

A persistent challenge and goal in ALE experiments is differentiating between causal mutations and genetic hitch-hikers. In these set of experiments alone, 72 unique mutations were identified across non-mutator strains. To aid in determining causal mutations, jumps in fitness were identified using a jump finding algorithm (see

methods). Clones were isolated that bracketed jump regions and sequenced in order to evaluate if jumps in growth rates could be linked to a genetic change which had been fixed in the population over the course of the jump (Figure 3.3). An analysis of key mutations is given in Table 3.2. The genes or genetic regions listed in Table 3.2 are those that were found mutated in multiple experiments, or which contained multiple unique mutations across the gene/genetic region. Figure 3.3 additionally show if a given mutation persisted, was found in multiple points of clonal analysis, or was no longer detected but another mutation in the same gene was identified. Mutations that were linked to fitness jumps are identified in Table 3.2.

Overall, 52 unique genetic regions (i.e., genes or intergenic regions between two genes) were mutated across all non-mutator clones sequenced, encompassing 72 total unique mutations. Of the 52 unique genetic regions, multiple unique mutations occurred in eight genetic regions (Table 3.2). 57% (30 of 53) of all mutations persisted in every subsequent clone examined until the experiment ended (mutations only observed in the last clone examined for each experiment were not considered). Some mutations were found in multiple subsequent clones from an experiment, but did not persist after first being observed. There were two such instances in experiment 10, where three distinct genotype lineages were observed in the various clones sequenced. Of the genes containing the 30 persistent mutations, only three have been reported in a similar glucose minimal media ALE experiment: *rpoB*, *ygiC*, and *ydhZ/pykF* (44). When considering the hypermutator population clones, an additional *pykF* mutation was also observed. It should be noted that the exact mutations were different than those previously reported and only *rpoB* was included in our analysis of key

mutations. Overall, there were 7 - 21 mutations identified in each experiment, with a median value of 13. Experiment 4 had the fewest genetic changes with seven unique mutations across all sequenced clones, and only four in the final clonal isolate. In comparison, experiment 10 had 21 unique mutations observed across all clones and 12 in the final clonal isolate. Similar continuous exponential growth-phase ALE experiments run for approximately 10^{11} CCDs (more than an order of magnitude fewer than in this study) on glycerol, lactic acid, and L-1,2-propanediol minimal media yielded 2-5, 1-8, and 5-6 mutations per independent experiment, respectively (23, 24, 45).

Several genes and genetic regions were identified that contained mutations across many of the independent ALE experiments, implying causality. The most frequent mutation targets were the intergenic region between *pyrE* and *rph*, the *rpoB* gene, and between *hns/tdk* via an insertion sequence (IS). An 82bp *pyrE/rph* deletion was observed in every sequenced clone. A K-12 specific defect has been previously described which is ameliorated by this mutation (23, 46). A subunit of RNA polymerase, *rpoB* was found to be mutated in every experiment and likely has a genome-wide impact on transcription given its vital role in the transcription process (47, 48). All of the mutations were single amino acid changes. Multiple unique mutationsg were found singly across clones which harbored *rpoB* mutations after the first jump in fitness. IS element mediated mutations were found in all experiments, typically after the second jump in fitness, except where a hypermutating phenotype was dominant. Three different IS elements (IS1, IS2, and IS5) were inserted in seven

different locations, and one identical IS5 mutation was detected using the described clonal analysis.

The clones sequenced after the second jump in experiment 7 exhibited hypermutator behavior. This was readily apparent from the 139 mutations it possessed, an order of magnitude greater than any other strain for a given number of CCDs. Additionally there was an IS element inserted into the *mutT* gene of this strain. Due to the large size of the insertion (777bp), it almost surely results in *mutT* loss-of-function. It has been shown, by knock-out, that defective MutT increases SNPs in the form of A:T to C:G conversions (49). Of all the mutations observed in the hypermutator strains, only 6 of 381 were not A:T to G:C conversions. When all four isolated and resequenced hypermutator clones were compared, 33 mutations were shared between all four. The overlap in genes or genetic regions between the hypermutators and non-mutators was analyzed, and it was found that the only identical shared mutation was the 82bp deletion in *pyrE/rph*. Only two (*iap*, *ydeK*) of the same genes or genetic regions were mutated in both the non-mutator and hypermutator lineages. Thus, these genes also indicate potential key mutations for the observed phenotypes.

3.3.3 Analysis of Reproducibility for Key Mutations Which Enable Increased Fitness Phenotypes

To analyze how reproducibly key mutations occur, the evolution process was repeated starting with strains that harbored three of the key mutations identified in this study: *rpoB* E546V, *rpoB* E672K, and *pyrE/rph* Δ82bp. The hypothesis which was tested was the expectation that key mutations would again occur and the approach

developed in this work could select for them when starting another ALE experiment with one of the key mutations already present (i.e., with different starting material). Consequently, the fitness increase associated with each mutation could also be tested. Each of these single mutants were reconstructed in the starting strain background and validated (see Methods). The conditions of this 'validation' ALE experiment were essentially identical to the first ALE experimental setup, but with the dilution ratio changed to 0.67% of the total culture volume (as compared to 5.0% in the initial experiment) in order to reduce clonal interference and genetic drift. The fitness trajectories of the validation evolution experiment are shown in Figure 3.4. The initial and physiologically-adapted growth rates of the three reconstructed strains demonstrated that their mutations were indeed causal for faster growth on minimal media. Key mutations detected in the validation ALE are given in Table 3.4. It is interesting to note that a different mutation between *pyrE/rph* was detected (a 1bp deletion) besides the ubiquitous 82bp deletion detected in the primary ALE. Furthermore, using PCR it was revealed that all populations showed evidence of obtaining the 82bp deletion, though the entire population did not harbor the mutation. Additionally, *metL* and *hns/tdk* mutations were also detected in the validation ALE. *metL* mutations are not as widespread, but two out of three mutations that did appear in *metL* are consistently loss of function suggesting that inactivation of the gene can increase growth rate in the minimal media conditions tested.

To examine the increase in fitness from key mutations identified, growth screens were performed for relevant single and double mutants (Figure 3.5). These strains were either reconstructed manually or were isolates of the validation ALE. The

results show that the mutation observed in *metL* and the IS1 insertion into *hns/tdk* also conferred a fitness advantage. The *metL* and *hns/tdk* were both shown in the presence of additional mutations, so their potential for epistasis is unknown. However, for the mutant with the IS1 insertion into the region between *hns/tdk*, it only harbors the 82bp deletion in *pyrE/rph* which has been previously shown to alleviate a known K-12 MG1655 specific defect (23, 46). Thus, it is highly likely that it is uniquely causal without epistasis. In the case of *metL*, mutations were only observed after a mutation in *rpoB* was present. This could either indicate epistasis between the two mutations or simply that *rpoB* confers a larger fitness advantage and thus was selected for before a mutation in *metL*. If the fitness advantage from the double mutant screens is assumed to be additive, the increase in fitness for the observed mutation in *metL* and between *hns/tdk* is $0.065\pm0.023hr^{1}$ and $0.045\pm0.035hr^{-1}$, respectively. Furthermore, the double mutant harboring both the *rpoB* E672K and Δ82bp *pyrE/rph* mutation follows this additive trend as each single mutant increased fitness $0.125\pm0.038hr^{-1}$ and $0.146\pm0.044hr^{-1}$, respectively, and when they were both present the increased fitness was $0.237\pm0.058hr^{-1}$. It should be noted that the growth rate measured from just the *rpoB* E672K and Δ82bp *pyrE/rph* mutations ($1.027\pm0.043hr^{-1}$) matches the highest growth rate measured from the populations that harbored both of these mutations ($1.01hr^{-1}$) in its 95% confidence interval.

3.3.4 Transcriptomic Analysis of Evolved Strains

Expression profiling was performed on endpoint strains using RNA-seq to identify system-wide changes in gene expression after evolution. For the eight strains profiled using RNA-seq, out of 4298 protein-coding ORFs, reads aligned to a total of

4189 genes (109 have no reads) in at least one strain, and 2922 genes in all strains (see sequencing methods), indicating a comprehensive/deep coverage of the transcriptome. Genes were identified that were differentially expressed in endpoint strains compared to the wild-type (see sequencing methods). In all strains, hundreds of genes significantly increased and decreased in expression, indicating large shifts in the transcriptome.

The common changes in gene expression across strains were analyzed to examine the heterogeneity of the different independent ALE experiments. As a null model, it was assumed that the expression changes in each gene are independent of each other. Using this null model, the expectation would be that no genes should be commonly differentially expressed across 6 or more strains. However, 448 genes commonly increased in expression and 383 genes commonly decreased in expression across 6 or more strains (Figure 3.6A), indicating largely consistent changes in expression (though there is also a significant amount of diversity in the expression changes). This commonly differentially expressed gene set was selected for further analysis to better understand the coordinated change in the transcriptomes of the evolved strains.

For a broad overview of the cellular processes with modulated expression, over-represented COG (Cluster of Orthologous Group) annotations (39) in the commonly differentially-expressed genes were identified. Overall, 79% (359) of the commonly increased and 65% (252) of the commonly decreased genes had annotated COGs (see Methods). While no COG annotation was enriched in the genes that decreased in expression, three categories were enriched in the increased genes. These

up-regulated COGs are translation, protein folding, and amino acid metabolism (Figure 3.6B). All of these COGs are related to protein synthesis, indicating that an increase in protein synthesis capacity is a common trend among evolved strains. These changes are consistent with previously described growth rate dependent increases in ribosomal and other protein synthesis machinery (50). At faster growth rates, the increased dilution of protein to daughter cells places a higher demand on protein synthesis, driving the increased expression.

In order to connect genotype to molecular phenotype where possible, a comparison was made between the identified common mutations (Table 3.4) and gene expression levels within or between the mutational loci. Paired mutation and expression data for 6 endpoint strains (numbers 3, 4, 6, 8, 9, and 10) along with two hypermutator isolates, 7A and 7B, were used in the analysis. The same *pyrE/rph* mutation occurred in all 6 endpoint strains; *pyrE* was significantly up-regulated in all strains whereas *rph* was significantly down-regulated in 5 out of 6 strains (with no significant differential expression in strain 6). The up-regulation of *pyrE* is consistent with the previously identified mechanism of the mutation as relieving a pyrimidine pseudo-auxotrophy (23, 46); the *rph* down-regulation, on the other hand, is likely not directly beneficial for fitness as the gene contains a frameshift and lacks RNase PH activity (46). An intergenic *hns/tdk* mutation also occurred in all 6 endpoint strains, and in all strains, *hns* is significantly up-regulated and *tdk* is significantly down-regulated (though not significantly in strain 9). Histone-like nucleoid structuring protein (H-NS) is a global transcription factor, which represses a wide array of stress responses (51); the benefit of the *hns/tdk* mutation may therefore be due to the up-

regulation of *hns* and subsequent down-regulation of many stress responses. Tdk down-regulation has no apparent benefit, but may ameliorate a potential imbalance in deoxyribonucleotide biosynthesis. A mutation occurred in *rpoB* in all 6 endpoint strains and *rpoB* was also up-regulated in all of these strains (though not significantly in strain 8). The mutation was intragenic within *rpoB* and likely does not directly affect its expression level, however *rpoB* was up-regulated (in addition to all other subunits of the sigma 70 holoenzyme) as a consequence of increases in growth rate (see section below). This growth-rate dependency is further corroborated in that the hypermutator clones did not have an *rpoB* mutation, but all of the RNAP holoenzyme subunits are upregulated in these strains as well. For the other key mutations that occurred repeatedly, there was no clear pattern between the occurrence of the mutation and differential expression of the related gene. Looking at an additional strain-specific intergenic IS element insertion between *uvrY/yecF* in endpoint strain 6, it was found that *uvrY* was significantly down-regulated, a shift experienced in three of the other strains as well (*yecF* expression was essentially the same as wild-type). Furthermore, there was an intragenic mutation in *uvrY* (W42G) in strain 7A, one of the other strains where it was differentially expressed. Thus, comparison of expression data and mutation data revealed potential links between genotype and molecular phenotype for the three intergenic IS element mutations identified in evolutions (those where one would most expect to see a change in transcription) (52-54).

3.3.5 Integrated Genome-scale Modeling

Constraint-based models are capable of predicting growth-optimizing phenotypes (15, 30, 55, 56). A recent genome-scale model of Metabolism and gene

Expression for *E. coli*, a ME-Model, extends predictions beyond metabolism to also include growth-optimization of gene expression phenotypes (38). To test the predictions of gene expression, categorize the transcriptomic data, and provide further insight into the expression data, model predictions were compared to the commonly differentially expressed genes from the analysis of evolved strains.

Utilizing the ME-Model of *E. coli*, growth rate optimizing phenotypes in glucose aerobic culture media conditions (i.e., the same conditions as the ALE experiments) were simulated. Based on these simulations, three groups of genes were identified: 1) genes utilized by the ME-Model in maximum growth rate conditions ('Utilized ME', n=540), 2) genes within the scope of the ME-Model, but not predicted to be utilized in a maximum growth phenotype ('Non-utilized ME', n=1014), and 3) genes outside the scope of the ME-Model ('Outside scope ME', n=2744) which have yet to be reconstructed in a constraint-based formalism (38).

If the *in silico* predicted Utilized ME genes are indeed important for an apparent optimal growth rate, one would expect them to be in the commonly differentially expressed set as determined through untargeted transcriptomics. To test this hypothesis, the three model-defined gene classes were compared to the commonly differentially expressed genes. Indeed, it was determined that the Utilized ME genes were more often commonly differentially expressed (Figure 3.7A top). Furthermore, of the Utilized ME genes that are differentially expressed, 85% were up-regulated, indicating that the transcriptome generally shifts towards these optimal growth-supporting genes (Figure 3.7A bottom). The Non-utilized ME genes form an intermediate category whose frequency of differential expression (and frequency of

increased differential expression) is between that of Utilized ME genes and Outside scope ME genes. Non-utilized ME genes, although not predicted to be utilized for purely growth-optimizing phenotypes, still contribute to increased growth; whereas many Outside scope ME genes do not. While differentially expressed Non-utilized ME genes have increased expression about half of the time, Outside scope ME genes more often show decreased expression, indicating a shift away from the Outside scope ME genes.

The COG and model-based gene categorizations were combined to provide further insight into the processes commonly differentially expressed among the endpoint clonal isolate strains. By dividing up the genes into Utilized ME and outside scope ME, new processes missed by just considering the COG annotations alone were identified, which also served to highlight important areas of model expansion.

As in the analysis of the transcriptomic data alone, amino acid metabolism, translation, and protein maturation were enriched in the commonly differentially expressed Utilized ME genes, indicating that the ME-Model correctly predicted a number of the genes in these processes that are important for increased growth rate. By further categorizing the COGs based on the Utilized ME genes, transcription was identified as an up-regulated process. This finding was missed by the categorization based on COGs alone as a result of the numerous genes annotated as related to Transcription. However, by further segmenting this COGs group by model-predicted genes essential for transcription, it is revealed as an up-regulated process.

Looking at the specific genes in the pared gene groups at the intersection of COGs annotations and modeling predictions revealed more details on the specific

processes and complexes that change in expression (Figure 3.7B). However, there are some clear pathway-level shifts worth mentioning here. Energy production and conversion was identified as a down-regulated process (again, energy production and conversion (C) is a broad COG category), but when it is pared-down to only consider model-predicted Utilized ME genes, it is identified as a category with significant changes in expression. Interestingly, genes that decrease in expression all belong to the TCA and glyoxylate cycles (*mdh, acnAB, aceAB, gltA, icd*). This concerted down-regulation is likely related to the increase in fermentative metabolism and acetate secretion of the evolved strains (Figure 3.2). Though aerobic respiration has higher energy yields than fermentative metabolism, it has been hypothesized that the flux through the respiratory reactions is limited by protein synthesis cost and capacity (38, 57, 58) (as TCA and the electron transport system require more proteins than glycolysis and acetate secretion) or limitations in membrane space (58) (for electron transport system enzymes). These gene expression and physiological changes may be driven by these key capacity constraints.

Many COG categories were revealed as enriched when combining this categorization with the Outside scope ME genes. COG categories with significantly increased expression indicate processes important for growth, but not yet encompassed by the ME-Model, whereas COG categories with decreased expression indicate processes important for growth, but not important for optimal growth in glucose-excess aerobic culture conditions (Figure 3.7B). The up-regulated Outside scope ME genes involved in intracellular trafficking and secretion are all involved with protein translocation from the cytosol to the membranes and periplasm. These include genes

in the Sec (*secA*, *secE*, *yajC*), Tat (*tatB*), and SRP (*ffh*, *ftsY*) translocation pathways. Similar to the common changes in gene expression and protein folding, this increased expression is likely driven by the increased need to synthesize a functional (and localized) proteome, as the dilution of these proteins to daughter cells increases their demand. Thus, categorization using both COGs and the ME-Model allows for an interpretation of the expression changes driving the observed growth increases in the evolved strains, and highlights areas of poor understanding to be further characterized and included in future genome-scale models.

3.4 Discussion

Adaptive laboratory evolution was utilized to explore optimal growth of *E. coli* K-12 MG1655 on glucose minimal media. This combination of organism and media conditions is arguably the most widely-used in basic science and biotechnology applications (59). Multiple parallel experiments were performed to use as comparison points for the overall process. The ALE was performed by propagating batch cultures during exponential growth phase where the passage volume was intentionally kept at a relatively large amount and held constant throughout the experiment. This is different from previous ALE studies where passage volume was generally decreased as the growth rate increased (45). The intent was to isolate the growth rate as the only selection pressure and remove any bottlenecks associated with a lower passage size. The results show that the large increases in growth rates observed here are achieved over a significantly shorter time-frame (44). As with stationary phase batch culture propagation, any fixed mutated genetic regions could very well be causal for a

secondary selection to growth rate (e.g., lag phase duration). The strains produced by this experiment were screened for their phenotype, genotype, and transcriptome. Genome-scale models were used to analyze the results of these screens. Accordingly, the major findings from this work are: i) passing larger volumes strictly in exponential phase batch culture can increase the rate of selection for improved fitness, ii.) the identification of key reproducibly-occurring mutations that enable higher growth rates for *E. coli* K-12 MG1655 under glucose minimal media conditions, iii.) apparent optimal phenotypes can be realized through modification of different mechanisms, and iv.) optimal phenotypic states, as probed through transcriptomic assays, are in good agreement with predicted cellular states from genome-scale modeling, and categorization with modeling results reveal drivers for the optimal phenotypes on a pathway level.

The growth rates achieved in this work surpass those from comparable studies. In a long-term evolution experiment (LTEE), in which *E. coli* have been evolving for over 50,000 generations in glucose minimal media, results at the 2,000 generation mark were used for comparison, as those were closest in evolutionary timeframe to the results of this work (60). It is important to note that in the LTEE, an *E. coli* B strain was used on glucose minimal media, as opposed to K-12 used here, and cells were always passed during stationary phase. Nonetheless, the LTEE observed a 1.29±0.10 (standard deviation) fold increase in growth rates of the populations, compared to the 1.42-1.59 fold increase achieved here. Further, the LTEE took 10,000-15,000 generations to reach an approximate 1.5 fold increase in growth rate, here this fold increase was achieved in approximately 2,000 generations.,. No identical mutations

were seen between the LTEE and this work, and only three mutated genetic regions were found in both: *rpoB*, *ygiC*, and *pykF*. The differences can presumably be attributed to the serial passage of cultures and/or the different starting *E. coli* strain. As another point of comparison, a different evolution study was performed on glucose minimal media for 50 days using the same K-12 strain and media conditions used here (3). In that experiment, a 1.1-fold increase in growth rate was observed, drastically lower than the increase found here. The only major difference between the two K-12 studies was that in the previous work the passage size was adjusted (i.e., reduced as the fitness increased) to keep the cultures out of stationary phase. Thus, these findings point to the importance of methodology used in an ALE experiment as highlighted by the differences in phenotypic and genotypic outcomes.

Key mutations were identified which enabled faster growth of *E. coli* K-12 MG1655 on glucose minimal media and these mutations did not appear in the identified hypermutating lineage. These key mutations were straightforward to identify as the given genetic regions were reproducibly mutated across multiple ALE experiments. The causality of select single and double mutants of these regions was shown (Figure 3.5). The reproducibility observed is likely due to the strict selection pressure that was maintained in the experiment, keeping the populations in constant exponential growth. However, in one experiment, a hypermutating population arose. The genotype of the hypermutator differed significantly from the non-mutators; the vast majority of the key mutations determined from the non-mutator set were not detected in the hypermutator clones sequenced. This indicated that there were multiple genetic changes capable of enabling a similar fitness increase, which is further

supported by the similarities in the transcriptome across all strains. Furthermore, the *rpoB* and *hns/tdk* mutations in the non-mutator strains likely affect global transcriptional levels. This would allow for single mutations to affect a multitude of reactions in the network. Compared to the hypermutator that did not have either of these, it was able to confer a similar effect on the network by fixing numerous mutations that presumably have similar, perhaps more local, individual effects. It should be mentioned however, that while the hypermutator did not have mutation in *rpoB,* it did have one in the *rpoC* subunit of the RNA polymerase holoenzyme, which could have a similar board impact on transcriptional levels in the cellular network.

The occurrence of the identified key mutations was highly reproducible. This conclusion was supported by the results of the validation ALE experiment which was started using clones already harboring single causal mutations (Figure 3.5, Table 3.4). Mutations in *pyrE/rph*, *rpoB*, *hns/tdk*, and *metL* all reappeared in these experiments, to varying extents. The ability of clonal analysis to capture population dynamics was also examined. Although clonal resequencing most often yielded agreement with the population-level analysis (analyzed with population PCR), it did not always capture the presence of a specific mutation shown to cause an increase in fitness (in this case, the 82 bp deletion between *pyrE/rph*). Thus, clonal analysis is useful and informative, but it has its limitations and ultimately ALE studies can benefit from a more population-centric analysis of mutations. Looking at the differences in mutations which occurred in a given gene, it appears that there are multiple specific mutations that can have a similar effect on fitness (Table 3.2). Of the specific mutations observed in *rpoB*, all conferred a fitness advantage but to varying degrees (Figure 3.5). More

than one mutation in *rpoB* was never observed in a single strain, suggesting that there could be negative epistasis between the different identified SNPs; their effects are non-additive. Nonetheless, this study presents a number of reproducibly occurring and causal genes which enable rapid growth of *E. coli* on glucose minimal media.

The physiological characterization of evolved strains indicated that there were multiple mechanisms through which to realize an increased growth rate. The clones isolated from the endpoints of the primary ALE experiments all increased in fitness to a relatively similar degree, yet the GUR and Y_{X/S_ss} varied between them (Figure 3.2). Of the three hypermutator clones isolated and characterized, two seemed to diverge from the others by having significantly lower GURs yet higher Y_{X/S_ss} (i.e., they are more efficient). The observed extremes in GUR, APR, and Y_{X/S_ss} show that the trajectory across the fitness landscape traversed by MG1655 on glucose minimal media is not a rigid, predetermined path. It should be noted that the growth rates of the two aforementioned hypermutators fell in between the range of growth rates of the other clones. Furthermore, this study has shown that there is a clear and distinct physiologically adapted growth state which is realized after several generations of continuous exponential growth (differing from growth started directly from a stationary phase culture). This observed phenomenon was reproducible using the quantitative approach in this study and puts an emphasis on critically evaluating previously reported "maximum" growth rates of strains.

Genome-wide analysis of the evolved strains using transcriptomics revealed a consistent evolved expression shift, and further categorization using genome-scale modeling revealed pathway-level shifts underlying the increased growth phenotypes.

Furthermore, transcriptomics was utilized to link genotype to phenotype when considering the effects of IS element mutations. The most apparent mutational effect was that of IS elements between *hns/tdk*, where the *hns* gene product was significantly up-regulated in all of the strains harboring these mutations. These *hns/tdk* insertions were shown to be causal for an increased growth rate and could be further utilized, along with other key mutations, to improve efficiency in biomass yield or GUR. The most highly conserved changes in the transcriptomes across the evolved strains were in good agreement with the predicted gene products whose differential expression would enable rapid growth, as determined through genome-scale modeling. When considering the coordinated changes in the transcriptomes of the evolved strains solely with a classification like COGs, enriched pathways became apparent which contributed to the shift in the functional state of the cells. The results of the genome-scale modeling classification changed this enrichment significantly and allowed a deeper examination into the physiological state and mutation-induced pathway expression changes of the evolved strains. Thus, it was useful to interpret the outcome of evolution in the context of an *in silico* analysis of optimal performance in this particular condition.

In summary, we have shown that ALE can be utilized to find reproducible causal mutations that optimize for a selectable phenotype using a controlled experimental setup and strict selection pressure. Whole-genome resequencing enabled the mutational discovery, and transcriptomic analysis coupled with genome-scale modeling uncovered the metabolic pathways underlying the evolved phenotypes. These findings and the general experimental approach we have laid out can be

extended to additional culture conditions, strains, and selection pressures for a variety of basic science and applied biotechnological purposes.

Chapter III, in full, is a reprint that the dissertation author was the principal researcher and author of. The material appears in *Applied and Environmental Microbiology*. (LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM. 2015. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol 81:17-30.)

3.5 Tables

Table 3.1 - Fitness properties of the evolved populations

Experiment	Population Growth Rate (hr^{-1})	Total CCD	Total Doublings	Ratio of Final Fitness to Wild Type	Total Number of Flasks
Wild-type K-12 MG1655	0.69±0.02	0	0	1	NA
3	1.01±0.16	$13.5x10^{12}$	1903	1.46	382
4	0.98±0.10	$10.2x10^{12}$	1440	1.42	288
5	1.01±0.08	$8.3x10^{12}$	1184	1.46	288
6	1.00±0.16	$11.3x10^{12}$	1630	1.46	327
7	1.11±0.10	$13.6x10^{12}$	1870	1.59	375
8	0.99±0.11	$10.5x10^{12}$	1542	1.43	309
9	1.01±0.09	$18.1x10^{12}$	2589	1.46	519
10	1.02±0.12	$18.3x10^{12}$	2582	1.48	518

CCD – Cumulative cell divisions, 95% Confidence interval for the wild-type strain was determined from biological triplicates, population growth rate were taken from the endpoint of the fitted spline.

Table 3.2 - Key Mutations

Gene	Mutation	Appearance Location	Replacing Mutation (within same experiment)	Appearance Location	Occurrences	Experiment(s)
pyrE/rph	Δ82 bp	Jump 1 pre-Jump 1			8	3, 5, 9, 10 4, 6, 7, 8
rpoB	E672K (GAA→AAA)	Jump 1			8	3, 5, 9
	P1100Q (CCG→CAG)	Jump 1				4, 8
	E546V (GAA→GTA)	Jump 1				10
	H673Y (CAC→TAC)	Jump 1	D785Y (GAC→TAC)	Jump 2		6
	L671P (CTG→CCG)	Jump 1	hypermutator	Jump 2		7
hns/tdk	intergenic (-114/-487) IS2	Jump 2			7	3
	intergenic (-110/-488) IS1	Jump 2				4
	intergenic (-274/-328) IS5	Jump 2				5
	intergenic (-86/-511) IS1	post Jump 2				6
	intergenic (-67/-531) IS1	Jump 2				8
	intergenic (-93/-505) IS1	Jump 3				9
	intergenic (-258/-344) IS5	Jump 2	intergenic (-274/-328) IS5	post Jump 2		10
corA	coding (726-728/951 nt) Δ3bp	Jump 1	coding (220-224/951 nt) Δ5 bp	Jump 1	3	4
	A206V (GCG→GTG)	Jump 1-2	coding (113-211/951 nt) Δ99bp	Jump 2		5
	coding (668/951 nt) duplication 21bp	Jump 2-3	wild type	Jump 3		10
ygaZ	coding (529-532/738 nt) IS5	Jump 2	coding (307-316/738 nt) Δ10 bp	post Jump 3	3	3
	E49* (GAA→TAA)	Jump 3				9
	2807900 19bp x 2	post Jump 3				
iap	coding (98‑101/1038 nt) IS5	post Jump 2/3				6, 9
metL	coding (1338/2433 nt) Δ1bp	Jump 2-3	A798E (GCG→GAG)	Jump 3	1	10
ygeW	S200R (AGC→CGC)	Jump 1			2	5, 9

Table 3.3 - Phenotypic data from clones isolated from the final flask of each experiment

Strain	Growth Rate (hr^{-1})	Glucose Uptake Rate (mmol gDW^{-1} hr^{-1})	Acetate Production Rate (mmol gDW^{-1} hr^{-1})	Biomass Yield (gDW $gGlc^{-1}$)	Fold Increase vs. wild-type	Population/Clone Growth Rate
Wild-type K-12 MG1655	0.69±0.02	8.59±1.42	3.91±1.14	0.44±0.07	-	-
Exp. 3	0.98±0.02	13.51±1.15	8.43±2.17	0.40±0.04	1.42	1.03
Exp. 4	0.96±<0.01	12.19±0.68	7.89±1.88	0.44±0.02	1.39	1.02
Exp. 6	0.93±0.01	12.77±0.85	7.11±1.51	0.40±0.03	1.34	1.07
Exp. 7*	1.01±0.04	13.13±1.29	5.12±0.57	0.43±0.06	1.46	1.10
Exp. 7A*	0.97±<0.01	11.01±0.79	3.97±0.98	0.49±0.03	1.41	1.14
Exp. 7B*	0.92±0.02	10.43±0.62	2.36±0.54	0.49±0.03	1.33	1.20
Exp. 8	0.89±0.01	12.59±1.01	5.05±0.40	0.39±0.03	1.29	1.11
Exp. 9	0.92±0.02	13.13±0.59	6.99±0.48	0.39±0.02	1.33	1.10
Exp. 10	0.95±0.01	13.98±1.11	9.27±1.76	0.38±0.03	1.38	1.07

* denotes hypermutator strain, Exp. – experiment

Table 3.4 - Key Mutations in Validation ALE

Genetic Region	Starting Strain	Mutation	Occurrences	Experiment(s)
pyrE/rph	*rpoB* E546V	Δ82bp deletion	1	2
		Δ1bp deletion	1	3
	rpoB E672K	Δ82bp deletion	3	4,5,6
rpoB	*pyrE/rph*	A679V (GCA→GTA)	1	8
		V857E (GTG→GAG)	1	9
hns/tdk	*pyrE/rph*	intergenic (-75/-522) IS1	1	9
metL	*rpoB* E546V	W424* (TGG→TAG)	1	1

3.6 Figures

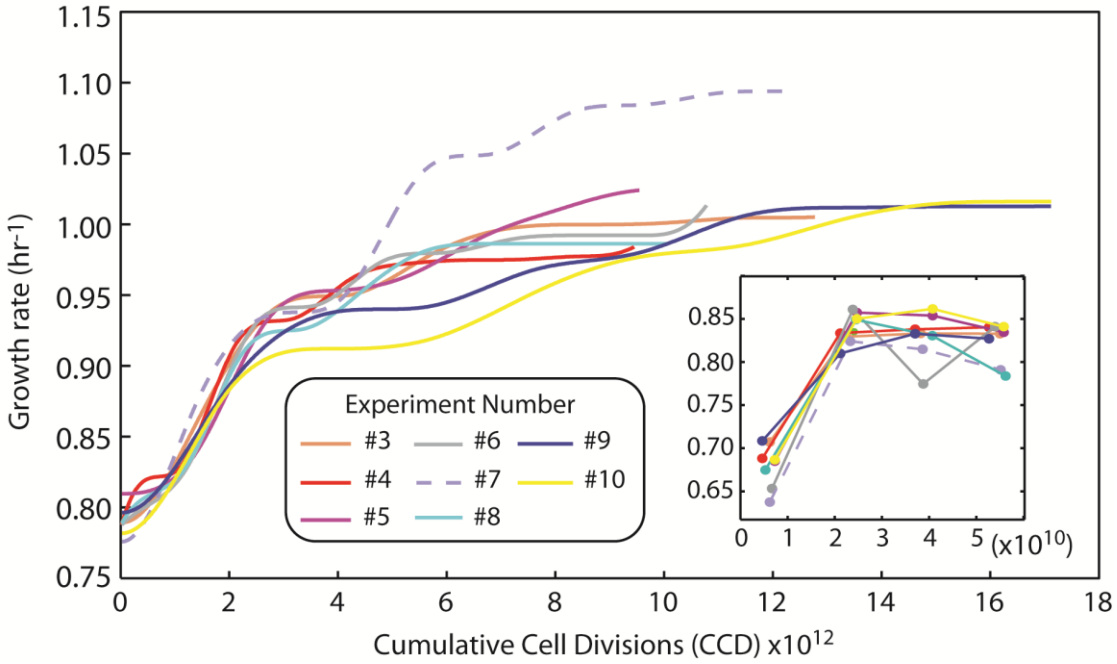

Figure 3.1: Fitness trajectories for E. coli populations evolved on glucose minimal media. Shown is a plot of the fitness (i.e., the growth rate) of the independently evolved experiments versus the number of cumulative cell divisions (CCD). The strain indicated with a dashed line was classified as a hypermutator. The insert shows the growth rates of the initial four flasks of batch growth in each experiment. Overall, the fitness of the hypermutator population outpaced the non-mutators.

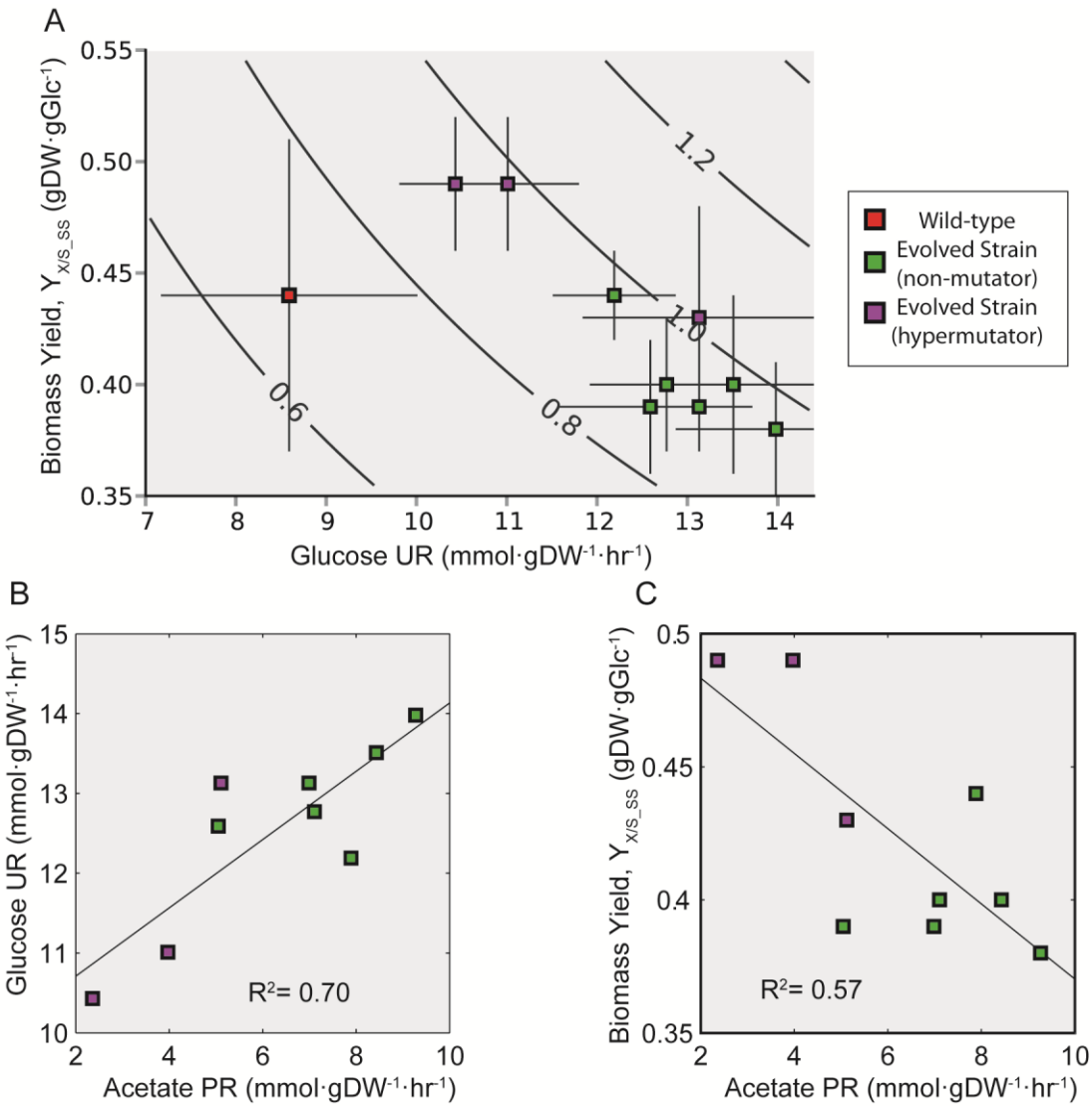

Figure 3.2: Phenotypic properties of evolved strains.
Clones isolated from the last flask of the experiments (i.e., endpoint strains of non-mutators) and three hypermutator strains were characterized phenotypically. (**A**) A plot of biomass yield versus glucose uptake rate (UR) (see Methods for calculations). The isoclines indicate different growth rates. Of all measured phenotypic traits for the evolved strains, the correlations between (**B**) glucose uptake rate and acetate production rate (PR), and (**C**) biomass yield and acetate production rate were the strongest. The percent of carbon from glucose being secreted in the form of acetate increased in all of the non-mutator endpoint strains (18-22%) except for one (13%), as compared to wild-type (15%). This percent decreased for all of the hypermutator strains (8-13%).

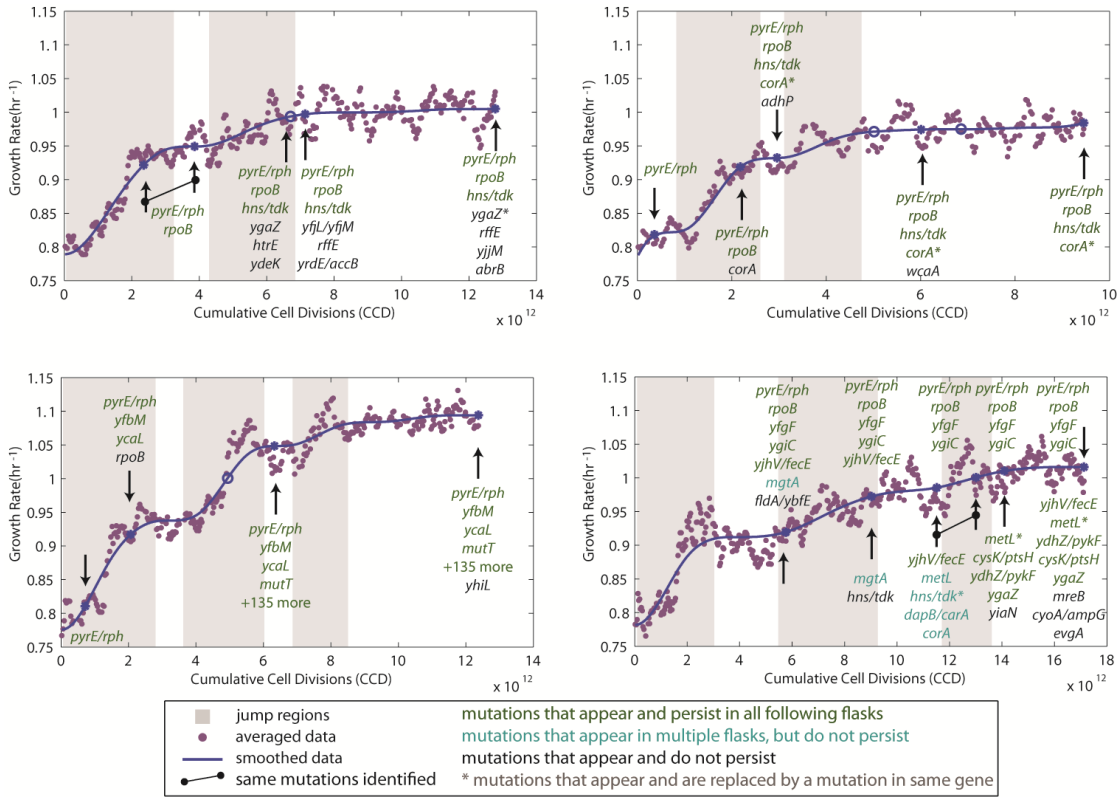

Figure 3.3: The fitness trajectories of ALE experiments 3, 4, 7, and 10 along with identified jump regions and resequencing data.
Shown is the fitness increases over the course of the evolution as a function of cumulative cell divisions (CCD) and the jump regions (grey boxes) identified using the outlined algorithm. Arrows indicate where colonies were isolated and resequenced. Mutations are categorized by color: those which occurred and were found in each subsequent colony resequencing (green), those which appear in colonies from multiple flasks but not consecutively (blue), and those which were only found in one particular clone and not in subsequent clones (black). Further, genetic mutations that replace a previously identified mutation in the same gene are marked with an asterisk. All of the mutations from the hypermutator strain that arose in experiment 7 are not shown (more than 135 total mutations).

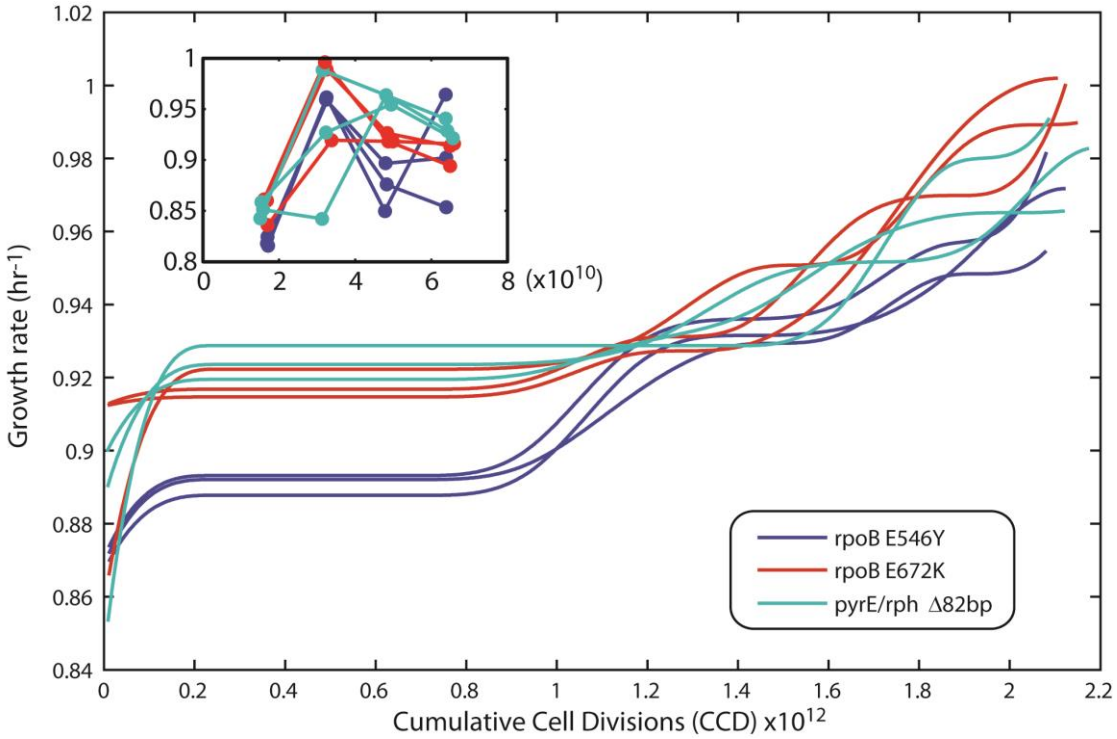

Figure 3.4: Fitness Trajectory for the Validation ALE
Shown is a plot of the validation ALE where three unique starting strains were evolved in biological triplicate, each harboring one of the following mutations: *rpoB* E546V, *rpoB* E672K, and *pyrE/rph* Δ82bp. The increase in fitness is shown as a function of the cumulative cell divisions (CCD). The insert shows the unsmoothed and filtered growth rates of the beginning of the experiment to show any possible physiological adaptation that is characteristic of ALE experiments. A smoothing spline will often obscure such abrupt changes.

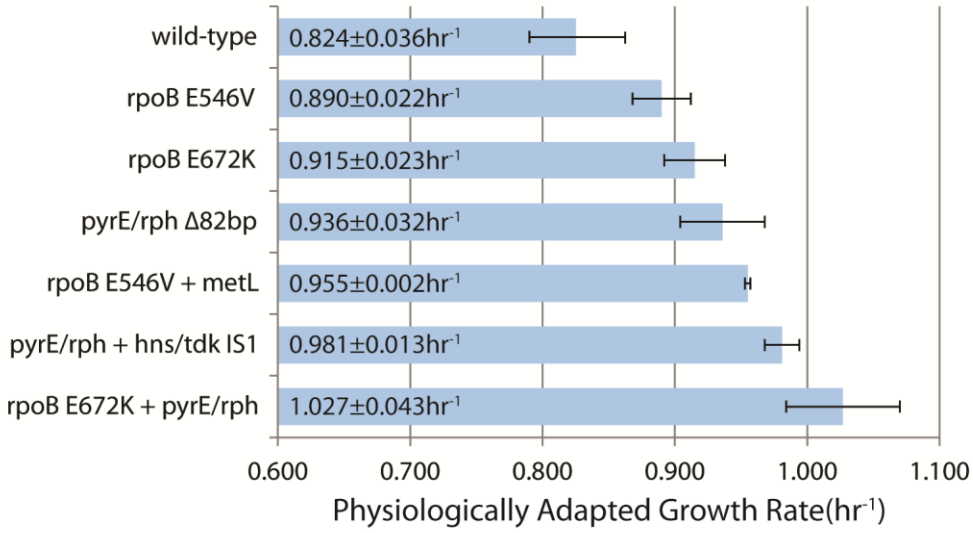

Figure 3.5: Causal Mutation Analysis
Shown is a bar graph of the physiologically adapted growth rates of strains harboring key mutations identified in this work. The error bars represent 95% confidence intervals from three biological replicates. This shows that the mutation in *metL* and the IS1 insertion between *hns/tdk* are causal in the presence of the additional mutations shown. The strain with *metL* also had one additional mutation, but this was not observed in any other sequenced *metL* mutant from the ALE experiment. It is clear from the fastest growing mutant, with growth 1.3 fold higher than the wild-type, how significantly the *pyrE/rph* and *rpoB* mutations can affect growth rate.

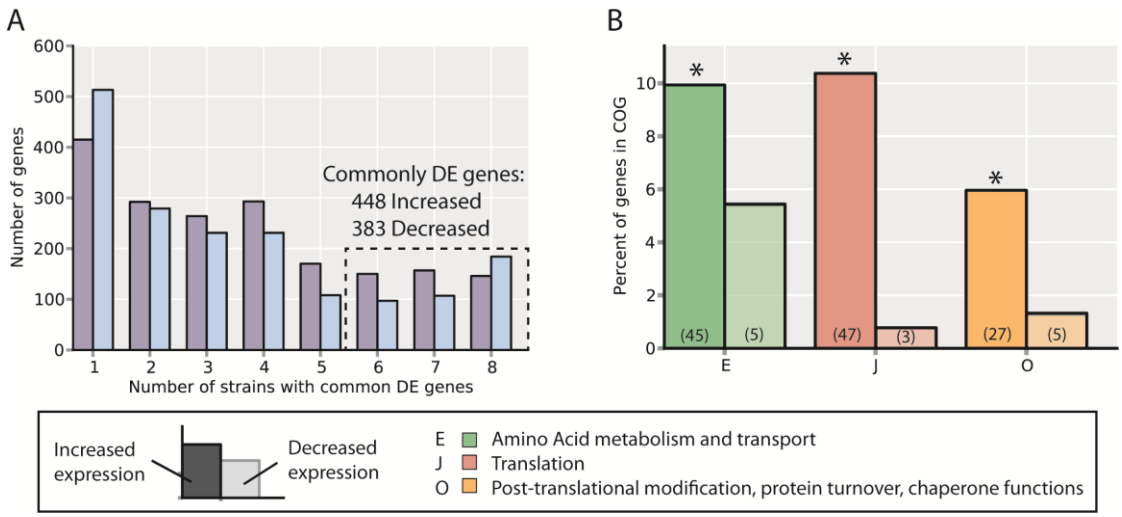

Figure 3.6: Commonly differentially expressed genes
(A) The number of differentially expressed genes (with respect to the wild-type strain) common across evolved strains is indicated. Increased and decreased expression genes are counted separately to ensure the direction of change is conserved across strains. The y-axis indicates the number of genes differentially expressed in exactly the number of strains indicated on the x-axis. From this, 448 increased and 383 decreased genes are identified as common to at least 6 strains, whereas one would expect no genes in common to all six by random chance. (B) The commonly differentially expressed genes' functions are interrogated using annotated Clusters of Orthologous Groups (COGs). COGs over-represented in either the up-regulated or down-regulated gene sets were identified with a hypergeometric test ($p<0.05$; see Methods). The percentage and number of genes for the identified COGs is indicated in the bar chart. Asterisk indicates over-represented.

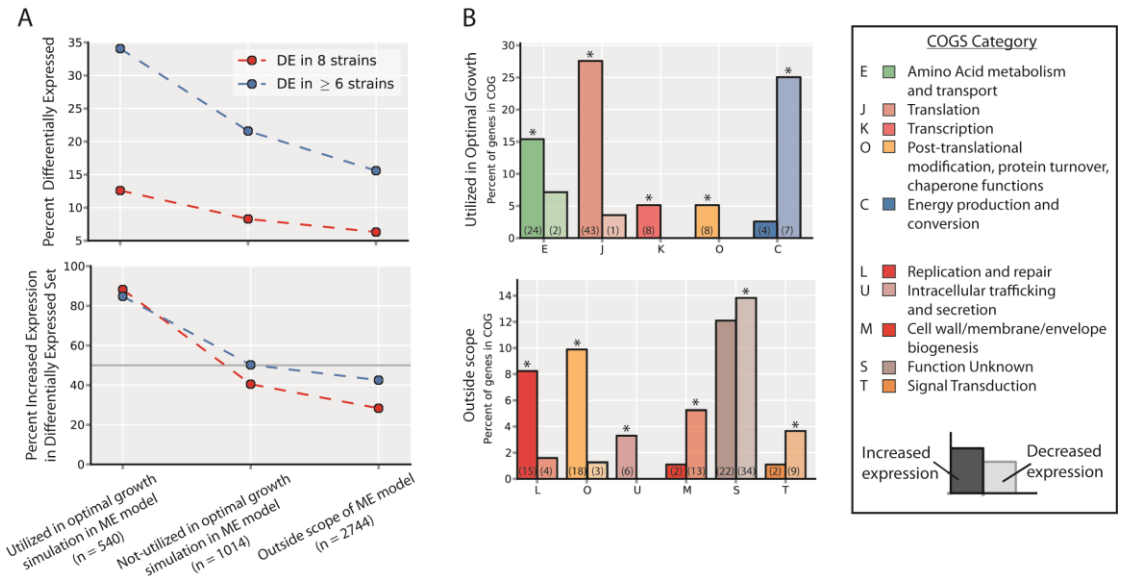

Figure 3.7: Comparison of genome-scale modeling predictions and categorization of commonly differentially expressed genes.
(A) The commonly differentially expressed genes were compared to a gene classification obtained by a genome-scale model of *E. coli* (38). Growth rate is optimized in the same glucose aerobic batch conditions as used in the ALE experiment. Simulation results can be used as an additional characterization of gene content (x-axis). Overall, differentially expressed genes are more enriched in the set of genes predicted to enable an optimal growth phenotype (top). Furthermore, within the differentially expressed set of genes, those which increased in expression versus wild-type are enriched within the predicted set of genes which enable an optimal growth phenotype (bottom). (B) Using the combination of *in silico* predicted genes and COGS for categorization, subsets of genes could be identified which enabled the observed optimal states of the evolved strains on the pathway level.

3.7 References

1. **Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS.** 2012. The molecular diversity of adaptive convergence. Science **335:**457-461.

2. **Dragosits M, Mozhayskiy V, Quinones-Soto S, Park J, Tagkopoulos I.** 2013. Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in Escherichia coli. Mol Syst Biol **9:**643.

3. **Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BO.** 2010. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet **6:**e1001186.

4. **Palsson B.** 2011. Adaptive Laboratory Evolution. Microbe **6:**6.

5. **Dragosits M, Mattanovich D.** 2013. Adaptive laboratory evolution -- principles and applications for biotechnology. Microb Cell Fact **12:**64.

6. **Reyes LH, Almario MP, Winkler J, Orozco MM, Kao KC.** 2012. Visualizing evolution in real time to determine the molecular mechanisms of n-butanol tolerance in Escherichia coli. Metab Eng **14:**579-590.

7. **Atsumi S, Wu TY, Machado IM, Huang WC, Chen PY, Pellegrini M, Liao JC.** 2010. Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. Mol Syst Biol **6:**449.

8. **Horinouchi T, Tamaoka K, Furusawa C, Ono N, Suzuki S, Hirasawa T, Yomo T, Shimizu H.** 2010. Transcriptome analysis of parallel-evolved Escherichia coli strains under ethanol stress. BMC Genomics **11:**579.

9. **Tremblay PL, Summers ZM, Glaven RH, Nevin KP, Zengler K, Barrett CL, Qiu Y, Palsson BO, Lovley DR.** 2011. A c-type cytochrome and a transcriptional regulator responsible for enhanced extracellular electron transfer in Geobacter sulfurreducens revealed by adaptive evolution. Environ Microbiol **13:**13-23.

10. **Jansen G, Barbosa C, Schulenburg H.** Experimental evolution as an efficient tool to dissect adaptive paths to antibiotic resistance. LID - S1368-7646(14)00004-1 [pii] LID - 10.1016/j.drup.2014.02.002 [doi].

11. **Conrad TM, Lewis NE, Palsson BO.** 2011. Microbial laboratory evolution in the era of genome-scale science. Mol Syst Biol **7:**509.

12. **Mozhayskiy V, Tagkopoulos I.** 2013. Microbial evolution in vivo and in silico: methods and applications. Integr Biol (Camb) **5:**262-277.

13. **Fong SS, Joyce AR, Palsson BO.** 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res **15:**1365-1372.

14. **Ibarra RU, Edwards JS, Palsson BO.** 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420:**186-189.

15. **Fong SS, Palsson BO.** 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet **36:**1056-1058.

16. **Conrad TM, Frazier M, Joyce AR, Cho BK, Knight EM, Lewis NE, Landick R, Palsson BO.** 2010. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. Proc Natl Acad Sci U S A **107:**20500-20505.

17. **Wiser MJ, Ribeck N, Lenski RE.** 2013. Long-term dynamics of adaptation in asexual populations. Science **342:**1364-1367.

18. **Farida Vasi MT, Richard E. Lenski.** 1994. Long-Term Experimental Evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. American Naturalist **144:**432-456.

19. **Vasi FK, Lenski RE.** 1999. Ecological Strategies and Fitness Tradeoffs in Escherichia coli Mutants Adapted to Prolonged Starvation. Journal of Genetics **78:**43-49.

20. **Deng Y, Fong SS.** 2011. Laboratory evolution and multi-platform genome re-sequencing of the cellulolytic actinobacterium Thermobifida fusca. J Biol Chem **286:**39958-39966.

21. **Quan S, Ray JC, Kwota Z, Duong T, Balazsi G, Cooper TF, Monds RD.** 2012. Adaptive evolution of the lactose utilization network in experimentally evolved populations of Escherichia coli. PLoS Genet **8:**e1002444.

22. **Herring CD, Glasner JD, Blattner FR.** 2003. Gene replacement without selection: regulated suppression of amber mutations in Escherichia coli. Gene **311:**153-163.

23. **Conrad TM, Joyce AR, Applebee MK, Barrett CL, Xie B, Gao Y, Palsson BO.** 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.

24. **Lee DH, Palsson BO.** 2010. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. Appl Environ Microbiol **76:**4158-4168.

25. **Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM.** 2009. Programming cells by multiplex genome engineering and accelerated evolution. Nature **460:**894-898.

26. **Hill SA, Little JW.** 1988. Allele replacement in Escherichia coli by use of a selectable marker for resistance to spectinomycin: replacement of the lexA gene. J Bacteriol **170:**5913-5915.

27. **Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO.** 2009. Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol **7:**129-143.

28. **Joyce AR, Palsson BO.** 2006. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol **7:**198-210.

29. **Schmidt BJ, Ebrahim A, Metz TO, Adkins JN, Palsson BO, Hyduke DR.** 2013. GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics **29:**2900-2908.

30. **McCloskey D, Palsson BO, Feist AM.** 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol **9:**661.

31. **Feist AM, Palsson BO.** 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol **26:**659-667.

32. **Shachrai I, Zaslaver A, Alon U, Dekel E.** 2010. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. Mol Cell **38:**758-767.

33. **Deatherage DE, Barrick JE.** 2014. Identification of Mutations in Laboratory-Evolved Microbes from Next-Generation Sequencing Data Using breseq. Methods Mol Biol **1151:**165-188.

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

35. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:**357-359.

36. **Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L.** 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28:**511-515.

37. **Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muniz-Rascado L, Ong Q, Paley S, Schroder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD.** 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res **41:**D605-612.

38. **O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO.** 2013. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol **9:**693.

39. **Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

40. **Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF.** 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res **17:**2503-2516.

41. **Lee DH, Feist AM, Barrett CL, Palsson BO.** 2011. Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of Escherichia coli. PLoS One **6:**e26172.

42. **Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM.** 2014. Evolution of Escherichia coli to 42°C and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations. Molecular Biology and Evolution.

43. **Portnoy VA, Herrgard MJ, Palsson BO.** 2008. Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient Escherichia coli strain. Appl Environ Microbiol **74:**7561-7569.

44. **Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF.** 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature **461:**1243-1247.

45. **Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

46. **Jensen KF.** 1993. The Escherichia coli K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. J Bacteriol **175:**3401-3407.

47. **Kobayashi M, Nagata K, Ishihama A.** 1990. Promoter selectivity of Escherichia coli RNA polymerase: effect of base substitutions in the promoter -35 region on promoter strength. Nucleic Acids Res **18:**7367-7372.

48. **Ayers DG, Auble DT, deHaseth PL.** 1989. Promoter recognition by Escherichia coli RNA polymerase. Role of the spacer DNA in functional complex formation. J Mol Biol **207:**749-756.

49. **Cox EC.** 1976. Bacterial mutator genes and the control of spontaneous mutation. Annu Rev Genet **10:**135-156.

50. **Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T.** 2010. Interdependence of cell growth and gene expression: origins and consequences. Science **330:**1099-1102.

51. **Wang W, Li GW, Chen C, Xie XS, Zhuang X.** 2011. Chromosome organization by a nucleoid-associated protein in live bacteria. Science **333:**1445-1449.

52. **Barker CS, Pruss BM, Matsumura P.** 2004. Increased motility of Escherichia coli by insertion sequence element integration into the regulatory region of the flhD operon. J Bacteriol **186:**7529-7537.

53. **Hall BG.** 1999. Transposable elements as activators of cryptic genes in E. coli. Genetica **107:**181-187.

54. **Umeda M, Ohtsubo E.** 1989. Mapping of insertion elements IS1, IS2 and IS3 on the Escherichia coli K-12 chromosome. Role of the insertion elements in formation of Hfrs and F' factors and in rearrangement of bacterial chromosomes. J Mol Biol **208:**601-614.

55. **Feist AM, Palsson BO.** 2010. The biomass objective function. Curr Opin Microbiol **13:**344-349.

56. **Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U.** 2012. Multidimensional optimality of microbial metabolism. Science **336:**601-604.

57. **Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN.** 2007. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc Natl Acad Sci U S A **104:**12663-12668.

58. **Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T.** 2012. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol **8:**e1002575.

59. **Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA.** 2005. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. EMBO Rep **6:**397-399.

60. **Lenski RE, Rose MR, Simpson SC, Tadler SC.** 1991. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. The American Naturalist **138:**1315-1341.

Chapter IV – Dynamics of Batch Culture Adaptive Laboratory Automation Experiments

4.1 Adaptive Laboratory Evolution

Adaptive laboratory evolution (ALE) has been performed *in vitro* for decades and the field is expanding. The basic principles governing ALE experiments are easily understood across a breadth of disciplines, which has led to its adoption in many laboratories (1, 2). The recent growth in the use of ALE can be attributed to the ease of access and decreasing costs of genome sequencing (3-5). Falling sequencing costs have led to the increased investigation of genomic, transcriptomic, and fluxomic changes over the course of evolution (5). While the analysis of ALE experiments has grown, the manner in which the ALE experiments themselves are performed has remained relatively *ad hoc*. The most commonly employed techniques are chemostat adaptation and batch culture adaptation, with batch culture adaptation being more popular as it is easily expanded and does not require setting up complex machinery (3, 6).

A primary attribute of any ALE experiment is the selection pressure imposed on the culture. The selection pressure (i.e., exponential growth, biomass yield, stationary phase, or lag phase) is responsible for the outcome of the evolution study (4, 7-10). For example, in a 24hr batch culture ALE experiment with fast growing bacteria, the culture is subjected to alternating environments of feast and famine. At the beginning of the batch there are excess nutrients but inevitably, within 24hrs, the nutrients are consumed and stationary phase is reached. Because of this alternating environment, the selection pressure is complex (e.g., stationary phase fitness, lag phase duration, and growth rate all contribute) (9). This complexity often confounds the analysis depending on the application. To alleviate complexity, the cells can be

kept in one phase (e.g. exponential phase) to mitigate most of the alternating selection and focus selection specifically on growth rate. The desired outcome of the experiment would dictate the ideal selection pressure to be imposed and thereby the experimental design, but the difference between the two designs is non trivial.

There are several parameters that affect the outcome of a batch culture ALE experiment. A primary parameter involved is the passage size (11-13). Specifically, passage size determines how much of the population is allowed to propagate to each subsequent batch culture. This is an integral parameter. If a beneficial mutation occurs, but is lost when the bottleneck is imposed, the rate of evolution can be slowed or even halted. Since smaller passage sizes can hinder the rate of evolution, it is often easier to perform a batch culture ALE under alternating environments of feast and famine where a change in passage size only effects the duration of growth and stationary phases. However, if the application requires exponential phase passaging, a change in passage size also changes the time when the culture must be passaged. Because of this, the passage size is often dictated by an individual's schedule. Typically, the time in between passaging can be no shorter than ~12hrs. Consequentially, as the culture adapts and begins to grow faster, the passage size must be decreased. As an example, a previous study adapting *E. coli* to glycerol started with a passage size of approximately 100μL and by experiment's end was less than 0.1μL (14). A more in-depth retrospective analysis revealed similar trends where passage amounts were significantly decreased (14-18). In these studies, the bottleneck became so small that the calculated number of cells being passed was on the order of 10^1 or even occasionally 10^0. The chance of capturing a beneficial mutation, when only passing

10^1 cells from a culture of millions, is practically null over a reasonable timeframe. At this point, continuing the experiment is futile. The question then becomes at what point is the passage size too low?

Passage size can have a large impact on the trajectory of an ALE experiment. This can be seen in the comparison of two studies that evolved wild-type *E. coli* K-12 MG1655 on M9 glucose minimal media (7, 18). One study (7) used a consistent passage size of 800µL on an automated platform. The second study (18) was done "by hand" and had widely varying passage sizes that were considerably smaller than the automated study. The outcomes of the ALE experiments were quite distinct. The final growth rates achieved were 1.00±0.24 hr^{-1} and 0.79±.01 hr^{-1} in the consistent and variable passage size studies, respectively. The apparent lack of fitness achieved in variable passage study was not due to a lack of available beneficial mutations (as the same strains and culturing conditions were used), but rather insufficient experimental design to find and fix them in a reasonable amount of time. Understanding why these two outcomes differ is imperative to the efficient design of ALE experiments.

Theoretical studies have looked at the effect of passage size on batch culture adaptation and resulted in varying predictions of an ideal passage size depending on the model used (19, 20). The ideal passage sizes calculated are ideal from a mathematical standpoint. This essentially gives the best chance for various mutations of different selective advantages to fix themselves in a population. The values calculated are relatively large (13.5% and 20%)(19, 20). As mentioned previously, a larger passage necessitates an increase in resources. More specifically, the resources required increase exponentially with passage size, yet the gains slowly diminish. This

work thus focuses on examining the diminishing returns in the context of the desired result and the resources available. We set out to examine the impact of the key ALE parameter: passage size. To address this, we created an *in silico* evolutionary model that simulates the dynamics of capturing and fixing beneficial mutations in the context of an exponentially-passed batch culture ALE experiment. After building the model, we parameterize it using a combination of 30 independent ALE experiments of *E. coli* on glycerol minimal media across five different passage sizes (10%, 1%, 0.1%, 0.01%, and 0.001%). Using the parameterized model, we investigated the biological consequences of changing passage sizes and how close to optimal a given experiment is. With this knowledge, an experiment can be designed to fit the desired outcome, giving consideration to the resources required to achieve it, and the feasibility of performing such an experiment.

4.2 Materials and Methods

4.2.1 Adaptive Laboratory Evolution

Adaptive laboratory evolutions were started from wild-type *E. coli* strain MG1655 (ATCC47076) glycerol frozen stock and grown up overnight in 15mL magnetically stirred 0.2% glycerol M9 minimal media supplemented with trace elements. The magnet was stirred at 1150rpm, sufficient for completely aerobic growth. 30 experiments were started from 150µL aliquots from the overnight pre-culture. The experiments were subsequently grown in identical vessels and media as the pre-culture. Culture optical densities at 600nm (OD) were monitored over the course of each batch culture. When the culture reached an OD of 0.300 (±10%) as measured by a plate-reader with 100µL sample volume in a 96 well flat bottom microplate, an aliquot was taken and passed to a new batch culture filled with sterile media. An OD of 0.300 was chosen to preclude reaching stationary phase in any of the cultures and ensures OD measurements have not begun to saturate. Growth rates of each culture were determined using OD measurements taken over the lifetime of each batch culture.

4.2.2 Media

All cultures were grown in 0.2% glycerol M9 minimal media. The media consisted of 0.2% glycerol by volume, 0.1mM $CaCl_2$, 2.0mM $MgSO_4$, Trace element solution and M9 salts. 4000X Trace element solution consisted of 27g/L $FeCl_3*6H_2O$, 2g/L $ZnCl_2*4H_2O$, 2g/L $CoCl_2*6H_2O$, 2g/L $NaMoO_4*2H_2O$, 1g/L $CaCl_2*H_2O$, 1.3g/L $CuCl_2*6H_2O$, 0.5g/L H_3BO_3, and Concentrated HCl dissolved in ddH_2O and sterile

filtered. 10x M9 Salts solution consisted of 68g/L Na_2HPO_4 anhydrous, 30g/L KH_2PO_4, 5g/L NaCl, and 10g/L NH_4Cl dissolved ddH_2O and autoclaved. Final concentrations in the media were 1x.

4.2.3 DNA Sequencing

Genomic DNA was isolated using Macherey-Nagel NucleoSpin® Tissue kit. The quality of DNA was assessed with UV absorbance ratios using a Nano drop. DNA was quantified using Qubit dsDNA High Sensitivity assay. Paired-end resequencing libraries were generated using Illumina's Nextera XT kit with 700 pg of input DNA total. Sequences were obtained using an Illumina Miseq with a MiSeq 600 cycle reagent kit v3. The breseq pipeline version 0.23 with bowtie2 was used to map sequencing reads and identify mutations relative to the *E. Coli* K12 MG1655 genome (NCBI accession NC_000913.2) (21). All samples had an average mapped coverage of at least 25x.

4.2.4 Computer Modeling

Modeling of simulations was computed using MATAB 2015b on a Windows 7 professional platform. The BMR was computed by a maximum likelihood estimation. It was calculated for making a transition from State 1 to State 2 and State 2 to State 3 for passage sizes of 0.01% and 0.001%. These passage size were chosen as they were the only ones that showed a distribution of states achieved. The transition from State 1 to State 2 was capped at 20 days to give a maximally distributed data set. The transition from State 2 to State 3 was started by assuming that State 2 was already achieved. Thus, the length of time simulated was started based of when State 2 was achieved. This was variable for different experiments.

A value of $1.55x10^{12}$ cells·L^{-1}· OD_{600nm}^{-1} was used to estimate the number of cells in a culture for a given OD_{600nm} with a 1 cm path length cuvette for the purposes of ALEsim. A standard curve relating the ODs measured in the plate reader with a 100µL sample volume in a 96 well flat bottom microplate to the OD measured with a 1 cm cuvette to obtain a ratio of 3.15 for equivalent measurements between the two. The biomass (grams of dry weight) per OD_{600nm} per volume was calculated by filtering known volumes of cultures at specific ODs though 0.22µm filters. The filters were weighed before and after filtering and drying to obtain the total dry weight of the culture. The differences in these values was used to calculate ratio of 0.45·gDW L^{-1}·OD_{600nm}^{-1}. The dry mass per cell has previously been reported as 2.9×10^{-13}gDW·cell (22). The quotient of these two values gives our final conversion factor of $1.55x10^{12}$ cells·L^{-1}·OD^{-1} to estimate the cell counts of cultures at various ODs and volumes.

4.3 Results

4.3.1 Modeling the ALE process

ALEsim is a model built on the basic principles of exponential growth in order to understand the dynamics of ALE. The scope of ALEsim is to predict the observed growth rate in each batch culture of an ALE experiment while allowing individual cells to change their growth rate when dividing (i.e., a proxy for receiving a beneficial mutation). The observed population growth rate is different from a clonal growth rate in that each batch culture of an ALE experiment is a population of multiple clones with varying growth rates. Figure 4.1 provides a workflow of the modeling process. Each *in silico* experiment begins with a clonal inoculation of a strain with a given growth rate. A population of mixed phenotypes can be used in this framework, but here the starting population will be assumed to be isogenic with the same phenotypic behavior. This organism is allowed to replicate according to an exponential growth function. During each cell division event, there is a probability that it will mutate and start a new lineage with a mutated growth rate. This new lineage is allowed to grow alongside the parent strain according to exponential growth, but with its mutated growth rate. The new lineage is itself allowed to continue mutating in the simulation.

Mutated growth rates in ALEsim must be constrained to remain biologically meaningful, i.e., growth rates that are of magnitudes that remain plausible. These rates are determined empirically by the user, as done here from the parameterization experiment (see section below). The growth rates can be constrained to allow various types of epistasis. For example, if two distinct growth rates are allowed, there is a

possibility that a single cell line could mutate twice and receive both of these mutations. ALEsim employs the flexibility to define the type of epistasis between these two mutations, if any epistasis at all is to occur. Similarly, an order to the mutations accumulated can be set, as certain mutations can be beneficial only in the presence of a pre-existing mutation. As the population of cells continues to replicate and mutate, their total cell count naturally increases. When the cell count reaches a given threshold, a simple random sample of cells is used to inoculate the next batch culture. The threshold corresponds to a target cell count at which to passage the cells to the next batch culture. The number of cells taken is determined by the passage size, which is a percentage of the total culture volume. After this sample is computed, a new batch culture is started with the chosen cells and corresponding growth rates. Figure 4.2 provides the key parameters of the model.

In using the basic principles of microbial growth and the brute for nature of forcing competition, many of the fundamental attributes of natural selection intrinsically contained in the simulation. This includes clonal interference which is integral to asexual evolution. ALEsim can be used to model a system where two local maxima are possible but the greater maximum can only be found by first acquiring a mutation that is initially suboptimal compared to other possible single beneficial mutations. The experimental parameters can be modulated to potentially find an experiment design that would find the desired optimum or both.

Given the stochastic nature of many steps in the model, the results are non-deterministic. Stochasticity is incorporated into the model in three ways: i) when a cell mutates its growth rate, ii) what growth rate a cell mutates to, and iii) what sample of

cells are propagated to a subsequent batch culture. The simulation is then run multiple times to capture the dynamics of the stochasticity (23).

For a simulation to be biologically meaningful using the developed model, there are three types of parameter sets that must be determined. The first set of parameters is experimental: batch culture size, passage size, passage optical density (or cell count), and length of experiment. These can be set based on the desired experimental setup. The optical density is used as a proxy for cell count when mimicking an experimental design in ALEsim. It is understood that for *E. coli* the relationship between OD and cell count is not constant over a range of growth rates. Though not constant the relationship is known (24). Using an OD to cell count factor as a function of growth rate is possible with ALEsim but incurs a marked increase in simulation time over a constant. Identical simulations but only varying the constant show that the difference between the most extreme differences in OB to cell count factors only yielded a 10% difference in outcomes. Thus using a constant average value for the range of growth rates expected was determine to be sufficient considering the benefit in computation time.The second set of parameters is the statistical parameters: random number seed and the number of identical experiments to run. The random number seed is set by the native random number generator. The number of parallel simulations to run is determined by the statistical power needed. Depending on the magnitudes and complexities of the parameters set, the number of simulations can vary drastically. For the results shown here, 500 simulations were computed unless otherwise stated. It was found that after 500 simulations there was no appreciable difference in the means or spread of the distribution of results calculated

when combined with another set of 500. The third set of parameters is biological: beneficial mutation rate (BMR) and allowed increases in growth rate. This set of parameters must be derived experimentally. Intuitively, these parameters can be different for different strains, conditions, and can even change along the course of a single experiment (25, 26). As long as the values determined are biologically meaningful, generalizations about the ALE process can be concluded.

Alternative models of evolution and adaption have been developed to understand the dynamics of evolution. These types of mathematical models capture various aspects of adaptation including selection, drift, and clonal interference (27-29). Classically, this has been a target of the field of population genetics (30-32). An expansion of the Fisher model was developed by Wahl et. al. which conceptually relates to ALEsim in that it targets the question of passage sizes (33). However, ALEsim deviates from the classical mathematical approach and employs the use of an *in silico* organism that can then replicate, mutate, and evolve. Simulations here are carried out in brute force where they are allowed to grow under the conditions laid out by the user. The advantage of such a method is that the experimental and biological parameters can be strictly controlled over the course of an experiment. The resulting simulation is able to more closely mimic the conditions of an actual laboratory evolution experiment in its entirety where parameters are not always constant throughout. This approach differs from the use of a digital organism in that it is an attempt to model specific biology instead of general evolutionary dynamics which allows for direct modeling of the ALE experiment as would be performed in a laboratory (34).

4.3.2 Parameterization of ALEsim by evolving E. coli on Glycerol Minimal Media

The two biological parameters, BMR and allowed increase in growth rate, were determined using 30 independent cultures of *Escherichia coli* K-12 MG1655 evolved in 15mL of 0.2% glycerol M9 minimal media until a stable growth rate was observed in most experiments (38 days). One experiment only lasted 23 days after it was restarted due to contamination. The 30 experiments were separated into five groups of six passage sizes and each group was evolved under identical conditions except for the passage size. The passage sizes used were 10%, 1%, 0.1%, 0.01%, and 0.001% of the culture size (15mL). The growth rate of each experiment was monitored over the course of the experiment using optical density measurements as a proxy for cell count (Figure 4.3).

Allowed increases in growth rate were determined by identifying jumps in growth rates from the fitness trajectories. A spline was fit to the growth rate of each experiment and significant increases in growth rate were identified as discussed previously (7). The resulting jumps in growth rates showed that the plateaus in growth occurred at specific values (Figure 4.3, 4.4). These plateaus are identified as State 1, 2, 3A, and 3B. State 3 was split into two sub-states since there is an obvious distinction between the lower and upper growth rates but there is no obvious gap between them. This gap is most likely obscured since the difference between the growth rates is fairly small and noise in the measurements can bleed into any gap that might exist. Figure 4.4 groups the jumps in fitness observed by their transition between states. Contrary to the conclusion of other ALE experiments, the largest jump in fitness was not observed first but actually followed a smaller jump. This yields an allowed increase in growth

rate that can be used to constrain ALEsim. In simulations run here, the growth rates allowed were set to the mean of the range of each state.

The BMR can be calculated by fitting ALEsim to the distribution of the end states. Passage sizes of 10% - 0.1% did not show any appreciable variation between states, thus only the experiments with passage sizes of 0.01% and 0.001% were used for fitting. ALEsim was fit by performing simulations that only allowed for a single jump from one state to another. Multi-state jumps and two sequential jumps were not allowed. This simplification skews the BMR calculation to only include beneficial mutations that were able to fix themselves. There is a potential that other beneficial mutations are possible, but were not observed due to either clonal interference or genetic drift (35). As observed in the fitness trajectories for passage sizes of 0.01% and 0.001%, not all experiments were able to make jumps to occupy all the states. For instance, with a passage size of 0.01%, only 4 of 6 experiments were able to make the transition from State 2 to State 3 by experiment's end. In simulation, the same propensity to distribute among the various end states is observed. The distribution observed in simulation is highly dependent on the supply of beneficial mutations captured by the BMR parameter. Thus, the BMR can be fit to yield the same distribution across states as observed experimentally. The BMR was computed using transitions from both State 2 to State 3 and from State 1 to State 2. Since all experiments made the transition from State 1 to State 2, the distribution was used at the day 20 mark where a distribution existed. The 95% confidence interval for the BMR was calculated by fitting the BMR to the 95% confidence interval of the experimental distribution of states. The results yielded a BMR of $10^{-6.9}$-$10^{-8.4}$

mutations per cell division. The confidence interval was determined by a maximum likelihood estimate as implemented in the binofit function in MATLAB.

4.3.3 Retrospective Validation of ALEsim

ALEsim and the derived parameters were analyzed using two previously performed ALE experiments on glucose (7, 18) and a legacy experiment on glycerol (14). The outcomes of the two glucose experiments yielded disparate final growth rates despite identical strains and media (*E. coli* K-12 MG1655 in M9 glucose minimal media), 1.00±0.24 and 0.79±.01, respectively. The only differences between the experiments were three experimental parameters: batch culture volumes (250 mL vs. 25 mL), optical densities when passed (variable vs. OD_{600nm} 1.2), and passage sizes (variable vs. 800µL) in the Charusanti et al. (18) and the LaCroix et al. (7) studies, respectively. ALEsim was constrained to allow only the jumps in growth rates observed in these studies and then simulated the expected fitness trajectories for the two different experimental parameters. The results showed that the difference in the final growth rates achieved can be sufficiently explained by the differences in experimental design only (Figure 4.5). Furthermore, when simulating a legacy dataset for evolving *E. coli* on glycerol minimal media, ALEsim was able to successfully predict that all experiments (n=4) should reach fitness state 3 for the given experimental parameters, as reported in the study (14). The largely different outcome in fitness (i.e., no fitness jumps vs. a significant increase) on glucose, as well as a consistent prediction of fitness on a legacy glycerol dataset, further highlights the importance of properly designing an experiment and validates ALEsim and its parameterization.

4.3.4 ALEsim Applications

Simulations of ALE experiments with the derived BMR and fitness states can allow statements to be made about optimality. The time required to see a given increase in fitness was simulated for a range of increases in growth rate over a range of passage sizes (Figure 4.6). The results show the average length of time needed to see a given increase in fitness fix in the population over a range of passage sizes. This accounts only for growth rate increases that occur from a single mutational event. Based on the passage size and length of time with no increase in growth rate, a conclusion about how close a population is to optimum can be made. For example, if a given evolution experiment has achieved a certain growth rate, μ, and has not shown an increase in growth rate with a passage size of 0.1% for 13 days, then there is no likely increase in growth rate available at greater than $0.10hr^{-1}$ with a single mutational event.

Increasing the passage size raises the probability of capturing a beneficial mutation however this also leads to an inflation in the resources needed to sustain the experiment (Figure 4.6). For example, if an ALE experiment with a passage size of 0.1% were being passed twice a day (every 12 hours), the same experiment with a passage size of 10% would need to be passed 6 times per day (every 4 hours). A single person can feasibly do an experiment passed every 12 hours whereas passing every 4 hours would require coordinated effort by multiple persons or an automated platform. Therefore, understanding what is gained with the larger passage size is important before committing to such a large expenditure of resources. ALEsim can quantify the

gains or losses achievable with different passage sizes to help identify the ideal experimental setup (Figure 4.6).

4.3.5 Mutation Frequency Analysis by Passage Size

Clones from the endpoint populations of each independent experiment were isolated and resequenced. Two clones showed hypermutating tendencies. This was identified by the number of mutations ($p<0.01$) and the presence of a mutation in *mutY* or *mutL*. Experiments with larger passage size led to an increase in the number of mutations found. Mutated genetic regions were therefore grouped by passage size. Clones isolated from larger passage size experiments, on average, had more genetic regions mutated (Figure 4.7). Of all mutations identified, those in *glpK* were specifically tracked. Mutations in *glpK* have previously been shown to be highly causal as well as ubiquitous, mutating more than any other genetic region under glycerol growth conditions (14). Thus *glpK* is a good indicator of the fixation strength of the varying passage sizes. Consequently, there is a positive relationship between the fixing of *glpK* mutations and the passage size until saturation is reached. With the passage size dropped to the lowest value (0.001%), the observed fraction that fixed was only 0.33 (2/6).

4.4 Discussion

The conceptual purpose of an ALE experiment is to move an organism towards a more optimal (fit) state in the presence of a selection pressure. Absolute optimality is difficult, if even possible, to define. It has been shown that even for a laboratory evolution, there is still room for evolution after 50,000 generations (36). The continual ability of organisms to evolve and innovate makes it difficult to analyze the results of an ALE experiment in the context of optimality. What is immediately apparent is that there are diminishing returns. As an ALE experiment progresses, the increase in growth rate or fitness tends to decrease in magnitude (1, 37-41). The smaller increases take longer lengths of time to occur and become fixed in the population. Given this property and the desire to understand and leverage the ALE process, ALEsim was built and validated through performing a control experiment. ALEsim was first parameterized with a set of control experiments using different passage sizes. Parameterization revealed a beneficial mutation rate of $10^{-6.9}$-$10^{-8.4}$ mutations per cell division, consistent with previously reported values and distinct fitness states (25, 26). Validation was then carried out using additional legacy experiments and ALEsim proved sufficient for explaining the differences in observed experimental outcomes (i.e., growth rates) based on the parameters employed in each study (i.e., passage size, passage OD, and culture volume) (Figure 4.5). Lastly, ALEsim was applied to quantify tradeoffs in experimental design considerations for desired outcomes and was used to demonstrate how it can be leveraged for determining the key aspect of experiment termination.

The ability to optimize and design ALE experiments is possible with the ALEsim computational framework. Given a certain amount of resources, ALEsim can calculate how best to deploy them at different stages of an experiment to shorten project timelines and achieve desired outputs. For example, near the beginning of the ALE experiments, the increases in growth rates found are typically quite large. Because of this, a large passage size does not have an additional benefit. This is evident in the experiment performed here in that passage sizes of 0.1%, 1%, and 10% mostly reached states 1, 2, and 3A at about the same time (Figure 4.3). In planning future ALE experiments, the added resource usage needed to maintain an experiment at a 10% passage size does not appear to be justified. However, the added benefits become apparent when looking at the transition from state 3A to 3B. It could then be suggested that if the absolute optimal state is desired, the added resources of maintaining a 10% passage size experiment only need to be maintained after initial large increases in growth rate or fitness are found. This would not eliminate the difficulty in maintaining such an experiment, but would at least reduce the length of time the experiment would need to be run at such a high resource 'burn' rate. With ALEsim, these types of resource/fitness tradeoff analyses can now be calculated and should be leveraged in experimental design. The approach of dynamic resource allocation opens the door for project optimization typical of engineering process design.

Knowing the distance to optimality can aid in determining when to terminate an ALE experiment. The typical method of determining when to stop an ALE experiment is to subjectively determine that no more increases in fitness are being

observed. However, this approach of waiting to observe a plateau in fitness can be artificial given a small passage size. An example of how this approach can be misleading is the observation that passage sizes of 0.1% and 1% showed no increase in growth rate after reaching state 3A for at least 15 days (Figure 4.3). However, given that slight increases in growth rates beyond state 3A to state 3B with a passage size of 10% were observed, it can be concluded that state 3A is not the optimal state. Thus, if only a 1% passage size was used, the experiment could be terminated before finding state 3B. Further, it would be incorrect to compare experiments with a 10% passage size to a 1% passage size without understanding the context of the effects of the different passage sizes. Perhaps the best example of this is provided through the analysis of legacy ALE experiments (Figure 4.5). Two experiments with the same strain and media conditions yielded vastly different fitness outcomes. This difference is subsequently explainable within the scope of ALEsim. Therefore, having access to a computational framework such as ALEsim can enable the researcher to make an informed decision about when to terminate an experiment given the capacity and resources of the experimental setup and the desired/acceptable outcome. This type of termination analysis is laid out in Figure 4.6 and can be calculated *de novo* for any experiment given the current growth rate and passage size. It also should be noted that this type of analysis could result in a standard for the ALE community as one could state the ALEsim generated $\Delta\mu$ at the time of termination.

The ability to design and carry out complicated and high resource burn ALE experiments is likely only feasible though automation of the ALE process. Automation was utilized here and in previous studies (4, 7, 42). Manual processes are often

hindered by researcher availability whereas machines can measure and pass around the clock (e.g., approximately 5-7 passages per day were performed in automated studies (4, 7, 42), compared to 1-2 per day manually (14, 15, 18). Thus, the ability to automate and optimize ALE is likely to accelerate adoption of the ALE experimental technique and broaden the application areas. Furthermore, the ALEsim framework and output can also be used as a basis for modeling much of the legacy data currently available for ALE experiments which include lag, exponential, stationary, and/or stressed phases. As the selection pressure in such experiments is more complex and growth is defined by more than the growth rate parameter (e.g. lag phase duration, stationary phase mutation rate, growth phase transistions, etc…), ALEsim in its current format would have to be expanded. Nonetheless, ALEsim and it parameterization here demonstrates the utility of using simulated design in the ALE process and establishes a portable code base.

The field of adaptive laboratory evolution is expanding, largely due to lower costs of next generation sequencing. Innovative applications are appearing and are being applied to a range of organisms (1, 3). This growth in ALE use has occurred without a standard operating procedure for performing and quantifying these experiments. Consequently, this leads to ill-defined endpoints of experiments and the inefficient use of resources. The ALEsim computational platform developed here would provide a basis with which to quantify experiments and aid in their design; matching the desired outcome with resources available.

Chapter IV, in full, is a reprint that the dissertation author was the principal researcher and author of. The material has been submitted to *Applied and Environmental Microbiology*. (LaCroix, RA, Palsson, BO, Feist AM. 2016. Designing Adaptive Laboratory Evolution Experiments).

4.5 Figures

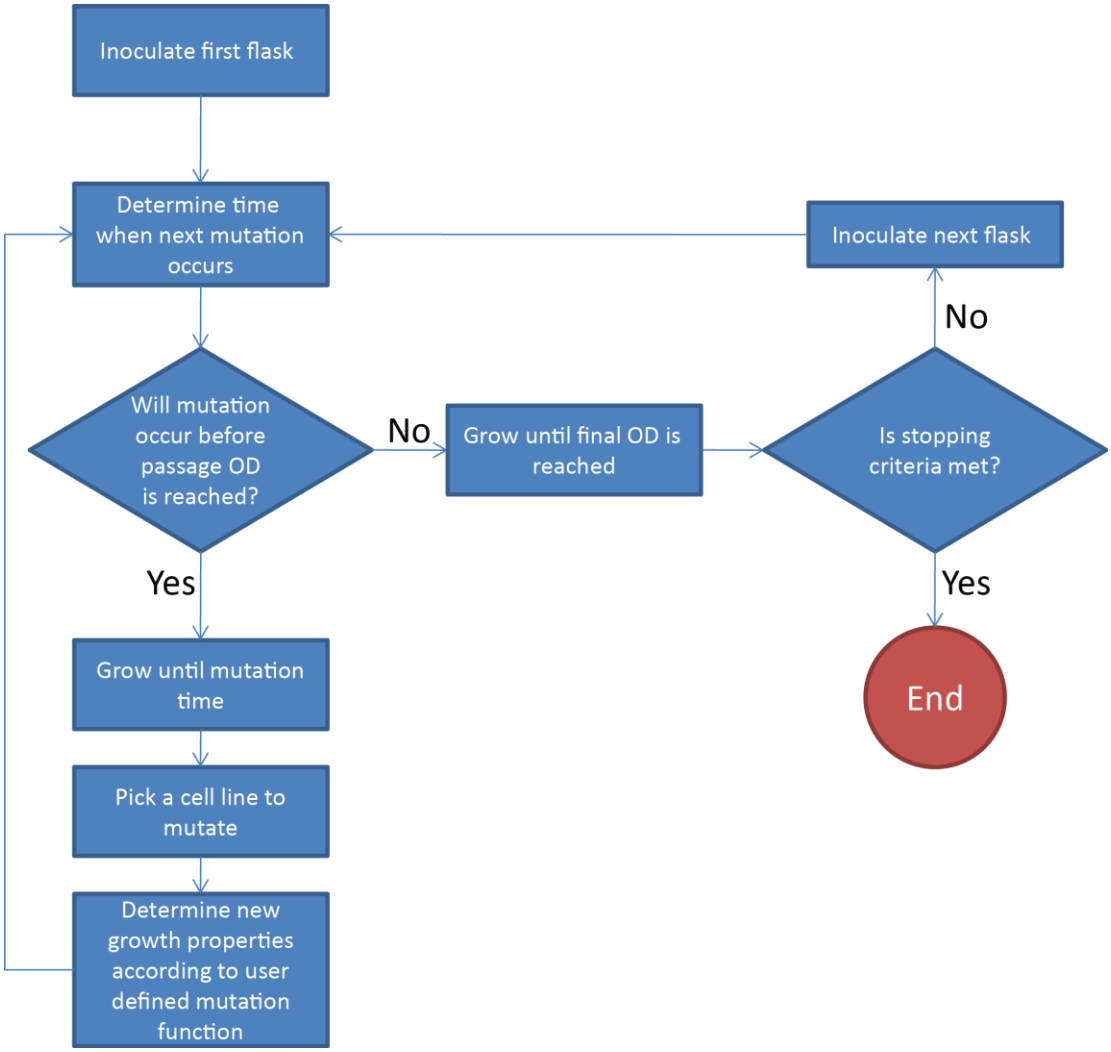

Figure 4.1 - ALEsim Flow Chart
A workflow outlining the logical steps the simulator takes when performing a single simulated ALE experiment. Due to the stochastic nature of ALE experiments, *in vivo* and *in silico*, multiple experiments are averaged together to identify general trends.

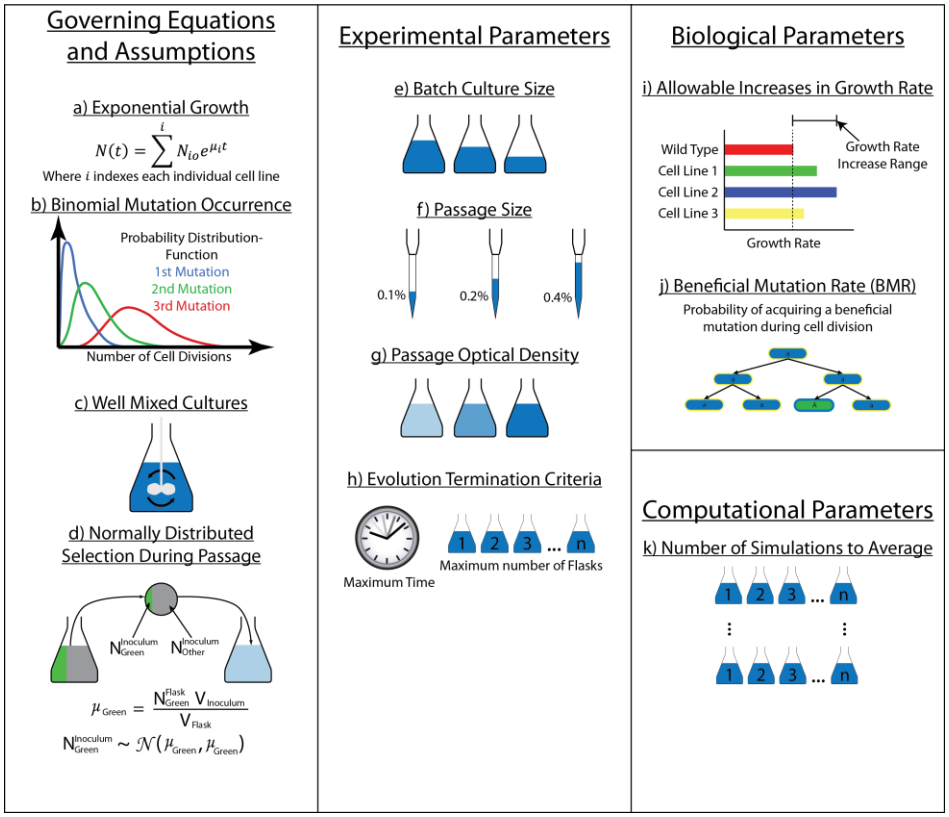

Figure 4.2 - Governing Equations, Assumptions, and Parameters for ALEsim
a) Microbe growth occurs according to an exponential growth curve where μ is the growth rate, t is the time elapsed, N_0 is the initial cell count at t=0, and N(t) is the cell count at a given time, t. No lag phase or stationary phase is modeled. The total cell count (N(t)) is determined by the summation of exponential growth curves for all individual cells lines. b) Favorable mutations occur during cell growth according to a binomial distribution where each cell division represents one Bernoulli trial with a probability of success equal to the beneficial mutation rate (BMR). c) Each flask is modeled as a completely homogenous culture. d) The number of cells represented for each cell line in each inoculum is randomly chosen according to a normal distribution with a mean and variance equal to the number of cells represented in the flask, N_{Green}^{Flask} times the ratio of the flask volume, V_{Flask}, to inoculum volume, $V_{Inoculum}$. e-g) The volume of media per flask, inoculum volume, and passage optical density can be altered. h) The simulated ALE experiment can be stopped after a specified amount of time or maximum number of flasks. i) Based on the relative growth rate increases seen in ALE experiments, a range of allowable growth rate increases is determined. j) Based on matching the evolution trajectory (plot of growth rate vs. flask #) with varying the beneficial mutation rate (BMR), the probability of a favorable mutation is obtained. k) Since each ALE is based on randomly generated mutations, multiple ALE simulations are averaged together to get repeatable results from the same parameters. The number of simulations is user controlled.

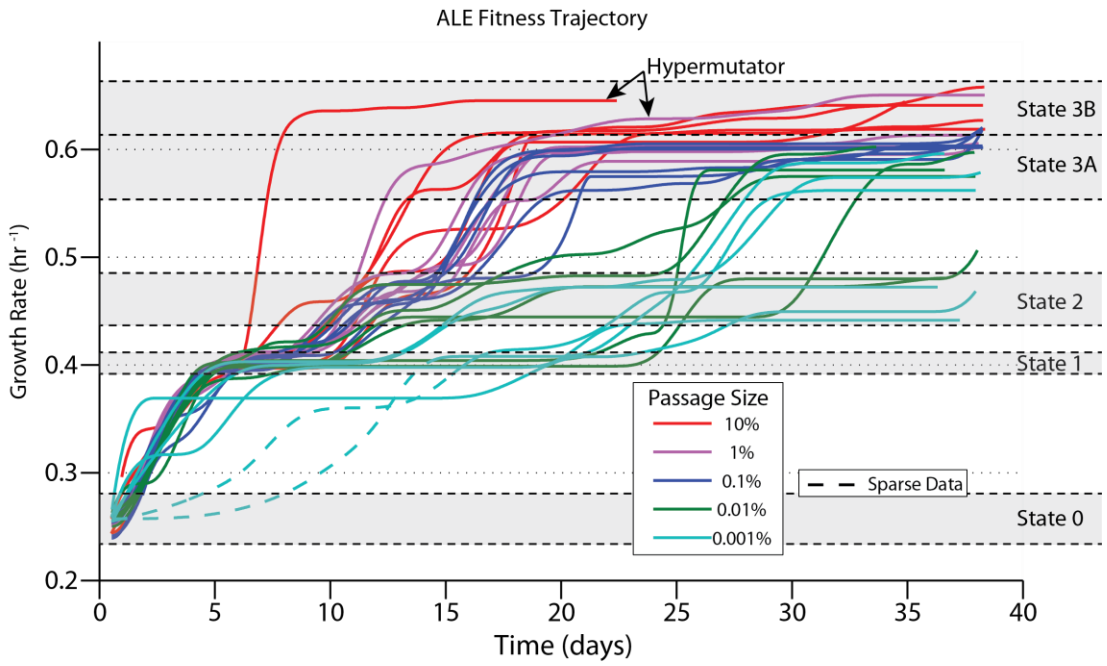

Figure 4.3 – Fitness Trajectory of *E. coli* evolved on Glycerol
The absolute growth rates of independently evolved cultures of *E. coli* as fitted by a cubic spline for all ALE experiments separated by the different passage sizes. Dashed lines represent regions where the spline fit is based on sparse data, and therefore not considered accurate. The small upturn in growth rates at the endpoint is an artifact of the spline interpolation and is ignored when determining endpoint growth rates. All except five ALE experiments reached fitness State 3. The rate at which the final growth rate was achieved varied. The hypermutating strain with a passage size of 10% reached State 3 significantly faster than all others (it possessed a mutation in *mutY*). The purple hypermutating strain was identified as a potential hypermutating strain based on the number of mutations fixed (p=0.003, FDR=0.087) and the presence of a frame shift insertion in *mutL*.

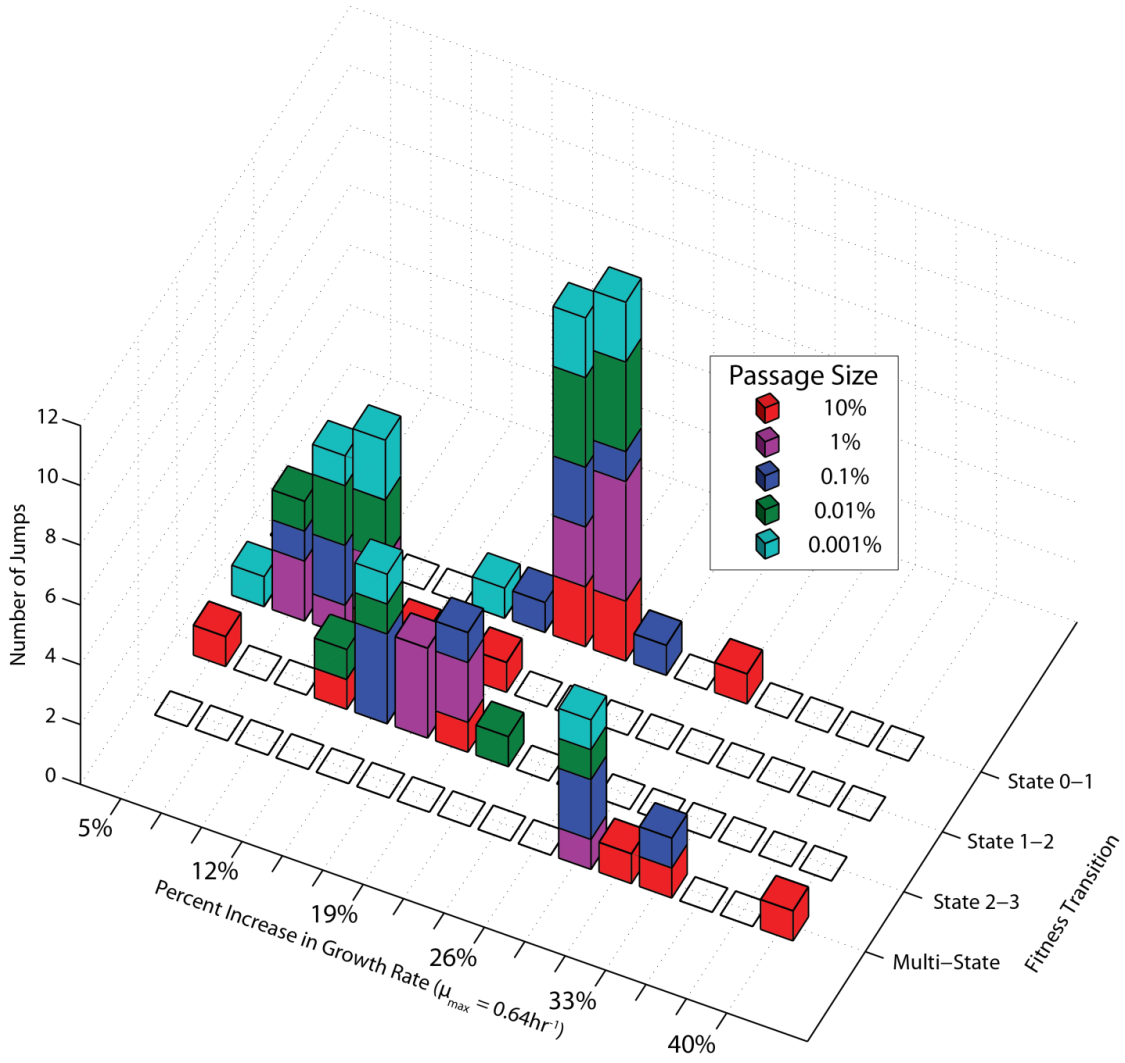

Figure 4.4 – Distribution of Fitness Increases in Glycerol ALE
A histogram of the normalized increases in growth rate (μ_{max} = 0.64 hr^{-1}) attributed to each jump for the different experiments. The fitness increases were categorized by which state transition was made. The different passage sizes (indicated by different colors) did not show any significant variance in the ability to fix distinct increases in growth rate. A few small jumps not shown are small observed increases in fitness that did not jump between any of the states identified.

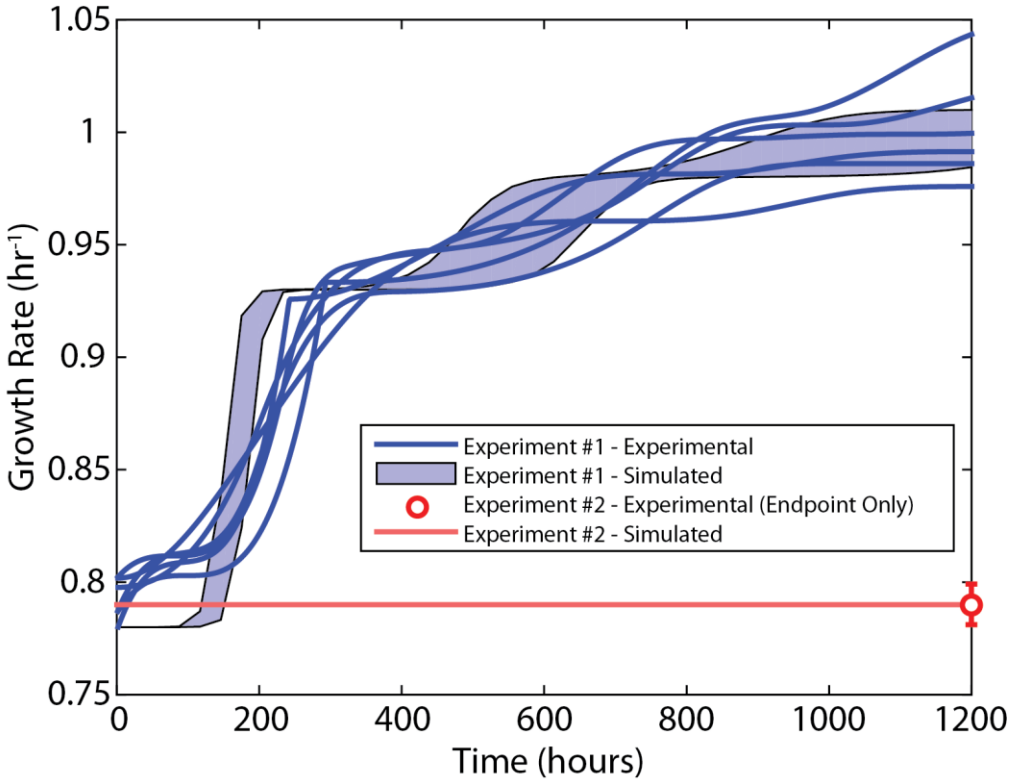

Figure 4.5 – Simulated vs Experimental Results with Large and Small Passage Sizes

Two ALE experiments of *E.coli* MG1655 in glucose M9 minimal media were simulated using ALEsim. The strain and media conditions were identical in the two experiments. The only differences were in the culture volume (25ml vs. 250mL), optical density when passed (variable vs. 1.2 OD_{600nm}), and passage volume (variable vs 800µL). The variable nature of the optical density when passed and the passage size in the latter experiment was a consequence of manually passing the culture each day. The former experiment employed an automated system of monitoring and passing the culture to maintain consistency. Despite being the same strain and conditions, the final fitness achieved in the two experiments were quite different. ALEsim was used to simulate these same experiments with the only differences being the three aforementioned parameters. Consequently, the ALEsim results showed that the differences in these parameters were sufficient to explain why the final growth rates were distinct.

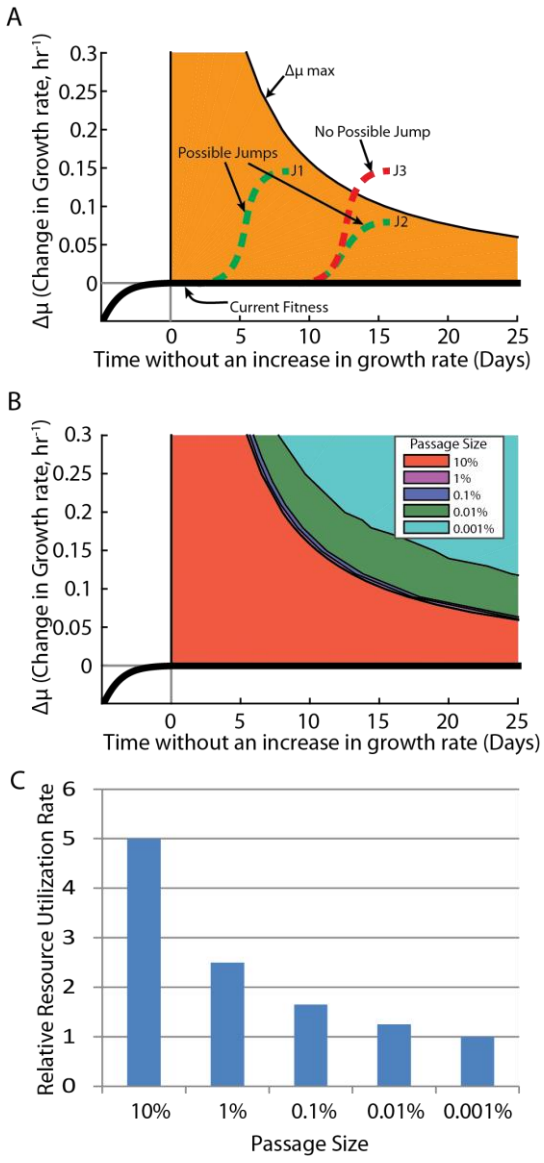

Figure 4.6 – Upper Bound on possible jumps in growth rates
A. Upper bounds on possible jumps in growth rates are shown. At a given point in time, a jump that reaches above the upper bound is statistically infeasible from a single mutation, whereas jumps that stay below the line are possible. B. The upper bound on jumps is shown for varying passage sizes. Increasing the passage size can have a significant impact on the upper bound. Consequently, the time required to eliminate jumps of certain magnitudes can take much longer to achieve. However, as the passage size increases there comes a point when the returns begin to diminish such that passage sizes between 0.1% and 10% did not show a large difference in the time required to find a given jump. C. Relative amount of resources needed to perform an ALE experiment normalized to the lowest passage size. As the passage size is increased the resource usage begins to increase greatly.

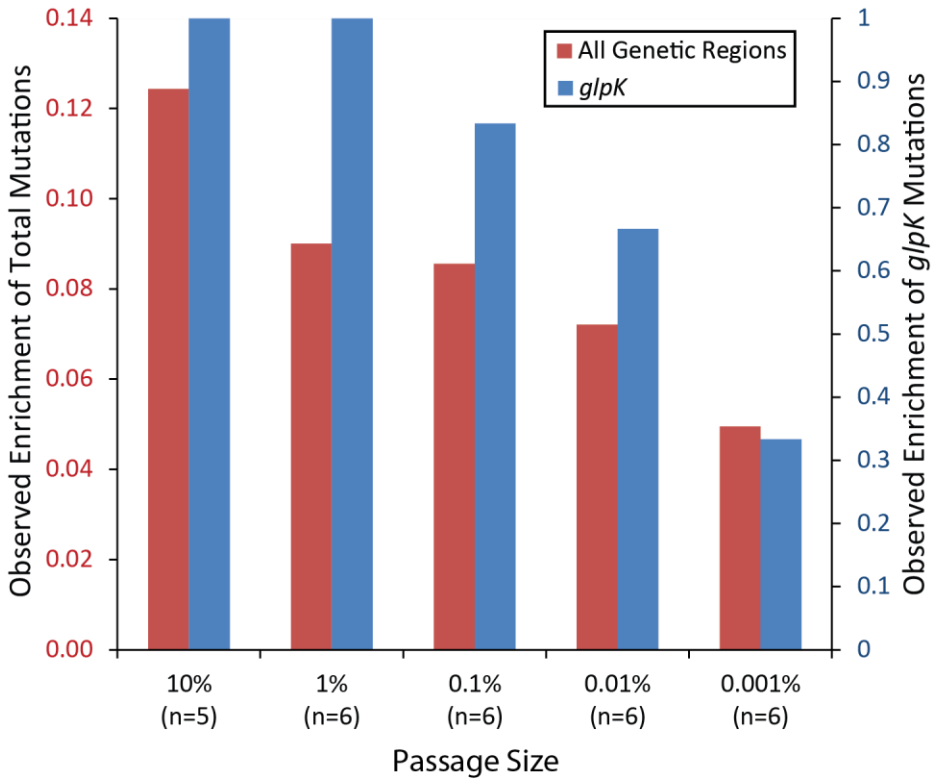

Figure 4.7 – Genetic Analysis – By Passage Size
A bar chart representing the observed fraction of mutations at a given passage volume. As a general trend, the larger the passage size, the greater the probability of a mutation in a given genetic region fixing in the population. A key mutation in the *glpK* gene is displayed as well as all mutations.

4.6 References

1. **Palsson B.** 2010. Adaptive laboratory evolution. Microbe.

2. **Conrad TM, Lewis NE, Palsson BO.** 2011. Microbial laboratory evolution in the era of genome-scale science. Mol Syst Biol **7:**509.

3. **Dragosits M, Mattanovich D.** 2013. Adaptive laboratory evolution -- principles and applications for biotechnology. Microb Cell Fact **12:**64.

4. **Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM.** 2014. Evolution of Escherichia coli to 42 degrees C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Mol Biol Evol **31:**2647-2662.

5. **Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ.** 2013. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. PLoS Comput Biol **9:**e1003091.

6. **Gresham D, Hong J.** 2015. The functional basis of adaptive evolution in chemostats. FEMS Microbiol Rev **39:**2-16.

7. **LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM.** 2015. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol **81:**17-30.

8. **Bacun-Druzina V, Cagalj Z, Gjuracic K.** 2007. The growth advantage in stationary-phase (GASP) phenomenon in mixed cultures of enterobacteria. FEMS Microbiol Lett **266:**119-127.

9. **Vasi F, Travisano M, Lenski RE.** 1994. Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. American Naturalist**:**432-456.

10. **Bachmann H, Fischlechner M, Rabbers I, Barfa N, Branco dos Santos F, Molenaar D, Teusink B.** 2013. Availability of public goods shapes the evolution of competing metabolic strategies. Proc Natl Acad Sci U S A **110:**14302-14307.

11. **Raynes Y, Halstead AL, Sniegowski PD.** 2014. The effect of population bottlenecks on mutation rate evolution in asexual populations. J Evol Biol **27:**161-169.

12. **Campos PR, Wahl LM.** 2010. The adaptation rate of asexuals: deleterious mutations, clonal interference and population bottlenecks. Evolution **64:**1973-1983.

13. **Campos PR, Wahl LM.** 2009. The effects of population bottlenecks on clonal interference, and the adaptation effective population size. Evolution **63:**950-958.

14. **Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:**1406-1412.

15. **Ibarra RU, Edwards JS, Palsson BO.** 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420:**186-189.

16. **Lee DH, Palsson BO.** 2010. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. Appl Environ Microbiol **76:**4158-4168.

17. **Conrad TM, Joyce AR, Applebee MK, Barrett CL, Xie B, Gao Y, Palsson BO.** 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.

18. **Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BO.** 2010. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet **6:**e1001186.

19. **Wahl LM, Gerrish PJ.** 2001. The probability that beneficial mutations are lost in populations with periodic bottlenecks. Evolution **55:**2606-2610.

20. **Hubbarde JE, Wahl LM.** 2008. Estimating the optimal bottleneck ratio for experimental evolution: the burst-death model. Math Biosci **213:**113-118.

21. **Deatherage DE, Barrick JE.** 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol **1151:**165-188.

22. **Neidhardt FC, Ingraham JL, Schaechter M.** 1990. Physiology of the bacterial cell : a molecular approach. Sinauer Associates, Sunderland, Mass.

23. **Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS.** 2012. The molecular diversity of adaptive convergence. Science **335:**457-461.

24. **Pramanik J, Keasling JD.** 1997. Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng **56:**398-421.

25. **Desai MM, Fisher DS, Murray AW.** 2007. The speed of evolution and maintenance of variation in asexual populations. Curr Biol **17:**385-394.

26. **Perfeito L, Fernandes L, Mota C, Gordo I.** 2007. Adaptive mutations in bacteria: high rate and small effects. Science **317:**813-815.

27. **Gerrish PJ, Lenski RE.** 1998. The fate of competing beneficial mutations in an asexual population. Genetica **102-103:**127-144.

28. **Uecker H, Hermisson J.** 2011. On the fixation process of a beneficial mutation in a variable environment. Genetics **188:**915-930.

29. **Lande R.** 2007. Expected relative fitness and the adaptive topography of fluctuating selection. Evolution **61:**1835-1846.

30. **Wright S.** 1929. Fisher's Theory of Dominance. American Naturalist **63:**274-279.

31. **Haldane JBS.** 1927, p 838-844. Mathematical Proceedings of the Cambridge Philosophical Society.

32. **Fisher RA.** 1930. The genetical theory of natural selection. The Clarendon press, Oxford,.

33. **Wahl LM, Zhu AD.** 2015. Survival probability of beneficial mutations in bacterial batch culture. Genetics **200:**309-320.

34. **Foster JA.** 2001. Evolutionary computation. Nat Rev Genet **2:**428-436.

35. **Reyes LH, Almario MP, Winkler J, Orozco MM, Kao KC.** 2012. Visualizing evolution in real time to determine the molecular mechanisms of n-butanol tolerance in Escherichia coli. Metab Eng **14:**579-590.

36. **Wiser MJ, Ribeck N, Lenski RE.** 2013. Long-term dynamics of adaptation in asexual populations. Science **342:**1364-1367.

37. **Kryazhimskiy S, Rice DP, Jerison ER, Desai MM.** 2014. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science **344:**1519-1522.

38. **Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF.** 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. Science **332:**1193-1196.

39. **Chou HH, Chiu HC, Delaney NF, Segre D, Marx CJ.** 2011. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science **332:**1190-1192.

40. **Barrick JE, Kauth MR, Strelioff CC, Lenski RE.** 2010. Escherichia coli rpoB mutants have increased evolvability in proportion to their fitness defects. Mol Biol Evol **27:**1338-1347.

41. **Perfeito L, Sousa A, Bataillon T, Gordo I.** 2014. Rates of fitness decline and rebound suggest pervasive epistasis. Evolution **68:**150-162.

42. **Sandberg TE, Long CP, Gonzalez JE, Feist AM, Antoniewicz MR, Palsson BO.** 2016. Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology. PLoS One **11:**e0151130.

Chapter V – Adaptive Laboratory Evolution Module Development

5.1 Introduction

The classical adaptive laboratory evolution (ALE) experiment involves subjecting a culture to a given environment and letting it continuously grow. During this time beneficial mutations serendipitously accumulate and fitness is increased in the given conditions. Other molecular biology techniques have been used to increase fitness in microorganisms using various genetic engineering techniques. A shortcoming of direct genetic manipulations is that is requires some *a priori* knowledge of how to manipulate the organism. Such knowledge is not always available nor does is it necessarily lead to an optimal solution. Adaptation through ALE is able to sample a wider range of possibilities than would be possible through molecular biology techniques and further select for those that confer the greatest benefit. A limitation of ALE is that the genetic diversity is created during cell replication. Since ALE relies on seemingly random mutations, it is important that large quantities of these mutations are created. Achieving these large numbers of mutations requires sustained and robust growth. Ultimately this means that the conditions that the cells are presumed to adapt must already be able to sustain growth. Unfortunately, this is an apparent limitation of the basic technology. A potential method of overcoming this limitation is using variable conditions from each batch culture. Using the already developed ALE machine platform modules can be developed to expand its applications to include pathway activation of latent enzymes (PALE ALE) and tolerization (TALE).

5.2 Pathway Activation of Latent Enzymes by Adaptive Laboratory Evolution

Adaptive laboratory evolution has proven to be a powerful tool in increasing the fitness of microorganisms in a given environment. What makes ALE such an effective tool is that it is able to compete a large number of possible genetic changes against each other in a single culture. Since the growth conditions are the target of adaptation, the strain or strains that ultimately dominate should be optimal or at least closer to optimal than the parent strain. This ability to select for such a large number of mutations in a single culturing vessel makes the process extremely efficient. A potential problem arises when the organism cannot natively grow in the targeted environment. An example would be evolving an organism to growth on a substrate that it natively cannot use (e.g. unique carbon, nitrogen, or sulfur source). A standard ALE experiment would fail in that it relies on growth to even get started.

The possibilities of using non-native substrates are vast and include many industrial and scientific applications. From an industrial standpoint many of the resources used in biological processing can be expensive. Having the ability to create an organism that can use a cheaper substrate has potential to yield significant increases in efficiency. As an example, a strain of *E. coli* was evolved to try and use cellulosic biomass as a carbon source (1). Given the renewable and cheap nature of plant based biomass it is industrially desired. From a scientific standpoint there are myriad potential new aspects to understand. For instance there has been significant inquiry into underground metabolism in microorganisms. Underground metabolism is classified as the reactions that enzymes catalyze that are secondary to their primary function (2, 3). It is often the case that substrates with similar chemical structure may

have enzyme reactivity but only at a very low rate. This low rate makes detection and understanding of the underground metabolism difficult to elucidate. However, if the organism were able to be evolved on a substrate that would require use of the secondary reaction, the flux might increase enough to properly identify and characterize it. The need to generate these novel metabolic characteristics exists in both industry and the sciences.

Previous attempts at evolving organisms to non-native substrates with varying degrees of success have been tried. This process has been named pathway activation of latent enzymes by adaptive laboratory evolution (PALE ALE). The process typically involves using a native substrate that is similar to the non-native substrate to induce growth and then over time increase the concentration of the non-native substrate while decreasing the concentration of the native substrate in hopes that enough genetic diversity is created to find this new functionality. The process can be logistically difficult in dealing with many different media types and has shown to have limited success when used (4).

Given the limited success with previous PALE ALE attempts creating a process where PALE ALE can be accomplished using the ALE machine is desirable. The first potential benefit from using the ALE machine for PALE ALE experiments is simply what comes with automation, more careful monitoring of growth and the ability to perform action at the ideal times, especially when the ideal time-frame is narrow. Implementing the module using the process described in other studies is sufficient for these gains. Since the machine was designed to accommodate new ALE processes this process was straightforward and easily obtainable. The second potential

benefit is in the algorithm itself. With automation, the algorithm is not bound by waking and sleeping hours, nor does it have to make decisions at the end of one day about what is expected by the beginning of these next. Because of this increased freedom, the actual algorithm to develop genetic diversity and selection can be amended to yield more desirable results in a shorter amount of time.

A primary attribute of the PALE ALE algorithm is imposing the correct selection pressure on the culture so that the strains that ultimately fix themselves in the population are those that can grow on the new substrate. Previous methods all employed some use of a native substrate to generate genetic diversity and then slowly decreased its concentration over the course of the experiment until it had disappeared. Though on the surface it may seem as if it could properly select for growth on the new substrate, it is most likely inhibitory. These experiments are passed during exponential phase. By virtue of the culture being in exponential phase, it implies that nutrients have not yet been depleted. This means that cells in the culture can increase their fitness by solely using the native substrate. Use of the non-native substrate can potentially be ignored. The only method by which cells could fix in the population is if they developed capabilities to grow and the non-native substrate and were able to do it at a faster growth rate than those still using the native substrate. It would probably be a reasonable assumption to assume that under many cases the non-native substrate will originally start out with limited growth and would be unlikely to exhibit growth that outpaces that of the native substrate. Using this method, while the native substrate is present, there is a limited chance of selecting and fixing a strain with the non-native substrate growth capabilities. In one study that applied this technique it was observed

that after two weeks the culture had very little of the native substrate left and was growing very poorly (5). As soon as the native substrate was removed growth on the non-native substrate was immediately observed. Either growth on the non-native substrate happened to have been conferred exactly when the native substrate was removed, or growth on the non-native substrate was already present and not allowed to fix itself in the population while the native was present. Due to these potential shortcomings amendments to the process would facilitate faster and more consistent selection for growth on the non-native substrate.

Another aspect of performing PALE ALE experiments is choosing the native substrate to initiate growth before transitioning to the non-native substrate. In many cases the native substrate was chosen due to its similarity to the non-native substrate. As an example is a study which evolved *E. coli* onto 1-2 propanediol. They chose glycerol as a native substrate since it appears chemically similar to 1-2 propanediol. Ultimately the study was successful but the pathways used to metabolize 1-2 propanediol were completely dissimilar to that of glycerol. This study is at least an example that it is possible to achieve success without the use of a similar native substrate but does not necessarily indicate whether glycerol was advisable or not. It is most likely the case that different strains and conditions lend themselves to different ideal non-native substrates. When choosing a non-native substrate it is important to keep in mind its purpose. The purpose is to generate genetic diversity by means of replication. Therefore an ideal substrate would be one that is easily obtainable and confers a significant amount of robust growth. The idea of the native substrate being similar is rooted in the idea that the native substrate is present while growth on the

non-native substrate is presumably taking place. As discussed above, this is non-ideal since it will disallow fixation of the strains that can grow on the non-native substrate. Ultimately the ideal substrate is one that simply creates large amounts of genetic diversity is a short amount of time.

Given the aforementioned issue with the current methods an algorithm was proposed and tested for implementation on the ALE machine. The method can be summarized in Figure 5.1. In an effort to continue putting selection pressure on the culture and generate genetic diversity. This new method uses two concurrent cultures. The first is used to create and enrich genetic diversity and the second is to test whether growth is sustainable on the non-native substrate. The first culture tube is given media that has both the native and non-native substrate. The non-native substrate is at the target concentration used for growth. This concentration is constant throughout the entire experiment. The native substrate however is variable. A key function of this culture is to create genetic diversity and enrich for those that can grow on the non-native substrate. This is accomplished by allowing the culture to reach what looks to be stationary phase. It is important that this stationary phase is a result of using all of the non-native substrate and deletion of another media component. If this is the case, this culture now has generated a tube full of genetic diversity and there is no non-native substrate left to consume. The next process is simply to wait for a pre-determined about of time in stationary phase. If any mutations arose that conferred growth on the non-native substrate they will begin to grow and those that are unable to utilize it will remain in stationary phase. Thus instead of needing the strain to have learned to grow on the non-native substrate and grow faster than the parent strain on

the native substrate. Any growth rate, no matter how slow, will allow enrichment. In each subsequent batch culture, the native substrate's concentration can be altered. This is important as the native substrate must be the limiting factor in the media and not some other component. It is expected that the culture will start evolving to these conditions while genetic diversity is being created. Being able to lower its concentration ensures that if it starts using the substrate more efficiently it will still be the limiting factor.

The length of time to leave the culture in this enrichment phase is a critical component. It is likely that once a strain is able to grow on the substrate that it will do so with a slow growth rate. Because of this, to see an observable enrichment in the culture through optical density measurement would require a considerable amount of time. The time may be well spent if there indeed is a strain that has conferred growth but there is always the possibility that this has not happened. During this enrichment phase, generation of genetic diversity is limited to that that is generated in stationary phase (6). Ideally it would be preferable to continue testing for non-native substrate growth while continuing to generate genetic diversity. To accomplish this, a second culture is used. After a modest enrichment phase the culture is first passed to a subsequent culture for another round of growth and enrichment, but it is also passed to the second culture where there is no native substrate available and only the non-native substrate. Thus any growth observed in this culture is indicative of successful adaptation to the non-native carbon source. It may take a considerable amount of time to observe this growth so it is set aside from the main culture so as not to inhibit the evolutionary process.

Overall, the previously employed methods of adaptation to non-native substrates created a weak selection on the phenotype desired. Because of this experiments took longer than necessary and failed experiments were potentially artifacts of the process. The newly proposed method implemented on the ALE machine ensure that there is a balance between generating genetic diversity, that is required for evolution, and imposing the selection pressure to enrich and ultimately fix strains that have adapted to the non-native substrate.

5.3 Tolerization Adaptive Laboratory Evolution

A step beyond adaptation to a constant condition, as is typical of ALE experiments, is adaptation to an ever increasing stressful environment. There are a wide range of stresses that can be imposed upon a microorganism that range from chemical stress to temperature, pH, and physical stress (1, 7-10). This is of particular interest to those involved with industrial bioprocessing applications. Metabolic engineering has created myriad strains that produce economically relevant byproducts, often non-native ones. A particular pitfall of the process is that industrially meaning concentrations of these products often induce stress in the cells that create them and can furthermore be completely toxic to them. The concentrations necessary to be economically viable are often toxic and molecular biology tools often prove to be insufficient. A strength of ALE is that it is a blindly operating tool that does not require *a priori* knowledge of anything pertinent to the desired outcome, nor does it require biological discovery or understanding to reach the end goal. An algorithm was developed for the ALE machine to tolerize cultures to a wide range of stressors.

Tolerizing microorganisms suffers from the same issue that PALE ALEs do in that the end goal is a condition that currently does not support any growth. This lack of growth inhibits traditional adaptive evolution. A common method of accommodating this is to periodically increase the stressor over the course of the experiment. This has proven to be successful (11). As such the module described herein does not deviate from the concept significantly but rather implements a series of checks to ensure that growth remains robust. A significant risk when tolerizing a strain is that if the stress becomes too great the culture will die off. On the other hand if the stress is minimal the experiment is inefficient and takes longer than necessary. Ideally there would be a balance. Generalizing for all stressors would be impossible but ideally the module created will allow the researcher the latitude to adjust to various stressors.

A preliminary version of the TALE module was implemented to ensure robust growth continued. It was observed that in a batch culture immediately after the stressor was increased growth would often be observed, however in the next batch the culture would die off. Intuitively, it was assumed that growth in one batch would be sufficient to identify it as being able to grow however this was not the case as this phenomenon was observed across multiple conditions and strains. As such a minimum rest parameter was added. The minimum rest parameter specifies the number of batch cultures that must show growth before it is considered growth and the stress increased again. This proved sufficient in ensuring robust growth in a wide range of studies. Additional parameters were defined to further ensure robust growth as well and a reasonable amount of stress. In addition to the minimum rest, these include a

maximum rest and upper and lower growth rate limits. These are outlined in Figure 5.2.

5.4 Figures

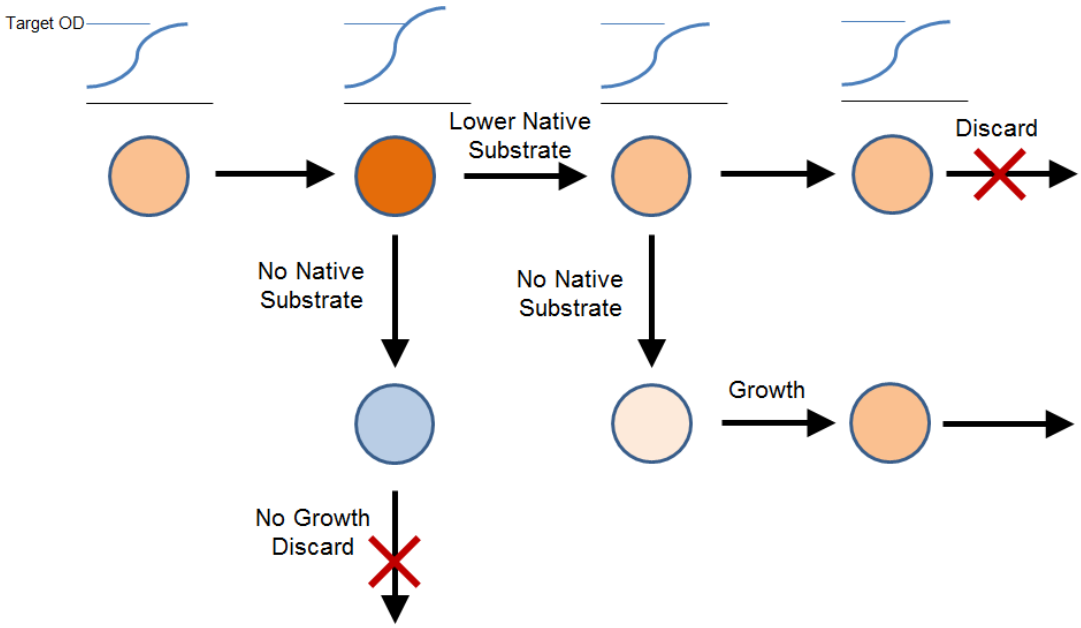

Figure 5.1 Pathway Activation of Latent Enzymes by Adaptive Laboratory Evolution (PALE ALE) Workflow. The goal of PALE ALE is to evolve an organism onto a non-native substrate. This workflow begins with a culture filled with media that has a native substrate that the organism can grow on as well as the non-native substrate of interest. A target optical density (OD) is defined. This is the ideal OD of stationary phase. This OD is important in that it must be high enough for ample genetic diversity to have occurred in the culture but low enough that only the native substrate has been depleted. If in a subsequent cultures the actual OD goes beyond the target OD, the native substrate concentration is reduced to accommodate. Reaching stationary phase with only the native substrate depleted allows for any strains that have mutated to use the non-native substrate to grow while all others remain stagnant. Ultimately this enriches them in the population. When the culture is passed to the next culturing vessel it is also passed to a second culture vessel where only the non-native substrate is available. If this tube is able to grow then it has been successfully evolved and the original culture is discarded. If it does not grow it is discarded and tired again with the next passage.

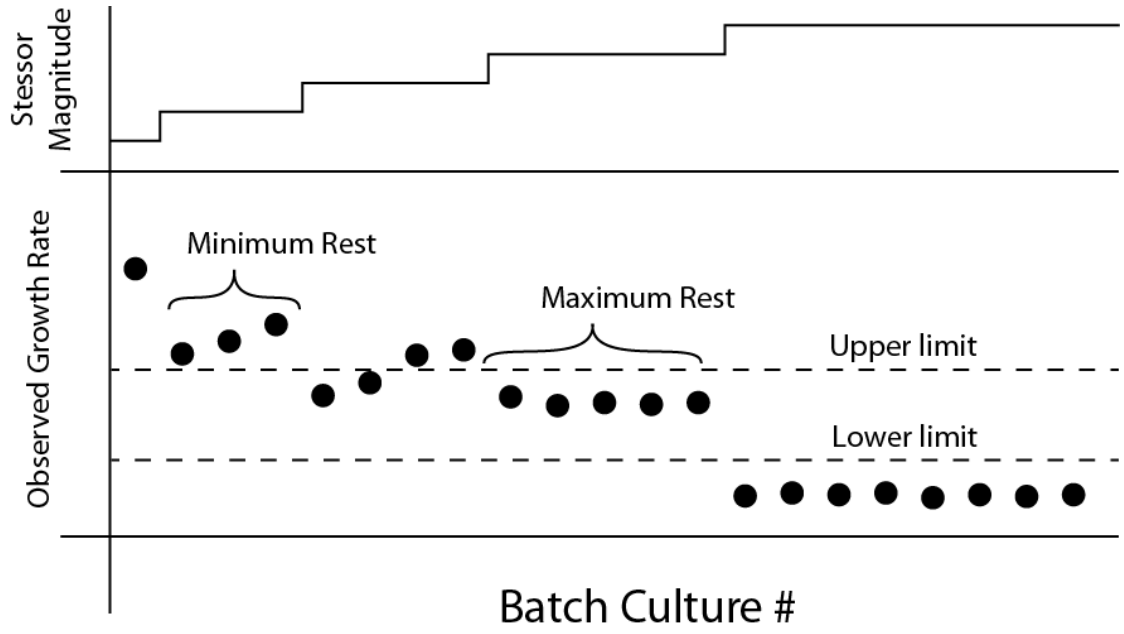

Figure 5.2 – TALE Algorithm Summary. The tolerization ALE employs four parameters to ensure that cultures are subjected to a reasonable amount of stress but do not crash. The first is the minimum rest. This specifies the number of batch cultures that must be passed to before stress can be increased again. There are no exceptions that allow stress to be increased until this condition is met. Next is the upper limit. This specifies the growth rate that if exceeded will immediately increase the stress as long as the minimum rest condition is met. Next is the lower limit. This specifies the minimum growth rate that must be met for stress to increase. Under no circumstances will the stress be increased if this condition is not met. Finally is the maximum rest. If the upper limit is not exceed after a specified number of batches, the stress will be increased as long as the lower limit condition is met.

5.5 References

1. **Applebee MK.** 2010. Dynamics of genetic adaptation in Escherichia coli K12 MG1655.

2. **D'Ari R, Casadesus J.** 1998. Underground metabolism. Bioessays **20:**181-186.

3. **Jensen RA.** 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol **30:**409-425.

4. **Orth JD, Palsson B.** 2012. Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. BMC Syst Biol **6:**30.

5. **Lee DH, Palsson BO.** 2010. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. Appl Environ Microbiol **76:**4158-4168.

6. **Finkel SE.** 2006. Long-term survival during stationary phase: evolution and the GASP phenotype. Nat Rev Microbiol **4:**113-120.

7. **Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BO.** 2010. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet **6:**e1001186.

8. **Dragosits M, Mattanovich D.** 2013. Adaptive laboratory evolution -- principles and applications for biotechnology. Microb Cell Fact **12:**64.

9. **Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS.** 2012. The molecular diversity of adaptive convergence. Science **335:**457-461.

10. **Atsumi S, Wu TY, Machado IMP, Huang WC, Chen PY, Pellegrini M, Liao JC.** 2010. Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. Molecular Systems Biology **6**.

11. **Reyes LH, Abdelaal AS, Kao KC.** 2013. Genetic determinants for n-butanol tolerance in evolved Escherichia coli mutants: cross adaptation and antagonistic pleiotropy between n-butanol and other stressors. Appl Environ Microbiol **79:**5313-5320.